Controllable EEG Editing Via Guided Diffusion Sampling For Affect Transposition

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we study single-trial EEG editing with diffusion models, using affect transposition as a case study: given a real seed signal and a target valence label, we generate an edited signal that shifts the recognized affect while preserving affect-unrelated brain activity patterns. For this purpose, we propose a controllable EEG-to-EEG diffusion framework that performs seed-conditioned generation and supports various control mechanisms, to balance the tradeoff between affect transposition strength and seed signal fidelity. In our evaluation protocol, we test for editing effectiveness, seed fidelity, and physiological plausibility; on affective EEG (DEAP dataset), we quantify successful affect transposition with a valence classifier while measuring spectral-temporal deviation (e.g., STFT-based distances) from the seed. We probe edit plausibility with Frontal Alpha Asymmetry (FAA) measurements, which is a known neurophysiological marker that can indicate high valence. Furthermore, we test the robustness of affect-unrelated signal preservation on out-of-distribution data on an SSVEP (steady-state visually evoked potential) task using FBCCA (filter-bank canonical correlation analysis)-based frequency recognition. Our experiments indicate that classifier guidance towards the desired affective state combined with minimal noising of the input samples enables targeted affect edits with measurable limits regarding seed-fidelity and affect-unrelated feature degradation. These results suggest that single-trial EEG edits can be viably achieved utilizing our diffusion-based editing framework.

1 Introduction

Electroencephalography (EEG) provides a non-invasive window into brain dynamics and is widely used in affective computing and brain–computer interface (BCI) research. In affective EEG, a central goal is to relate neural activity to affective states such as valence and arousal, enabling applications ranging from human–computer interaction to personalized neurofeedback (Craik et al., 2019). At the same time, EEG signals are highly structured: they contain subject- and session-specific characteristics, task-related rhythms, and stereotyped spectral–temporal patterns. This multifaceted structure is precisely what makes EEG useful—but the complex underlying latent processes which independently influence the aggregated signal measured by the surface electrodes on the scalp.

In recent years, several approaches have been tried to generate synthetic EEG data, either as an implicit model regularization mechanisms to improve general model accuracy (Eilts & Putze, 2022) or as a means for generating data of rare or unseen cases Xu et al. (2022). Most generative approaches on the EEG domain focus on synthesizing samples from scratch, or on augmenting datasets with new trials that resemble the training distribution. However, many scientific and practical use-cases require a different capability: counterfactual editing. Given a real seed trial, we would like to generate a minimally modified version that changes a specific attribute—here, as an example, affective valence—while preserving the remaining aspects of the signal. Such edits can support controlled analysis ("what would this trial look like if the subject were in a different affective state?") or data augmentation/transformation that preserves trial identity. Editing affect in EEG is challenging for three reasons: First, the mapping from affect labels to EEG patterns is subtle and noisy; supervision is weak and often dataset-dependent. Second, EEG is multi-channel and non-stationary, so naive perturbations can introduce artifacts or violate known spectral/physiological constraints. Third,

controllability requires a tunable mechanism to trade off edit strength against seed retention: pushing too hard may achieve the target label but collapse or distort trial structure, while preserving the seed too strongly may yield imperceptible edits. In this work, we address these challenges with a controllable EEG-to-EEG diffusion approach, using affect editing as our testbed for high-level EEG attribute editing. Diffusion models provide a flexible framework for conditional generation, and importantly, allow several complementary ways to steer sampling at inference time. We formulate high-level, single trial EEG editing as seed-conditioned generation with a desired affective state: given a real EEG (low or natural valence) seed trial and a target label (high valence in our setting), we generate an edited signal intended to express the target affect while keeping non-affective structure close to the seed. For affect transposition beyond standard diffusion network label conditioning, we study guidance strategies that explicitly amplify conditioning signals, including classifier-free guidance and classifier guidance to find the most performant mechanism for our purpose. We also introduce a simple inference-time control, seed starting-noise, that acts as an intuitive knob to control seed–fidelity within the generated samples: starting sampling from a less-noised state constrains edits toward the seed, while starting from a more-noised state enables stronger modifications. Furthermore, signal conditioning mechanism also is compared against similar guidance methods, such spectral difference (STFT-L1) driven and classifier-free seed signal guidance. To evaluate whether high-level edits are both effective and meaningful, we adopt an expansive protocol centered on the affect-transposition case. On an affective EEG dataset (DEAP, Koelstra et al. (2011)), we quantify affect transposition using a valence classifier and measure seed fidelity via spectral–temporal distances (STFT-L1). We further probe physiological plausibility using frontal alpha asymmetry, an established marker for positive valence in EEG (Coan & Allen, 2004). Finally, to test whether editing preserves affect-unrelated structure under distribution shift, we evaluate on an out-of-distribution (OOD) dataset (Wang et al., 2016) containing SSVEP trials (Steady-State visually evoked potentials, a phase-locked neural response to periodic stimulation) and quantify degradation of frequency-specific structure using an FBCCA-based recognizer (Chen et al., 2015). Here, FBCCA (filter-bank canonical correlation analysis) identifies the attended stimulus by correlating EEG with sine/cosine reference signals across multiple frequency sub-bands, providing a robust measure of frequency-specific structure. Because SSVEPs are stimulus-locked EEG responses to flickering visual targets, their frequency-specific structure provides a task signal that is largely independent of affective valence and cognitive-load variations in typical BCI settings (Evain et al., 2017).This combination allows us to grasp how changing the recognized affect label of samples relates to other unrelated structures in it. Our results show that diffusion-based EEG editing can be made controllable in practice: edit strength and start-noise induce a predictable trade-off between the targeted semantic (here affective) structure and retention of seed structure. In particular, classifier guidance is the most effective affect guidance, while initializing the synthesis with minimal start-noise helps constrain edits so that affect-unrelated features degrade in a limited way. This paper makes the following contributions:

