EXPLAINING TYPICAL SUBJECT-INVARIANT EEG PATTERNS OF EMOTION THROUGH DEEP LEARNING COUNTERFACTUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual explanations of electroencephalography (EEG) deep learning models may provide actionable insights for downstream applications such as neuroadaptive brain-computer interfaces, user experimentation, and psychological assessment. However, counterfactual changes in EEG affective state classification are often incompatible with target-class distributions and may produce unrealistic EEG patterns. This complicates the interpretation of EEG-based affect models, particularly when they must generalize across participants. We propose a framework for generating counterfactual explanations under distributional and physiological constraints. We start by introducing a joint variational classifier-decoder architecture that combines affect prediction and EEG feature reconstruction with adversarial regularization to reduce subject-unique information. Then, we propose a counterfactual objective that combines a target-class prediction loss with class-conditional density regularization, proximity penalties in latent and reconstructed feature spaces, and physiological constraints on decoded EEG features. This objective explicitly favors target-class typicality while expressing changes in interpretable EEG frequency-band features. In a strict subject-independent setting, our classifier achieves highly competitive 56.04% valence and 57.50% arousal **balanced accuracies** on the DREAMER dataset. Typicality-constrained optimization improves counterfactual alignment with the learned target-class distribution relative to an objective without this constraint. All counterfactual EEG patterns generated for held-out subjects had distances to the target class within the 95th percentile of those observed for real target-class samples across subjects, supporting subject-independent generation of counterfactual patterns. These results suggest that typicality optimization alongside prediction change can produce more interpretable EEG affect counterfactuals that generalize across participants.

1 INTRODUCTION

Electroencephalography (EEG) supports applications ranging from clinical and psychological assessment to brain-computer interfaces (BCIs) that infer cognitive, affective, or motor states. Neuroadaptive BCIs are applications that use estimated user states to modify an interface in real time, with potential benefits for adaptive games, education, and mixed-reality experiences (Ewing et al., 2016; Kosmyna & Maes, 2019; Bernal et al., 2022). As wearable and consumer EEG systems become easier to integrate with everyday devices, EEG-based affect models may also support user experimentation informed by physiological measurements (Gaspar-Figueiredo et al., 2023). These applications motivate the understanding of how EEG features relate to model predictions while accurately classifying affective states across participants. Counterfactual explanations (CFX), which identify the changes in features required to alter a given model's prediction, offer an approach to identify the relationship between EEG features and model predictions (Wachter et al., 2018). However, such changes must be physiologically plausible and compatible with the target-class distribution to support meaningful interpretation. This challenge is particularly relevant when models must generalize across subjects using noisy consumer-grade and sparse electrode setups (Katsigiannis & Ramzan, 2018).

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

While a large body of work has applied deep learning (DL) architectures to publicly available EEG datasets for cognitive and affective state classification, comparatively less work has addressed the interpretability of these models. In this paper, we focus on strict subject-independent settings that support generalization to unseen participants (Khan et al., 2025; Del Pup et al., 2025). Efforts to improve predictive performance in these settings have produced increasingly complex architectures (Liu et al., 2021; Prabowo et al., 2023), making it difficult to determine which learned features and representations contribute to a model's decisions. Explainability methods can help characterize these predictive dependencies and inform how model outputs are used in downstream applications (Kindermans et al., 2019; Ravindran & Contreras-Vidal, 2023; Pinto et al., 2023). However, methods developed for other DL domains do not always transfer reliably to EEG because neural recordings are noisy, high-dimensional, and strongly influenced by preprocessing and feature construction (Ravindran & Contreras-Vidal, 2023; Khan et al., 2025). Gradient-based saliency methods are particularly susceptible to these issues because their attributions may reflect properties of the input representation or preprocessing pipeline rather than stable predictive dependencies learned by the model (Kindermans et al., 2019; Haufe et al., 2014). The reliability of backpropagation-based explanations also depends on how the explanation method propagates information through the network, further complicating their interpretation (Sixt et al., 2020).

These limitations motivate complementary explanation methods that characterize how changes in interpretable EEG features would alter a model's prediction. Counterfactual explanations provide a framework for examining such changes (Guidotti, 2024) and aim to offer actionable insights (Mothilal et al., 2020; Pinto et al., 2023), with potential applications in neurofeedback and user experimentation. We investigate this approach through reconstructed latent representations learned for EEG valence and arousal prediction on DREAMER under strict subject-independent evaluation. Beyond prediction changes, we examine how counterfactual representations align with learned class distributions and explore subject-invariant typicality across the dataset.

The contributions of this study are listed as follows:

1. a combination and adaptation of multiple state-of-the-art techniques to create a competitive subject-invariant EEG valence-arousal variational prediction model while advocating for the use of balanced accuracy (BA) as a primary metric (see Appendix A.4.1),
2. a distributional typicality counterfactual optimizer that targets class changes in latent space and reconstructs them into EEG feature space through decoder traversal, and
3. a proximally and physiologically constrained generative optimization target function for faithful, valid, and plausible features.

To the best of our knowledge, prior work has not combined variational learning for EEG emotion prediction with typicality-aware counterfactual optimization (CFO) under subject-independent evaluation. We interpret the generated counterfactuals as model-level explanations and hypothesis-generating descriptions of electrode frequency and temporal patterns, rather than as direct evidence of causal changes in cortical generators. Establishing the latter would require additional neurophysiological and interventional validation.

Implementations of our DL architectures and counterfactual optimization method are released through the open-source *anonymized* library.

2 Related Work

2.1 Deep Learning and EEG

Autoencoding and Feature Fusion. Traditional EEG pipelines use handcrafted features such as band power, power spectral density, and entropy (Liu et al., 2021), whereas recent models learn spatio-temporal and spatio-spectral representations through autoencoding and feature fusion (Li et al., 2023b;a; Hilali et al., 2025). This motivates our jointly trained encoder–decoder architecture, which learns representations while preserving a reconstruction path for subsequent traversal.

Classification Targets. Meta-learning for Domain Generalization (MLDG) treats participants as separate domains and favors updates that generalize from meta-training to meta-test subjects (Li

et al., 2018). We use its first-order formulation without accessing data from the participant that is held out for evaluation. Then, variational classification (VC) combines discriminative prediction with class-conditional latent regularization, providing a probabilistic representation of class structure and uncertainty (Dhuliawala et al., 2023). While Dhuliawala et al. (2023) suggest using adversarial learning for discrimination, we assume Gaussian distribution of EEG features (see Appendix A.1). Later, we decompose the classification target for our counterfactual target to distinguish target-class validity from distributional typicality (see Section 3.4.1). Additionally, focal loss may downweight easily classified observations and direct optimization toward difficult or ambiguous examples (Lin et al., 2017; Li et al., 2023a). Unlike Li et al. (2023a), we choose to discard the class-weighting term α given MLDG promotes balanced learning.

Subject-Invariant Representation Learning. We use **subject-invariance** to describe the aim of learning emotion-relevant representations with reduced participant-specific structure. Operationally, we apply subject-adversarial training to encourage reduced recoverability, but neither the training objective nor a low score from one subject head establishes complete independence from subject-specific structure. Subject-adversarial learning discourages representations from retaining participant-identifying information that may reflect individual spectral profiles, recording conditions, or artifacts rather than generalizable affective structure (Özdenizci et al., 2020; Del Pup et al., 2025). Our inverse subject adversarial head applies this principle to the recurrent trial representation through gradient reversal (Ganin et al., 2016). We complement this objective with supervised contrastive learning, which draws same-class embeddings together while separating different classes (Khosla et al., 2020). We select positive pairs to share an emotion label but originate from different participants. Together with MLDG, these objectives organize the latent space around affective structure that transfers across subjects and is subsequently traversed by our counterfactual method.

2.2 Explainability in Deep Learning

DL models often achieve strong predictive performance at the cost of transparency, making it difficult to determine which aspects of the input drive a given prediction (Kindermans et al., 2019). Explainability methods aim to address this problem by producing descriptions of model behavior. In practice, two major families are receiving attention in EEG predictive modeling: **saliency-based** methods, which assign importance scores to input features for a fixed prediction (Simonyan et al., 2013; Sundararajan et al., 2017), and **counterfactual** methods, which search for a nearby alternative input that would change the model output (Wachter et al., 2018; Mothilal et al., 2020). Saliency-based methods are useful for identifying which regions, channels, or features influence a prediction, while counterfactual methods are useful for characterizing how the input would need to change in order to obtain a desired prediction. This makes CFX particularly interesting for psychological diagnostics where it may show desired changes, motor imagery where it may map brain to motor divergences (Yin et al., 2024), user experimentation where it may guide design decisions on iterations, and neurofeedback where it may find the desired EEG patterns for approximation.

Counterfactual Optimization. Wachter et al. (2018) formulate CFO as the search for the smallest change to an input x that produces a desired output y_{o}. While their proposed target balances validity and proximity (Guidotti, 2024), these metrics alone do not guarantee a useful counterfactual. Unconstrained optimization can produce off-manifold, unjustified, or unstable examples (Laugel et al., 2019; Jiang et al., 2024). Practical formulations consequently augment Wachter et al. (2018)'s counterfactual target with regularizers for properties such as diversity, sparsity, plausibility, robustness, and temporal coherence (Mothilal et al., 2020; Wang et al., 2024; Sevellec et al., 2025; Refoyo & Luengo, 2026; Kostrzewa et al., 2026). Lastly, Lim et al. (2025) define counterfactual typicality as the proportion of observations from the target class contained within a hyperball centered on the counterfactual, providing a proximity measure instead of distributional typicality.

