
Dual-Objective Diffusion Models for EEG

Mengchun Zhang*
University of Pittsburgh
mengchuz@andrew.cmu.edu

Eddie Guo*
Carnegie Mellon University
yuzhiguo@andrew.cmu.edu

Yucheng Shao*
Carnegie Mellon University
yshao3@andrew.cmu.edu

Parusha Pradhan*
University of Pittsburgh
pap203@pitt.edu

Abstract

Electroencephalography (EEG) decoding remains challenging due to noisy signals, inter-subject variability, and data scarcity. We explore whether a compact latent representation z , learned through a dual-objective diffusion model, is sufficient for robust EEG classification and reconstruction. We introduce three variants: *DualDiff-Latent*, which relies on z ; *DualDiff-Mixup*, which incorporates input/output mixup strategies; and *DualDiff-Hybrid*, which combines both latent and mixup techniques. Our results across five EEG task paradigms, including SSVEP, Motor Imagery, and speech-related tasks (imagined, spoken, and heard speech from the FEIS dataset), show that *DualDiff-Latent* consistently outperforms both the EEGNet baseline and its mixup-enhanced counterparts, particularly in cross-subject generalization. On SSVEP, *DualDiff-Latent* achieves 87.82% unseen accuracy (vs. 81.03% with EEGNet), and 88.14% when combining z with the denoised output $\hat{x}$. Neither input nor output mixup alone (*DualDiff-Mixup*) improved performance, and the hybrid model yielded comparable results to *DualDiff-Latent*, suggesting that z may already capture the relevant task features.

These findings support the hypothesis that a well-trained latent space can serve as a joint backbone for both discriminative and generative objectives in EEG modeling. Future work will test the statistical significance of these findings over multiple seeds and explore where in the encoder embedding mixup might add value.

1 Introduction

Electroencephalography (EEG) offers valuable insights into brain activity for applications like Brain-Computer Interfaces (BCIs) and neural decoding, but its effectiveness is often limited by inherent data challenges. EEG signals suffer from noise and high inter-subject variability, which hinders the ability of machine learning models to learn robust representations that generalize across subjects and tasks. To address these limitations, this project investigates whether diffusion models designed with a dual objective (simultaneously optimizing for signal reconstruction and classification) can enhance EEG representations and improve generalization performance across diverse neural decoding paradigms.

We proposed and evaluated three dual-objective diffusion architectures: *DualDiff-Latent*, *DualDiff-Mixup*, and *DualDiff-Hybrid*. These models were tested on multiple EEG datasets. Performance was benchmarked against EEGNet, a widely-used compact convolutional neural network for EEG classification. Our results indicate that the dual-objective diffusion approach holds considerable promise. Nevertheless, our findings also highlight ongoing challenges, as the models still struggled to achieve robust subject-independent generalization on the more complex and variable FEIS phoneme decoding tasks.

2 Related Work

EEG-based machine learning models face challenges due to limited, noisy, and highly variable data, which restricts generalization across subjects and tasks. Traditional data augmentation techniques, such as time-domain transformations (jittering, flipping) and frequency-based augmentations (filtering, spectral modifications), fail to introduce sufficient diversity for robust model training. To address these limitations, we explored deep generative Diffusion Models for EEG data synthesis and augmentation.

Classical EEG decoding architectures

Among classical deep learning approaches, EEGNet Lawhern et al. [2018] is a widely adopted compact Convolutional Neural Network (CNN) architecture specifically designed for EEG-based BCI tasks. It is known for its efficiency and versatility across various paradigms, including SSVEP, P300, Motor Imagery, and speech decoding. EEGNet uses depthwise and separable convolutions to capture crucial spatio-temporal features with significantly fewer parameters than standard CNNs. Due to its established performance and suitability for diverse EEG analyses, we utilized EEGNet as the primary baseline to benchmark our proposed dual-objective diffusion models in this study.

