Neural-Driven Image Editing

Pengfei Zhou[1*†], Jie Xia[2*†], Xiaopeng Peng[3†], Wangbo Zhao[1], Zilong Ye[2], Zekai Li[1], Suorong Yang[1,4], Jiadong Pan[2], Yuanxiang Chen[5], Ziqiao Wang[1], Kai Wang[1], Qian Zheng[2], Xiaojun Chang[5,6], Gang Pan[2], Shurong Dong[2‡], Kaipeng Zhang[7,8‡], Yang You[1]
[1]NUS [2]Zhejiang University [3]RIT [4]NJU [5]USTC [6]MBZUAI [7]Shanghai AI Lab [8]SII
zpf4wp@outlook.com, {jiexia, dongshurong}@zju.edu.cn, zhangkaipeng@pjlab.org.cn

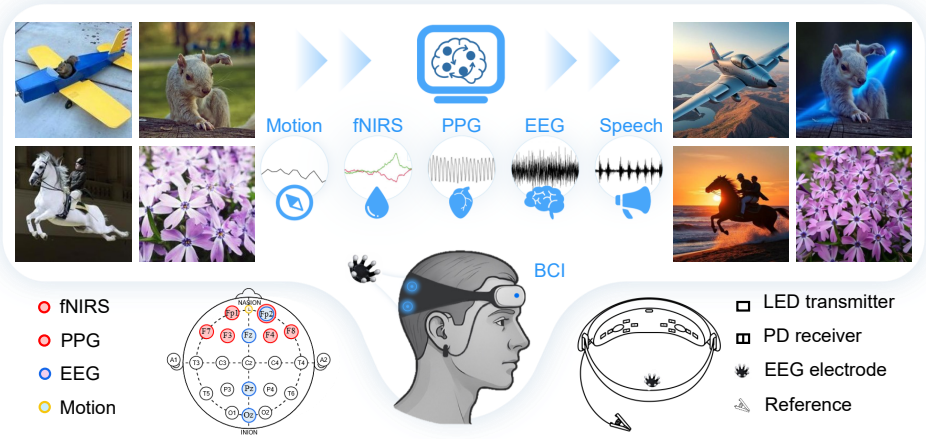

Figure 1: Illustration of LoongX for hands-free image editing via multimodal neural signals.

Abstract

Traditional image editing typically relies on manual prompting, making it labor-intensive and inaccessible to individuals with limited motor control or language abilities. Leveraging recent advances in brain-computer interfaces (BCIs) and generative models, we propose **LoongX**, a hands-free image editing approach driven by multimodal neurophysiological signals. LoongX utilizes state-of-the-art diffusion models trained on a comprehensive dataset of 23,928 image editing pairs, each paired with synchronized electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), photoplethysmography (PPG), and head motion signals that capture user intent. To effectively address the heterogeneity of these signals, LoongX integrates two key modules. The cross-scale state space (CS3) module encodes informative modality-specific features. The dynamic gated fusion (DGF) module further aggregates these features into a unified latent space, which is then aligned with edit semantics via fine-tuning on a diffusion transformer (DiT). Additionally, we pre-train the encoders using contrastive learning to align cognitive states with semantic intentions from embedded natural language. Extensive experiments demonstrate that LoongX achieves performance comparable to text-driven methods (CLIP-I: 0.6605 vs. 0.6558; DINO: 0.4812 vs. 0.4636) and outperforms them when neural signals are combined with speech (CLIP-T: 0.2588 vs. 0.2549). These results highlight the promise of neural-driven generative models in enabling accessible, intuitive image editing and open new directions for cognitive-driven creative technologies. The code and dataset are released on the project website: https://loongx1.github.io.

*Equal contribution; †Core contributor; ‡Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Image editing involves manipulating digital visuals to achieve desired effects, significantly impacting fields like advertising, entertainment, and scientific visualization [1]. Traditionally, this task demands extensive manual effort and technical expertise. Advances in generative models have streamlined instruction-based image editing through automated pipelines [2–4]. Nevertheless, these methods still heavily depend on intensive user inputs, such as text prompts [5, 6], visual references like masks or sketches [7, 8], and physical operations like dragging [9–12]. Such reliance limits efficiency and accessibility, especially for users with motor or communication impairments.

To address these challenges, alternative input modalities have been explored [13–15] for image editing. Among these, brain–computer interfaces (BCIs) provide a promising possibility with their recent advancement in hardware precision [16, 17]. Starting from early attempts in passive tasks such as mental state recognition [18] and neural activity analysis [19, 20], BCIs have begun to be involved in more active generative tasks such as neural-driven chat [21] and visual content creation [22, 23].

However, existing approaches remain limited to the use of single-modality data such as electroencephalography (EEG) [22, 24] or functional magnetic resonance imaging (fMRI) [25], which is insufficient to capture nuanced user intentions for enabling complex editing tasks. In practice, physiological signals from different modalities can offer complementary insights into cognitive states such as attention, motivation, and emotional regulation [26–32], underscoring the need for multimodal neural information integration.

Given the limited exploration of this emerging area, here we ask three key research questions (RQs):

RQ1. Can neural signals alone drive instruction-based image editing?
RQ2. If yes, what kind of information do multimodal neural signals contribute?
RQ3. How do neural-signal conditions compare and complement natural-language instructions?

To answer these questions, we construct **L-Mind**, a comprehensive multimodal dataset comprising 23,928 image pairs with synchronously collected EEG, functional near infrared spectroscopy (fNIRS) [33], photoplethysmography (PPG) [34], head motion, and speech signals from 12 participants conceiving image editing tasks. Captured using a wireless, lightweight BCI system that supports unconstrained head movements and speech [35], L-Mind offers higher ecological validity under natural real-world conditions and supports robust training of brain-supervised generative models.

Building on L-Mind, we propose **LoongX**, a hands-free image editing approach that innovatively integrates the proposed multimodal neural signal fusion strategy with a diffusion transformer (DiT) to translate neural intent into image edits. Unlike prior single-modal methods, LoongX integrates EEG, fNIRS, PPG, and head motion signals, extracting explicit user intentions from EEG signals across multiple scalp regions, incorporating cognitive load and emotional valence data from fNIRS, and capturing stress and engagement indicators through PPG and motion signals. We introduce two new modules to manage diverse multimodal input: a cross-scale state space (CS3) encoder for robust feature extraction and a dynamic gated fusion (DGF) module for comprehensive multimodal integration. These encoders are pretrained via contrastive learning on combined large-scale datasets and L-Mind to align neural features with semantic text embeddings.

Extensive experiments qualitatively and quantitatively demonstrate the feasibility of neural-driven image editing. Integrated multimodal neural signals achieve performance comparable to text-driven baselines (CLIP-I: 0.6605 vs. 0.6558; DINO: 0.4812 vs. 0.4637). Combined neural signals with speech instructions surpass text prompts alone (CLIP-T: 0.2588 vs. 0.2549). Ablation studies verify the effectiveness of proposed modules and further explore the contribution of each signal, showing that EEG + fNIRS contribute most among signals, and the Oz and Fpz sites, as EEG input channels, represent the key brain region. These findings underscore LoongX's potential to facilitate intuitive, inclusive image editing and inspire future human–AI interaction.

Our main contributions are summarized as follows:

1) **L-Mind**, a multimodal dataset with 23,928 image-editing pairs featuring synchronized EEG, fNIRS, PPG, motion, and speech signals collected in natural settings.
2) **LoongX**, a novel neural-driven editing method with CS3 and DGF modules for effective feature extraction and multimodal integration (see the effect in Fig. 1).
3) Extensive experiments validate multimodal neural signals' effectiveness and provide insights into modality-specific contributions and their synergy with speech-based inputs.

2 Related Works

2.1 Brain-supervised Generation

As an emerging technology, brain–computer interface (BCI) builds direct communication between the brain and devices by decoding neural signals [36, 37]. Advances in machine learning have improved its accuracy, enabling brain-guided generative methods for visual content creation. Several recent methods integrate neurophysiological data (e.g., fMRI, EEG, or fNIRS) with generative models [22, 38, 39]. For instance, CMVDM aligns fMRI features with semantics for image synthesis [40], and the MindEye series further lifts the resolution of generated images from decoded fMRI [41, 42]. OneLLM leverages the large fMIR dataset for multimodal alignment in a multimodal large language model [16]. DreamDiffusion produces images from EEG via temporal masked modeling [22]. EEG2Video extends the idea to dynamic video content [43]. While Davis *et al.* [24] initially explore brain-guided semantic image editing using a generative adversarial network, this work is limited to facial images and EEG signals. Moreover, Adamic *et al.* [44] reconstructs visual images from brain activity measured by fNIRS.

Unlike previous studies, our data are collected using a wireless BCI system (Fig. 2) as participants conceive instruction-based image edits. Compared with fMRI methods, our framework combines lightweight EEG, fNIRS, PPG, and head-motion signals, which can support greater portability and broader real-world applicability. To the best of our knowledge, this is the first work to fully leverage all these signals for instruction-based image editing, focusing on improved neural feature encoding and optimized multimodal fusion strategies.

2.2 Instruction-based Image Editing

Recent generative models like GPT-4o [45] and Gemini [46] have evolved from basic question answering to advanced image editing by interpreting user instructions. Modern instruction-based image editing agents integrate multimodal inputs, including text, images, and videos, to accurately identify and apply visual edits [13]. Leveraging learned multimodal representations, these agents interpret instructions from input, localize relevant regions, and perform targeted modifications [47–49]. Recent approaches, such as InstructPix2Pix [1], UltraEdit [2], MagicBrush [50], MIGE [51], and ACE [52] improve region-specific edits guided by natural language prompts. Speech-driven image editing [14] was also explored, highlighting the feasibility of hands-free interaction but still limited by linguistic expressiveness in recorded speech.

Despite these advancements, achieving efficient, delicate, prompt-free image editing remains challenging. Our work addresses this gap, exploring neural-signal-driven editing agents to decode cognitive intent directly for image manipulation. It is believed that the findings in this work can significantly enhance accessibility and interaction efficiency in future BCI-enabled applications, particularly benefiting individuals with physical disabilities.

3 Dataset

3.1 Data Collection

We collect 23,928 editing samples (22,728 training, 1,200 testing) from 12 participants using the setup depicted in Fig. 2. Participants wear our multimodal sensor while viewing image-text pairs sourced from SEED-Data-Edit [53] on a 25-inch monitor (resolution: 1980×1080). The measured EEG, fNIRS, and PPG physiological signals are streamed in real-time via Bluetooth 5.3, synchronized and aligned via lab streaming layer by the proprietary *Lab Recorder* software [54]. Participants simultaneously read displayed editing instructions aloud, providing audio speech signals.

Experiments are conducted in a quiet, temperature-controlled room (24°C, consistent humidity), starting at 9 AM daily. EEG signals are collected via non-invasive hydrogel electrodes, replaced every five hours to maintain signal quality. The experimental room is shielded from sunlight to prevent interference with fNIRS and PPG signals. Sessions start and end with participant-controlled audio recording and are marked by image names. Data from inactive intervals are excluded.

Each session (Fig. 2) starts and ends with user-initiated audio recording and is labeled by the paired image. A 1-second cross-fixation follows each image pair, with breaks every 100 images. Twelve

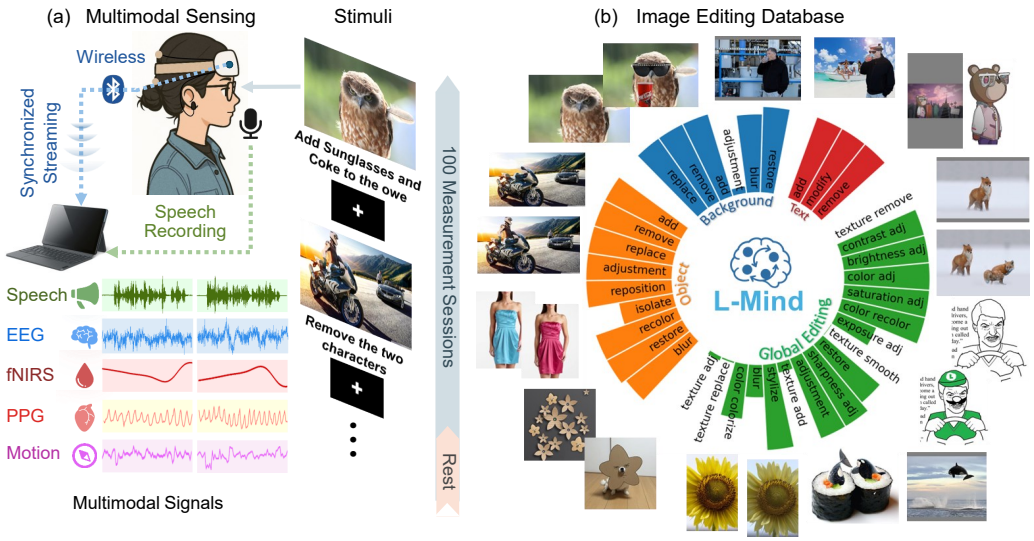

Figure 2: The L-Mind dataset comprises 23,928 multimodal editing samples, each including an original image, a ground truth text editing instruction, a ground truth edited image, as well as measured EEG, fNIRS, PPG, motion and speech signals. (a) Multimodal data collection pipeline; (b) Illustration and statistics of 35 types of image editing tasks.

healthy college students (6 female, 6 male; mean age: 24.5 ± 2.5 years old) with normal or corrected vision participated. All participants gave informed consent and received financial compensation. The study was officially approved by the corresponding institute's ethics committee.

3.2 Data Preprocessing

EEG. Four EEG channels (Pz, Fp2, Fpz, Oz; sampled at 250 Hz) undergo band-pass filtering (1–80 Hz) and notch filtering (48–52 Hz) to remove drifts, noise, and powerline interference. Ocular artifacts in Fp2 and Fpz are retained to capture eye movements.

fNIRS. Six-channel fNIRS signals (735 nm, 850 nm) are converted to relative hemoglobin concentration changes (HbO, HbR, HbT) using the Modified Beer–Lambert law. Optical density change is computed as $\Delta A(\lambda) = \log\left(I_0(\lambda)/I(\lambda)\right)$. Concentration changes are calculated as:

$$\begin{bmatrix}\Delta\text{HbO}\\ \Delta\text{HbR}\end{bmatrix} = \frac{1}{\text{DPF}\cdot L}\cdot\begin{bmatrix}\varepsilon^{\lambda_1}_{\text{HbO}} & \varepsilon^{\lambda_1}_{\text{HbR}}\\ \varepsilon^{\lambda_2}_{\text{HbO}} & \varepsilon^{\lambda_2}_{\text{HbR}}\end{bmatrix}^{-1}\cdot\begin{bmatrix}\Delta A(\lambda_1)\\ \Delta A(\lambda_2)\end{bmatrix} \tag{1}$$

Hemodynamic signals (ΔHbO, ΔHbR, and ΔHbT, where ΔHbT = ΔHbO + ΔHbR) are band-pass filtered (0.01–0.5 Hz) to isolate relevant neural responses, averaged per hemisphere to reflect task-related brain activity.