Counterfactuals in EEG Modeling. CFX remains relatively uncommon in EEG, but the literature spans different tasks and representations. Pinto et al. (2023) generated feature-level counterfactuals for seizure prediction by identifying minimal feature changes that reversed a prediction. Yin et al. (2024) replaced time–frequency patches with features from an alternative motor-imagery class until the classifier decision changed. More recent work has used causal counterfactual analysis to distinguish causal from associative EEG patterns in autism evaluation using adversarial generative

CFO (Wang et al., 2026). Another study modeled seizure recordings as spatio-temporal graphs to obtain plausible and temporally robust explanations (Zannotti et al., 2026). With respect to affective modeling, Zhu et al. (2026) used counterfactual training combined with causal discovery to enhance multimodal emotion recognition. Zhu et al. (2026) also provide a counterfactual physiological constraint that is methodologically different from the one we propose in Section 3.4.

3 METHODOLOGY

3.1 DATASET DETAILS AND DATA PREPARATION

We evaluate the proposed framework on DREAMER, an affective-computing dataset comprising 128 Hz EEG recordings from 23 participants viewing 18 emotion-eliciting audiovisual stimuli with a 14-channel Emotiv EPOC headset (Katsigiannis & Ramzan, 2018). After each stimulus, participants rated valence, arousal, and dominance on a five-point scale. We retain valence and arousal and define separate binary tasks, assigning ratings ≥ 3 to the high class and ratings < 3 to the low class. DREAMER was selected because its moderate cohort supports comprehensive leave-one-subject-out cross-validation (LOSOCV) and repeated CFO, while its extensive and recent usage provides a prediction benchmark (Liu et al., 2021; Prabowo et al., 2023; Hilali et al., 2025; Hua et al., 2026; Zhang et al., 2026). It also utilizes a sparse, consumer-grade EEG headset that is typically more susceptible to noise, placement, and movements, making it ideal for uncertainty estimation. Finally, emotion recognition is a demanding test of typicality-aware explanation because affective states are abstract, and self-reported labels are subjective and highly variable across subjects (Özdenizci et al., 2020; Del Pup et al., 2025).

For each subject and trial, samples are chronologically ordered and checked for continuity, finite values, and label consistency. The 14 channels are detrended, notch-filtered at 50 Hz, and decomposed with fourth-order zero-phase Butterworth filters into θ (4–8 Hz), α (8–13 Hz), and β (13–30 Hz) waveforms. Filtering is applied to the full contiguous stimulus before extracting the central 60 s to reduce boundary artifacts, yielding 7,680 samples per channel at 128 Hz. Treating each electrode–band pair as a separate input produces 42 channel-major features and an EEG tensor of size $23 \times 18 \times 42 \times 7{,}680$. Each trial is then divided into non-overlapping 1-s windows of 128 samples, while trial identities and labels are preserved only for trial level aggregation and subject-independent evaluation. Trial identities are not included in training.

3.2 CROSS-VALIDATION SCHEME

We use subject-independent LOSOCV with 23 outer folds. In each fold k, one participant is excluded from model optimization, while the remaining 22 participants form the source pool. Subject-level separation prevents windows from the held-out participant from entering the source-training episodes (Khan et al., 2025; Del Pup et al., 2025). Each source epoch comprises multiple MLDG episodes, with episode sizes limited by memory constraints. For each episode, meta-test participants are selected from the source pool, followed by disjoint meta-training participants. Participants omitted from an episode remain eligible for subsequent episodes. Each source epoch contains 12 episodes, each comprising 12 meta-training participants and six meta-test participants, with two complete trials per participant. Trial sampling balances the two classes within each group where each participant contributes one trial from each class. Each episode performs one temporary inner gradient update using the meta-training participants, followed by a persistent outer update incorporating both meta-training and meta-test gradients. Source training occurs for ten epochs. For each affective dimension $d \in \{v, a\}$, latent space features are concatenated at the trial-level for full-trial zero-shot predictions evaluated on the held-out subject.

3.3 SPATIO-SPECTRAL-TEMPORAL ENCODING TO MULTI-HEAD ARCHITECTURE

As illustrated in Fig. 1, each 60-s trial is represented as an ordered sequence $\mathbf{X}_{s,t} \in [\mathbf{x}_1, \ldots, \mathbf{x}_w]$, where s, t, and w denote the subject, trial, and 1-s window. We use a (1) Bidirectional Long Short-Term Memory (BiLSTM) model to produce 126 spatio-temporal features and we use a (2) Graph Convolutional Network-Gated Recurrent Unit (GCN-GRU) inspired by Li et al. (2023a) to produce 384 spatio-spectral features. Each encoder maps each window to the parameters of a latent posterior

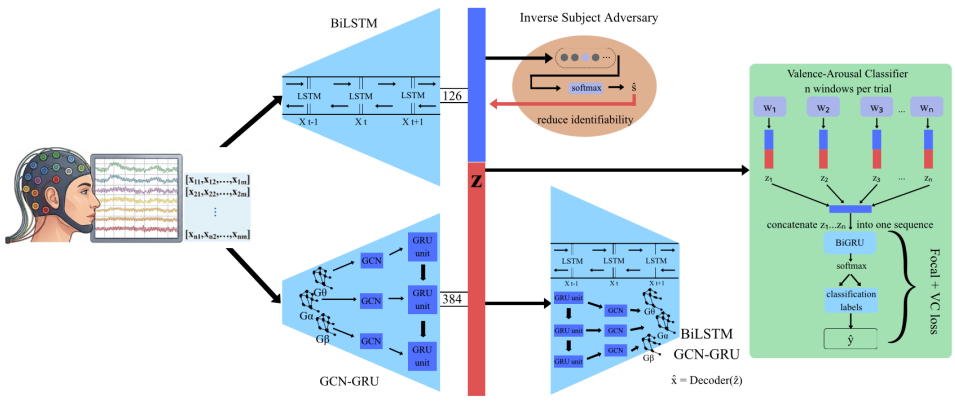

Figure 1: Deep learning model workflow. Each trial is divided into ordered EEG windows that are encoded into variational latent embeddings. The embeddings are processed sequentially by the variational classifier in full trials, while the shared decoder reconstructs the channel–band waveforms and enables counterfactual decoder traversal. MLDG controls batch sampling for fair optimization.

$q_\phi(\mathbf{z}_w \mid \mathbf{x}_w)$. Latent embeddings are then concatenated across full trials and passed to a bidirectional gated recurrent unit (BiGRU) so that classification incorporates full temporal resolution of fluctuating emotions across a trial. The classifier predicts the probability of high valence or arousal, and multiple latent samples are averaged during inference to account for posterior uncertainty.

The decoder, classifier, and ISA are optimized jointly according to

$$\mathcal{L}_{\text{joint}} = \left(\lambda_{\text{AE}}\mathcal{L}_{\text{AE}} + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}}\right) + \left(\lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{VC}}\mathcal{L}_{\text{VC}}\right) + \lambda_{\text{ISA}}\mathcal{L}_{\text{ISA}}, \tag{1}$$

where $\mathcal{L}_{\text{AE}}$ preserves autoencoder (AE) reconstructive structure while including the supervised contrastive loss $\mathcal{L}_{\text{sup}}$ (described in Section 2.1), $\mathcal{L}_{\text{focal}}$ and $\mathcal{L}_{\text{VC}}$ promote discriminative trial-level predictions, $\mathcal{L}_{\text{ISA}}$ promotes subject-invariance, and λ terms refer to non-negative weight hyperparameters. Joint optimization encourages embeddings that are sufficiently structured for reconstruction, predictive of affective state, and subject-invariant.

The VC regularizer may be decomposed as $\mathcal{L}_{\text{VC}} = \beta\mathcal{L}_{\text{posterior}} + \lambda\mathcal{L}_{\text{prior}}$, where

$$\begin{aligned}\mathcal{L}_{\text{posterior}} &= \mathbb{E}_{p(y)}\left[D_{\text{KL}}\left(q_\phi(z \mid y) \,\|\, p_\theta(z \mid y)\right)\right] = \\ &- \mathbb{E}_{p(y)q_\phi(z|y)}\left[\log p_\theta(z \mid y)\right] - \mathbb{E}_{p(y)}\left[\mathcal{H}\left[q_\phi(z \mid y)\right]\right].\end{aligned} \tag{2}$$

Weights β and λ are non-negative. This term aligns the empirical class-conditional latent distribution with the distribution anticipated by the classifier (see Appendix A.1). Not only does this term allow modeling uncertainty, but it will later promote typicality analysis in CFO. The expected entropy term encourages the learned distribution q_ϕ supported by the anticipated distribution p_θ, while the entropy term $\mathcal{H}$ discourages collapse of within-class variation. The remaining term, $\mathcal{L}_{\text{prior}} = D_{\text{KL}}(p(y)\|p_\pi(y))$, supports learning. The target classifier estimates the binary affective state associated with the EEG window $\hat{p}_w = p_{\text{C}}\left(y_w = 1 \mid \mathbf{Z}_{s,t}\right)$. Here, $y_w \in \{0, 1\}$, where $\{0 = low, 1 = high\}$, denotes the binary valence or arousal target inherited from the corresponding trial-level rating. $\mathbf{Z}_{s,t}$ denotes the trial-level concatenated feature embedding.