1. We formulate seed-conditioned EEG-to-EEG generation with diffusion models and use affect transposition as a concrete case to investigate whether high-level semantic attributes can be edited while preserving non-affective structure.
2. Utilizing a diffusion-based, conditioned EEG generation framework, we compare several control mechanisms for high-level EEG attribute editing—including classifier-free and classifier guidance for attribute steering, start-noise control for seed retention, and spectral-loss-driven sampling—for achieving seed signal retention.
3. We present a comprehensive evaluation protocol centered on the affect-transposition case study, including in-distribution metrics (valence adherence, seed fidelity, physiological plausibility) and out-of-distribution robustness tests (SSVEP task-structure preservation), demonstrating both the potential and limits of diffusion-based high-level EEG editing.

2 Related Work

EEG-based affect recognition often targets valence and arousal and relies on learning subtle spectral–temporal patterns from multichannel recordings, but the field also repeatedly emphasizes that EEG is noisy, subject-specific, and difficult to model robustly, which motivates both strong evaluation protocols and data-efficient

learning strategies (Craik et al., 2019). A long-standing direction for improving learning under limited data has been synthetic data generation and augmentation. Earlier work in this space has often used GAN-based generators, including methods designed specifically to mitigate class imbalance in BCI settings such as RSVP; for example, BWGAN-GP generates minority-class trials and reports downstream gains with EEGNet-style classifiers (Xu et al., 2022). More recently, Transformer-based generation and representation learning have also been explored to improve cross-subject generalization by decomposing EEG into time-invariant and subject-invariant factors (Wan et al., 2024). Alongside these developments, diffusion models have emerged as a compelling alternative for generative modeling due to their stable training and flexible conditional sampling. In the EEG domain, Tosato et al. (2023) demonstrate that DDPMs can synthesize affective EEG by operating on STFT-derived representations, providing evidence that diffusion can capture affect-relevant structure and can be useful for augmentation. Moving to native time-domain biomedical signals, Li et al. (2024) propose Biodiffusion, a unified diffusion framework for multivariate physiological time series, and show that diffusion-based models can outperform common time-series GAN baselines across modalities, including EEG. Klein et al. (2024) further support time-domain conditional diffusion for EEG by synthesizing ERP responses and evaluating them with paradigm-specific temporal metrics, strengthening the case that diffusion models can reproduce fine-grained electrophysiological dynamics. A key advantage of diffusion models is controllability at inference time via guidance mechanisms. These mechanisms allow us to steer the denoising process toward a desired attribute, while allowing us to set the strength of this editing effect. This is particularly relevant when the objective is to transform an existing sample. Architecturally, many diffusion implementations build on encoder–decoder backbones such as U-Net (Ronneberger et al., 2015), often enhanced with attention mechanisms (Vaswani et al., 2017) and conditioning layers such as FiLM (Perez et al., 2018), which provide practical tools for injecting timestep and label context in a parameter-efficient way. For downstream evaluation and for learning affect-related targets, compact EEG classifiers such as EEGNet (Lawhern et al., 2018) and end-to-end extensions like E2ENNet (Han et al., 2022) remain strong baselines in affective EEG and BCI applications. Finally, because affect editing should preserve affect-irrelevant signal structure, it is useful to evaluate under conditions where the "to-be-preserved" structure is well defined. In SSVEP-based BCIs, frequency-specific structure can be quantified with established datasets and recognizers; the Benchmark SSVEP dataset provides a standard reference point (Wang et al., 2016), and FBCCA is a widely used method for robust frequency recognition across sub-bands (Chen et al., 2015). This line of work motivates evaluating whether generative interventions preserve stimulus-locked periodic structure even when the model is primarily trained for affect-related objectives.