Diffusion models for EEG

Recent advancements in Denoising Diffusion Probabilistic Models (DDPMs) have introduced a promising new approach for EEG generation. Unlike GANs, which struggle with mode collapse, diffusion models learn the true data distribution through a stepwise denoising process, allowing them to generate higher-fidelity EEG signals. Tosato et al. [2023] applies diffusion models to EEG synthesis, demonstrating that diffusion-generated EEG is more physiologically accurate than GAN-generated EEG, improves classifier performance when used for data augmentation, and closely resembles real EEG data, as confirmed by Fréchet Inception Distance (FID) and Sliced Wasserstein Distance (SWD) evaluations. Building on this, An et al. [2024] introduces reinforcement learning (RL) for adaptive EEG generation. The RL agent dynamically adjusts diffusion model parameters to optimize EEG synthesis, ensuring that generated signals maintain strong spectral and temporal fidelity. The combination of diffusion models and reinforcement learning further improves classification accuracy in downstream BCI applications. Torma and Szegletes [2025] further refine diffusion models for EEG synthesis by incorporating Implicit Sampling (DDIMs) to reduce inference steps while maintaining sample quality and leveraging Progressive Distillation to accelerate training and enable single-step EEG generation. Their study demonstrates that diffusion models outperform GANs and VAEs in EEG augmentation tasks, particularly for Motor Imagery decoding in BCI applications.

While these works focus on generation and augmentation, Diff-E [Kim et al., 2023] is a key step toward decoding imagined speech using DDPMs. Our work builds directly on this foundation—we use the Diff-E codebase and extend it to multi-paradigm EEG decoding with a novel dual-objective design. Unlike Diff-E, which is tailored to imagined speech, our approach generalizes across diverse paradigms by jointly optimizing reconstruction and classification. As such, Diff-E is not treated as a separate baseline but rather as the architectural backbone from which our models are developed.

Mixup and sample augmentation for EEG

To implement the EEG generated using DDPMs and other generated data, a temporal mixup strategy was introduced. In a recent work by Luo and Cai [2024], they reported a change in the ERD/ERS characteristics from generated EEG data. The original data helps preserve the temporal information. Other mixup strategies such as time-frequency transformation and spatial mixup, can also help incorporate generated data.

3 Methodology

3.1 Model Description

3.1.1 EEGNet

In this study, we employ EEGNet, a compact convolutional neural network (CNN) designed for EEG-based BCI [Lawhern et al., 2018]. EEGNet is a commonly used baseline model for EEG signal classification [Cichy et al., 2023, Ravipati et al., 2024], offering the following advantages: (1) applicability across multiple BCI paradigms, (2) effective training with limited data, and (3) extraction

of neurophysiologically interpretable features. It utilizes depthwise and separable convolutions to efficiently capture spatiotemporal information while minimizing the number of trainable parameters. Its architecture consists of temporal convolutions for frequency filtering, depthwise convolutions for spatial feature extraction, and separable convolutions for classification, ensuring both efficiency and interpretability. A visualization and detailed description of EEGNet are provided below.

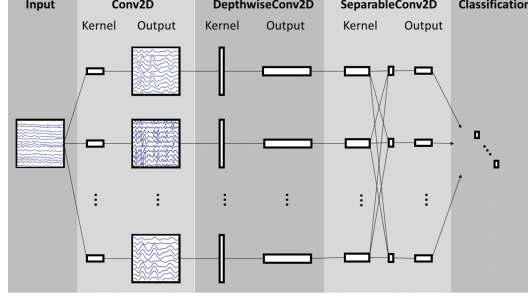


Figure 1: Overall visualization of the EEGNet architecture [Lawhern et al., 2018]. Lines denote the convolutional kernel connectivity between inputs and outputs (called feature maps). The network starts with a temporal convolution (second column) to learn frequency filters, then uses a depthwise convolution (middle column), connected to each feature map individually, to learn frequency-specific spatial filters. The separable convolution (fourth column) is a combination of a depthwise convolution, which learns a temporal summary for each feature map individually, followed by a pointwise convolution, which learns how to optimally mix the feature maps together. Full details about the network architecture can be found in Table 1.