3.4 Constrained Counterfactual Optimization

In our architecture, optimization is performed in latent space, and a jointly trained decoder maps the optimized representations into reconstructed EEG features for interpretation. To do this, we propose the joint architecture in order to preserve a decoder that allows the reconstruction of the subject-invariant counterfactual features (see Appendix B.1). The CFO process described in Equation 4 moves a latent representation toward a target class probability, after which decoder traversal maps the optimized latent point into reconstructed EEG channel–band waveform space for interpretation. Therefore, reconstructed counterfactual feature representations must be measured both for domain-wise fidelity and for physiological realism, motivating the integration of validity, faithfulness, and plausibility constraints (Mothilal et al., 2020; Wang et al., 2024; Sevellec et al., 2025).

Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_W) \in \mathbb{R}^{W \times T \times m}$ denote the complete encoded trial, where $\mathbf{z}_w \in \mathbb{R}^{T \times m}$ contains the T encoded timesteps of window w. The windows are concatenated in chronological order along the time axis before being passed to the frozen BiGRU:

$$C_\theta(\mathbf{Z}) = \mathrm{BiGRU}_\theta\left(\begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\\ \vdots\\ \mathbf{z}_W\end{bmatrix}\right) \in \mathbb{R}^d. \tag{3}$$

For binary valence or arousal classification, the counterfactual target is defined as the class opposite to the model's original prediction. Beginning from the starting latent space $\mathbf{Z}$, gradient-based optimization produces a counterfactual sequence $\mathbf{Z}^{\mathrm{cf}}$ that increases the probability of the target class while limiting its deviation from the original representation:

$$\begin{aligned}\mathcal{L}_{\mathrm{cfo}} = \lambda_o \mathcal{L}_{\mathrm{o}}\left(f(\mathbf{Z}'), y_{\mathrm{o}}\right) + \lambda_z d_z(\mathbf{Z}, \mathbf{Z}') + \lambda_R d_R\left(R(\mathbf{Z}), R(\mathbf{Z}')\right) + \lambda_{\mathrm{ph}} \mathcal{P}_{\mathrm{ph}}\left(R(\mathbf{Z}')\right),\\ \mathbf{Z}^{\mathrm{cf}} = \arg\min_{\mathbf{Z}'} \mathcal{L}_{\mathrm{cfo}}.\end{aligned} \tag{4}$$

In Equation 4, $\mathbf{Z}$ denotes the observed latent sequence, $\mathbf{Z}'$ the optimization candidate, and $\mathbf{Z}^{\mathrm{cf}}$ the selected counterfactual, respectively. The fixed classifier f and decoder R map $\mathbf{Z}'$ to the predicted class and EEG signal space. The target loss $\mathcal{L}_{\mathrm{o}}$ promotes the desired class y_{o}, while $d_z(\mathbf{Z}, \mathbf{Z}')$, $d_R(R(\mathbf{Z}), R(\mathbf{Z}'))$, and $\mathcal{P}_{\mathrm{ph}}(R(\mathbf{Z}'))$ preserve latent proximity, signal-level reconstruction similarity, and physiological plausibility. Their contributions are controlled by λ_o, λ_z, λ_R, and λ_{ph}. Only $\mathbf{Z}'$ is updated and all model components remain fixed. Optimization terminates when the target-confidence threshold or iteration limit is reached.

We then assess two essential metrics for post-hoc explainability. (1) **Validity**, following Wang et al. (2024) and Sevellec et al. (2025), is the proportion of counterfactuals that the model assigns to y_{o}. (2) **Decoder faithfulness** is measured before intervention using root mean-squared error (RMSE) $E_{\mathrm{rec}} = \mathrm{RMSE}(R(\mathbf{Z}), \mathbf{X})$. For each counterfactual, $E_{\mathrm{cf}} = \mathrm{RMSE}(R(\mathbf{Z}^{\mathrm{cf}}), \mathbf{X})$ measures its total signal-space displacement, whereas $\Delta_{\mathrm{dec}} = \mathrm{RMSE}(R(\mathbf{Z}^{\mathrm{cf}}), R(\mathbf{Z}))$ isolates the change induced by latent optimization from baseline reconstruction error.

Physiological Plausibility. We further constrain the target through volume-conduction spatial-coherence (VCSC), which promotes a reference coherence and weighted phase-lag-index (wPLI) between nearby electrodes (Vinck et al., 2011). Consequently, anomalously low short-range coherence is physiologically atypical (Nunez et al., 1997; Srinivasan et al., 1998). Let $\mathbf{C}^b$ and $\mathbf{\Omega}^b$ denote the matrices of magnitude-squared coherence and debiased squared wPLI for band b, respectively, with their construction given in Appendix B.2. Both are standardized against reference matrices $(\widehat{C}^b, \widehat{\Omega}^b, \sigma_C^b, \sigma_\Omega^b)$ estimated per pair and band, and combined into each single deviation per pair, so that a pair registers as anomalous on either the raw coherence or the volume-conduction robust axis:

$$\begin{aligned}A &= \left(\sum_b \left[\left(\frac{C^b - \widehat{C}^b}{\sigma_C^b}\right)^2 + \left(\frac{\Omega^b - \widehat{\Omega}^b}{\sigma_\Omega^b}\right)^2\right]\right)^{1/2},\\ \mathcal{P}_{\mathrm{ph}}(R(\mathbf{Z}')) &= \frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} w(d_{ij})\,\phi(\mathbf{A}_{ij}(R(\mathbf{Z}'))).\end{aligned} \tag{5}$$

All divisions and powers in the expression for $\mathbf{A}$ are elementwise, and $\mathbf{A}_{ij}$ is its (i, j) entry. With positive reference standard deviations, $\mathcal{P}_{\mathrm{ph}}$ is mostly differentiable and can therefore be optimized directly, using appropriate subgradients at nondifferentiable points. Because coherence is also sensitive to referencing, signal-to-noise ratio, and common input, we treat it as evidence of plausibility rather than proof of physiological validity (Bastos & Schoffelen, 2016).

3.4.1 Distributional Typicality

We consider a predicted class 0, and generate counterfactuals targeting class 1 using the CFO procedure above. The VC head uses these embeddings to compute $p_\theta(y \mid \mathbf{Z})$. The learned class-1 Gaussian is defined in a d-dimensional trial-embedding space:

$$p_{C_1}(u) = p_\theta(u \mid y = 1) = \mathcal{N}\big(u \mid \mu_{C_1}, \mathrm{diag}(\sigma_{C_1}^2)\big), \qquad u \in \mathbb{R}^d. \tag{6}$$

Its parameters are taken from the trained fold checkpoint and remain fixed during CFO. With variance floor $\varepsilon > 0$, we assign a complete trial the discrepancy

$$D_{C_1}(\mathbf{Z}) = \frac{1}{d}\sum_{j=1}^{d} \frac{\left(C_{\theta,j}(\mathbf{Z}) - \mu_{C_1,j}\right)^2}{\max(\sigma^2_{C_1,j}, \varepsilon)}. \tag{7}$$

Thus, D_{C_1} is a function of the whole encoded trial, evaluated through the embedding used for its prediction. Then, for each LOSO fold, we calibrate τ_1 as the empirical 95th percentile of discrepancies for source trials with true class-1 labels. The typical class-1 region is defined by $\mathcal{T}_{C_1} = \{Z : D_{C_1}(\mathbf{Z}) \leq \tau_1\}$, where the held-out subject does not contribute to calibration. A counterfactual achieves joint latent target attainment and typicality when

$$\boxed{p_\theta(y = 1 \mid \mathbf{Z}^{\mathrm{cf}}) \geq p_{\mathrm{o}} \quad \wedge \quad \mathbf{Z}^{\mathrm{cf}} \in \mathcal{T}_{C_1}.} \tag{8}$$

Therefore, after we finish optimizing the CFO objective from Equation 4, we can activate a normalized typicality penalty given by

$$\mathcal{L}_{\mathrm{cfo}}^{\mathcal{T}_{C_1}} = \mathcal{L}_{\mathrm{cfo}} + \lambda_{\mathcal{T}_{C_1}} \frac{D_{C_1}(\mathbf{Z}')}{\max(\tau_1, \varepsilon)}. \tag{9}$$

Normalization by τ_1 scales the penalty across folds without changing the evaluation criterion $D_{C_1}(\mathbf{Z}') \leq \tau_1$. The penalty remains active within the typicality region, encouraging further reduction in discrepancy.

4 Experimental Results

4.1 Emotion Recognition on DREAMER

Table 1 summarizes trial-level performance across the 23 held-out subjects. Valence classification shows relatively balanced learning between both classes. Arousal achieves higher Area Under the Receiver Operating Characteristic (AUROC) but more uneven class-specific recall, showing the ability to learn high arousal patterns but worst performance in learning low arousal. This distinction highlights the value of examining class-specific performance alongside aggregate metrics. Performance also varies across participants (Appendix A.4), consistent with participant-level differences reported in affect recognition studies (Katsigiannis & Ramzan, 2018; Hilali et al., 2025; Del Pup et al., 2025). Differences in physiological responses, subjective ratings, and recording quality may contribute to this variation, although the present evaluation does not isolate their respective effects. With that said, even minuscule gains in balanced discrimination between classes are important to the advancement of this technology. Decoder MSE provides a complementary measure of reconstruction error between X and $R(\mathbf{Z})$.