3 Methodology

3.1 Datasets

The main data set for training and evaluation is the DEAP data set (Koelstra et al., 2011), which is widely used as a benchmark for emotion recognition from EEG data. During its trials, participants (N=32) watched forty one-minute music videos and rated their valence and arousal on a nine-point scale while EEG was recorded from thirty-two scalp electrodes at 512 Hz (downsampled to 128 Hz in the public preprocessed release). For our analysis, each trial in the dataset was segmented into non-overlapping one-second windows (128 samples), producing 60 windows per trial and 76,800 segments in total across participants. Valence was binarized (rating > 5 as "high", otherwise "low"). This dataset is used for in-distribution training and evaluation for both the valence classifier and the diffusion generator. For out-of-distribution testing we use the Benchmark SSVEP dataset (Wang et al., 2016) and evaluate frequency recognizability with Filter Bank Canonical Correlation Analysis (FBCCA) as described by Chen et al. (2015). Steady-State Visual Evoked Potentials are brain responses that oscillate at the stimulus frequency, strongest over occipital electrodes (O1/Oz/O2, PO3/PO4, Pz). We resampled the dataset's 64-channel, 250 Hz recordings to 128 Hz and extract a 1.0 second to match the generator's input and output. FBCCA builds sinusoid/cosine references at the target frequencies and harmonics, filters EEG into multiple sub-bands, and computes per-band canonical correlations with the references. Specifically, a weighted sum of squared correlations yields a score per frequency, and the argmax is the prediction.

3.2 EEG editing framework

We formulate affect editing as a seed-conditioned generation problem. Given a real low or neutral valence EEG seed window x_{seed} (with C channels and T time samples) and a target affect label y* (high valence), our goal is to generate an edited window x' that is recognized as the target affect while preserving the seed's affect-unrelated structure. Instead of synthesizing from scratch, we keep the seed in the loop during inference and expose controllable knobs that trade off affect edit strength against seed retention. The underlying framework for this approach is illustrated in Figure 1: We train a valence classifier on the DEAP data set as well as the unconditioned EEG diffusion generator. To this generator, we add two parallel modes for conditioning: A class guidance mode for generating EEG data of a specific class, as well as a signal guidance mode for which the model is given a sample of raw EEG data and is incentivized to retain as much of this seed signal as possible.

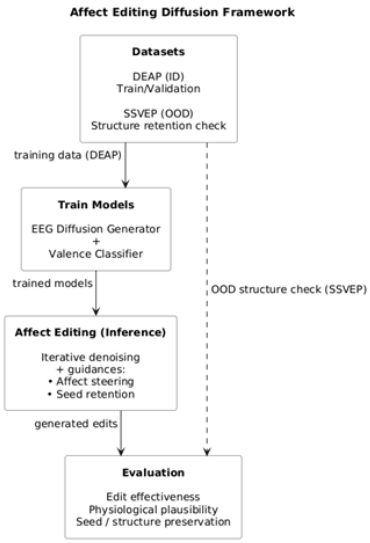

Figure 1: Block diagram of the affect editing diffusion framework.

3.3 Diffusion backbone

We train a denoising diffusion probabilistic model (DDPM) on 1-second non-overlapping windows created from the DEAP dataset, represented as an electrode×time grid (32x128; 76,800 windows), using a 90/10 train/validation split. From numerical labeling, samples are relabeled to binary (high/low valence) with a threshold of 5.0. We use a 1000-step noise schedule and optimize the standard noise-prediction objective (Ho et al., 2020) with AdamW (lr 1x10-4, batch size 64) and early stopping (patience 20). Although more modern diffusion variants (e.g., latent diffusion models (Rombach et al., 2022)) are prevalent in image generation, we use the standard DDPM formulation of Ho et al. (2020) because it is a simple, stable, and widely used baseline for non-image signals like EEG, enabling clearer attribution of gains to our conditioning/guidance components rather than to a more complex diffusion backbone. At inference, we run the reverse process (optionally initialized from a noised seed for editing) to produce edited EEG. Upon visual inspection generated samples exhibit realistic waveforms and preserve inter-channel correlation structure that follows patterns of the real distribution.

3.4 Network architecture and conditioning

The generator is a spatiotemporal 2D U-Net that treats an EEG window as a two-dimensional field (electrodes by time). The U-Net has an encoder-decoder structure with skip connections to preserve fine temporal detail. To capture longer-range dependencies across electrodes and time, attention blocks are inserted at intermediate resolutions. More specifically, the generator's input shape is batch size $\times$ 1 (channel) $\times$ 32 (EEG

channels) × 128 (1-second samples). It uses a diffusion model trained with a noise-prediction objective, where time embeddings modulate features through affine transformations. A 3×3 convolution expands the input channel to a learned feature width. The encoder has resolution levels with channel multipliers (1, 2, 4), each containing residual blocks with GroupNorm, Swish, and 3×3 convolutions. Strided convolutions halve temporal resolution while preserving electrode topology. Fourier features encode timesteps, combined with optional embeddings to form a conditioning vector (see Section 3.4.2). This vector modulates activations via FiLM-style transformations, enabling context-aware convolutions. Attention blocks augment convolutional processing at select resolutions. The decoder mirrors the encoder, using transposed convolutions and skip connections to preserve high-frequency details. The output head includes GroupNorm, Swish, and a 3×3 convolution to match the input shape, with optional interpolation to ensure dimension alignment for diffusion updates.