PPG and motion Four-channel PPG signals (735 nm, 850 nm) are averaged per hemisphere via adaptive average pooling and filtered (0.5–4 Hz) to extract cardiac-related hemodynamic signals that reflect heart rate variability. Motion data from a six-axis sensor (12.5 Hz), capturing triaxial linear acceleration and angular velocity, characterizes head movements. See supplement for more details.

4 Method

As illustrated in Fig. 3, LoongX extracts multimodal features from diverse neural signals and fuses them into a shared latent space in a pair-wise manner. Using Diffusion Transformer (DiT), the original image is translated into an edited image conditioned on the fused features. Following three research questions, we conduct a multi-label classification experiment (Sec. A.2) showing that EEG outperforms noise by 20%, and fusing all signals yields the highest F1 score. Combining neural signals with text achieves the best mAP, confirming modality complementarity. An input length of 8,192 gives the best performance with higher computational cost, motivating our framework's design: a cross-scale state-space encoder for long sequences and dynamic gated fusion for feature integration.

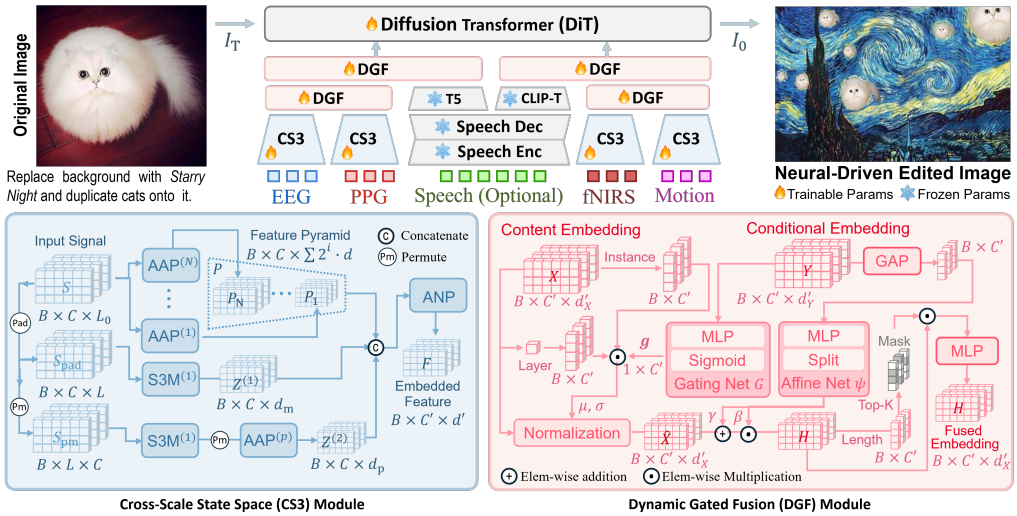

Figure 3: Overview of our proposed LoongX method for hands-free image editing. Receiving an input image, LoongX outputs an edited image using neural signals (and optional speech) as conditions.

4.1 Cross-Scale State Space Encoding

CS3 encoder extracts multi-scale features from diverse signals using an adaptive feature pyramid. To further capture dynamic spatio-temporal patterns beyond the fixed pyramid, CS3 uses a structured state space model (S3M) [55] for efficient long-sequence encoding with linear complexity. To manage cost, it uses a cross-feature mechanism that separately encodes temporal and channel information.

Pyramid Encoding. A single modality input signal $\mathbf{S} \in \mathbb{R}^{C \times L_0}$ is fed into an N-layer adaptive average pooling (AAP) module:

$$\{\mathbf{P}_i | i = 1, ..., N\} = \mathrm{AAP}^{(i)}_{L \to s_i}(\mathbf{S}), \quad s_i = d \cdot 2^i \tag{2}$$

where we set $N = 5$ and $d = 64$ for EEG. The extracted embedding is computed as the concatenation of the feature pyramid $\mathbf{P} = \mathrm{Concat}(\{\mathbf{P}_i\})$.

State Space Encoding. To fully exploit both temporal and channel-wise dependencies in neural signals, we design a cross-shaped spatiotemporal encoding scheme, where one axis focuses on temporal patterns and the other on channel-wise dynamics.

Specifically, the input signal $\mathbf{S}$ is padded from length L_0 to L, where $\mathbf{S}_{\text{pad}} \in \mathbb{R}^{C \times L}$ with signal intensity normalized to $[-1, 1]$. The padded signals and its permuted version $\mathbf{S}_{\text{pm}} \in \mathbb{R}^{L \times C}$ are passed to two parallel S3M blocks, $\mathrm{S3M}^{(1)}$ and $\mathrm{S3M}^{(2)}$, respectively:

$$\mathbf{Z}^{(1)} = \mathrm{S3M}^{(1)}(\mathbf{S}_{\text{pad}}), \quad \mathbf{Z}^{(2)} = \mathrm{S3M}^{(2)}(\mathbf{S}_{\text{pm}}), \tag{3}$$

where each S3M block uses the continuous-time diagonal state-space model:

$$\dot{\mathbf{e}}(t) = \hat{\mathbf{A}}\mathbf{e}(t) + \hat{\mathbf{B}}s(t), \quad \mathbf{z}(t) = \hat{\mathbf{C}}\mathbf{e}(t) + \hat{\mathbf{D}}s(t), \tag{4}$$

where $\mathbf{e}(t)$ denotes the latent state at time t, and $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$, $\hat{\mathbf{D}}$ are diagonal matrices that parameterize state transitions, input injection, state-to-output mapping, and direct input-to-output mapping, respectively. Due to the diagonal parameterization, the S3M block admits efficient computation with linear complexity $\mathcal{O}(L \log L)$. Through the S3M blocks, $\mathbf{Z}^{(1)}$ is down-sampled from length L to d_m, yielding $\tilde{\mathbf{Z}}^{(1)} \in \mathbb{R}^{C \times d_m}$; $\mathbf{Z}^{(2)}$ is permuted and down-sampled via an AAP, giving $\tilde{\mathbf{Z}}^{(2)} \in \mathbb{R}^{C \times d_p}$.

Cross-Pyramid Aggregation. The encoder merges multi-scale and temporal streams along the channel dimension, resulting in:

$$\mathbf{F} = \mathrm{ANP}\big(\,\mathrm{cat}_c(\tilde{\mathbf{Z}}^{(1)}, \mathbf{P}, \tilde{\mathbf{Z}}^{(2)}) \in \mathbb{R}^{C \times d'}\big). \tag{5}$$

where $d' = d_m + d_p + \Sigma_i^N 2^i \cdot d$. The concatenated feature was projected via Adaptive Nonlinear Projection (ANP), which consists of two fully-connected layers, LayerNorm, ReLU, and dropout. The final embeded feature $\mathbf{F} \in \mathbb{R}^{C' \times L'}$ is obtained.

4.2 Dynamic Gated Multimodal Fusion

We propose the Dynamic Gated Fusion (DGF) module to dynamically bind a pair of content and condition embeddings to a unified latent space, which is further aligned with text embeddings. Our DGF includes gate mixing, adaptive affine modulation, and a dynamic masking block.

Gated Mixing. We calculate instance-wise and layer-wise mean μ and variance σ from input **content** embedding (e.g., EEG) $\mathbf{X} \in \mathbb{R}^{C' \times L'_X}$ and **condition** embedding (e.g., PPG) $\mathbf{Y} \in \mathbb{R}^{C' \times L'_Y}$ for further fusion into $\tilde{\mathbf{H}} \in \mathbb{R}^{C' \times L'_X}$ while emphasising informative channels and suppressing noise:

$$\underbrace{\begin{bmatrix} \mu_{\text{inst}} & \sigma_{\text{inst}} \\ \mu_{\text{layer}} & \sigma_{\text{layer}} \end{bmatrix}}_{\text{each entry} \in \mathbb{R}^{C \times 1}} = \begin{bmatrix} \dfrac{1}{L'_X} \displaystyle\sum_t \mathbf{X}_{:,t} & \sqrt{\dfrac{1}{L'_X} \displaystyle\sum_t \left(\mathbf{X}_{:,t} - \mu_{\text{inst}}\right)^2 + \varepsilon} \\ \dfrac{1}{C'L'_X} \displaystyle\sum_{c,t} \mathbf{X}_{c,t}\, \mathbf{1}_{C'} & \sqrt{\dfrac{1}{C'L'_X} \displaystyle\sum_{c,t} \left(\mathbf{X}_{c,t} - \mu_{\text{layer}}\right)^2 + \varepsilon} \end{bmatrix}, \qquad \varepsilon = 10^{-3}. \tag{6}$$

where ε is a regularization term for numerical stability and $\mathbf{1}_C$ is a unit vector. A 1-D Gating Network $G(\cdot) = \sigma(\text{Conv}-\text{ReLU}-\text{Conv})$ is used to compute per-channel weights $\mathbf{g} \in [0,1]^{C \times 1}$ from $\mathbf{C}$, adaptively mixing statistics:

$$\mu = \mathbf{g} \odot \mu_{\text{inst}} + (1 - \mathbf{g}) \odot \mu_{\text{layer}}, \quad \sigma = \mathbf{g} \odot \sigma_{\text{inst}} + (1 - \mathbf{g}) \odot \sigma_{\text{layer}}. \tag{7}$$

The content feature is then normalized by the adaptively gated mean μ and standard deviation σ as: $\hat{\mathbf{X}} = (\mathbf{X} - \mu)/\sigma$.

Adaptive Affine Modulation. The conditional feature was averaged by a global average pooling (GAP) as $\bar{\mathbf{Y}} = \frac{1}{L} \sum_t \mathbf{Y}_{:,t}$. This averaged feature is then passed to the Affine Network ψ, which consists of a multi-layer perceptron (MLP). The output is split into two affine coefficients γ and β:

$$[\gamma, \beta] = \psi(\bar{\mathbf{Y}}), \quad \mathbf{H} = (1 + \gamma) \odot \hat{\mathbf{X}} + \beta. \tag{8}$$

Dynamic Masking. Channel importance scores $s_c = \frac{1}{L} \sum_t |\mathbf{Y}_{c,t}|$ are computed to select the top-k channels ($k = \lfloor \rho Y \rfloor$, $\rho = 0.7$) among the modulated features. Additionally, a binary mask $\mathbf{M} \in \{0,1\}^C$ is applied:

$$\tilde{\mathbf{H}}_{c,:} = \mathbf{M}_c\, \mathbf{H}_{c,:}. \tag{9}$$

Finally, the fused latent feature $\tilde{\mathbf{H}}$ is residually fused with the original prompt/text embeddings before being fed into a DiT decoder. Because DGF operates on arbitrary (C, L) tensors, it handles four types of modality fusion in LoongX: EEG-PPG, fNIRS-Motion, neural-prompt, and neural-pooled-prompt.

4.3 Conditional Diffusion

The fused latent representation conditions a DiT backbone [56] for image editing. The DiT model accepts the encoded input image $\mathbf{I}$ and fused latent feature $\tilde{\mathbf{h}}$ and outputs the edited image aligned with the semantic intention via fine-tuning.

Specifically, DiT predicts a velocity $\boldsymbol{v}_\theta(\mathbf{I}_t,\, t,\, \tilde{\mathbf{h}})$ that is used to iteratively refine the latent image in T uniform steps,

$$\mathbf{I}_{t-\frac{1}{T}} = \mathbf{I}_t \;-\; \frac{1}{T}\, \boldsymbol{v}_\theta(\mathbf{I}_t,\, t,\, \tilde{\mathbf{h}}), \qquad \forall\, t \in \{1/T, 2/T, \ldots, 1\}. \tag{10}$$

At inference time we apply (10) until $t = 0$, yielding the edited image $\hat{\mathbf{I}}_0$.

4.4 Pre-training and Finetuning

We adopt a two-phase process: 1) neural signal encoders (EEG is the most important one) are pretrained on neuro-text corpora, compressing public data and L-Mind, 2) The full stack is optionally fine-tuned with paired original images and ground truth edited images.

Pretraining. Signal encoders are pretrained to align with semantic embeddings using large-scale cognitive datasets [57, 58] and L-Mind. CS3 encoders (EEG + PPG and fNIRS + Motion, respectively) are aligned to frozen text embeddings via symmetric NT-Xent loss:

$$\mathcal{L}_{\text{con}} = \frac{1}{2M} \sum_{i=1}^{M} \Big[-\log \frac{e^{s_{ii}}}{\sum_j e^{s_{ij}}} - \log \frac{e^{s_{ii}}}{\sum_j e^{s_{ji}}} \Big]. \tag{11}$$

where $s_{ij} = (\mathbf{z}_i^\top \mathbf{q}_j)/\tau$, $\mathbf{z}_i$ and $\mathbf{q}_j$ are neural and text embeddings, and M is the number of neural modalities. During pretraining, signal encoders are learned while text encoders stay frozen.

Finetuning. Encoders and the DiT are finetuned jointly on L-Mind, mapping user neural patterns to editing target following a standard diffusion objective that minimizes the mean-squared velocity error. For an input image $\mathbf{I}_0$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\bar{\alpha}_t$ is the cumulated noise schedule:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{t, \mathbf{I}_0, \boldsymbol{\epsilon}} \Big\| \boldsymbol{v}_\theta(\mathbf{I}_t, t, \tilde{\mathbf{h}}) - \underbrace{\big(\sqrt{\bar{\alpha}_t}\, \boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_t}\, \mathbf{I}_0\big)}_{\boldsymbol{v}(\mathbf{I}_t, t, \boldsymbol{\epsilon})} \Big\|_2^2. \tag{12}$$

5 Experiment

To answer the research questions (RQs) asked in Sec. 1, we conduct a comprehensive evaluation to validate the effectiveness of LoongX on the test set of L-Mind. This section first describes the experimental setup, evaluation metrics, and implementation details, and presents results of comprehensive quantitative evaluations, detailed breakdown analyses, and qualitative assessments.

5.1 Experimental Setup

Implementation Details. All models are trained on eight NVIDIA H100 GPUs. Text prompts are embedded by T5-XXL [59] and CLIP [60]; neural signal streams are encoded by the proposed CS3. Unless stated otherwise, EEG montage (Fz, Fp2, O2, Pz, Cz) is sampled at 256 Hz and down-sampled to 32 Hz after band-pass filtering. Inference runs at 8 steps with classifier-free guidance $w = 4$. We choose OminiControl [48] as our baseline as it supports the text-conditioned image-editing based on DiTs. We also implement LoongX using only neural signals (EEG, fNIRS, PPG and Motion) and using both text prompts and neural signals. We load the pretrained weights from FLUX.1-dev[2] and use low-rank approximation (LoRA) for fine-tuning (learning rate 1.0, weight decay 0.01).

Evaluation Metrics. We mainly use five metrics for quantitative assessment following [50]:

1) **L1 Distance (Mean Absolute Error)**: Calculating the average absolute difference between corresponding pixels in edited and ground truth tar images.
2) **L2 Distance (Mean Squared Error)**: Computing the average squared difference between pixels. Penalizes large errors more heavily than L1.
3) **CLIP-I Score**: Evaluating semantic similarity between model-edited images and ground truth target images, which focuses on global semantics of editing results.
4) **DINO Score**: Assessing feature similarity between editing results and ground truth. Compared with CLIP, it is believed that DINO features capture fine-grained structural similarity that correlates with perceived preservation of identity, pose, and local geometry [61].
5) **CLIP-T Score**: Evaluating semantic similarity between image and textual prompts.