Block	Layer	# Filters	Size	# Params	Output	Activation	Options
1	Input	-	-	-	(C, T)	-	-
	Reshape	-	-	-	$(1, C, T)$	-	-
	Conv2D	F_1	$(1, 64)$	$64 \times F_1$	(F_1, C, T)	Linear	mode=same
	BatchNorm	-	-	$2 \times F_1$	(F_1, C, T)	-	-
	DepthwiseConv2D	$D \times F_1$	$(C, 1)$	$C \times D \times F_1$	$(D \times F_1, 1, T)$	Linear	depth=D, mode=valid, max_norm=1
	BatchNorm	-	-	$2 \times D \times F_1$	$(D \times F_1, 1, T)$	-	-
	Activation	-	-	-	$(D \times F_1, 1, T)$	ELU	-
	AvgPool2D	-	$(1, 4)$	-	$(D \times F_1, 1, T/4)$	-	-
	Dropout*	-	-	-	$(D \times F_1, 1, T/4)$	-	p = 0.25 or 0.5
	Dropout*	-	-	-	$(D \times F_1, 1, T/4)$	-	-
2	SeparableConv2D	F_2	$(1, 16)$	$16 \times D \times F_1 + F_2 \times (D \times F_1)$	$(F_2, 1, T/4)$	Linear	mode=same
	BatchNorm	-	-	$2 \times F_2$	$(F_2, 1, T/4)$	-	-
	Activation	-	-	-	$(F_2, 1, T/4)$	ELU	-
	AvgPool2D	-	$(1, 8)$	-	$(F_2, 1, T/32)$	-	-
	Dropout*	-	-	-	$(F_2, 1, T/32)$	-	p = 0.25 or 0.5
	Dropout*	-	-	-	$(F_2, 1, T/32)$	-	-
	Flatten	-	-	-	$F_2 \times (T/32)$	-	-
Classifier	Dense	$N \times (F_2 \times T/32)$	-	-	N	Softmax	max_norm = 0.25

Table 1: EEGNet architecture [Lawhern et al., 2018], where C = number of channels, T = number of time points, F_1 = number of temporal filters, D = depth multiplier (spatial filters), F_2 = number of pointwise filters, and N = number of classes. Dropout: $p = 0.5$ (within-subject) or $p = 0.25$ (cross-subject).

3.1.2 Dual-Objective Diffusion Model

Proposed model 1: *DualDiff-Latent*

DualDiff-Latent is a dual-objective architecture designed to jointly optimize signal reconstruction and classification from EEG data. The model consists of three key components: a Denoising Diffusion Probabilistic Model (DDPM), a discriminative encoder, and a generative decoder (see Figure 2). The summary of key variables, shapes, and module-level parameter counts are shown in Table 2.

Given an EEG sample $x \in \mathbb{R}^{C \times T}$, where C and T are the number of channels and timepoints, the model performs the following operations.

A conditional UNet-based DDPM processes the input EEG signal x by learning a reverse diffusion trajectory that gradually removes noise and produces a refined signal $\hat{x}$. During this process, intermediate multi-scale features (dn11, dn22, dn33) are extracted and retained as skip connections for later reconstruction. In parallel, the same input is passed through a modified EEGNet encoder consisting of three temporal downsampling blocks and an attention pooling layer. This encoder yields a compact latent representation $z \in \mathbb{R}^D$, along with downsampled features dn1, dn2, and dn3.

To reduce subject-specific variability and improve class separability, the latent vector z is locally normalized and then projected through a lightweight head to obtain z_{proj} , which serves as input to a supervised contrastive loss. The same latent vector is also mapped via a fully connected layer to produce class logits $\hat{y}$, optimized using a classification loss (see Table 6).

The decoder reconstructs the EEG signal by integrating information from both the generative and discriminative branches. It takes as input the raw signal x , the DDPM-refined output $\hat{x}$, and skip features from the DDPM (dn11, dn22, dn33) and encoder (dn3). Two L1 reconstruction losses are defined: one between $\hat{x}$ and x , used to guide the DDPM, and another between the decoder output x_{dec} and $\hat{x}$, which is the only reconstruction loss included in the final training objective to preserve the learned diffusion trajectory.

Component	Code Name	Symbol	Shape	Description
Input				
	x	x	$B \times C \times T$	Raw EEG input
DDPM (265K parameters)				
	$\hat{x}$	$\hat{x}$	$B \times C \times T$	Denoised signal
	dn11	dn11	$B \times C_1 \times T_1$	DDPM downsampled feature (level 1)
	dn22	dn22	$B \times C_2 \times T_2$	DDPM downsampled feature (level 2)
	dn33	dn33	$B \times C_3 \times T_3$	DDPM downsampled feature (level 3)
Encoder (EEGNet-based) (61K parameters)				
	dn1	dn1	$B \times C'_1 \times T'_1$	Encoder downsample (level 1)
	dn2	dn2	$B \times C'_2 \times T'_2$	Encoder downsample (level 2)
	dn3	dn3	$B \times C'_3 \times T'_3$	Encoder downsample (level 3)
	z	z	$B \times D$	Latent vector
Classifier (410K parameters)				
	z_{proj}	z_{proj}	$B \times D'$	Projected vector for SupCon
	fc_out	$\hat{y}$	$B \times N$	Class logits
Decoder (156K parameters)				
	decoder_out	x_{dec}	$B \times C \times T$	Reconstructed output
Total: ~627K parameters				

Table 2: Summary of key variables, shapes, and module-level parameter counts in *DualDiff-Latent*.