The model was trained on the DEAP dataset, where the trials in their experiments were recorded at 128 Hz with the standard 32-channel montage from 32 different participants. During preprocessing, the trials were segmented into fixed-length (128 samples, 1 second long), non-overlapping windows by a sliding window to align the generator with the downstream evaluation pipeline. No further preprocessing was required, as the dataset is already published in a cleaned format. Each sample has an assigned binary valence label based on the corresponding valence label coming from the dataset. As the dataset provides labels on a 1-9 range, 5 was chosen as the threshold for the high/low valence threshold point. The resulting dataset consisted of 76800 samples with a 44.7 vs. 55.3% low/high valence distribution, which was then randomly subdivided into a 90-10% training - validation set.

Conditioning is injected with feature-wise affine modulation (FiLM-style Perez et al. (2018)). We build a conditioning vector from (1) an embedding of the diffusion timestep, (2) an optional embedding of the class label, and (3) an optional embedding of the seed signal produced by a lightweight signal encoder. The fused conditioning vector produces per-channel scale and shift parameters that modulate normalized activations inside residual blocks. This design lets one model support unconditional sampling, class-conditional sampling, and seed-conditional sampling.

3.4.1 Valence classifier model

To steer edits toward a target affect label at inference time, we additionally use a separate valence classifier trained on the same 1-second DEAP windows. The classifier takes a 32×128 EEG segment and outputs a sigmoid probability. Architecturally, it is the E2ENNet model (Han et al., 2022): an EEGNet-style temporal–spatial CNN front-end coupled with stacked LSTMs, i.e., a compact CNN+LSTM hybrid suited for EEG classification. More specifically, it processes inputs of shape (32 channels × 128 samples) per segment and produces a single probability via a sigmoid output. The network begins with a temporal convolutional block that applies a 2D convolution (kernel 1×64, 8 filters) independently across channels, followed by batch normalization, to capture short-range temporal dynamics. A subsequent spatial filtering stage employs a depthwise convolution across channels (kernel 32×1, grouped by the previous feature maps) that expands to 16 channels, accompanied by batch normalization, ELU activation, average pooling with a factor of four along time, and dropout (p=0.5). This design separates early temporal and spatial processing to explicitly learn channel-wise spatial projections relevant for EEG decoding. A third stage applies a separable temporal convolution to enlarge the receptive field with minimal cost: a depthwise 1×16 convolution and a pointwise projection reduce to 16 channels, followed by batch normalization, ELU activation, further temporal average pooling (factor eight), and dropout. The resulting feature map is rearranged into a sequence of length determined by the pooled time dimension, with 16 features per step. Two stacked LSTMs then model remaining wider scale temporal dependencies: the first LSTM has a hidden size of 64 and processes the full sequence; its final timestep is reshaped and passed to a second LSTM with hidden size 32. A linear layer maps the final hidden state to a single logit, and a sigmoid converts it to a high-valence probability.

The classifier is used in two roles: 1) evaluation, where we quantify controllability by measuring the high-valence rate of generated/edited samples, and 2) classifier guidance, where we backpropagate the gradient of the target log-probability and inject it into the sampling updates to nudge the diffusion trajectory toward the target valence. Because guidance can amplify classifier-specific cues, we complement it with physiology-grounded checks (e.g., FAA) and explicitly report controllability–fidelity trade-offs.

Data for the model training were obtained from the DEAP dataset, the preprocessing procedure for the trials was unchanged compared to the pipeline implemented for the generator model training. The train-test-validation data split was different here however. The original 76800 samples were divided into a 60/20/20% split, where 20% of the samples were used as a validation set to track overfitting during training and to help select the best model when using early stopping. Another held-out 20% were used as an unbiased test dataset that was utilized to perform the final evaluations for the fully trained model.

3.4.2 Affect editing: class conditioning, classifier-free guidance, and classifier guidance

We investigate multiple ways to enforce the target affect label during sampling: direct class conditioning, classifier-free guidance, and classifier guidance. In direct class conditioning, the diffusion network is conditioned on the target label throughout sampling, with direct FiLM-style Perez et al. (2018) adaptation during training. This provides baseline controllability but can be limited when labels are weak (as we only provide the binary target class with no further guiding information). For classifier-free guidance (CFG), the model is trained with conditioning dropout so it can run in conditional and "unconditional" modes. During sampling, we combine the conditional and unconditional predictions using a guidance scale w_{CFG} as follows:

$$\epsilon_{guided} = \epsilon_{uncond} + w_{CFG} \cdot (\epsilon_{cond} - \epsilon_{uncond}) \tag{1}$$

In our experiments w is evaluated from 0 to 3.0. Increasing w_{CFG} strengthens adherence to the target label but can reduce fidelity to the seed. Finally, for classifier guidance, we also steer sampling using the valence classifier from Section 3.4.1. At each step, we compute the gradient (g_t) of the target-label log-probability with respect to the current sample and apply a scaled update (w_{CG}) that pushes the sample toward high valence. This can provide strong affect edits but may increase distortion if applied too aggressively.