5.2 Reliability of Neural Signals

Answer RQ1: Neural signals can serve as reliable indicators to drive image editing, outperforming text-instructed baselines on key metrics.

As shown in Table 1, neural-signal-only LoongX outperforms the text-based OminiControl baseline in semantic discriminability (CLIP-I: 0.6605 vs. 0.6558) and robustness (DINO: 0.4812 vs. 0.4636),

[2] https://huggingface.co/black-forest-labs/FLUX.1-dev

Table 1: Comparison between baseline methods and two LoongX paradigms: (i) neural signals only and (ii) neural signals enhanced by speech. Mean ± 95% confidence interval (CI) over three runs.

Methods	L1 (↓)	L2 (↓)	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)
OminiControl (Text)	0.2632 ±0.006	0.1161 ±0.010	0.6558 ±0.010	0.4636 ±0.017	0.2549 ±0.008
OminiControl (Speech)	0.2714 ±0.006	0.1209 ±0.008	0.6146 ±0.009	0.3717 ±0.013	0.2501 ±0.008
LoongX (Neural Signals)	**0.2509** ±0.006	**0.1029** ±0.009	**0.6605** ±0.009	**0.4812** ±0.015	0.2436 ±0.008
LoongX (Signals+Speech)	0.2594 ±0.006	0.1080 ±0.009	0.6374 ±0.009	0.4205 ±0.014	**0.2588** ±0.009

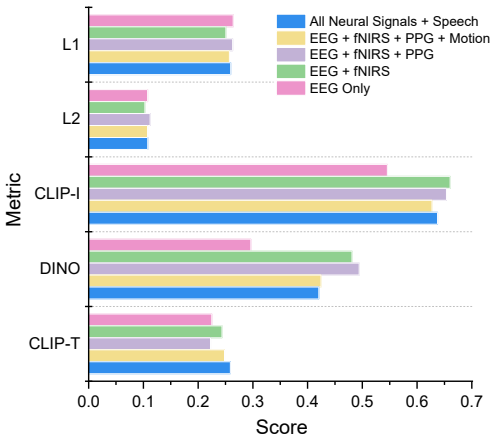

Figure 4: Evaluation of different signal combinations on the proposed DGF module.

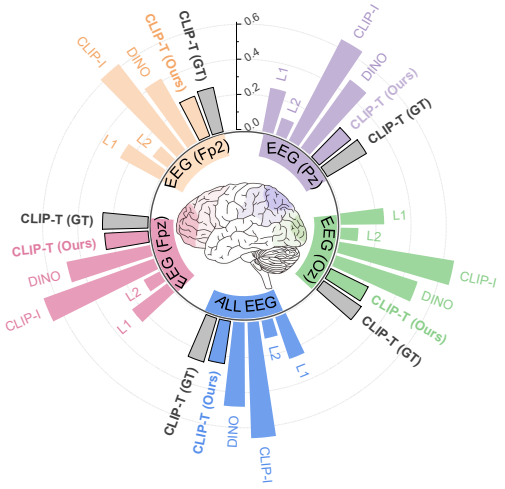

Figure 5: Evaluation results on different brain region signals where LoongX is trained and tested on each respective EEG channel.

which demonstrates the potential of neural signals as a standalone modality that carries rich semantic information for image editing. The slightly higher L1 and L2 errors indicate better preservation of semantic fidelity over pixel-level accuracy. Combining speech cues with neural signals boosts semantic alignment, reaching the highest CLIP-T score of 0.2588 and demonstrating their joint effectiveness in capturing nuanced user intentions for hands-free image editing.

5.3 Ablation Studies on Modality Contribution

Answer RQ2: Different neural signal modalities contribute complementary strengths, enhancing discriminability, robustness, and semantic precision, respectively.

Modality contributions are compared in Fig. 4. EEG signals alone enable basic high-level semantic editing, supported by the semantic discriminability of the extracted features (CLIP-I: 0.5457). Integrating fNIRS significantly improves feature robustness (DINO: from 0.2963 to 0.4811), highlighting the complementary nature of hemodynamic responses in enhancing signal completeness and structural fidelity. Including PPG and Motion improves global physiological awareness and indicates sensitivity to subtle engagement patterns (e.g., heart rate and user movements) that express editing intent. They both contribute to the features' robustness and completeness to ensure stable CLIP-T score gains.

We show the contribution of each individual EEG channel in Fig. 5, where each channel corresponds to a specific scalp region as detailed in Table 3. The occipital cortex channel (Oz), which is in charge of visual processing, emerges as dominant in global editing effect (CLIP-I: 0.6619) and robustness (DINO: 0.4873) to finer details, affirming its critical role in basic visual perception and processing tasks. Conversely, the frontopolar cortex (Fpz) provides superior semantic alignment (CLIP-T: 0.2481), consistent with its association with more complex cognitive processes. Specifically, Fpz provides decision control and attention regulation compared with basic visual perception provided by Oz, which precisely confirms the discovery patterns in medical anatomy. This channel-specific analysis provides insights valuable for targeted applications or constrained hardware settings.

5.4 Breakdown Analysis: Neural vs. Language-based Conditions

Answer RQ3: Neural signals excel in low-level visual edits, while language excels in high-level semantics; combining both yields comprehensive and optimal control.

The analysis of text and neural-driven image editing is shown in Fig. 6. Using pure neural signals (N) is particularly effective for global texture editing, with higher CLIP-I scores highlighting their strong visual and structural consistency. Neural signals also outperform other modalities in several tasks like object removal and background blur, reflecting their strength in conveying intuitive intent, though they remain limited in handling complex semantics like text editing. Text instructions (T) are inherently stronger in high-level semantic tasks (e.g., image restoration), which indicates their advantage in describing instruction details. Combined neural and speech (N+S) signals achieve the highest semantic alignment (CLIP-T: 0.2588), showcasing the superior effectiveness of hybrid conditioning in capturing complex user intentions. Overall, neural signals are more effective for low-level visual edits, while neural and text-based approaches each provide complementary advantages.

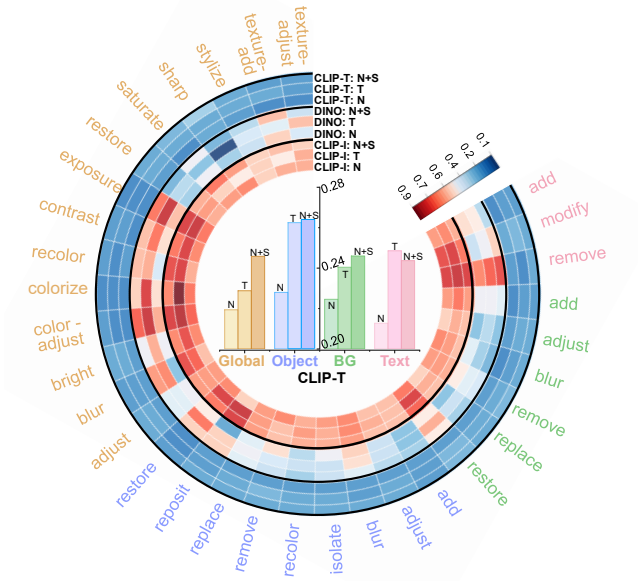

Figure 6: Breakdown results of text and neural-driven image editing. BG: background.

5.5 Ablation Studies on Model Architecture

Each component in the LoongX architecture contributes uniquely, and their composition (especially with pretraining) maximizes the performance potential, leading to improved editing accuracy across metrics. The ablation study in Table 2 is conducted on the fused neural signals and speech to evaluate the impact of each proposed module. It is found that CS3 encoder enhances feature completeness and smoothness, leading to a 5% reduction in L2 error. The DGF module improves semantic alignment with textual instructions, yielding a 3.5% increase in CLIP-T. Supplemented by pre-training, LoongX reaches optimal performance, indicating the important role of robust and representation learning and multimodal alignment in maximizing editing performance.

Table 2: Ablation studies on the architecture of LoongX. Mean ± 95% CI over three runs.

Pretrain	CS3	DGF	L1 (↓)	L2 (↓)	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)
			0.2645 ±0.005	0.1099 ±0.003	0.5966 ±0.009	0.3948 ±0.011	0.2584 ±0.010
	✓		**0.2567** ±0.006	**0.1047** ±0.004	**0.6408** ±0.010	**0.4588** ±0.012	0.2248 ±0.007
		✓	0.2629 ±0.005	0.1106 ±0.003	0.6025 ±0.009	0.3992 ±0.011	**0.2620** ±0.007
	✓	✓	0.2648 ±0.006	0.1124 ±0.004	0.6319 ±0.009	0.4162 ±0.012	0.2534 ±0.008
✓	✓	✓	**0.2594** ±0.006	**0.1080** ±0.004	**0.6374** ±0.009	0.4205 ±0.012	**0.2588** ±0.009

5.6 Qualitative Analysis and Discussion

Qualitative examples confirm LoongX's intuitive editing capabilities, though performance may vary slightly in scenarios with complex or ambiguous intentions. Qualitative results presented in Fig. 7 demonstrate that neural-driven LoongX can successfully achieve various visual and structural modifications, such as background replacement and global adjustments. However, the fused neural-language method better captures nuanced instructions involving abstract semantics (e.g., "modify the text information" in Fig. 7(d)).

Limitations and future work. It is noted that while the neural signals and combined methods perform better in multiple tasks such as object manipulation (e.g., letting the cat look down in Fig. 7(a)), the text-based method handles spatial manipulation more effectively (e.g., "place the cat

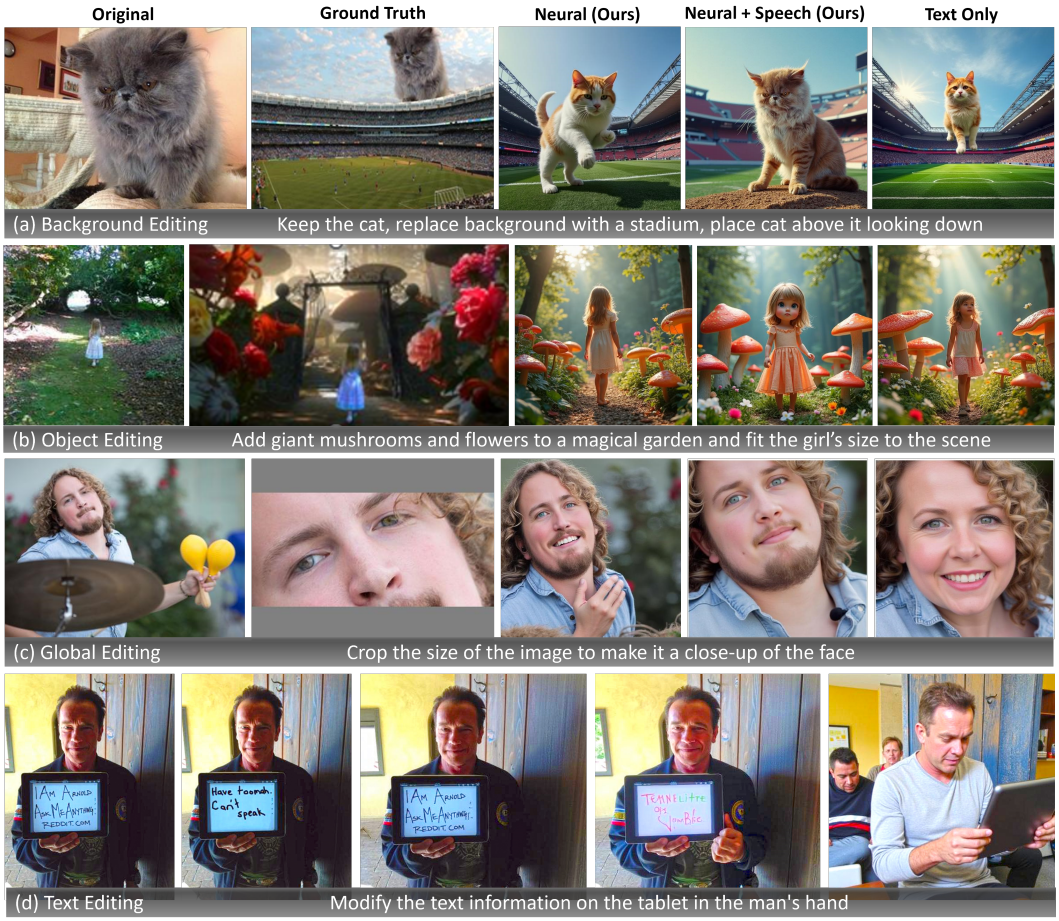

Figure 7: Qualitative evaluation comparing the text prompt-based method and our neural-driven methods for four editing categories: (a) background, (b) object, (c) global, and (d) text editing.

above" in Fig. 7(a)). Despite significant advancements with multimodal fusion, entity consistency (e.g., the style of the little girl in Fig. 7(b)) remains a challenge, which is limited by the capabilities of backbone image editing models at the time this work was mainly completed. Moreover, highly abstract or ambiguous instructions occasionally still pose challenges (e.g., "winged white animal in Fig. 11(f)"). Several more failure cases are shown in Fig. 14, indicating areas where further refinement in entity interpretation and neural-data-based disambiguation remains necessary.

Overall, the experiments validate LoongX's efficacy as a robust, intuitive interface leveraging neural signals for image editing. Multimodal integration, particularly neural and linguistic fusion, emerges as essential for capturing comprehensive user intent. These findings advocate for future research focusing on improving semantic interpretation fidelity from neural signals and exploring adaptive methodologies for further enhancing accessibility and precision in cognitive-driven creative technologies. Incremental fine-tuning for new unseen users is also worth exploration in practice.

6 Conclusion

We presented LoongX, a novel framework for hands-free image editing by conditioning diffusion models on multimodal neural signals, achieving performance comparable to or superior to traditional text-driven baselines. Looking ahead, the portability of our wireless setup opens exciting possibilities for real-world applications in immersive virtual environments. Future directions include integrating LoongX with VR/XR to support intuitive cognitive interaction and aligning neural representations with emerging world models [62, 63] to project human intention into an interactive synthetic world, enabling mind-driven control in virtual realities.

Acknowledgments and Disclosure of Funding

We sincerely thank Prof. Tat-Seng Chua (National University of Singapore) for his valuable comments and suggestions during the drafting of this work. We also thank Dr. Zhiwei Tang (Alibaba) for his constructive guidance during the rebuttal. This work is sponsored by the NUS Startup Grant (Presidential Young Professorship), Singapore MOE Tier-1 Grant, ByteDance Grant, NUS ARTIC Grant, Apple Grant, Alibaba Grant, and Google Grant for TPU usage, and also supported by the STI2030-Major Projects (No. 2021ZD0200401), Zhejiang Province Key R&D Programs (No. 2024C03001, 2024C03007, 2024C01031, 2025C01137, 2025C02165), Zhejiang Province High-Level Talent Special Support Plan (No. 2022R52042), and the Fundamental Research Funds for the Central Universities (No. 2025ZFJH01).

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[2] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

[3] Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Diffusion-based visual art creation: A survey and new perspectives. *ACM Computing Surveys*, 57(10):1–37, 2024.

[4] OpenAI. Introducing 4o image generation, 2025.