Table 1: Trial-level emotion-recognition performance across the 23 LOSOCV folds. Values are pooled across all trial predictions.

Task	Bal. acc. (%)	Recall 0 (%)	Recall 1 (%)	AUROC	Decoder MSE
Valence	56.04	57.14	54.94	0.56	0.87
Arousal	57.50	50.00	65.00	0.60	0.59

4.2 Single Subject-Trial Case Analysis

Figure 2 illustrates optimization for a correctly classified class-0 arousal trial from held-out subject 0, selected as the successful example nearest the median latent displacement. Across optimization, the class-1 probability increased from 0.336 to 0.998, while the class-1 discrepancy decreased from $D_{C_1}(\mathbf{Z}) = 1.012$ to $D_{C_1}(\mathbf{Z}^{\mathrm{cf}}) = 0.839$, below the fold-specific threshold $\tau_1 = 1.026$. The two-stage schedule is visible in the trajectory, in which the optimizer operates on the target and physiological objectives alone until the confidence threshold p_{o}. Then, the typicality penalty activates, reducing D_{C_1} while the class-1 probability is maintained. The counterfactual required a

latent displacement of $d_z = 0.977$ and a displacement of $\Delta_{\text{dec}} = 0.221$ in reconstructed feature space, relative to a baseline reconstruction error of $E_{\text{rec}} = 0.528$ and a total signal-space displacement of $E_{\text{cf}} = 0.761$. Reclassifying $C_\theta(g_\phi(R(\mathbf{Z})))$ (where g_ϕ denotes the fused encoder) recovered the original prediction, with the class-1 probability moving from 0.336 to 0.309, showing that the reconstruction path preserves the decision for this trial.

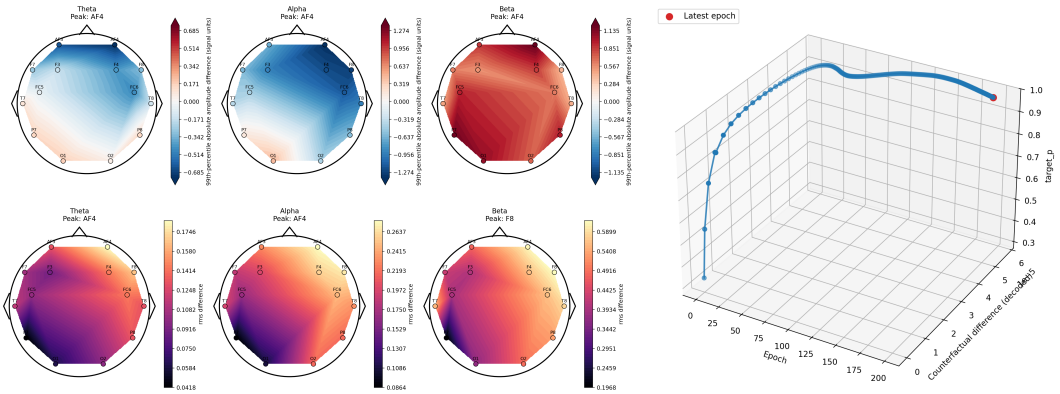

Figure 2: Topographies of mean channel amplitude difference (top left) and root mean squared activity difference (bottom left) between $R(\mathbf{Z})$ and $R(\mathbf{Z}^{\text{cf}})$ per band frequency across the trial. Counterfactual optimization trajectory (right).

The original EEG sample scores a VCSC of $\mathcal{P}_{\text{ph}} = 0.191$ while the reconstructed starting embeddings $g_\phi(R(\mathbf{Z}))$ score $\mathcal{P}_{\text{ph}} = 1102.31$. The large numeric difference is due to the exploding nature of the term (see Equation 5). After CFO, $R(\mathbf{Z}^{\text{cf}})$ scores $\mathcal{P}_{\text{ph}} \approx 0.0$. The decoded signal passed 2 of the 4 evaluable checks, with coherence (99.3% of pairs in range) and debiased squared wPLI (100%) satisfied while amplitude (69.0%) and spectral power (52.4%) were not. We hypothesize that subject-invariant learning may remove significant information and act as an impediment for physiological plausibility in reconstruction.

Examining which decoded features carry the counterfactual change situates the explanation with respect to the affective EEG literature. Relative to the reconstruction of the same trial, the change is concentrated in frontal alpha: alpha-band amplitude falls by 27.0%, 25.7%, 25.3%, 24.3% and 23.0% at AF4, F4, F8, F3 and FC6, respectively, against a median absolute change of 12.2% across all 42 electrode band features. The reduction is regionally specific, averaging -22.5% across six frontal electrodes while leaving the four posterior electrodes essentially unchanged ($+0.4\%$); theta amplitude is approximately flat (-1.5%) and beta amplitude increases ($+17.4\%$). Alpha desynchronization accompanied by a relative beta increase is the direction reported for increased arousal and alertness in event-related desynchronization (Pfurtscheller & Lopes da Silva, 1999; Klimesch, 1999), so the optimizer moves the decoded signal along an axis that prior work associates with the targeted class 1, i.e., high emotional arousal.

4.3 Evaluation of Counterfactual Optimization

Validity, Proximity, and Physiological Plausibility. Table 2 compares the four counterfactual objectives across 57 eligible arousal trials and 92 eligible valence trials. Every optimized latent counterfactual flipped the prediction to class 1, meeting the task-specific confidence threshold ($p_1 \geq 0.80$ for arousal and $p_1 \geq 0.50$ for valence). After typicality optimization, all trials satisfied the calibrated class-1 distribution criterion. Adding $\mathcal{T}_{C_1}$ to CFO reduced the median paired normalized class-1 discrepancy for valence ($\Delta D/\tau = -0.202$). While there was no median arousal change, we see a difference in Q1 of -0.13. VCSC passage remained at 49.12% for arousal and 81.52% for valence. Removing the physiological penalty yielded larger reductions in discrepancy after typicality regularization (-0.155 and -0.224, respectively). However, we observe that removing physiological calibration not only failed our plausibility evaluation but increased the latent distance $d_z(\mathbf{Z}, \mathbf{Z}^{\text{cf}})$ by $\sim$62% in arousal and $\sim$5% in valence.

Table 2: Counterfactual displacement and physiological diagnostics across eligible class-0 trials. Distances are median [Q1, Q3]. VCSC denotes passage of the archived raw-penalty tolerance.

Task	**Objective**	d_z	Δ_{dec}	VCSC (%)	$\Delta D/\tau$ (paired)
Arousal	$\mathcal{L}_{\text{base}}$	0.29 [0.24, 0.33]	0.00 [0.00, 0.00]	0.00	+0.043 [0.04, 0.05]
Arousal	$\mathcal{L}_{\text{cfo}}$	0.70 [0.67, 0.75]	0.49 [0.41, 0.52]	49.12	0.000 [0.00, 0.00]
Arousal	$\mathcal{L}_{\text{cfo}}^{\mathcal{T}_{C_1}-\lambda_{\text{ph}}}$	1.39 [1.37, 1.43]	0.01 [0.01, 0.01]	0.00	-0.155 [-0.17, -0.14]
Arousal	$\mathcal{L}_{\text{cfo}}^{\mathcal{T}_{C_1}}$	0.86 [0.71, 1.27]	0.49 [0.41, 0.51]	49.12	0.000 [-0.13, 0.00]
Valence	$\mathcal{L}_{\text{base}}$	0.13 [0.09, 0.20]	0.00 [0.00, 0.00]	0.00	+0.010 [0.00, 0.02]
Valence	$\mathcal{L}_{\text{cfo}}$	0.28 [0.20, 0.36]	0.10 [0.09, 0.11]	81.52	0.000 [0.00, 0.00]
Valence	$\mathcal{L}_{\text{cfo}}^{\mathcal{T}_{C_1}-\lambda_{\text{ph}}}$	1.45 [1.42, 1.48]	0.00 [0.00, 0.00]	0.00	-0.224 [-0.23, -0.21]
Valence	$\mathcal{L}_{\text{cfo}}^{\mathcal{T}_{C_1}}$	1.38 [0.93, 1.42]	0.10 [0.09, 0.11]	81.52	-0.202 [-0.22, -0.12]

Class-Conditional Typicality We report the median paired change in class-1 discrepancy described in Appendix B.4, using each fold's source-calibrated threshold. Negative values indicate that an arm moved its optimized embeddings closer to the learned class-1 reference than the CFO objective in Equation 4. The decrease in discrepancy after typicality optimization indicates closer agreement with the model's learned class-1 embedding.

Subject-Invariant Alignment. We then iteratively compare the discrepancy of held-out subject's real original class-1 EEG samples, $\mathbf{X}_k^{C_1}$, and reconstructed class-0-to-1 counterfactuals $R(\mathbf{Z}_k^{\text{cf}})$ against corresponding source subjects' typical class-1 original $\mathbf{X}_S^{C_1}$ and encoded $R(\mathbf{Z}_S^{C_1})$ patterns, aiming at evaluating the prevalence of subject-invariant patterns. Each fold uses the discrepancy in Equation 7, fitted to the data of the source subjects. While all counterfactuals fall within the 95th percentile region of source-calibrated typicality, only 68.1% of arousal and 62.1% of valence $\mathbf{X}_k^{C_1}$ samples fall within the 95th percentile their respective $\mathbf{X}_S^{C_1}$ source-calibrated distribution. These results show distributional alignment of reconstructed generated counterfactual class-1 features across subjects, supporting the goal of subject-invariant counterfactual patterns.