$$\epsilon_{guided} = \epsilon_\theta(x_t, t, \cdot) - w_{CG} \cdot g_t, w_{CG} \in [0, 3]. \tag{2}$$

3.5 Seed retention guidance: start-noise control, signal CFG, and STFT-L1 guidance

To preserve seed structure, we use three complementary retention mechanisms: Signal conditioning with optional signal CFG, STFT-L1 guidance, and start-noise control. Signal conditioning with optional signal CFG: When a seed embedding is provided as conditioning, the model can learn to preserve aspects of the seed. Similar to class CFG, we optionally amplify the effect of signal conditioning during sampling with a signal-guidance scale $w_{sig} \in [0, 3]$ (0: no extra retention; larger values here imply a stronger pull toward the seed), providing a retention lever without explicit reconstruction losses. STFT-L1 guidance: In initial tests, we saw that signal conditioning minimized the error in time domain, but did not lead to a good fit in the spectral domain. This motivated the introduction of a loss specifically focused on spectral similarity. We define an explicit spectrotemporal retention loss between the current sample and the seed using magnitude STFTs. For each channel we compute STFT with a Hann window (window length=64 samples, hop length=32, nfft=64, sampling rate=128 Hz), and define

$$L_{STFT}(\hat{x}_0, x_{seed}) = \text{mean}_{c,f,\tau} ||\text{FTFT}(x_0) \cdot c(f, \tau)| - |\text{FTFT}(x_{seed}) \cdot c(f, \tau)|| \tag{3}$$

averaged over channels and time–frequency bins. At each denoising step t we first form a provisional clean estimate x_0 from the U-Net prediction, then compute

$$g_{STFT} = \nabla_{xt} L_{STFT}(x_0(x_t), x_{seed}) \tag{4}$$

by backpropagating through x_0, and apply it as an additional guidance term:

$$\epsilon_{guided} = \epsilon_{pred} - w_{STFT} \cdot g_{STFT}, w_{STFT} \in [0, 3]. \tag{5}$$

This encourages preservation of bandpower and temporal modulation while being relatively insensitive to phase compared to pointwise waveform losses.

Finally, for start-noise control, we initialize the reverse process from a noised version of the seed instead of pure noise. A start-noise parameter s in $[0, 1]$ selects the starting diffusion step t_{start} (approximately s times the maximum step). We forward-noise the seed to that step and then denoise back to step 0. Small s keeps sampling close to the seed (strong retention, weaker edits). Large s allows stronger modifications (weaker retention, stronger edits). This provides a simple and interpretable retention knob that can be combined with any affect-steering method.

3.6 Combined control during sampling

The final editing pipeline uses classifier guidance for affect steering together with start-noise control for seed retention, since these two controls performed best in the individual evaluations. We evaluate the combined editor by sweeping the start-noise level (with the classifier-guidance scale fixed at the most effective setting) and reporting target-label adherence with the valence classifier alongside seed-fidelity metrics (STFT-L1 and time-domain MSE).

Figure 2 shows the output of the final model, demonstrating how a given seed signal is edited.

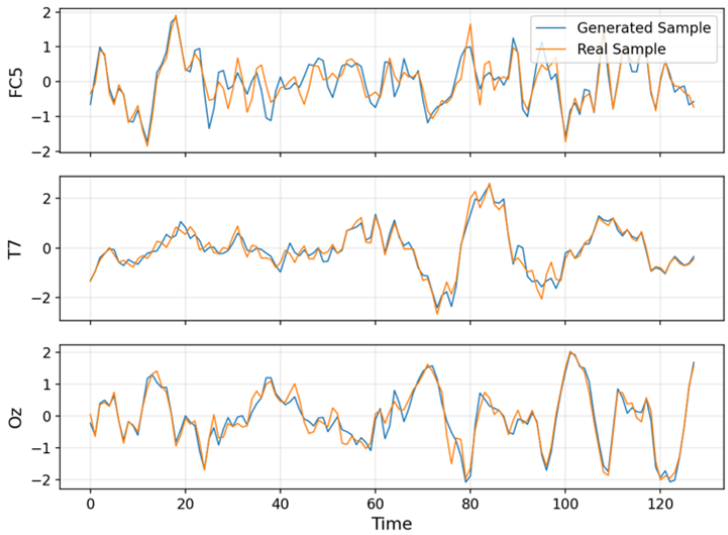

Figure 2: Comparison of three channels of a seed (orange) and a corresponding, successfully edited, generated sample (blue)

4 Results

We report results for low→high valence EEG affect editing as a case study, emphasizing the controllability–retention trade-off for high-level edits. We measure 1) target-label adherence as the percentage of generated/edited samples classified as high valence by a valence classifier, 2) seed fidelity using STFT-L1 distance and time-domain MSE to the seed, and 3) out-of-distribution task-structure retention on SSVEP using an FBCCA frequency recognizer. To study neural plausibility, we consider the Frontal Alpha Asymmetry (FAA). FAA measures the difference in alpha-band power between the left and right frontal hemispheres of the brain and has been associated with emotional processing. All results are computed over 1000 generated samples per setting.