[5] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

[6] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.

[7] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, pages 4055–4075, 2023.

[8] Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5951–5961, 2022.

[9] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024.

[10] Joonghyuk Shin, Daehyeon Choi, and Jaesik Park. InstantDrag: Improving interactivity in drag-based image editing. In *SIGGRAPH Asia*, pages 1–10, 2024.

[11] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. *arXiv preprint arXiv:2411.09703*, 2024.

[12] Ziqi Jiang, Zhen Wang, and Long Chen. CLIPDrag: Combining text-based and drag-based instructions for image editing. In *International Conference on Learning Representations*, 2025.

[13] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

[14] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021.

[15] Yue Yang, Kaipeng Zhang, Yuying Ge, Wenqi Shao, Zeyue Xue, Yu Qiao, and Ping Luo. Align, adapt and inject: Audio-guided image generation, editing and stylization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3475–3479, 2024.

[16] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024.

[17] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

[18] Yiming Wang, Bin Zhang, and Lamei Di. Research progress of EEG-based emotion recognition: a survey. *ACM Computing Surveys*, 56(11):1–49, 2024.

[19] Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for EEG to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918, 2023.

[20] Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. In *International Conference on Learning Representations*, 2024.

[21] Dünya Baradari, Nataliya Kosmyna, Oscar Petrov, Rebecah Kaplun, and Pattie Maes. NeuroChat: A neuroadaptive AI chatbot for customizing learning experiences. *arXiv preprint arXiv:2503.07599*, 2025.

[22] 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*, pages 472–488, 2024.

[23] 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.

[24] Keith M. Davis, III, Carlos de la Torre-Ortiz, and Tuukka Ruotsalo. Brain-supervised image editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18480–18489, 2022.

[25] Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pages 56–73, 2024.

[26] Jing Cai, Alex E. Hadjinicolaou, Angelique C. Paulk, Daniel J. Soper, Tian Xia, Alexander F. Wang, John D. Rolston, R. Mark Richardson, Ziv M. Williams, and Sydney S. Cash. Natural language processing models reveal neural dynamics of human conversation. *Nature Communications*, 16(1):3376, 2025.

[27] Niklas Brake, Flavie Duc, Alexander Rokos, Francis Arseneau, Shiva Shahiri, Anmar Khadra, and Gilles Plourde. A neurophysiological basis for aperiodic eeg and the background spectral trend. *Nature Communications*, 15(1):1514, 2024.

[28] Yuxuan Li, Jesse K. Pazdera, and Michael J. Kahana. EEG decoders track memory dynamics. *Nature Communications*, 15(1):2981, 2024.

[29] Yipeng Yu, Cunle Qian, Zhaohui Wu, and Gang Pan. Mind-controlled ratbot: A brain-to-brain system. In *IEEE International Conference on Pervasive Computing and Communication Workshops*, pages 228–231, 2014.

[30] Xuyang Si, Hao He, Jun Yu, and Dong Ming. Cross-subject emotion recognition brain-computer interface based on fNIRS and DBJNet. *Cyborg and Bionic Systems*, 4:0045, 2023.

[31] Muhammad Arsalan and Muhammad Majid. Human stress classification during public speaking using physiological signals. *Computers in Biology and Medicine*, 133:104377, 2021.

[32] Amir Tazarv, Shervin Labbaf, Steven M. Reich, Nikil Dutt, Amir M. Rahmani, and Marco Levorato. Personalized stress monitoring using wearable sensors in everyday settings. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 7332–7335, 2021.

[33] Lu Cao, Dandan Huang, Yue Zhang, Xiaowei Jiang, and Yanan Chen. Brain decoding using fNIRS. In *AAAI Conference on Artificial Intelligence*, pages 12602–12611, 2021.

[34] Chang-Hee Han, Euijin Kim, and Chang-Hwan Im. Development of a brain–computer interface toggle switch with low false-positive rate using respiration-modulated photoplethysmography. *Sensors*, 20(2):348, 2020.

[35] Boyu Li, Mingjie Li, Jie Xia, Hao Jin, Shurong Dong, and Jikui Luo. Hybrid integrated wearable patch for brain EEG-fNIRS monitoring. *Sensors*, 24(15):4847, 2024.

[36] Bradley J. Edelman, Shuailei Zhang, Gerwin Schalk, Peter Brunner, Gernot Müller-Putz, Cuntai Guan, and Bin He. Non-invasive brain-computer interfaces: State of the art and trends. *IEEE Reviews in Biomedical Engineering*, 18:26–49, 2025.

[37] Seif Eldawlatly. On the role of generative artificial intelligence in the development of brain–computer interfaces. *BMC Biomedical Engineering*, 6:4, 2024.

[38] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[39] Kalyan Tripathy, Zachary E Markow, Andrew K Fishell, Arefeh Sherafati, Tracy M Burns-Yocum, Mariel L Schroeder, Alexandra M Svoboda, Adam T Eggebrecht, Mark A Anastasio, Bradley L Schlaggar, et al. Decoding visual information from high-density diffuse optical tomography neuroimaging data. *NeuroImage*, 226:117516, 2021.

[40] 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 *AAAI Conference on Artificial Intelligence*, pages 6935–6943, 2024.

[41] Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 2023.

[42] Paul S. Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *International Conference on Machine Learning*, pages 44038–44059, 2024.

[43] 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*, 2024.

[44] Michel Adamic, Wellington Avelino, Anna Brandenberger, Bryan Chiang, Hunter Davis, Stephen Fay, Andrew Gregory, Aayush Gupta, Raphael Hotter, Grace Jiang, et al. Progress towards decoding visual imagery via fNIRS. *arXiv preprint arXiv:2406.07662*, 2024.

[45] OpenAI. GPT-4 technical report, 2024.

[46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[47] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.

[48] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

[49] Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. OminiControl2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*, 2025.

[50] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[51] Xueyun Tian, Wei Li, Bingbing Xu, Yige Yuan, Yuanzhuo Wang, and Huawei Shen. MIGE: A unified framework for multimodal instruction-based image generation and editing. *arXiv preprint arXiv:2502.21291*, 2025.

[52] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. ACE++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025.

[53] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. SEED-Data-Edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024.

[54] Tristan Stenner, Chadwick Boulay, Matthew Grivich, David Medine, Christian Kothe, Tobias Herzke, Christian Schausner, Giso Grimm, X Loem, Arthur Biancarelli, Boris Mansencal, Paul Maanen, Jérémy Frey, Jidong Chen, Kyle Crane, Samuel Powell, Pierre Clisson, and Paul Fix. Sccn/liblsl: v1.16.2, 2023.

[55] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[56] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[57] Nicolás Nieto, Victoria Peterson, Hugo Leonardo Rufiner, Juan Esteban Kamienkowski, and Ruben Spies. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Scientific Data*, 9(1):52, 2022.

[58] Matthew Ning, Sudan Duwadi, Meryem A Yücel, Alexander Von Lühmann, David A Boas, and Kamal Sen. fnirs dataset during complex scene analysis. *Frontiers in Human Neuroscience*, 18:1329086, 2024.

[59] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[61] Yang Zhou, Zichong Chen, and Hui Huang. Deformable one-shot face stylization via dino semantic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7787–7796, 2024.

[62] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

[63] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1559–1569, 2025.

[64] Peter Praamstra, Luc Boutsen, and Glyn W. Humphreys. Frontoparietal control of spatial attention and motor intention in human eeg. *Journal of Neurophysiology*, 94(1):764–774, 2005. PMID: 15744008.

[65] Ankita Sengupta, Sanjna Banerjee, Suhas Ganesh, Shrey Grover, and Devarajan Sridharan. The right posterior parietal cortex mediates spatial reorienting of attentional choice bias. *Nature Communications*, 15(1):6938, 2024.

[66] Naomi P. Friedman and Trevor W. Robbins. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, 47(1):72–89, 2022.

[67] Jae-Hyun Kim, Dong-Hyun Ma, Eunji Jung, Ilsong Choi, and Seung-Hee Lee. Gated feed-forward inhibition in the frontal cortex releases goal-directed action. *Nature Neuroscience*, 24(10):1452–1464, 2021.

[68] Bülent Y. Özkara. An introduction to the event-related potential technique. *Journal of Consumer and Consumption Research*, 11(2):437–441, 2019.

[69] Kavya Agrawal, Shashwat Mishra, Shashwat Sinha, Vineeta Khemchandani, Sushil Chandra, and Nachiket Milind Wadalkar. Chapter 29 - machine learning-based workload identification using functional near-infrared spectroscopy (fnirs) data. In M.A. Ansari, R.S. Anand, Pragati Tripathi, Rajat Mehrotra, and Md Belal Bin Heyat, editors, *Artificial Intelligence in Biomedical and Modern Healthcare Informatics*, pages 299–312. Academic Press, 2025.

[70] Syed Hammad Nazeer, Noman Naseer, Muhammad Jawad Khan, and Keum-Shik Hong. Noninvasive brain–computer interfaces using fNIRS, EEG, and hybrid EEG-fNIRS. In Ayman S. El-Baz and Jasjit S. Suri, editors, *Brain-Computer Interfaces*, pages 297–326. Elsevier, 2025.

[71] Zhihan Lv, Liang Qiao, Qingjun Wang, and Francesco Piccialli. Advanced machine-learning methods for brain-computer interfacing. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(5):1688–1698, 2020.

[72] Xiang Zhang, Lina Yao, Shuai Zhang, Salil Kanhere, Michael Sheng, and Yunhao Liu. Internet of things meets brain–computer interface: A unified deep learning framework for enabling human-thing cognitive interactivity. *IEEE Internet of Things Journal*, 6(2):2084–2092, 2018.

[73] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, June 2023.

[74] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. Generative adversarial networks conditioned by brain signals. In *IEEE International Conference on Computer Vision*, pages 3430–3438, 2017.

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

[76] Yasheng Sun, Bohan Li, Mingchen Zhuge, Deng-Ping Fan, Salman Khan, Fahad Shahbaz Khan, and Hideki Koike. Connecting dreams with visual brainstorming instruction. *Visual Intelligence*, 3(1):1–18, 2025.

[77] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[78] 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.

[79] Xinjin Li, Yu Ma, Kaisen Ye, Jinghan Cao, Minghao Zhou, and Yeyang Zhou. Hy-facial: Hybrid feature extraction by dimensionality reduction methods for enhanced facial expression classification, 2025.

[80] Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Weijie Wang, Haoyun Li, Guosheng Zhao, Jie Li, Wenkang Qin, Guan Huang, and Wenjun Mei. Wonderturbo: Generating interactive 3d world in 0.72 seconds. *arXiv preprint arXiv:2504.02261*, 2025.

[81] Yi Xin, Le Zhuo, Qi Qin, Siqi Luo, Yuewen Cao, Bin Fu, Yangfan He, Hongsheng Li, Guangtao Zhai, Xiaohong Liu, et al. Resurrect mask autoregressive modeling for efficient and scalable image generation. *arXiv preprint arXiv:2507.13032*, 2025.

[82] Jiuming Liu, Zheng Huang, Mengmeng Liu, Tianchen Deng, Francesco Nex, Hao Cheng, and Hesheng Wang. Topolidm: Topology-aware lidar diffusion models for interpretable and realistic lidar point cloud generation. *arXiv preprint arXiv:2507.22454*, 2025.

[83] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. *ICLR*, 2024.

[84] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025.

[85] Rongchao Zhang, Yu Huang, Yiwei Lou, Yi Xin, Haixu Chen, Yongzhi Cao, and Hanpin Wang. Exploit your latents: Coarse-grained protein backmapping with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1111–1119, 2025.

[86] Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[87] Zhen Xu, Kunyuan Ma, Yunbo Liu, Wenying Sun, and Youzhu Liu. Causal representation learning for robust anomaly detection in complex environments. 2025.

[88] Sibei Liu, Yuanzhe Zhang, Xiang Li, Yunbo Liu, Chengwei Feng, and Hao Yang. Gated multimodal graph learning for personalized recommendation. *INNO-PRESS: Journal of Emerging Applied AI*, 1(1), 2025.

[89] Yuyuan Li, Yizhao Zhang, Weiming Liu, Xiaohua Feng, Zhongxuan Han, Chaochao Chen, and Chenggang Yan. Multi-objective unlearning in recommender systems via preference guided pareto exploration. *IEEE Transactions on Services Computing*, 2025.

[90] Yiming Zeng, Wanhao Yu, Zexin Li, Tao Ren, Yu Ma, Jinghan Cao, Xiyan Chen, and Tingting Yu. Bridging the editing gap in llms: Fineedit for precise and targeted text modifications, 2025.

[91] Xiaoling Zhou, Wei Ye, Zhemg Lee, Lei Zou, and Shikun Zhang. Valuing training data via causal inference for in-context learning. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[92] Wangbo Zhao, Jiasheng Tang, Yizeng Han, Yibing Song, Kai Wang, Gao Huang, Fan Wang, and Yang You. Dynamic tuning towards parameter and inference efficiency for vit adaptation. *NeurIPS*, 37:114765–114796, 2024.

[93] Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*, 2025.

[94] Xinjin Li, Yu Ma, Yangchen Huang, Xingqi Wang, Yuzhen Lin, and Chenxi Zhang. Synergized data efficiency and compression (sec) optimization for large language models. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pages 586–591, 2024.

[95] Zhenjun Zhao. Balf: Simple and efficient blur aware local feature detector. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3362–3372, 2024.

[96] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Zhikai Li, Yibing Song, Kai Wang, Zhangyang Wang, and Yang You. A stitch in time saves nine: Small vlm is a precise guidance for accelerating large vlms. In *CVPR*, pages 19814–19824, 2025.

[97] Yujia Wang, Fang-Lue Zhang, and Neil A Dodgson. Target scanpath-guided 360-degree image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8169–8177, 2025.

[98] Chaocan Xue, Bineng Zhong, Qihua Liang, Haiying Xia, and Shuxiang Song. Unifying motion and appearance cues for visual tracking via shared queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[99] Yi Xin, Juncheng Yan, Qi Qin, Zhen Li, Dongyang Liu, Shicheng Li, Victor Shea-Jay Huang, Yupeng Zhou, Renrui Zhang, Le Zhuo, et al. Lumina-mgpt 2.0: Stand-alone autoregressive image modeling. *arXiv preprint arXiv:2507.17801*, 2025.

[100] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

[101] Jiuming Liu, Guangming Wang, Weicai Ye, Chaokang Jiang, Jinru Han, Zhe Liu, Guofeng Zhang, Dalong Du, and Hesheng Wang. Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15109–15119, 2024.

[102] Rongchao Zhang, Yu Huang, Yiwei Lou, Weiping Ding, Yongzhi Cao, and Hanpin Wang. Synergistic attention-guided cascaded graph diffusion model for complementarity determining region synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[103] Chaojun Ni, Jie Li, Haoyun Li, Hengyu Liu, Xiaofeng Wang, Zheng Zhu, Guosheng Zhao, Boyuan Wang, Chenxin Li, Guan Huang, et al. Wonderfree: Enhancing novel view quality and cross-view consistency for 3d scene exploration. *arXiv preprint arXiv:2506.20590*, 2025.