To assess the role of different decoder inputs, we conducted an ablation study on the SSVEP task. As shown in Table 3, the default configuration (using the combination $x + \hat{x} + \text{skips}$) was initially selected based on strong baseline performance across both seen and unseen subjects. Building on this setup, we systematically evaluated alternative input combinations. Among them, the use of $z + \hat{x}$ led to the best results, achieving 84.27% seen-subject accuracy and 88.14% unseen-subject accuracy, suggesting that combining latent representations with denoised signal features yields the most effective decoder input.

The results further suggest that different input sources contribute differently depending on the evaluation setting. The latent vector z plays a dominant role in promoting generalization to unseen subjects, likely due to its compact, class-discriminative nature. The denoised signal $\hat{x}$, derived from the DDPM, consistently improves performance across both domains, offering refined structural information. In contrast, the raw signal x contributes more noticeably to performance on seen subjects, potentially capturing subject-specific patterns. Configurations that exclude latent or denoised components (such as those relying only on skip features) show significant performance drops, underscoring the importance of integrating both semantic and cleaned signal-level information during reconstruction.

Taken together, these findings indicate that $z + \hat{x}$ offers the best trade-off between generalization and discriminability. The default setting remains a competitive and balanced choice, but the ablation results highlight a more optimal configuration. *DualDiff-Latent* effectively combines generative and discriminative modeling: the encoder extracts class-discriminative latent features, while the decoder reconstructs structured EEG signals by integrating denoised outputs and hierarchical skip connections.

This hybrid design supports improved robustness and cross-subject generalization, particularly under the variability inherent in EEG decoding tasks.

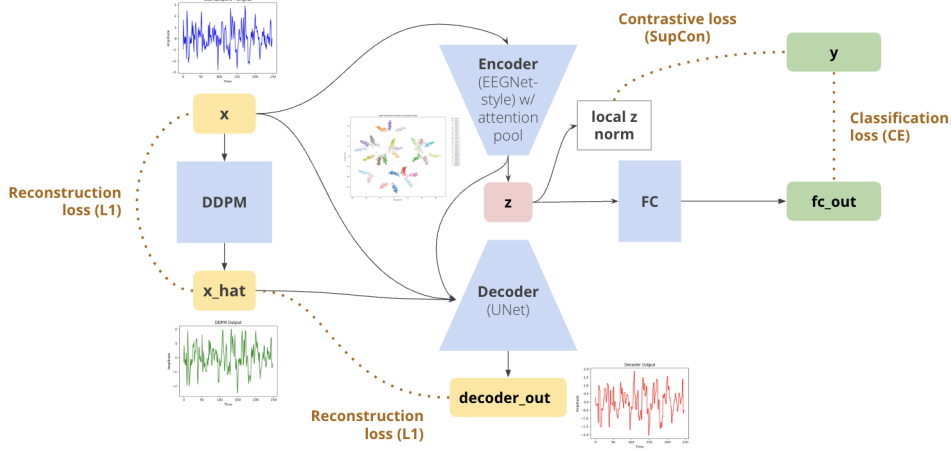


Figure 2: *DualDiff-Latent v2* model diagram

Table 3: Ablation study of decoder input combinations on the SSVEP task using *DualDiff-Latent v2*. Accuracy is reported for both seen-subject and unseen-subject test splits.

Task Paradigm	Model	Decoder Input	Seen Acc. (%)	Unseen Acc. (%)
SSVEP	<i>DualDiff-Latent v2</i> (attention)	x + x_hat + skips (default)	84.27	87.82
		x + x_hat	83.80	87.18
		x_hat + skips	80.19	87.82
		x + skips	83.10	86.54
		skips only	81.70	84.94
		z only	82.87	88.14
		z + x	83.80	87.82
		z + x_hat	84.27	88.14
		z + skips	82.05	86.22

Proposed model 2: *DualDiff-Mixup*

This section aim to test data augmentation using the DDPM and DiffE generated data. In *DualDiff-Mixup* (Figure 3), the model used was similar to the *DualDiff-Latent* model except the encoder for the CAE was the original UNet instead of EEGNet-based. The training stage for this model was split into two stages. First, the two generators (DDPM and DiffE) were trained using classification loss (α), contrastive loss (β), and reconstruction loss (γ). A few different combinations of the three losses were tested (4). In stage two, the generators were only used to generate $\hat{x}$ and decoder output. Both were used to for the mixup training of the classifier, which was an EEGNet. To do the augmentation, two approaches were tested. The first was to mix the input by a simple weighted addition of the original input x and each of the generated ($\hat{x}$ and decoder output) then normalized by the sum of weights. The augmented input was used to generate a single y to calculate the classification loss. The second was to pass all three different inputs through the encoder and classifier to generate three predicted outputs, y . All three y 's were used to calculate three different classification loss, and the final loss was a weighted sum of the three different losses. (4).