4.1 Affect steering (class guidance)

Direct class conditioning alone provides only moderate controllability. When sampling from pure noise, internal class conditioning produces 67.1% high-valence samples, compared to a 57.0% unconditional baseline. Classifier-free class guidance provides a small improvement over this baseline when tuned: across guidance scales w in $\{0.0, 0.5, \ldots, 3.0\}$, high-valence adherence ranges from 65.7% to 71.9%, with the best result of 71.9% at $w = 0.5$. In comparison, classifier guidance is substantially stronger and largely saturates. Sweeping classifier guidance scales in $\{0.0, 0.1, \ldots, 3.0\}$ yields 95.1%–99.3% high-valence adherence, with the best observed value 99.3% at guidance scale 0.7. Based on these results, classifier guidance is selected as the default affect-steering mechanism in the combined editor.

Regarding physiological plausibility (FAA) checks during editing, on DEAP, the average dataset-level low-vs-high valence difference is a +0.0087 mean FAA shift. In our low→high editing runs, FAA shifts were negative across class-steering methods: standard internal class conditioning -0.072, classifier-free class guidance ($w = 0.5$) -0.058, and classifier guidance (w=0.7) -0.061. These FAA shift measurements challenge the results coming from the valence classifier model, as they oppose the natural and expected directional shift present across the real DEAP samples.

4.2 Seed retention guidance (signal guidance)

From a fully noised start, direct signal conditioning yields limited retention of the original signal: average STFT-L1 = 1.98 and MSE = 1.52 relative to the seed. Classifier-free signal guidance improves retention slightly. Sweeping signal-guidance strength in $\{0.0, 0.5, \ldots, 3.0\}$, the best reconstruction is achieved at strength 0.5 with STFT-L1 = 1.77 and MSE = 1.40. As expected, STFT-L1 gradient guidance yields the best spectral fidelity at strength 3.0 (STFT-L1 = 0.431), but at the cost of higher waveform error (MSE = 2.181).

Start-noise control provides the most predictable retention knob. With start-noise $s = 0.0$ (one diffusion step), retention is near-identity: STFT-L1 = 0.01 and MSE = 0.004. Increasing start-noise yields a smooth, monotonic increase in both STFT-L1 and MSE, enabling direct control over how strongly the edited output is anchored to the seed, as seen in Figure 3.

4.3 Combined editing on DEAP (in-distribution)

For the final evaluation of all components, we evaluate combined editing on DEAP low valence samples using classifier guidance for affect steering (scale 0.7) together with start-noise control for retention. Even with minimal noising, the editor achieves strong affect transposition while remaining close to the seed. At 2% noising level (20 diffusion steps), edited samples are classified as high valence in 98.6% of cases, while preserving high seed fidelity with STFT-L1 = 0.07 and MSE = 0.01. As start-noise increases further, seed fidelity degrades as expected, while valence adherence remains high (i.e., improvements saturate because classifier guidance is already strong). Figure 3 summarizes these results for the combined edit evaluation on DEAP.

4.4 Out-of-distribution (OOD) editing on SSVEP (task-structure costs)

To quantify how affect editing impacts non-affective structure under distribution shift, we apply the same combined editor as identified in the previous section to the Benchmark SSVEP dataset. Here, the model is trained on DEAP data exclusively and the SSVEP data is used as seed signal for evaluation. For a successful editing of valence

To measure whether the non-affective components in the seed signal are preserved by the editing of valence, we evaluate stimulus frequency preservation with Filter Bank Canonical Correlation Analysis (FBCCA). This is a classification approach for identifying the stimulated frequency in SSVEP BCI and we use the accuracy of FBCCA as a proxy of signal retention, similarly to how we use the accuracy of valence classification as a proxy for the editing success. Beforehand, we resampled the data to 125Hz and reduced the data set to channels that matched the DEAP formatting. Afterwards, we evaluated the FBCCA recognizer on original

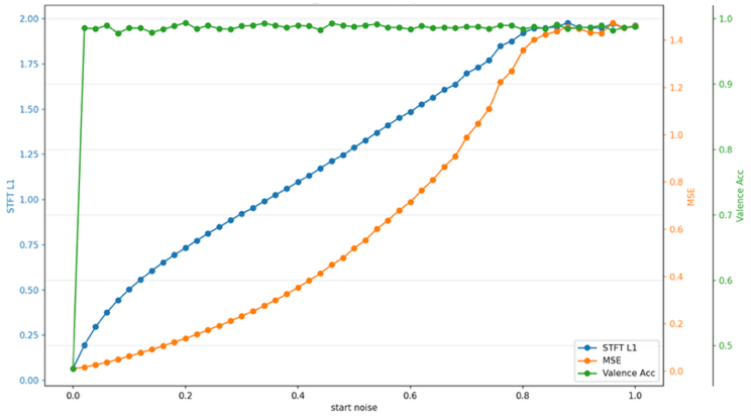

Figure 3: Combined edit evaluation on DEAP samples. X axis: start noise setting, Y axes: STFT-L1 distance (blue), MSE (orange), High Valence Classification Ratio (green).