[104] Chaocan Xue, Bineng Zhong, Qihua Liang, Yaozong Zheng, Ning Li, Yuanliang Xue, and Shuxiang Song. Similarity-guided layer-adaptive vision transformer for uav tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6730–6740, 2025.

[105] Zhangquan Chen, Xufang Luo, and Dongsheng Li. Visrl: Intention-driven visual perception via reinforced reasoning. *arXiv preprint arXiv:2503.07523*, 2025.

[106] Nasim Raufi and Leonardo Longo. Evaluating eeg alpha-to-theta and theta-to-alpha band ratios as indexes of mental workload. *arXiv preprint arXiv:2202.12937*, 2022.

[107] Malek Adjouadi, Mercedes Cabrerizo, Ilker Yaylali, and Prasana Jayakar. Interpreting eeg functional brain activity. *IEEE Potentials*, 23(1):8–13, 2004.

[108] Isabelle Constant and Nada Sabourdin. The eeg signal: a window on the cortical brain activity. *Pediatric Anesthesia*, 22(6):539–552, 2012.

[109] Claudio Babiloni, Raffaele Ferri, Davide V Moretti, Andrea Strambi, Giuliano Binetti, Gloria Dal Forno, Florinda Ferreri, Bartolo Lanuzza, Claudio Bonato, Flavio Nobili, et al. Abnormal fronto-parietal coupling of brain rhythms in mild alzheimer's disease: A multicentric eeg study. *European Journal of Neuroscience*, 19(9):2583–2590, 2004.

[110] Jane E Adcock and Chrysostomos P Panayiotopoulos. Occipital lobe seizures and epilepsies. *Journal of clinical neurophysiology*, 29(5):397–407, 2012.

[111] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.

[112] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[113] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024.

[114] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8861–8870, 2024.

A Supplementary Key Information

A.1 Preliminary: Analysis of Brain Region Function and Mechanism

We employ a noninvasive multimodal sensing system that synchronously records neural, hemodynamic, and peripheral vascular signals, including EEG, fNIRS, and PPG (Fig. 8(a)). This setup enables comprehensive monitoring of brain activity during human–computer interaction. fNIRS employs near-infrared light in the 690–930 nm range to penetrate scalp and skull tissues. Neural activation leads to increased local cerebral blood flow and metabolic demand, resulting in a rise in oxygenated hemoglobin (HbO) and a reduction in deoxygenated hemoglobin (HbR). These changes cause detectable variations in light absorption at the detector. fNIRS thus provides second-level sensitivity to the slow hemodynamic responses associated with cortical processing. EEG is recorded via hydrogel-based electrodes placed over the scalp, capturing millisecond-scale voltage fluctuations resulting from synchronized postsynaptic potentials in cortical pyramidal neurons. This modality offers high temporal precision for observing rapid fluctuations in cortical excitability and sensorimotor responses. PPG detects volumetric changes in blood flow via near-infrared light, enabling continuous measurement of pulse rate and vascular compliance. The sensor is co-located with fNIRS optodes, sharing similar optical pathways but tuned for peripheral cardiovascular features.

This multimodal approach provides a synergistic view of neural function, combining the high temporal resolution of EEG with the metabolic and vascular insights from fNIRS and PPG. Such cross-modal sensing is crucial for modeling the perception-decision-action cycle in neuroadaptive systems.

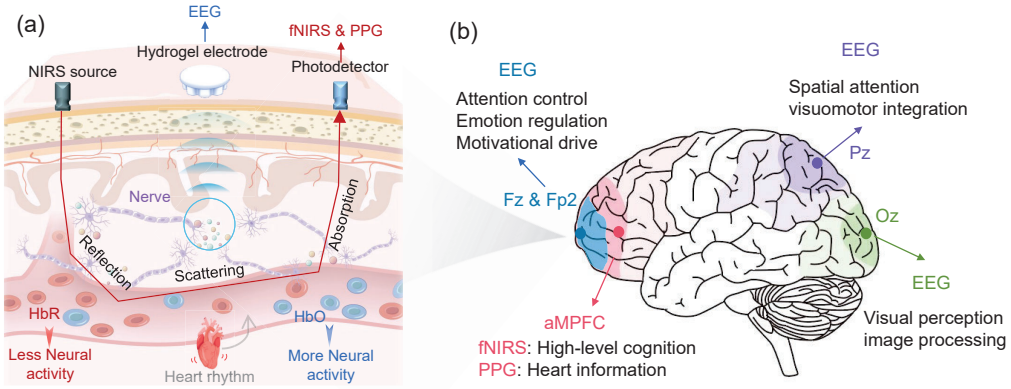

Figure 8: (a) Signal acquisition mechanism. (b) Cognitive function of different brain regions.

Table 3: Functional Roles of Multimodal Neural Signals in Hands-Free Image Editing.

Signal (Channel)	Cortical Region	Primary Function	Roles in Image Editing
EEG Ch 0 (Pz)	Parietal cortex	Spatial attention, visuomotor integration	Focuses on specific image areas, targets object localization [64]
EEG Ch 1 (Fp2)	Prefrontal cortex	Emotion regulation, motivational drive	Generates and regulates intentional editing actions [65]
EEG Ch 2 (Fpz)	Frontopolar cortex	Attention control, task initiation	Triggers editing intention, starts/stops editing operations [66, 67]
EEG Ch 3 (Oz)	Occipital cortex	Visual perception, image processing	Perceives visual changes, evaluates whether edits meet expectations [68]
fNIRS (aMPFC)	Anterior medial prefrontal cortex	High-level cognition, motivation, emotional valence	Indicates editing intent intensity, emotional confidence, and mental workload [69]
PPG (aMPFC)	Cardiovascular or autonomic system	Heart rate variability, arousal, stress	Monitors cognitive stress or emotional arousal during editing [31, 32]

To enable hands-free image editing, we integrate multimodal neural and physiological signals to decode user intent and cognitive state in real time. The functional roles of multimodal neural signals are provided in Table.3. Specifically, midline parietal EEG (Pz) reflects spatial attention and visuomotor integration, supporting the allocation of attention to target areas and coordination of motor plans during editing tasks [64]. Right prefrontal EEG (Fp2) is linked to emotional regulation and motivational drive, aligning with findings from frontal alpha asymmetry studies [65]. Frontopolar EEG (Fz/Fpz) tracks attention control and task initiation, facilitating the onset and modulation of editing operations [66, 67]. Occipital EEG (Oz) encodes visual perception and image processing load, enabling evaluation of visual changes and edit quality [68]. fNIRS over anterior medial prefrontal cortex (aMPFC) measures cognitive load, motivation, and emotional valence through hemodynamic activity [69]. PPG signals from the same region capture heart rate variability and autonomic arousal, which reflect user stress and engagement [31, 32]. This multimodal mapping (Fig. 8(b)) enables the system to continuously adapt to users' cognitive focus, affective state, and mental workload, thereby supporting a more responsive and intuitive editing experience.

A.2 Preliminary Analysis: Editing Type Classification Experiment

LoongX is proposed to address three key research questions:

1. Can neural signals serve as reliable conditions for image editing? (Does it really work?)
2. What kind of information is conveyed by multimodal neural signals? (What do they actually contribute to image editing?)
3. How do neural-based and language-based conditions differ in image editing? Can we combine their strengths to enable hands-free editing more effectively?

In response to these problems, we conduct a premise exploration based on a classification experiment and design the LoongX model based on the findings. Finally, we present a modular architecture comprising unified multimodal encoding, dynamic multimodal data fusion, and diffusion-based conditional generation. Based on these, LoongX can perform robust hands-free image editing by translating user neural states into structured conditions for a diffusion model.

To examine whether neural signals can reliably encode semantic conditions for image editing, we perform an exploratory classification experiment where the task is to predict editing types from neural signals or text. We use the 22,691 training instances and 1,200 test instances in L-Mind. As each editing instance can involve multiple editing types, we implement a simple multilayer perception (MLP) with three nonlinear-activated linear layers and conduct a multi-label classification experiment that recognizes all involved editing type labels for an instance, via text embeddings or neural signals, as a condition. We compare models trained with random noise, only text prompts, EEG, fNIRS, PPG, combinations of multimodal neural signals, and fusion of both text and signals.

As shown in Fig. 9, the classification results validate the informativeness and complementarity of neural signals for image editing conditions. Fig. 9(a) shows that using neural signals, especially the EEG signal itself, can achieve significantly better classification performance than random noise (as seen from over 7% mAP improvement). fNIRS contributes to recall performance gain compared with random noise since it provides more complete and robust information, which is more important for recall. As image editing requires discriminant semantic features, it shows that neural networks can still effectively recognize the editing types (over 60% precision) based on EEG signals and can even achieve comparable performance with text prompts in some cases. Results can also initially respond to question 2. As precision depends on the discriminability of features, EEG contributes more discriminability with less noise. For recall, robustness and completeness contribute more. Therefore, fNIRS becomes more stable with less fluctuation. Though PPG and motion do not affect classification performance significantly as their volatility is small, they are expected to provide richer background information to stabilize performance. Now, what will the performance change if we combine these signals?

Fig. 9(b) shows that binding all neural signals (via simple concat and MLP) achieves the best performance, and fusing only EEG and fNIRS can achieve comparable performance with only a recall reduction. Moreover, though neural signals are not more discriminant and robust than text embeddings, fusing both can achieve even better performance than text only. That is, the two types of data can complement the missing key information to achieve better category recognition in image

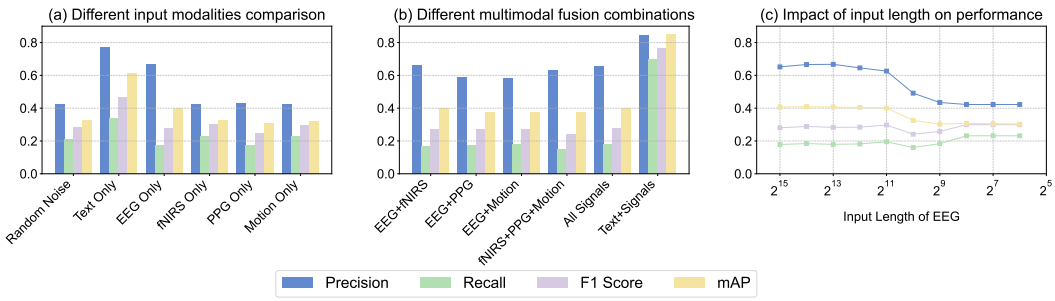

Figure 9: Multi-label classification result under different settings: (a) different single input modalities; (b) different modality fusion combinations; (c) different EEG input sequence lengths.

editing tasks, which shows higher performance as a more powerful condition for further generative models.

As the neural signals all have different shapes, unifying the input sequence into a unified length is necessary. Simple padding and truncating to a fixed length is one of the most reliable methods for preserving most information in raw signals. We also investigate the influence of input sequence length for EEG as a neural signal representative. Results in Fig. 9(c) show that truncating and padding EEG signal sequences to a unified 8,192 can achieve the best performance. While the longer sequence ensures a more reliable performance, the computation cost will become a burden as the sequence length increases. A trade-off method is needed to encode the most valid information in signals while not bringing unbearable computational costs. Therefore, we design the Cross-Scale State Space (CS3) model to ensure the best trade-off between performance and computational cost.

A.3 More Limitations Discussion

While LoongX demonstrates strong performance in neural-driven image editing, several limitations remain. First, the current dataset was collected from a relatively homogeneous group of 12 healthy young adults. Although the model performs well within this cohort, generalization to broader populations (e.g., different age groups or individuals with neurological conditions) is not yet fully validated. Moreover, the neural signals were acquired using low-density EEG and fNIRS systems, which, despite their practicality and portability, offer limited spatial resolution compared to high-density or invasive setups. This constraint may affect the system's ability to capture fine-grained neural representations.

The robustness of LoongX under varying data distributions and noisy conditions has not been systematically explored. Although some resilience is expected from the multimodal design and the DUAF fusion strategy, comprehensive stress testing against motion artifacts, sensor dropout, or environmental interference is a necessary next step for real-world deployment. Furthermore, while the BCI system we use is designed to work across participants, it may benefit from test-time adaptation or few-shot user-specific finetuning to account for individual variability in neural signatures, which can differ significantly across users.

Finally, the interpretability of the learned neural representations remains limited. While the CS3 encoder effectively distills signal patterns for downstream editing, it is not yet clear how these latent features relate to interpretable cognitive states or intentions. Improving the transparency and explanatory power of the system will be essential for broader acceptance and responsible deployment, especially in domains involving sensitive user data.

A.4 Broader Societal Impacts

The proposed neural-driven image editing technique has the potential for significant positive societal impact by improving accessibility for individuals with motor or communication impairments, enabling more inclusive creative workflows, and lowering the barrier to content creation. By enabling hands-free and intuitive interaction with generative models, this technology could foster greater participation in digital art, design, and communication, particularly for users with physical limitations.

However, as with other generative and brain-computer interface technologies, there are potential negative societal impacts. Malicious or unintended uses may include the creation of manipulated or deceptive media, privacy violations stemming from the misuse of neurophysiological data, or unauthorized surveillance. There are also concerns regarding fairness, as disparities in device availability or neural signal quality across user populations could exacerbate existing inequities.

To mitigate these risks, we recommend the implementation of safeguards such as user authentication, audit trails for sensitive editing actions, and transparent communication regarding data usage and model limitations. Responsible deployment should include ongoing monitoring and mechanisms to address feedback and misuse.

A.5 Safety and Ethics

All participants signed consent forms before the experiment, which was approved by the Institutional Review Board's Ethics Committee for Human Research Protections at Zhejiang University (Protocol ID: No. 067 (2019)).

Recognizing the potential for misuse of proposed models and multimodal neural datasets, we are committed to responsible release practices. We will require users requesting access to the models or data to undergo an application and review process, ensuring alignment with ethical guidelines and legitimate research or clinical objectives. Usage agreements will prohibit malicious activities, and access may be revoked in the event of violations.

For dataset release, we apply stringent filtering to remove personally identifiable, sensitive, or potentially harmful content. For model access, we will implement usage restrictions, safety filters, and ongoing monitoring to prevent abuse. Additionally, clear documentation outlining acceptable use cases, known limitations, and recommended best practices will be provided to all users.

We will continuously evaluate the impact and usage of the released resources and remain open to community feedback to improve the effectiveness of these safeguards over time.

B Supplementary Literature Review

B.1 Brain Computer Interface

BCI is an emerging technology that establishes a direct communication pathway between the brain and external devices by interpreting neural signals. BCIs have broad applications in healthcare, communication, gaming, and assistive technologies [36, 37]. The primary goal is to help individuals, especially those with motor impairments, interact with digital systems without the constraints of traditional input methods.

Non-invasive BCIs, such as EEG and fNIRS, are widely studied for their safety and real-time signal acquisition capabilities [70]. EEG captures the brain's electrical activity and is particularly effective in detecting cognitive states such as attention and intention. While fMRI provides high-resolution whole-brain imaging, its use is constrained by high cost and limited mobility. In contrast, fNIRS measures hemodynamic responses and offers a portable, cost-effective solution for monitoring brain activity in real-world environments (as illustrated in Fig. 2 and Fig. 8).