5 Conclusion

This paper presents a framework for a human-interpretable counterfactual explanation of how DL models learn the relationship between affective states and their associated subject-invariant EEG patterns. We constrain CFO to learn physiologically plausible features, while preserving domain-wise proximity (Wachter et al., 2018). By adapting VC, this paper translates learned class-conditional density into a CFO typicality constraint. We first train a subject-invariant variational DL model to achieve highly competitive EEG valence-arousal classification balanced accuracies, jointly trained with a decoder. While we perform CFO in high-dimensional latent space, we traverse the decoder to attain original EEG feature dimensionality. Our constrained CFO shows increased distributional alignment against a baseline implementation but increased displacement in reconstructed space, mainly due to the added requirements of increased physiological plausibility. We observe an increase in distributional alignment between held-out generative counterfactual trials and learned cross-subject class-1 distributions through our proposed density regularization, supporting the goal of learning subject-invariant patterns. Because CFO enables the identification of an EEG pattern corresponding to an affective class flip, a future direction for EEG CFO would be to integrate it directly into neuroadaptive and neurofeedback control loops, guiding software updates. Future work may apply this framework toward model enhancement, feature attribution, interpretation of cognitive and affective patterns in the brain, and for other use cases that may benefit from actionable explainability (Pinto et al., 2023; Meulemeester et al., 2026).

REPRODUCIBILITY

The code repository for these experiments has been pruned, reviewed, and added to *anonymized (link to repository)*. Instructions on how to run the experiments with the best found hyperparameters are included in markdown files across the repository.

AI USE STATEMENT

Generative AI tools (ChatGPT6-Astra, ChatGPT5.6-Sol, Claude Fable 5.1) were used to assist with manuscript review, literature discovery, and the implementation of this study. This included assistance with code generation and debugging. The authors reviewed and refined all AI-assisted contributions.

REFERENCES

André M. Bastos and Jan-Mathijs Schoffelen. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in Systems Neuroscience*, 9:175, 2016. doi: 10.3389/fnsys.2015.00175.

Guillermo Bernal, Nelson Hidalgo, Conor Russomanno, and Pattie Maes. Galea: A physiological sensing system for behavioral research in virtual environments. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 66–76. IEEE, 2022. doi: 10.1109/VR51125.2022.00024.

Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Alessandra Bertoldo, and Manfredo Atzori. The more, the better? evaluating the role of EEG preprocessing for deep learning applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 33:1061–1070, 2025. doi: 10.1109/TNSRE.2025.3547616.

Shehzaad Dhuliawala, Mrinmaya Sachan, and Carl Allen. Variational classification. *CoRR*, abs/2305.10406, 2023. doi: 10.48550/arXiv.2305.10406.

Kate C. Ewing, Stephen H. Fairclough, and Kiel Gilleade. Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop. *Frontiers in Human Neuroscience*, 10:223, 2016. doi: 10.3389/fnhum.2016.00223.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Daniel Gaspar-Figueiredo, Silvia Abrahão, Emilio Insfrán, and Jean Vanderdonckt. Measuring user experience of adaptive user interfaces using EEG: A replication study. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pp. 52–61, 2023. doi: 10.1145/3593434.3593452.

Riccardo Guidotti. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024. doi: 10.1007/s10618-022-00831-6.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biessmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. doi: 10.1016/j.neuroimage.2013.10.067.

Manal Hilali, Abdellah Ezzati, and Said Ben Alla. CMHFE-DAN: A transformer-based feature extractor with domain adaptation for EEG-based emotion recognition. *Information*, 16(7):560, 2025. doi: 10.3390/info16070560.

Chengyu Hua, Hui Cao, Zhaolong Li, and Zhiqiang Huang. A dual-branch network with brain region-constrained attention for EEG emotion recognition. *Frontiers in Neuroscience*, 20:1810609, 2026. doi: 10.3389/fnins.2026.1810609.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8086–8094, 2024. doi: 10.24963/ijcai.2024/894.

Stamos Katsigiannis and Naeem Ramzan. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018. doi: 10.1109/JBHI.2017.2688239.

Nazmun N. Khan, Taylor Sweet, Chase A. Harvey, Calder Knapp, Dean J. Krusienski, and David E. Thompson. The role of review process failures in affective state estimation: An empirical investigation of the DEAP dataset, 2025. URL https://arxiv.org/abs/2508.02417.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer International Publishing, Cham, 2019. doi: 10.1007/978-3-030-28954-6_14.

Wolfgang Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2–3):169–195, 1999. doi: 10.1016/S0165-0173(98)00056-3.

Nataliya Kosmyna and Pattie Maes. AttentivU: An EEG-based closed-loop biofeedback system for real-time monitoring and improvement of engagement for personalized learning. *Sensors*, 19(23):5200, 2019. doi: 10.3390/s19235200.

Marcin Kostrzewa, Krzysztof Galus, and Maciej Zięba. Towards plausibility in time series counterfactual explanations, 2026. URL https://arxiv.org/abs/2603.08349.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2801–2807, 2019. doi: 10.24963/ijcai.2019/388.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3490–3497, 2018. doi: 10.1609/aaai.v32i1.11596.

Rui Li, Chao Ren, Yiqing Ge, Qiqi Zhao, Yikun Yang, Yuhan Shi, Xiaowei Zhang, and Bin Hu. MTLFuseNet: A novel emotion recognition model based on deep latent feature fusion of EEG signals and multi-task learning. *Knowledge-Based Systems*, 276:110756, 2023a. doi: 10.1016/j.knosys.2023.110756.

Rui Li, Chao Ren, Sipo Zhang, Yikun Yang, Qiqi Zhao, Kechen Hou, Wenjie Yuan, Xiaowei Zhang, and Bin Hu. STSNet: A novel spatio-temporal-spectral network for subject-independent EEG-based emotion recognition. *Health Information Science and Systems*, 11(1):25, 2023b. doi: 10.1007/s13755-023-00226-x.

Suryani Lim, Henri Prade, and Gilles Richard. Ranking counterfactual explanations, 2025. URL https://arxiv.org/abs/2503.15817.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. IEEE, 2017. doi: 10.1109/ICCV.2017.324.

Haoran Liu, Ying Zhang, Yujun Li, and Xiangyi Kong. Review on emotion recognition based on electroencephalography. *Frontiers in Computational Neuroscience*, 15:758212, 2021. doi: 10.3389/fncom.2021.758212.

Bjorge Meulemeester, Raphael Mazzine Barbosa de Oliveira, Yifan He, and David Martens. Calculating and visualizing feature importance values for counterfactual explanations. *IEEE Access*, 14:50691–50703, 2026. doi: 10.1109/ACCESS.2026.3679213.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617. ACM, 2020. doi: 10.1145/3351095.3372850.

Paul L. Nunez, Ramesh Srinivasan, Andrew F. Westdorp, Ranjith S. Wijesinghe, Don M. Tucker, Richard B. Silberstein, and Peter J. Cadusch. EEG coherency I: Statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and Clinical Neurophysiology*, 103(5):499–515, 1997. doi: 10.1016/S0013-4694(97)00066-7.

Ozan Özdenizci, Ye Wang, Toshiaki Koike-Akino, and Deniz Erdoğmuş. Learning invariant representations from EEG via adversarial inference. *IEEE Access*, 8:27074–27085, 2020. doi: 10.1109/ACCESS.2020.2971600.

Gert Pfurtscheller and Fernando H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999. doi: 10.1016/S1388-2457(99)00141-8.

Mauro F. Pinto, Joana Batista, Adriana Leal, Fábio Lopes, Ana Oliveira, António Dourado, Sulaiman I. Abuhaiba, Francisco Sales, Pedro Martins, and César A. Teixeira. The goal of explaining black boxes in EEG seizure prediction is not to explain models' decisions. *Epilepsia Open*, 8 (2):285–297, 2023. doi: 10.1002/epi4.12748.

Dwi Wahyu Prabowo, Hanung Adi Nugroho, Noor Akhmad Setiawan, and Johan Debayle. A systematic literature review of emotion recognition using EEG signals. *Cognitive Systems Research*, 82:101152, 2023. doi: 10.1016/j.cogsys.2023.101152.

Akshay Sujatha Ravindran and Jose Contreras-Vidal. An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth. *Scientific Reports*, 13:17709, 2023. doi: 10.1038/s41598-023-43871-8.

Mario Refoyo and David Luengo. Multi-SpaCE: Multi-objective subsequence-based sparse counterfactual explanations for multivariate time series classification. *Neurocomputing*, 700:134526, 2026. doi: 10.1016/j.neucom.2026.134526.

Paul Sevellec, Elisa Fromont, Romaric Gaudel, Laurence Rozé, and Matteo Sammarco. Plausible conditional generation-based counterfactual explanations for multivariate times series classification. In *ECAI 2025: 28th European Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications, pp. 2009–2016. IOS Press, 2025. doi: 10.3233/FAIA251037.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified BP attributions fail. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9046–9057. PMLR, 2020.