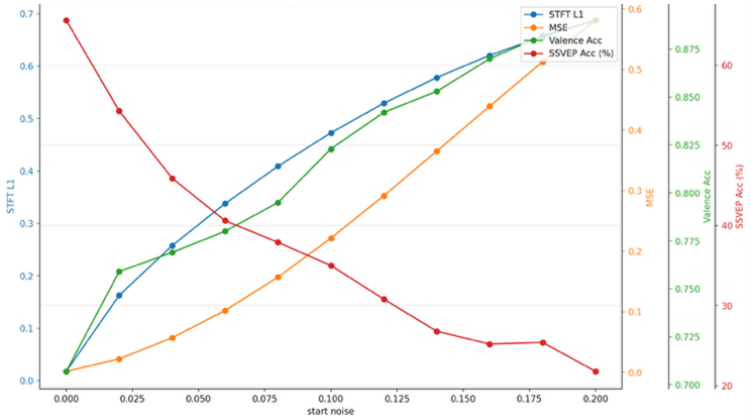

Figure 4: Combined edit evaluation on SSVEP samples X axis: start noise setting Y axes: STFT-L1 distance (blue), MSE (orange), Valence Classification Accuracy (green), SSVEP Classification Accuracy (red)

samples with integer frequency labels, which yielded a baseline accuracy of 68%. Under minimal edits (low start-noise), SSVEP structure is largely preserved. With 20 noising/denoising steps, edited samples reach 70.7% high-valence recognition while FBCCA exact accuracy remains 65.6% (close to baseline). Figure 4 summarizes the results and illustrates that with stronger edits, a clear trade-off emerges: when increasing the starting noise further, high-valence recognition increases to 89.0%, while FBCCA accuracy drops to 21.8%, indicating substantial degradation of stimulus-locked periodic structure.

5 Discussion

This work uses affect transposition as a case study to probe a broader question: can diffusion models perform controllable, high-level edits on single-trial EEG while retaining the seed's non-target structure? Across in-distribution (DEAP) and out-of-distribution (SSVEP) evaluations, the results support a consistent picture: classifier guidance can effectively steer the edited samples toward the target semantic attribute, while diffusion start-noise control can provide a simple lever to trade off edit strength against seed fidelity. However, achieving a dataset-agnostic editor remains challenging: the same editing pipeline performed substantially better on the in-distribution DEAP dataset than on the Benchmark SSVEP dataset. Under all conditions, when edits are constrained by low start-noise (i.e., initializing denoising close to the seed), the generated trials remain near-identity under spectral–temporal distances, and affect-unrelated structure is comparatively well preserved. Increasing start-noise yields stronger freedom to modify the signal and therefore stronger

attribute changes, but also increases spectral deviation and eventually in the case of OOD samples, induces measurable degradation of non-affective task structure (as FBCCA frequency recognizability shows). This behavior is consistent with the intuition that larger diffusion trajectories provide more "creative capacity" but reduce the guarantee of preserving structured, stimulus-locked patterns. Among the affect steering methods, classifier guidance most effectively shifts the recognized valence label, outperforming model-internal label conditioning and classifier-free guidance.

As for seed-fidelity guidances, STFT-L1 guidance improves spectral fidelity but can increase waveform error (MSE), while start-noise provides a simple, model-agnostic knob that smoothly trades off seed retention against edit strength. Evidence for affect transposition beyond classifier-based evaluation was inconclusive. In particular, frontal alpha asymmetry (FAA) did not show consistent shifts in the expected direction during editing, even when samples were strongly steered toward the target label by the valence classifier. One plausible explanation is that the classifier may rely on features that are correlated with the dataset labels but not uniquely tied to affective valence itself. Regardless, classifier guidance proved effective as a control mechanism: it reliably steered edited samples toward the target classifier output, demonstrating that high-level semantic control—as defined by the learned decision boundary—can be exerted within the diffusion framework. To evaluate preservation of non-affective signal structure, we applied the same editing procedure to an out-of-distribution SSVEP dataset with clearly defined task-related dynamics. Using FBCCA-based frequency recognition, we observed that stimulus-locked structure is largely preserved under minimal editing, while stronger edits progressively degrade recognition accuracy. This behavior highlights a controllable trade-off between semantic steering and signal retention: classifier guidance enables effective label manipulation, but preserving unrelated neural structure requires explicit constraints on the amount of noising-denoising step during synthesis.