With the rapid development of artificial intelligence, especially deep learning, BCI systems have seen significant improvements in decoding accuracy and robustness [71, 72]. Advanced models enable tasks such as mental spelling, prosthetic control, and gaming interfaces. Recent trends include integrating BCIs with generative models such as diffusion networks to synthesize visual content directly from brain activity, as demonstrated in work applying stable diffusion to fMRI decoding [73]. The integration of EEG and fNIRS holds great promise for improving the efficiency of human–AI interaction without compromising the portability that makes non-invasive systems suitable for real-world applications.

B.2 Brain-supervised Generation

Advances in machine learning have significantly enhanced BCI decoding accuracy and robustness, opening new possibilities for brain-supervised generative methods that transform neural signals into visual content [74, 75].

Several recent methods integrate neurophysiological data (e.g., fMRI, EEG, or fNIRS) with generative models [22, 38, 39]. For instance, CMVDM aligns fMRI features with semantics for image synthesis [40], and the MindEye series further lifts the resolution of generated images from decoded fMRI [41, 42]. DreamConnect translates brain signals into images based on fMRI, which is less accessible for everyday interaction [76]. DreamDiffusion produces images from EEG via temporal masked modeling [22]. EEG2Video extends the idea to dynamic video content [43]. While Davis *et al.* [24] initially explore brain-guided semantic image editing using a generative adversarial network, this work is limited to facial images and EEG signals. Moreover, Adamic *et al.* [44] reconstructs visual images from brain activity measured by fNIRS.

Unlike previous studies, our data were collected via a wireless BCI system, as shown in Fig. 2, from participants performing instruction-based image editing tasks. Compared with fMRI-based methods, this multimodal setup is portable and suited to daily use, which can support greater portability and broader real-world applicability. To the best of our knowledge, our work is the first to comprehensively utilize the full information of *EEG*, *fNIRS*, *PPG*, and head-motion signals for image editing. We specifically delve deeper into strategies of extracting neural features and fusing multimodal data, optimizing their integration to serve image editing needs.

B.3 Instruction-based Image Editing

Instruction-based editing tasks include adjusting color, contrast, and brightness; retouching objects or backgrounds; and applying filters or artistic effects. This progress ranges from global editing (e.g., style transfer [1]) to region-based image editing [2], which serves a wide range of applications in advertising, entertainment, and social media [2].

The surge in multimodal data has fostered the development of large models capable of complex creative tasks [77, 78]. Recent models such as GPT-4o [45] and Gemini [46] have evolved from basic data analysis to advanced image editing agents. These agents leverage generative methods to interpret user instructions for precise image manipulation.

Current instruction-based image editing agents integrate multimodal inputs, including text, images, and videos, to accurately identify and apply visual edits [13]. Leveraging learned multimodal representations, these agents interpret instructions from input, localize relevant regions, and perform targeted modifications [47–49]. Recent approaches, such as InstructPix2Pix [1], UltraEdit [2], MagicBrush [50], MIGE [51], and ACE [52] improve region-specific edits guided by natural language prompts. It is noticed that Jiang *et al.* [14] explore speech-driven image editing, highlighting the feasibility of hands-free interaction but still limited by linguistic expressiveness in recorded speech.

Despite these advancements, achieving efficient, delicate, prompt-free image editing remains challenging. Our work addresses this gap, exploring neural-signal-driven editing agents to decode cognitive intent directly for image manipulation, significantly enhancing accessibility and interaction efficiency.

B.4 Future Research Prospects

Recent studies have expanded the frontiers of multimodal representation learning and diffusion-based generation, which benifits the future improvement of this work [79–85]. For example, Zhou et al. [86], Xu et al. [87] and Liu et al. [88] extended causal and gated frameworks toward robust multimodal understanding. Li et al. [89] further introduced multi-objective unlearning, and Zeng et al. [90] advanced precise text editing for controllable LLM adaptation. These developments collectively reflect a broader shift toward interpretable, data-efficient, and causally grounded multimodal methods, which contribute to future neural-driven AI systems [91–94].

Complementary progress has also been made in visual information processing, multimodal perception and generative modeling [86, 89, 95–98]. Xin et al. [99, 100] who introduced scalable autoregressive and omni-diffusion architectures that can be further used in advanced image editing. Liu et al. [101] enhanced scene flow estimation via iterative diffusion, while Zhang et al. [102] demonstrated the potential of diffusion processes in molecular and structural synthesis. Other representative efforts [103, 104] continue to enrich multimodal benchmarks and architectures. It is noted that intention-driven visual reasoning [105] further reveal the emerging synergy between structured reasoning and perceptual modeling, inspiring the future direction of neural-driven methods toward reasoning models.

Collectively, these works point a bright future research path, which leverages diffusion-driven priors, causal reasoning, and behaviorally grounded learning for more interpretable and controllable multimodal intelligence based on advanced neural-driven approach and updated BCI devices.

C Supplementary Dataset Details

C.1 Background Details

Overall Mechanism. As illustrated in Fig. 8, EEG captures electrical activity along the scalp with high temporal resolution, enabling fine-grained monitoring of neural dynamics. Neural signals from Fpz, Fp2, Pz, and Oz electrodes serve complementary roles in facilitating hands-free image editing. Fpz and Fp2, located in the frontal cortex, are primarily responsible for attentional control and intentional decision-making, respectively. Notably, we deliberately preserve blink-related signals at Fpz and Fp2, allowing the system to retain ecologically valid user states and incorporate implicit ocular cues without requiring additional eye-tracking hardware. Pz supports spatial attention and target selection, while Oz reflects visual perception and validation of image modifications. Importantly, we avoid relying on conventional motor imagery regions (e.g., C3/C4), which are associated with imagined limb movements and commonly used in traditional BCI paradigms. These approaches often require extensive training, suffer from high inter-subject variability, and may be inaccessible to users with motor impairments. Instead, LoongX prioritizes frontal and parietal EEG channels that reflect universal cognitive processes such as attention, intention, and visual evaluation. This design enables plug-and-play usability without the need for motor calibration, ensuring a more intuitive, low-burden, and inclusive experience across diverse user populations. To further enhance the decoding of user states, we incorporate fNIRS signals from the left and right anterior medial prefrontal cortex (aMPFC), which provide critical hemodynamic insights into cognitive load, emotional valence, and motivation intensity. In addition, peripheral PPG signals are used to capture autonomic responses such as heart rate variability and arousal levels, enabling the system to monitor stress and engagement during editing tasks. It offers complementary physiological information, such as heart rate, blood oxygen saturation (SpO2), and peripheral blood flow variations. Given the susceptibility of these biosignals to motion-induced artifacts, particularly those arising from head movements, a triaxial accelerometer and a triaxial gyroscope are employed to capture translational and rotational motion of the head. This motion data is subsequently used for signal quality assessment and artifact mitigation, thereby enhancing the reliability of the acquired physiological measurements. Overall, this multimodal integration allows LoongX to decode user intent, attention, and emotional context in a holistic and adaptive manner, resulting in more precise, reliable, and user-aware editing commands.

C.2 Subject Information

12 healthy college students (6 females and 6 males) were recruited as the subjects for data collection. They have a mean age of 24.5$\pm$2.5 years and normal (or corrected-to-normal) vision. All volunteers were informed of the experimental process and received financial compensation. All volunteers signed the consent forms prior to the experiment, which was approved by the Ethics Committee of the ZJU Review Board for Human Research Protections (Protocol ID: No. 067 (2019)). The attention score in this study is objectively computed using the ratio between the power of the EEG alpha band (8–12 Hz) and the theta band (4–8 Hz), as shown below:

$$\text{Attention Score} = \frac{\text{Alpha Band Power}}{\text{Theta Band Power}} \tag{13}$$

This ratio has been widely used in cognitive neuroscience research as a neurophysiological index of attentional control and mental workload. For instance, Raufi and Longo [106] demonstrated that the alpha-to-theta and theta-to-alpha band ratios are reliable indicators of self-reported mental workload levels in EEG-based studies. The attention scores and the neural-signal-based image editing errors of 15 subjects in our experiment are summarized in Table 4. Subjects 13-17 are regarded as unseen subjects, and the L-Mind training set does not include these unseen data for cross-subject evaluation.

C.3 Data Collection Details

Using the setup shown in Fig. 2, a total of 23,928 pairs of effective data were collected from 12 participants. The specific details are given as follows.

Table 4: Attention and neural-signal-based image editing scores of individual subject

Subject	Gender	Age	#Samples	Attention	L1	L2	CLIP-I	DINO	CLIP-T
1	Female	25	2003	0.0887	0.2657	0.1109	0.6370	0.4890	0.2196
2	Female	29	2000	0.0817	0.2416	0.0950	0.6575	0.5021	0.2249
3	Female	26	2001	0.1340	0.2448	0.0963	0.6660	0.4878	0.2337
4	Female	22	1999	0.0739	0.2533	0.1005	0.6394	0.4606	0.2270
5	Female	28	1992	0.1218	0.2552	0.1031	0.6144	0.4157	0.2260
6	Female	29	1964	0.0822	0.2511	0.1000	0.6449	0.4564	0.2213
7	Male	22	1988	0.0851	0.2711	0.1160	0.6515	0.4634	0.2234
8	Male	23	1993	0.1105	0.2528	0.1017	0.6638	0.4833	0.2242
9	Male	22	1988	0.1500	0.2497	0.0998	0.6355	0.4571	0.2212
10	Male	24	2000	0.1298	0.2657	0.1144	0.6194	0.4240	0.2220
11	Male	24	2000	0.0954	0.2744	0.1151	0.6386	0.4339	0.2299
12	Male	22	2000	0.0971	0.2551	0.1034	0.6213	0.4323	0.2250
13 (unseen)	Male	35	500	0.1210	0.2681	0.1174	0.6022	0.4418	0.2594
14 (unseen)	Female	30	500	0.0775	0.2688	0.1179	0.6051	0.4405	0.2553
15 (unseen)	Male	13	200	0.0727	0.2618	0.1001	0.6055	0.4472	0.2576
16 (unseen)	Female	62	100	0.0441	0.2660	0.1141	0.6196	0.4611	0.2472
17 (unseen)	Male	63	100	0.0520	0.2610	0.1133	0.6017	0.4588	0.2595

Multimodal Device and Signal Collection Pipeline. LoongX employs non-invasive BCI technologies to acquire multimodal neurophysiological signals, integrating data from four EEG channels (Fpz, Fp2, Pz, Oz) sampled at 250 Hz, eight fNIRS channels located in the medial prefrontal cortex (MPFZ) zone sampled at 25 Hz, and eight PPG channels also within the MPFZ zone sampled at 25 Hz. Additionally, six channels of head motion data—comprising triaxial linear acceleration and triaxial angular velocity—are recorded at 12.5 Hz to capture head movement dynamics. EEG, fNIRS, PPG, and motion signals are synchronized and transmitted from the device to the PC-side software application via Bluetooth 5.3. The application streams these signals in real time to the Lab Recorder software through the Lab Streaming Layer (LSL). Meanwhile, event markers generated by the image-stimulus paradigm on the PC are also sent via LSL to Lab Recorder. These markers indicate the start and stop times of user-triggered recordings. As a result, we obtain well-aligned EEG, fNIRS, PPG, motion, and audio signals. There was a 1-second cross-fixation interval between each pair of pre- and post-edited images (Fig. 2). Data collection was organized in batches of 100 images, after which participants were given a rest period. A total of 2,000 images were collected from each participant. A total of 23,928 pairs of effective data were collected, comprising participants' speech, EEG, fNIRS, PPG, and head motion information.

Our developed computer software used in this non-invasive BCI device streams the data in real time to the *Lab Recorder* software through the Lab Streaming Layer (LSL) framework. Meanwhile, event markers generated by the image-stimulus paradigm on the PC are also sent via LSL to *Lab Recorder*. These markers indicate the start and stop times of user-triggered recordings. Neural signals are collected simultaneously while the participant uses speech to describe the content of the image editing. As a result, we obtain well-aligned EEG, fNIRS, PPG, motion, and audio signals. There was a 1-second cross-fixation interval between each pair of pre- and post-edited images to distinguish between different image-editing pairs clearly. 2,000 images are collected from each participant. In all, 23,928 valid data pairs are gathered, encompassing participants' speech, EEG, fNIRS, PPG, and head motion information.

Participants. 12 healthy college students (6 female and 6 males) were recruited as the subjects for data collection. They have a mean age of 24.5$\pm$2.5 years and normal (or corrected-to-normal) vision. All volunteers were informed of the experimental process and received financial compensation (at an hourly rate exceeding the minimum wage in the region). All volunteers signed the consent forms prior to the experiment, which was approved by the Ethics Committee of the Institutional Review Board at Zhejiang University for Human Research Protections (Protocol ID: No. 067 (2019)). To ensure data consistency during extended EEG acquisition, the experiment was conducted in a quiet room (3m $\times$ 5m) maintained at a constant temperature of 24°C and a constant humidity. EEG signals were recorded using the latest non-invasive hydrogel electrodes, which are known to provide one of the highest signal quality among available electrode types. To maximize signal integrity and minimize impedance-related artifacts, the electrodes were replaced every five hours or sooner if necessary. Data

acquisition for each participant starts at 9 AM daily, and the room was shielded from direct sunlight to reduce its impact on light-sensitive signals such as fNIRS and PPG. Example stimulus sessions that were displayed to participants are shown in Fig. 2.

C.4 Data Preprocessing

To make use of the most relevant information and reduce noise and artifacts, each type of multimodal neurophysiological signal was preprocessed based on its unique characteristics. The first step for our proposed neural-based image editing method is extracting and encoding data into a structured format based on collected metadata. Techniques such as interpolation for missing data and normalization methods to standardize signal amplitudes are applied to clean and normalize data to remove noise, fill in missing values, or correct errors. Advanced filtering skills based on machine learning are implemented to select data aligned with the model's predefined objectives, such as identifying visual patterns correlated with specific cognitive states or processing visual-related neural signals. Specifically, the data processing steps include:

EEG Preprocessing. Signals were band-pass filtered (1–80 Hz) and notch-filtered (48–52 Hz) to remove noise and powerline artifacts. The EEG channels near the eyes (Fpz and Fp2) retained ocular signals for intentional blink detection. Specific procedures are as follows:

- A band-pass filter was applied in the range of 1–80 Hz to remove low-frequency drifts and high-frequency noise.
- A notch filter was applied in the frequency range of 48–52 Hz to eliminate powerline interference, which is specific to the 50 Hz electrical grid in China.
- Ocular artifacts from the AF8 and FPz channels were preserved intentionally, as these channels were specifically designed to capture eye movements, including blinking.

fNIRS Preprocessing. The eight-channel fNIRS signals were processed to extract concentration changes of oxygenated hemoglobin (HbO), deoxygenated hemoglobin (HbR), and total hemoglobin (HbT). The following preprocessing steps were performed:

- The HbO, HbR, and HbT signals were band-pass filtered in the range of 0.01–0.5 Hz to isolate brain spontaneous and stimulus-induced hemodynamic responses associated with neural activity. These signals correspond to brain activity evoked by the image-editing paradigm.
- The processed HbO, HbR, and HbT signals were averaged to obtain two signals per hemisphere: left hemisphere HbO, right hemisphere HbO, left hemisphere HbR, right hemisphere HbR, left hemisphere HbT, and right hemisphere HbT.
- Moreover, light intensities at 735 and 850 nm were converted to HbO/HbR/HbT using the Modified Beer–Lambert Law, then filtered (0.01–0.5 Hz) to isolate hemodynamic responses.