Ramesh Srinivasan, Paul L. Nunez, and Richard B. Silberstein. Spatial filtering and neocortical dynamics: Estimates of EEG coherence. *IEEE Transactions on Biomedical Engineering*, 45(7): 814–826, 1998. doi: 10.1109/10.686789.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Martin Vinck, Robert Oostenveld, Marijn van Wingerden, Francesco Battaglia, and Cyriel M. A. Pennartz. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage*, 55(4):1548–1565, 2011. doi: 10.1016/j.neuroimage.2011.01.055.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018.

Yaodong Wang, Yiping Zuo, Dan Chen, Albert Y. Zomaya, Rajiv Ranjan, Jingying Chen, and Tengfei Gao. Counterfactual causal inference of biomedical signals: Unveiling causal EEG patterns for ASD evaluation. *IEEE Transactions on Artificial Intelligence*, 7(4):2073–2087, 2026. doi: 10.1109/TAI.2025.3610394.

Zhendong Wang, Isak Samsten, Ioanna Miliou, Rami Mochaourab, and Panagiotis Papapetrou. GLACIER: Guided locally constrained counterfactual explanations for time series classification. *Machine Learning*, 113:4639–4669, 2024. doi: 10.1007/s10994-023-06502-x.

Kang Yin, Hye-Bin Shin, and Hee-Dong Kim. Pioneering EEG motor imagery classification through counterfactual analysis. In *2024 12th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–4. IEEE, 2024. doi: 10.1109/BCI60775.2024.10480506.

Martina Zannotti, Bardh Prenkaj, Marco Piangerelli, Flavio Corradini, and Gjergji Kasneci. COunterfactual reasoning for temporal EXplanations: Plausible and robust explanations for EEG-based seizure detection. *Transactions on Machine Learning Research*, March 2026. ISSN 2835-8856.

Xin Zhang, Ye Li, Fei Pi, and Xiu Zhang. RGB-style input representations for EEG: Evaluating spatial concatenation versus band-wise stacking in deep emotion recognition. *Brain Sciences*, 16 (7):716, 2026. doi: 10.3390/brainsci16070716.

Kexin Zhu, Jian Shen, Suran Xu, Bingkun Ren, Xiaolin Tan, Zhuohan Sun, Chenxu Guo, Haoran Gao, Yanan Zhang, and Bin Hu. CausalCEM: Self-improved causal counterfactual emotion modeling for multimodal physiological signals. *Information Fusion*, 136:104489, 2026. doi: 10.1016/j.inffus.2026.104489.

A Deep Learning Model

A.1 Adaptation of Variational Classification Objective

We adapt the variational classification objective of Dhuliawala et al. (2023) by representing the empirical and anticipated class-conditional latent distributions as diagonal Gaussians:

$$q_\phi(z \mid y) = \mathcal{N}(m_y, \mathrm{diag}(v_y)), \qquad p_\theta(z \mid y) = \mathcal{N}\big(\mu_y, \mathrm{diag}(\sigma_y^2)\big). \tag{10}$$

For each trial i, let Z_i denote its complete encoded sequence and let $z_i = C_\theta(Z_i) \in \mathbb{R}^d$ denote the terminal embedding supplied to the variational classifier. The distributions $q_\phi(z \mid y)$ and $p_\theta(z \mid y)$ are defined in this trial-embedding space. For each class label $y \in \mathcal{Y} = \{0, 1\}$, let B_y denote the indices of trials with label y in the training minibatch. The coordinate-wise empirical Gaussian parameters are, for $j = 1, \ldots, d$,

$$m_{y,j} = \frac{1}{|B_y|} \sum_{i \in B_y} z_{i,j}, \qquad v_{y,j} = \frac{1}{|B_y|} \sum_{i \in B_y} (z_{i,j} - m_{y,j})^2. \tag{11}$$

This replaces discriminatory posterior density estimation with an analytic Gaussian KL divergence. Using a dimension-normalized posterior term, the resulting objective is

$$\mathcal{L}_{\mathrm{VC}} = \frac{\beta}{2d} \sum_{y \in \mathcal{Y}} p(y) \sum_{j=1}^{d} \left[\log \frac{\sigma_{y,j}^2}{v_{y,j}} + \frac{v_{y,j} + (m_{y,j} - \mu_{y,j})^2}{\sigma_{y,j}^2} - 1 \right] + \lambda \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{p_\pi(y)}. \tag{12}$$

Here, d is the latent dimensionality, $p(y)$ is the empirical frequency of label y, and $p_\pi(y)$ is its learned categorical prior. The $-\mathcal{H}(q)$ contribution in Equation 2 accounts for the negative entropy of the posterior, while the remaining Gaussian terms align its mean and variance with the learned class-conditional distribution. In practice, all variances are lower-bounded by a small constant $\varepsilon > 0$ for numerical stability.

A.2 Inverse Subject-Adversarial

To enforce cross-subject generalizability, we introduce a subject-invariant adversarial loss target inspired by Özdenizci et al. (2020). The adversarial loss objective can be divided into two components: the task classification likelihood, and the subject-identity likelihood recovered by the adversary,

$$(\hat{\theta}_e, \hat{\theta}_c, \hat{\theta}_a) = \arg \min_{\theta_e, \theta_c} \max_{\theta_a} \left\{ -\mathbb{E}_{(X,y)}[\log q_{\theta_c}(y \mid z)] + \lambda \mathbb{E}_{(X,s)}[\log q_{\theta_a}(s \mid z)] \right\}, \tag{13}$$

where $z = f(X; \theta_e)$ is the encoder's latent representation, $q_{\theta_c}(y \mid z)$ is the classifier's estimate of task label y, $q_{\theta_a}(s \mid z)$ is the adversary's estimate of subject identity s, and $\lambda > 0$ controls the strength of the invariance penalty.

The encoder and classifier parameters θ_e, θ_c are optimized by minimizing $\mathcal{L}_{\mathrm{adv}}$ while the adversary θ_a is optimized by maximizing it, in alternating steps. This opposition forces the encoder to retain features that help predict y while eliminating features related to the subject-specific structure which the adversary could have used to identify s. This reduces patterns that might allow the adversary to identify s from common differences in multi-subject EEG but do not reflect the underlying signal (Özdenizci et al., 2020; Del Pup et al., 2025). This is particularly relevant for emotion recognition, in which a 23-subject sample size can make overfitting to subject-specific artifacts a risk (Katsigiannis & Ramzan, 2018).

Our method differs from Özdenizci et al. (2020) in two ways. First, we apply subject-identification to the encoded latent state $\mathbf{Z}$ through a dense layer with a softmax function over S subjects. Then, our adversarial loss is incorporated through a gradient reversal layer, allowing ISA to be optimized through minimization (Ganin et al., 2016).

A.3 Supervised Contrastive Learning

To encourage emotion-related representations that are shared across subjects during autoencoding, we adapt the supervised contrastive objective of Khosla et al. (2020). The objective uses emotion

labels to identify positive pairs and subject identities to restrict these pairs to observations from different participants. For a mini-batch indexed by $\mathcal{M} = \{1, \ldots, B\}$, let z_i denote the learned representation of sample i, with emotion label y_i and subject identity s_i. We normalize the representations as $u_i = z_i/\|z_i\|_2$, so that $u_i^\top u_j$ measures cosine similarity. For each anchor i, the positive and eligible comparison sets are

$$\begin{aligned} P(i) &= \{p \in \mathcal{M} \setminus \{i\} : y_p = y_i,\ s_p \neq s_i\}, \\ J(i) &= P(i) \cup \{j \in \mathcal{M} : y_j \neq y_i\}. \end{aligned} \tag{14}$$

Thus, same-label samples from the same subject are excluded from both the positive set and the denominator, while different-label samples remain eligible negatives regardless of subject identity. The anchor itself is always excluded.

Writing $\mathcal{M}_+ = \{i \in \mathcal{M} : |P(i)| > 0\}$ for the valid anchors, the supervised contrastive loss is

$$\mathcal{L}_{\text{sup}} = -\frac{1}{\sum_{i \in \mathcal{M}_+} w_i} \sum_{i \in \mathcal{M}_+} \frac{w_i}{|P(i)|} \sum_{p \in P(i)} \log \left[\frac{\exp(u_i^\top u_p/\tau)}{\sum_{j \in J(i)} \exp(u_i^\top u_j/\tau)} \right], \tag{15}$$

where $\tau > 0$ is the temperature parameter and $w_i \geq 0$ is an optional anchor weight. Setting $w_i = 1$ recovers the unweighted mean over valid anchors. If no valid anchors exist, or their total weight is zero, the loss is defined as zero.

Minimizing $\mathcal{L}_{\text{sup}}$ encourages same-emotion samples from different participants to have similar representations relative to samples with different emotion labels. Our adaptation differs from the standard supervised contrastive objective by conditioning positive-pair selection on subject identity and excluding same-subject, same-label comparisons entirely. This prevents repeated windows from a single participant from supplying the positive-pair signal and avoids treating those excluded windows as negatives. Because cross-subject alignment alone does not guarantee that all subject-specific information has been removed, this term is complementary to subject-adversarial learning.