5.1 Limitations

First, we use low→high valence transposition as a convenient case study for "high-level" EEG editing, but success on valence does not necessarily imply that the same controls will transfer to other semantic attributes whose signatures are structurally different (e.g., motor imagery or movement-related activity, attentional engagement, or task-intention), and which may require different notions of fidelity and plausibility than spectral proximity to a seed. Second, preservation of task structure was quantified only in an out-of-distribution setting (SSVEP with FBCCA), so we do not yet characterize "retention" systematically across in-distribution tasks or multiple well-defined task paradigms, where different invariances might matter and degradation could manifest differently. Third, our strongest edits rely on classifier guidance and are primarily evaluated with that same type of discriminative signal (plus a single physiology-inspired check via FAA); this leaves open the possibility of classifier-specific shortcut features, and the observed FAA trends suggest that achieving target-label adherence may not always align with established valence correlates, motivating broader neurophysiological validation and additional, task-specific retention metrics in future work. Fourth, editing a target attribute may induce unintended changes in correlated affective dimensions; future work should measure cross-attribute leakage (e.g., monitor valence/arousal while editing workload).

6 Conclusion

Overall, the paper's main contribution is a diffusion-based EEG-to-EEG editor that can be made controllable and evaluated against non-target preservation criteria, with affect transposition serving as a concrete case study where editing viability appears promising. The combination of 1) explicit controllability knobs, 2) seed-fidelity metrics, 3) physiological sanity checks, and 4) OOD task-structure tests provides a template for studying high-level EEG editing beyond this single case study.

7 Code and Data availability

All code and its documentation are available on Github [GitHub repository—anonymized] as well as in the Supplementary Materials. Both datasets used are publicly available.

References

Xiaogang Chen, Yijun Wang, Shangkai Gao, Tzyy-Ping Jung, and Xiaorong Gao. Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface. 12(4):046008, 2015. URL https://iopscience.iop.org/article/10.1088/1741-2560/12/4/046008/meta. Publisher: IOP Publishing.

James A. Coan and John JB Allen. Frontal EEG asymmetry as a moderator and mediator of emotion. 67(1):7–50, 2004. URL https://www.sciencedirect.com/science/article/pii/S0301051104000316. Publisher: Elsevier.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: a review. 16(3):031001, 2019. ISSN 1741-2552. doi: 10.1088/1741-2552/ab0ab5. URL https://doi.org/10.1088/1741-2552/ab0ab5. Publisher: IOP Publishing.

Hendrik Eilts and Felix Putze. Is that real? a multifaceted evaluation of the quality of simulated eeg signals for passive bci. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2639–2644. IEEE, 2022.

Andéol Evain, Ferran Argelaguet, Nicolas Roussel, Géry Casiez, and Anatole Lécuyer. Can i think of something else when using a bci? cognitive demand of an ssvep-based bci. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 5120–5125, 2017.

Zhichao Han, Hongli Chang, Xiaoyan Zhou, Jihao Wang, Lili Wang, and Yongbin Shao. E2ennet: An end-to-end neural network for emotional brain-computer interface. 16:942979, 2022. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2022.942979/full. Publisher: Frontiers Media SA.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 33:6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Guido Klein, Pierre Guetschel, Gianluigi Silvestri, and Michael Tangermann. Synthesizing EEG signals from event-related potential paradigms with conditional diffusion models. 2024. doi: 10.3217/978-3-99161-014-4-077. URL http://arxiv.org/abs/2403.18486. Place: 9th Graz BCI conference.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. 3(1):18–31, 2011. URL https://ieeexplore.ieee.org/abstract/document/5871728/. Publisher: IEEE.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. 15 (5):056013, 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c. Publisher: IOP Publishing.

Xiaomin Li, Mykhailo Sakevych, Gentry Atkinson, and Vangelis Metsis. Biodiffusion: A versatile diffusion model for biomedical signal synthesis. 11(4):299, 2024. URL https://www.mdpi.com/2306-5354/11/4/299. Publisher: MDPI.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671. Number: 1.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pp. 234–241. Springer International Publishing, 2015. ISBN 978-3-319-24573-7 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28. URL http://link.springer.com/10.1007/978-3-319-24574-4_28. Series Title: Lecture Notes in Computer Science.

Giulio Tosato, Cesare M. Dalbagno, and Francesco Fumagalli. EEG synthetic data generation using probabilistic diffusion models, 2023. URL http://arxiv.org/abs/2303.06068.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zhijiang Wan, Qianhao Yu, Wujie Dai, Siyue Li, and Jin Hong. Data generation for enhancing EEG-based emotion recognition: Extracting time-invariant and subject-invariant components with contrastive learning. 2024. URL https://ieeexplore.ieee.org/abstract/document/10556575/. Publisher: IEEE.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for SSVEP-based brain–computer interfaces. 25(10):1746–1752, 2016. URL https://ieeexplore.ieee.org/abstract/document/7740878/. Publisher: IEEE.

Meng Xu, Yuanfang Chen, Yijun Wang, Dan Wang, Zehua Liu, and Lijian Zhang. BWGAN-GP: An EEG data generation method for class imbalance problem in RSVP tasks. 30:251–263, 2022. URL https://ieeexplore.ieee.org/abstract/document/9690467/. Publisher: IEEE.