Hemodynamic Conversion: fNIRS signals are calculated by raw 735 nm and 850 nm near-infrared light intensity. The raw fNIRS signals are measured as light intensity changes at different wavelengths. To convert these optical signals into concentrations of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR), we apply the Modified Beer–Lambert Law (MBLL). The logarithmic optical density change ΔA is calculated as:

$$\Delta A(\lambda) = \log\left(\frac{I_0(\lambda)}{I(\lambda)}\right) \tag{14}$$

where $I_0(\lambda)$ is the initial light intensity, $I(\lambda)$ is the measured intensity at wavelength λ, and $\Delta A(\lambda)$ is the optical density change.

Then, the concentration changes of HbO and HbR are derived using the MBLL as follows:

$$\begin{bmatrix}\Delta\text{HbO}\\ \Delta\text{HbR}\end{bmatrix} = \frac{1}{\text{DPF}\cdot L}\cdot\begin{bmatrix}\varepsilon_{\text{HbO}}^{\lambda_1} & \varepsilon_{\text{HbR}}^{\lambda_1}\\ \varepsilon_{\text{HbO}}^{\lambda_2} & \varepsilon_{\text{HbR}}^{\lambda_2}\end{bmatrix}^{-1}\cdot\begin{bmatrix}\Delta A(\lambda_1)\\ \Delta A(\lambda_2)\end{bmatrix} \tag{15}$$

Here, ε denotes the molar extinction coefficients of HbO and HbR at wavelengths λ_1 and λ_2, L is the source-detector separation, and DPF is the differential pathlength factor that accounts for scattering in tissue. Additionally, the total hemoglobin concentration change ΔHbT is calculated as the sum of the changes in HbO and HbR:

$$\Delta\text{HbT} = \Delta\text{HbO} + \Delta\text{HbR} \tag{16}$$

This process yields time-series concentration changes ΔHbO, ΔHbR, and ΔHbT. These eight channels of ΔHbO, ΔHbR, and ΔHbT were band-pass filtered (0.01–0.5 Hz) to isolate brain hemodynamic responses related to neural activity evoked by the image-editing paradigm. Finally, the HbO, HbR, and HbT signals were then averaged to obtain left and right hemisphere values for each.

PPG Preprocessing. Optical signals were band-pass filtered (0.5–4 Hz) to extract cardiac rhythms for heart rate variability and arousal estimation. The raw light intensity data for the four-channel PPG signals were acquired at wavelengths of 735 nm and 850 nm. The detailed preprocessing steps for these signals were as follows:

- The raw optical intensity data at both 735 nm and 850 nm were averaged to obtain the following four signals: left hemisphere 735 nm, right hemisphere 735 nm, left hemisphere 850 nm, and right hemisphere 850 nm.
- A band-pass filter with a frequency range of 0.5–4 Hz was applied to these four signals to extract hemodynamic changes associated with cardiac rhythms, reflecting the blood flow pulsations due to heartbeats.

Motion Tracking. The raw signals from the six-axis motion sensors (accelerometer + gyroscope) were retained to ensure accurate monitoring of head movements during the experiment, as these movements could potentially affect signal quality. These signals were also used to assess signal quality and reduce movement-related artifacts.

Session Segmentation.: Each editing session was defined by a speech-triggered onset and offset, linked to pre- and post-edited image pairs. Only active segments were retained for training and testing. Specifically, the image stimulus is used to mark the beginning of a session when the user starts recording audio. The session ends when the user stops the audio recording. During this time, the system synchronizes the recording of the participant's speech, EEG, fNIRS, PPG, and head motion data. Data recorded during periods of silence, when the audio recording is not active, is excluded from training and analysis.

D Supplementary Method Details

D.1 Background Information

To enable hands-free image editing, the integration of multimodal signals facilitates the interpretation of the user's intent and mental state: EEG activity at frontal sites such as Fpz reflects attention control and task initiation [107]; signals from the right prefrontal cortex (Fp2) are associated with emotional regulation and motivational drive, as evidenced by frontal alpha asymmetry [108]; EEG at the midline parietal site (Pz) captures spatial attention and visuomotor integration [109]; and occipital EEG (Oz) provides information on visual perception and image processing load [110]. Additionally, fNIRS measurements over the left and right anterior medial prefrontal cortex (aMPFC) reveal cognitive load and emotional valence through blood oxygenation patterns [30], and peripheral photoplethysmography (PPG) signals monitor heart rate variability as a proxy for autonomic arousal, enabling the system to track user stress levels and engagement during interaction [31, 32]. Integrating these modalities allows the system to adapt to the user's focus, emotional state, and workload, making hands-free image editing more intuitive and responsive.

Midline parietal EEG (fpz) is involved in spatial attention allocation and visuomotor integration, contributing to the coordination of visual and motor aspects of the editing task [107]. Occipital EEG at Oz captures visual perception dynamics and image processing load, providing insight into how visual information is being processed by the user's brain [110].

D.2 Theoretical Derivation of Flow-Aware Inversion

We defined inversion as a trajectory that transports a clean sample $x_0 \sim p_0(x)$ to a noisy latent $x_t \sim p_t(x)$. Within the DDPM framework, the forward process is described as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 \; + \; \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \qquad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i, \;\; \alpha_i = 1-\beta_i, \;\; \varepsilon \sim \mathcal{N}(0, I).$$

First, we formulate a pure stochastic SDE that follows the forward diffusion to gradually add noise, and then run the time-reversed SDE to retrieve an editable reconstruction, similar to the philosophy of SDEdit [111].

Second, a probability-flow ODE treats diffusion via the score-based velocity field $v(x_\tau)$, replacing the random noise with a deterministic velocity field $v(x_\tau)$ proportional to the score $\nabla_{x_\tau} \log p_\tau(x_\tau)$:

$$x_0 = x_t - \int_0^t v(x_\tau)\, \mathrm{d}\tau, \qquad x_t = x_0 + \int_0^t v(x_\tau)\, \mathrm{d}\tau = x_0 - \int_t^0 v(x_\tau)\, \mathrm{d}\tau.$$

A continuum between these two extremes is obtained by interpolating the stochastic and deterministic contributions with a parameter $\eta \in [0, 1]$:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \big[\eta\, \varepsilon + (1-\eta)\, u_t\big], \quad u_t = \int_0^t \frac{c_\tau\, v(x_\tau, \tau)}{\sqrt{1-\bar{\alpha}_\tau}}\, \mathrm{d}\tau,$$

where $\varepsilon \sim \mathcal{N}(0, I)$ and c_τ is a schedule-dependent factor that aligns the units of the velocity term with standard DDPM dynamics [112]. Choosing $\eta = 0$ recovers the deterministic ODE path, whereas $\eta = 1$ yields the fully stochastic SDE path, and intermediate values trade deterministic guidance for stochasticity.

Our flow-aware inversion belongs to the deterministic end. As Flux.1-Dev predicts rectified-flow velocity rather than a DDPM score, we insert a lightweight rank-128 LoRA adapter W that maps the frozen backbone's predicted velocity $\boldsymbol{\epsilon}_\phi(x_\tau, \tau)$ into the DDPM score domain through:

$$v(x_\tau) = \sigma_\tau W\big(\boldsymbol{\epsilon}_\phi(x_\tau, \tau)\big).$$

The time-dependent coefficient σ_τ helps bridge the rectified-flow velocity and the DDPM score scale, while the linear bridge preserves the benefits of flow pre-training and enables faithful one-to-one reconstructions, in a similar spirit to edit-friendly DDPM [113] or LEDITS++ [114].

D.3 More Method Explanations

To validate neural signals as reliable semantic conditions, we conduct a preliminary multi-label classification experiment detailed in Fig. 9. Results show that EEG signals yield over 7% mean Average Precision (mAP) gain compared to random noise, and fNIRS improved recall through robustness. Combining EEG and fNIRS shows stronger performance, and integration with textual prompts further enhances outcomes than using text only, confirming modality complementarity.

To handle diverse input shapes of raw signals, we pad/truncate inputs and find that EEG sequences of length 8,192 offer a good trade-off between performance and efficiency, which is summarized in Fig. 9(c). As longer input sequences bring unbearable computational costs, we seek an optimal solution that keeps key information in the long sequence data while not significantly increasing computational costs. This motivates our CS3 encoder, which captures both temporal and channel-wise patterns efficiently while achieving a better trade-off between information density and computing efficiency.

To effectively extract structured representations from diverse signals, we propose the CS3 encoder. CS3 utilizes the linear computational complexity and efficiency of the structured state space model (S3M) [55] for encoding long sequences into channel representations. Recognizing that increasing latent dimensions can still significantly raise computational costs, CS3 implements a cross-feature extraction mechanism that separately encodes temporal and channel information.

While the fixed latent dimension of the S3M can not fully capture the dynamic information in signals, we further apply an adaptive feature pyramid based on adaptive average pooling. Each signal is processed through modality-specific neural encoders to produce latent features. Taking a C-channel EEG sequence $\mathbf{X} \in \mathbb{R}^{C \times L}$ as an example, we first normalize it to $[-1, 1]$ and pass it through two parallel S3M blocks capturing complementary dynamics.

To align and integrate multi-source signals, we further introduce DGF, which dynamically processes and fuses features across modalities. DGF can model inter-modality interaction modelling to form a unified latent space, which is also optionally usable for alignment with text embeddings to support a more hybrid conditioning.

In summary, CS3 captures multi-scale temporal and structural patterns in neural signals, consistent with findings that multi-band EEG features improve intent decoding. DGF performs selective multimodal fusion through dual gating, which follows prior successes of gating and normalization strategies in multimodal learning. On pairing EEG plus PPG with T5, and fNIRS plus Motion with CLIP. Rationale is also not arbitrary. T5 provides fine-grained token-level semantics that help precise instruction following, which complements the fast neural dynamics in EEG and the lightweight hemodynamics from PPG. CLIP provides robust global semantics that align with slower cortex-wide fNIRS signals and intentional head Motion.

E Supplementary Experimental Details

E.1 Cross-subject Experiment

Our original dataset was collected from 12 participants (6 female, 6 male, mean age 24.5 ± 2.5 years), each contributing around 2,000 paired samples under carefully controlled experimental conditions. While our initial split ensured training/test separation, we acknowledge that the possibility of subject overlap could limit generalizability. To address this, we performed additional cross-subject evaluations with 5 new unseen participants (3 male, 2 female, ages 13–63, see Table 4). The results are presented in Table 5. It confirms that the model maintains strong generalization when applied to unseen individuals, with performance trends on CLIP-I, DINO, and CLIP-T remaining consistent with those from the original test set. This provides evidence that our approach is not overly reliant on subject-specific neural signatures, but instead captures transferable semantic representations.

Table 5: Performance comparison of baseline and our proposed LoongX on the original test set (12 subjects) and unseen test set (5 new subjects).

Test Dataset	Methods	Conditioning	L1 (↓)	L2 (↓)	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)
Original	OmniControl	Text	0.2632	0.1161	0.6558	0.4636	0.2549
	OmniControl	Speech	0.2714	0.1209	0.6146	0.3717	0.2501
	LoongX (OmniControl)	Neural Signals	0.2509	0.1029	0.6605	0.4812	0.2436
	LoongX (OmniControl)	Signals + Speech	0.2594	0.1080	0.6374	0.4205	0.2588
Unseen	OmniControl	Text only	0.2581	0.1133	0.6528	0.4655	0.2553
	OmniControl	Speech	0.2779	0.1271	0.6221	0.3942	0.2508
	LoongX (OmniControl)	Neural Signals	0.2574	0.1090	0.6019	0.4037	0.2403
	LoongX (OmniControl)	Signals + Speech	0.2668	0.1146	0.6049	0.4447	0.2568

E.2 Abalation Studies Breakdown

Table 6 and Table 7 correspond to the detailed ablation studies illustrated in Fig. 4 and Fig. 5 in the main manuscript.

From a neuroscience perspective, the superiority of fNIRS and EEG integration aligns well with our understanding of brain physiology: EEG captures rapid electrical oscillations reflecting millisecond-level neuronal activity, while fNIRS provides complementary information about slow hemodynamic responses, reflecting regional brain activation. Their fusion exploits both temporal and spatial dynamics of cognition, which is particularly relevant for decoding the complex neural basis of visual and semantic processing required by image editing tasks. The limited effect of PPG and motion

signals may stem from its primary focus on peripheral patterns, which are less directly involved in cortical information processing but still can affect the robustness of the performance.

Drilling down into Table 7, the channel-wise analysis offers intriguing support for established functional specialization in the human brain:

- The Oz channel (occipital cortex) stands out in global image alignment and robustness metrics, mirroring its neuroanatomical role as the hub for early-stage visual perception. The occipital lobe, and especially the Oz electrode position, is known for processing visual stimuli, edge detection, and scene analysis. The strong performance observed here suggests that even in a data-driven, deep learning context, the fundamental dominance of the visual cortex in image-based tasks persists.
- In contrast, the Fpz channel (frontopolar cortex) exhibits heightened performance in metrics linked to semantic understanding and higher-order cognitive alignment. The prefrontal regions are responsible for executive functions such as attention, planning, and integrating multimodal information, which are essential for aligning generated content with textual or conceptual prompts.

These results not only validate long-standing neuroscientific theories, such as the hierarchical processing streams in the brain (from occipital "what is seen" to frontal "what does it mean/what to do"), but also provide practical guidelines: in settings where only a limited number of electrodes or sensors are available, prioritizing signals from functionally specialized regions (e.g., Oz for vision, Fpz for semantic or cognitive control) can maximize decoding efficiency for targeted tasks.

Furthermore, the convergence of these findings with classical brain science underscores the translational value of deep learning in cognitive neuroscience. It highlights the potential for future brain-computer interfaces to be not only data-driven but also "anatomy-aware," leveraging our evolving map of the brain to design more effective and interpretable multimodal AI systems.

Table 6: Evaluation results for different signal combinations.

Metric	Pure EEG	EEG + fNIRS	EEG + fNIRS + PPG	All Signals	All Signals + Text
L1	0.2641	0.2508	0.2631	0.2571	0.2594
L2	0.1078	0.1029	0.1123	0.1076	0.1080
CLIP-I	0.5457	0.6604	0.6536	0.6274	0.6374
DINO	0.2963	0.4811	0.4942	0.4245	0.4205
CLIP-T	0.2251	0.2436	0.2226	0.2481	0.2588

Table 7: Results using different brain region signals. GT refers to ground truth. Ch means the channel.