A.4 LOSOCV Emotion Recognition

A.4.1 Balanced Accuracy as the Primary Metric

Ordinary accuracy can overstate performance in imbalanced EEG emotion recognition tasks. The DREAMER dataset is composed of a majority class rate of 72.5% for high arousal and 61.1% for high valence. Many studies do not report balanced metrics. For example, MTLFuseNet reports LOSOCV accuracies of 80.43% for valence and 83.33% for arousal (Li et al., 2023a), but the absence of BA and an obtainable public implementation as well as the inability to replicate from the paper itself limits independent assessment. We therefore treat these values as reported results rather than a reproducible benchmark. Similarly, CMHFE-DAN reports 64.73% valence and 79.83% arousal accuracy, only 3.63% and 7.33% above the corresponding majority-class baselines (Hilali et al., 2025). Because CMHFE-DAN also accesses unlabeled target-subject data during training, it represents transductive adaptation rather than strictly zero-shot evaluation (Hilali et al., 2025). More broadly, performance is highly sensitive to preprocessing, potential artifact exploitation, and validation settings (Del Pup et al., 2025). Meanwhile, methodological problems such as leakage, biased feature selection, and model collapse were found in nearly 87% of 101 reviewed studies and inflated accuracy by up to 46 percentage points (Khan et al., 2025). Under zero-shot LOSOCV on DREAMER, an RGB-style EEG model achieved 75.15% accuracy but reported 58.73% BA for arousal, and 65.55% versus 60.05% for valence (Zhang et al., 2026). This large gap is consistent with a strong majority-class bias and possible near-collapse of low arousal recognition. The corresponding valence metrics, while less pronounced, also suggest the possibility of class imbalance. Because the reported LOSOCV metrics are averaged across subjects and class-specific recalls are not provided, the extent of minority-class under-recognition cannot be determined exactly. Therefore, we select BA as the main metric then report separate class recalls and AUROC.

A.4.2 Predictive Model Evaluation Metrics

Due to significant class imbalance, we evaluate the model's ability to learn both high and low valence-arousal patterns. We define true positive rate (TPR) and false positive rate (FPR) as

$$\mathrm{TPR} = \frac{TP}{TP + FN}, \qquad \mathrm{FPR} = \frac{FP}{FP + TN}. \tag{16}$$

Then we evaluate the recall of each class:

$$\mathrm{Recall}_0 = 1 - \mathrm{FPR}, \qquad \mathrm{Recall}_1 = \mathrm{TPR}. \tag{17}$$

Finally, we evaluate BA and AUROC:

$$\mathrm{BA} = \frac{1}{2}\left(\mathrm{Recall}_0 + \mathrm{Recall}_1\right), \qquad \mathrm{AUROC} = \int_0^1 \mathrm{Recall}_1(u)\, du, \tag{18}$$

where $u = \mathrm{FPR}$ and $\mathrm{Recall}_1(u)$ represents the corresponding TPR along the ROC curve.

A.4.3 Discussion on Results

The separate subject evaluations reveal substantial variation in performance across the 23 DREAMER participants. Each subject contributes 18 trial-level predictions, with subject indices following the zero-based numbering in the figures. Individual BAs range from 0.29–0.89 for arousal and 0.31–0.75 valence, respectively.

Both classes are present for every subject in both tasks. However, several subjects have too few minority-class trials to support a stable estimate of class-specific performance. For example, subjects 5, 8, and 18 each have only one arousal class 0 trial and 17 class 1 trials. Subjects 6 and 12 have two class 0 trials, and subject 17 has three. The remaining subjects have between four to nine class 0 trials and comparatively stronger representation of both classes. Every arousal subject has at least nine class 1 trials. On the other hand, valence has less severe imbalance. Subjects 2, 7, 9, and 21 have the smallest class 0 counts, with four trials each. All remaining subjects have at least five class-0 trials. Subjects 10 and 14 are exactly balanced at nine trials per class. Therefore, changing the correctness of one class c prediction changes BA by $1/(2c)$. Consequently, one class 0 prediction changes arousal BA by 50 percentage points for subjects 5, 8, and 18, and by 25 points for subjects 6 and 12. BA gives equal weight to class recalls, but this weighting cannot compensate for having very few observations within a class.

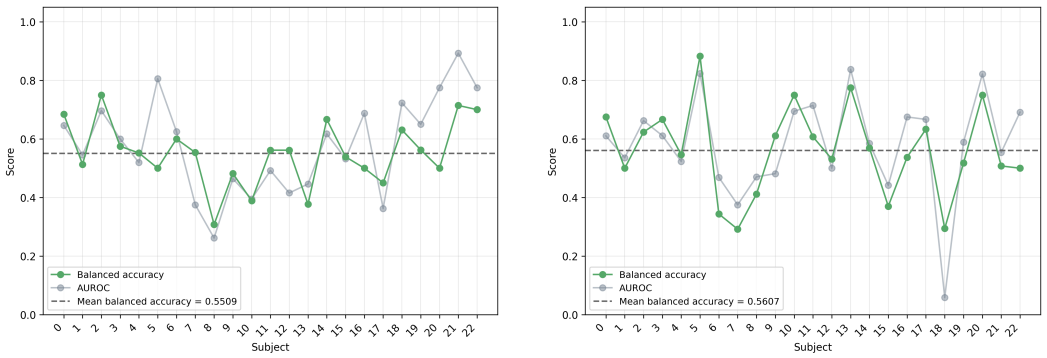

Figure 3: LOSOCV performance across subjects for valence (left) and arousal (right) classification.

Subject 8's 14 correctly classified class 1 trials and incorrectly classified class 0 trial produce ordinary accuracy of 0.78 but BA of 0.41. AUROC is also sensitive to sparse class support as we see with subject 18's arousal AUROC of 0.059, which represents one correctly ordered positive–negative pair out of 17, all involving the same class 0 trial. Decision-threshold effects provide another explanation for the differences between metrics. Across the valence evaluations, class 1 probabilities occupy a narrow interval of approximately 0.49–0.51, making threshold predictions sensitive to small changes around 0.5. Class imbalance nevertheless cannot explain all variation. The observed differences

therefore combine limited evaluation samples, unequal class support, and variation in model predictions across held-out subjects. Differences in EEG characteristics, individual responses and rating behavior, and the source-subject composition of each fold may contribute to these results.

B Counterfactual Optimization

B.1 Non-increasing Subject Identifiability in Counterfactual Optimization

Let $s \in \mathcal{S}$ denote participant identity and $\mathbf{Z}$ the complete latent sequence for a trial. The ISA head predicts identity from a pooled trial representation $H(\mathbf{Z})$ by minimizing $\ell_a(\mathbf{Z}, s) = -\log q_a(s \mid H(\mathbf{Z}))$. Gradient reversal makes the encoder oppose this objective while it also learns reconstruction and affective classification. This encourages reduced subject identification but does not guarantee independence between $\mathbf{Z}$ and s.

For each fixed trained model, assume CFO uses only $\mathbf{Z}$ and shared fixed quantities in its objective, target selection, stopping, and candidate selection. In particular, reconstruction proximity compares $R(\mathbf{Z}')$ with $R(\mathbf{Z})$, without supplying the original EEG, $\mathbf{X}$, as an additional input. The resulting counterfactual can therefore be written $\mathbf{Z}^{\mathrm{cf}} = T(\mathbf{Z})$.

Define the optimal subject-identification accuracy of a representation U as

$$\mathrm{Acc}_s^*(U) = \sup_b \Pr[b(U) = s], \tag{19}$$

where the supremum is over all measurable subject predictors. Any predictor b applied to $T(\mathbf{Z})$ induces a predictor $b \circ T$ applied to $\mathbf{Z}$. Hence,

$$\mathrm{Acc}_s^*(\mathbf{Z}^{\mathrm{cf}}) = \sup_b \Pr[b(T(\mathbf{Z})) = s] \le \sup_c \Pr[c(\mathbf{Z}) = s] = \mathrm{Acc}_s^*(\mathbf{Z}). \tag{20}$$

Because the fixed decoder receives only $\mathbf{Z}^{\mathrm{cf}}$, the same argument gives

$$\mathrm{Acc}_S^*(R(\mathbf{Z}^{\mathrm{cf}})) \le \mathrm{Acc}_S^*(\mathbf{Z}^{\mathrm{cf}}) \le \mathrm{Acc}_S^*(\mathbf{Z}). \tag{21}$$

These comparisons use the same population of trials under the same trained model. Intuitively, the fixed decoder, which receives only the optimized latent representation, cannot make the counterfactual reconstruction more subject-identifiable than the original latent representation. Thus, any reduction in optimal subject identifiability already achieved in the latent representation is preserved through CFO and decoding. The argument does not itself establish that ISA achieved low optimal subject identifiability.

Empirical Check. In the no-GRL ablation, the trained subject head identified participants with accuracy exceeding 20%. With GRL, its mean final source-training accuracy across 23 LOSO folds was 4.65%, close to the 4.55% uniform baseline for the 22 source participants in each fold. These observations provide evidence of reduced identification by the trained head; they do not estimate the optimal accuracy $\mathrm{Acc}_S^*(\mathbf{Z})$ used in the bound above.