The results showed a plunge in performance with lower accuracy in both seen and unseen test subjects. It is likely that the decoders magnified the between-subjects variance (individual specific noise), which could explain the extra drop in unseen data accuracy (extra overfitting to the signals from between-subjects variance and not between labels variance).

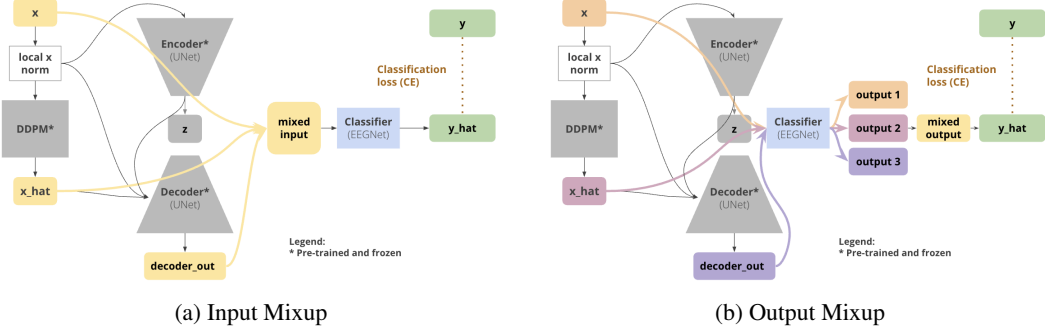


Figure 3: *DualDiff-Mixup* model diagram with input vs output mixup strategies

Table 4: Ablation study on loss and input/output strategies. Each configuration is denoted as $\alpha/\beta/\gamma$, representing weights for classification (CE), contrastive (SupCon), and reconstruction (L1) losses, respectively. Rows show first-stage loss combinations, and columns indicate the weighted composition of x , $\hat{x}$, and decoder output used in the second stage. Accuracy is reported as Seen / Unseen, and it's measured in %.

Strategy	First vs Second Stage	$x = 1, \hat{x} = 1, \text{decoder_out} = 1$	$x = 1, \hat{x} = 2, \text{decoder_out} = 2$	$x = 0, \hat{x} = 1, \text{decoder_out} = 1$
3*Input Mixup	$\alpha = 1, \beta = 1, \gamma = 1$	74 / 33	72 / 26	70 / 30
	$\alpha = 0, \beta = 1, \gamma = 1$	74 / 31	70 / 27	69 / 30
	$\alpha = 0, \beta = 0, \gamma = 1$	73 / 31	73 / 29	69 / 29
3*Output Mixup	$\alpha = 1, \beta = 1, \gamma = 1$	63 / 29	74 / 38	73 / 36
	$\alpha = 0, \beta = 1, \gamma = 1$	75 / 37	75 / 36	74 / 38
	$\alpha = 0, \beta = 0, \gamma = 1$	75 / 42	73 / 35	72 / 38

Proposed model 3: *DualDiff-Hybrid*

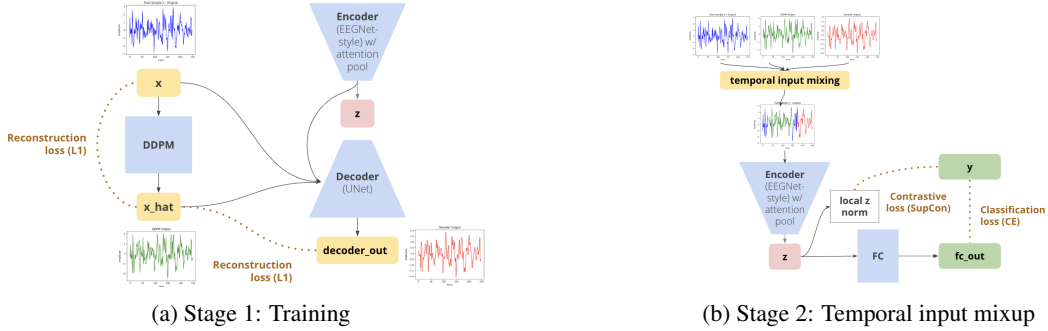


Figure 4: *DualDiff-Hybrid* with a two-stage training strategy and a temporal input mixup

In *DualDiff-Hybrid*, the model used was the same as the *DualDiff-Latent* v2 and a temporal mixup was implemented on the original plus the generated inputs instead of a simple weighted sum. The temporal mixup was done by creating a mask with a random number between -1 and 1, then up-sampling the mask based on the maximum allowed augmentation block with a Gaussian centered at each up-sampled point. The final input was a combination of x , $\hat{x}$, and decoder output depending on the corresponding value of the mask created. This is an approximation to traditional masking which was not efficient when for masking with generated inputs and differential masking for each channel.