Condition	L1	L2	CLIP-I	DINO	CLIP-T (Ours)	CLIP-T (GT)
EEG (All channels)	0.2508	0.1029	0.6604	0.4811	0.2436	0.2594
EEG (Ch 0, Pz)	0.2509	0.1028	0.6486	0.4787	0.2314	0.2594
EEG (Ch 1, Fp2)	0.2581	0.1070	0.6178	0.4150	0.2421	0.2594
EEG (Ch 2, Fpz)	0.2486	0.1022	0.6669	0.4846	0.2481	0.2594
EEG (Ch 3, Oz)	0.2475	0.1003	0.6619	0.4873	0.2367	0.2594

E.3 More Qualitative Results

More qualitative comparisons are presented in Figures 10-13, corresponding to the four broad editing categories: Global, Background, Object, and Text Editing. For clarity and conciseness in the figures, the original lengthy instructions have been distilled into single-sentence descriptions without altering their intended meaning. The editing results of our neural-driven and neural-speech fusion methods consistently outperform text-prompt-based editing results, demonstrating superior alignment with human intent and greater editing precision. Notably, Text Editing presents a more complex challenge compared to other categories. Given the current limitations of backbone models (with the exception of commercial models like GPT-4o), text-based edits remain difficult. As evidenced by the examples, neural-driven approaches exhibit a stronger ability to align with human intent, making the editing

process more intuitive and effective. It is foreseeable that in the near future, as reliable image-editing backbones become more accessible, neural-driven image editing will further stabilize and mature, evolving into an indispensable tool for everyday creative workflows.

Fig. 14 specifically analyzes three characteristic failure modes: (1) cases involving overly imaginative descriptions that deviate significantly from the training data distribution (e.g., "long-legged space creature"), (2) ambiguous instructions with insufficient semantic details (particularly evident in case (b) where background retention specifications were omitted), and (3) challenges posed by non-standard input image dimensions (such as panoramic aspect ratios). These failure cases provide valuable insights into the current limitations of neural-based editing systems.

E.4 More Failure Cases

Figure. 14 specifically illustrates three failure cases, where overly exaggerated imagination, vague instructions, or uncommon input image sizes may contribute to failed results. It demonstrates three characteristic failure modes: (1) cases involving overly imaginative descriptions that deviate significantly from the training data distribution (e.g., "long-legged space creature"), (2) ambiguous instructions with insufficient semantic details (particularly evident in case (b) where background retention specifications were omitted), and (3) challenges posed by non-standard input image dimensions (such as panoramic aspect ratios). These failure cases provide valuable insights into the current limitations of neural-based editing systems.

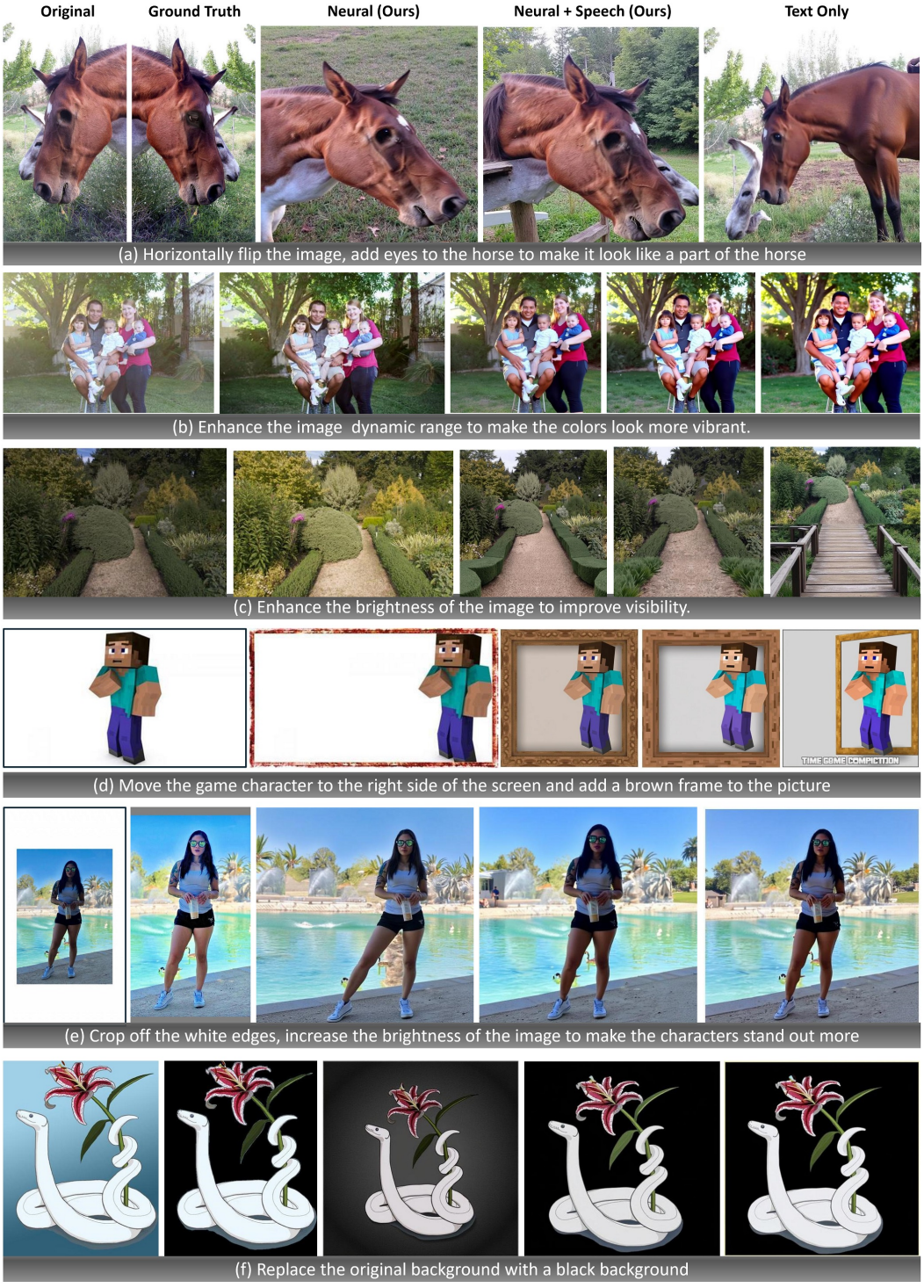

Figure 10: Qualitative comparison of our neural-driven and speech-neural fusion methods and text-prompt baseline for Global Editing category.

Figure 11: Qualitative comparison of our neural-driven and speech-neural fusion methods and text-prompt baseline for Background Editing category.

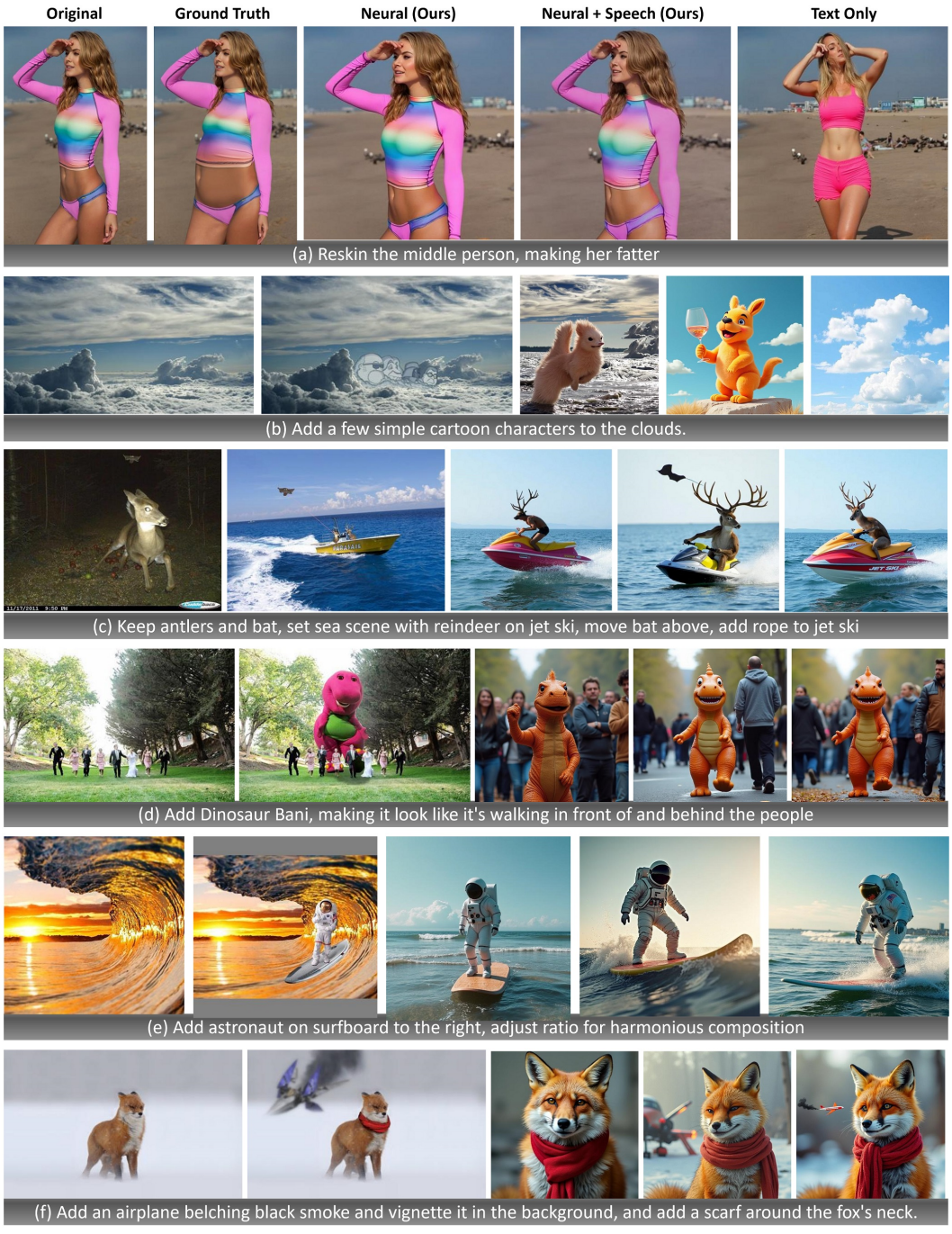

Figure 12: Qualitative comparison of our neural-driven and speech-neural fusion methods and text-prompt baseline for Object Editing category.

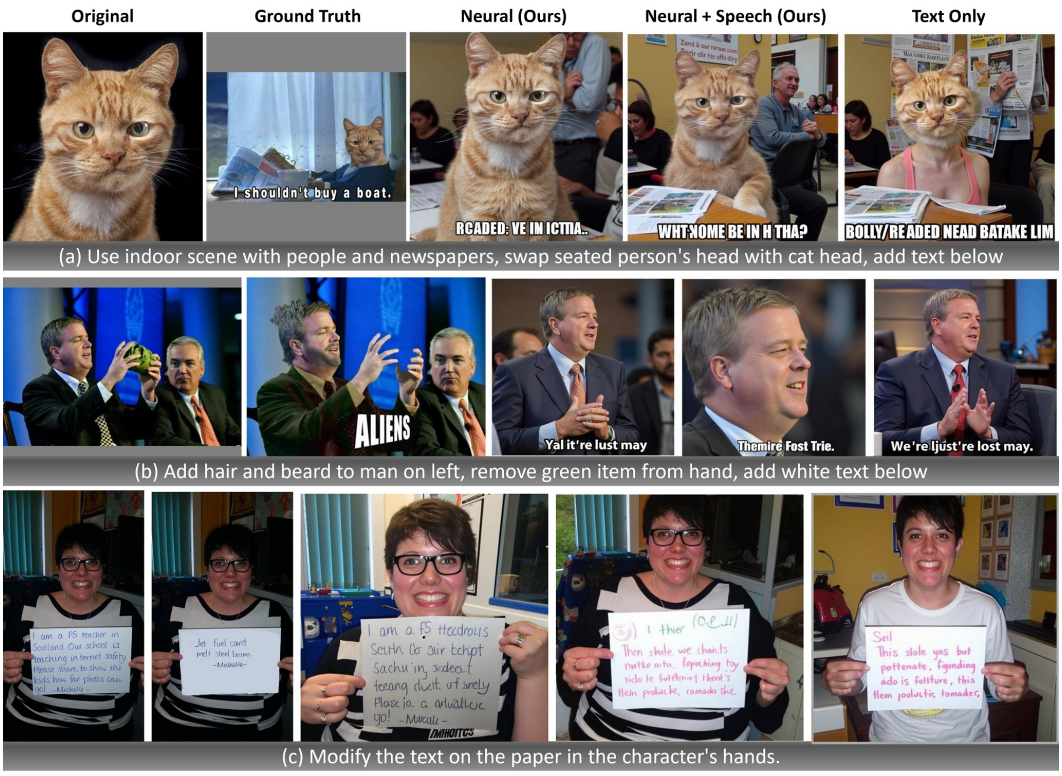

Figure 13: Qualitative comparison of our neural-driven and speech-neural fusion methods and text-prompt baseline for Text Editing category.

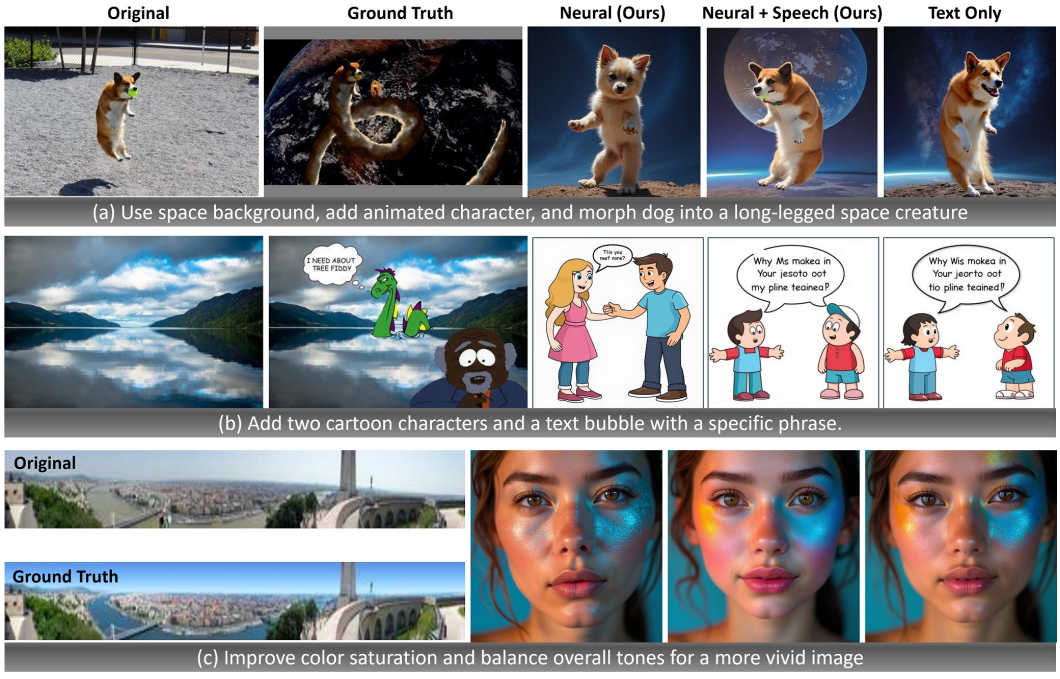

Figure 14: Qualitative analysiscomparison onf our neural-driven and speech-neural fusion methods and text-prompt baseline for three failure cases: (a) Overly exaggerated descriptions, e.g., "long-legged space creature"; (b) Vague instructions lacking detail, such as omitting whether to retain the background; (c) Uncommon image dimensions, e.g., panoramic input images.