B.2 Construction of Volume-Conduction Spatial Coherence

We show the construction of C^b and Ω^b from Equation 5. Let $X_{i,b}^w(f)$ denote the discrete Fourier transform (DFT), over window w, of the band b filtered signal at channel i, let $S_{ij,b}^w(f) = X_{i,b}^w(f)\overline{X_{j,b}^w(f)}$ be the corresponding cross-spectrum, and let $\mathcal{I}_b^w(f) = \mathrm{Im}\, S_{ij,b}^w(f)$. Writing $\langle\cdot\rangle_w$ and $\langle\cdot\rangle_f$ for averages over windows and over the DFT bins of that band, we form magnitude-squared coherence and the debiased squared wPLI (Vinck et al., 2011).

$$C_{ij}^b = \left\langle \frac{\left|\langle S_{ij,b}^w(f)\rangle_w\right|^2}{\langle |X_{i,b}^w(f)|^2\rangle_w \langle |X_{j,b}^w(f)|^2\rangle_w} \right\rangle_f, \qquad \Omega_{ij}^b = \left\langle \frac{\left(\sum_w \mathcal{I}_b^w(f)\right)^2 - \sum_w \mathcal{I}_b^w(f)^2}{\left(\sum_w |\mathcal{I}_b^w(f)|\right)^2 - \sum_w \mathcal{I}_b^w(f)^2} \right\rangle_f. \tag{22}$$

Here, $\mathcal{E}$ is the set of 91 electrode pairs, d_{ij} is the three-dimensional chord distance between electrode positions, and $w(d) = \exp\left((d_c - d)^+/\tau\right)$, with $\tau > 0$, upweights pairs within $d_c = 12\,\mathrm{cm}$ while assigning unit weight to more distant pairs. For each pair, let $a = \mathbf{A}_{ij} \ge 0$ and define

$$\phi(a) = \exp\big(\kappa \min(a, \mathbf{A}_{\max})\big) - 1, \qquad \kappa > 0, \quad \mathbf{A}_{\max} > 0. \tag{23}$$

The parameter κ controls how steeply the penalty increases, while $\mathbf{A}_{\max}$ caps the deviation entering the penalty. Thus, every nonzero deviation receives a penalty, which remains constant at $\exp(\kappa\mathbf{A}_{\max}) - 1$ once $a \geq \mathbf{A}_{\max}$. For each fixed electrode pair, the debiased squared wPLI denominator satisfies

$$\left(\sum_w |\mathcal{I}_b^w(f)|\right)^2 - \sum_w \left(\mathcal{I}_b^w(f)\right)^2 = \sum_{w \neq v} |\mathcal{I}_b^w(f)|\,|\mathcal{I}_b^v(f)| \geq 0. \tag{24}$$

The coherence denominator is also nonnegative, and both denominators are guarded by a positive additive constant in the implementation.

B.3 Architecture-Agnostic Implementation

We provide a reference implementation of the counterfactual optimizer through the *anonymized* Python library. The implementation separates the optimization procedure from model-specific architecture through a common adapter interface. For a given trained model, the adapter exposes the operations required to (i) obtain the latent representation, (ii) evaluate target-class probabilities from a candidate latent sequence, and (iii) decode that sequence into the input feature domain. Model-specific adapters remain responsible for tensor-shape conventions, preprocessing and inverse transformations. Thus, the implementation is architecture-agnostic subject to the differentiable latent-classification-decoding interface assumed by our method. It is not intended to universally support black-box predictors.

B.4 Paired Optimized Discrepancy

For each eligible trial i, let $D_{C_1}(\mathbf{Z}_i^{\mathrm{cf}})$ denote the counterfactual's discrepancy from the class-1 distribution, and let τ denote the calibrated class-1 typicality threshold. Let cfo be the procedure described in Equation 4 and cfo_m an alternative CFO procedure. We compare their counterfactuals on the same trial using

$$\frac{\Delta D_{C_1}}{\tau} = \frac{D_{C_1}^{(\mathrm{cfo_m})}\left(\mathbf{Z}_i^{\mathrm{cf}}\right) - D_{C_1}^{(\mathrm{cfo})}\left(\mathbf{Z}_i^{\mathrm{cf}}\right)}{\tau}. \tag{25}$$

A negative value indicates lower class-1 discrepancy under the alternative procedure. For example, in the valence comparison in which cfo_m adds $\mathcal{T}_{C_1}$, the observed $\Delta D_{C_1}^{(m)}/\tau = -0.202$ indicates a median paired decrease of 0.202τ.

B.5 Comparison of Reconstructed Bandwave Continuity

We qualitatively assess decoded counterfactuals and compare their temporal structure with their original EEG samples. We compare four signals for the held-out subject 0 arousal case study of Section 4.2: the observed class-0 trial, its decoded counterfactual $R(\mathbf{Z}^{\mathrm{cf}})$, a real class-1 trial from the same subject, and the decoded reconstruction of the original $R(\mathbf{Z})$ as a baseline that isolates decoder behavior from the counterfactual change. All signals are placed on a common scale for visual comparison.

In the time domain, $R(\mathbf{Z}^{\mathrm{cf}})$ remains continuous and oscillatory across the full trial, while staying within a physiological amplitude range, and preserving continuity across window boundaries. It is, however, visibly less smooth than the observed signals. With respect to VCSC (defined in Equation 5), its coherence and wPLI fall within the range spanned by the observed class-0 and class-1 trials across all three bands rather than departing sharply, while the reconstruction baseline $R(\mathbf{Z})$ shows worse connectivity relative to all observed signals. The latter is consistent with the reconstruction scoring $\mathcal{P}_{\mathrm{ph}} = 1102.31$ in Section 4.2, which we attribute to the loss of subject-specific structure through subject-invariant encoding rather than to the counterfactual intervention.

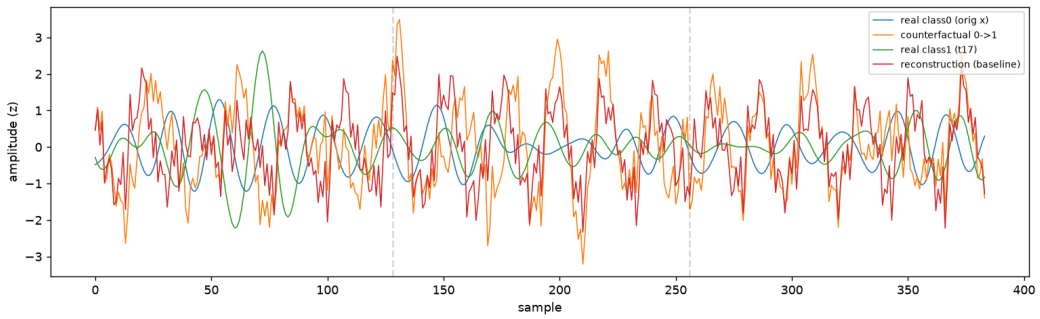

Figure 4: Reconstructed bandwave continuity for the subject-0 arousal case study (channel AF3, θ band, first three 1-s windows).

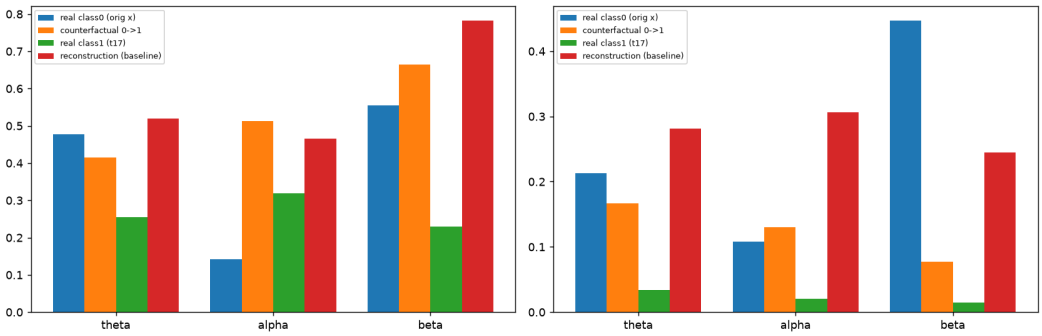

Figure 5: Mean magnitude-squared coherence (left) and debiased squared wPLI (right) per band, averaged over the 91 electrode pairs.

These observations agree with the per-component physiological checks of the same trial, in which coherence and debiased squared wPLI are satisfied while amplitude and spectral power are not. The excess high-frequency content visible in the waveform is the time-domain counterpart of the unmet spectral-power criterion. Consistent with the ablations of Section 4.3, the connectivity criteria pass under objectives that retain the physiological term and fail when it is removed. We report these as qualitative, single-trial observations. Population-level connectivity statistics follow the calibrated evaluation referenced against the LOSO-complement source subjects.

B.6 Limitations

The subject-adversarial objective may encourage the model to suppress features that identify individual participants, but it does not guarantee that all subject information is removed. Consequently, the model loses significant information throughout encoding, creating a significant challenge for reconstruction. Moreover, the prediction model must be trained across LOSOCV folds for each affective dimension and counterfactual generation requires a separate iterative optimization for each evaluated trial. Therefore, reconstruction, classification, subject-adversarial training, and physiological constraints add significant computational cost. This limits the number of repeated runs, hyperparameter settings, ablations, and datasets that we can evaluate within this study. Self-report data introduces a significant challenge in modeling. DREAMER's valence and arousal labels are based on subjective five-point self-reports. Individual differences in interpreting the scale and the conversion of ratings into binary classes introduce uncertainty that the model cannot resolve. While we may attribute specific EEG patterns to emotional classes, we are limited to the information and quality of the dataset. While the methodology suggested here may be applied to other datasets, the performance and explanation quality may differ and require significant tuning.