Two similar strategies were tested, with one training the two generators first with only the reconstruction loss. Then, mixup the original and generated inputs into a single input to train the classifier with only the contrastive and classification losses (Figure 4). The second approach (cold-start, not shown in the figure) was similar to the first one, except that the generator and the classifier was trained together. The generated inputs was mixed to original input using the same temporal mixup after some epochs when the generators achieved reasonable performance, and the whole network continues to train together.

DualDiff-Hybrid did not achieve better performance than the original *DualDiff-Latent v2* model, suggesting that the latent z already captured most of the information from the original input and the generated inputs did not contribute extra information or improve stability.

3.2 Dataset

To evaluate the robustness and generalizability of our dual-objective diffusion model, we tested across five EEG datasets, each representing a distinct neural decoding task. These tasks span a spectrum of difficulty levels, ranging from stimulus-driven responses like SSVEP to highly variable imagined speech decoding, allowing us to benchmark performance under diverse signal dynamics. Table 5 summarizes the paradigm, dataset origin, task structure, and preprocessing steps. Our main development and ablation experiments were conducted on the SSVEP dataset, while transferability was validated on Motor Imagery, P300, and FEIS benchmarks. We also evaluated both raw time-domain and spectrogram-based representations to assess input robustness across domains.

For each dataset, we partitioned the data into training, validation, and two distinct test sets: a seen-subject (intra-subject) split and an unseen-subject (cross-subject) split. This setup allows us to rigorously evaluate both within-subject performance and the model’s ability to generalize to entirely new subjects.

Table 5: EEG Datasets Used for Evaluation, Ranked by Task Difficulty

Paradigm (Rank)	Dataset	Task Details	Preprocessing
SSVEP (Easy) -> also used to refine the methods	Tsinghua SSVEP [Wang et al., 2017]	26-class frequency-tagged visual stimulus decoding	Bandpass 6–90 Hz, downsampled to 250 Hz, 64 channels
P300 (Medium)	2 datasets: BCI Competition III [Blankertz et al., 2004] P300 Speller [Carzaniga and Gualniera, 2019]	ERP-based 36-character speller using binary classification	Bandpass 0.1–20 Hz, 650 ms epochs, Z-score normalization, class balancing
Motor Imagery (Hard)	BCI Competition IV 2a	4-class motor intention decoding (left/right hand, foot, tongue)	Bandpass 4–40 Hz, 22 channels, 2–6s epochs, one-hot labels
Brain-to-Speech (Hardest)	FEIS [Wellington and Clayton, 2019]	16-phoneme decoding during imagined, heard, and spoken speech	Z-score normalization, 14 channels, 256-timepoint epochs

3.3 Loss Functions

The model is trained using a weighted combination of three loss components, each targeting a distinct learning objective and applied to a different part of the architecture:

$$\mathcal{L}_{total} = \underbrace{\alpha \cdot \mathcal{L}_{CE/MSE}(\hat{y}, y)}_{classification} + \underbrace{\beta \cdot \mathcal{L}_{L1}(x_{dec}, \hat{x})}_{reconstruction} + \underbrace{\gamma \cdot \mathcal{L}_{SupCon}(z_{proj}, y)}_{contrastive}$$

To ensure training stability and balanced optimization, we adopt a progressive weighting strategy over the 500 training epochs. The classification loss weight α is fixed at 1.0 throughout training. In contrast, the weights of the reconstruction and contrastive terms are gradually increased during the early stages to avoid overwhelming the initial learning dynamics.

Specifically, the L1 reconstruction loss weight is defined as:

$$\beta = \min \left(1.0, \frac{epoch}{100} \right) \times 0.05,$$

and the supervised contrastive loss weight is scheduled as:

$$\gamma = \min \left(1.0, \frac{epoch}{50} \right) \times 0.2.$$

This schedule allows the model to first focus on classification and then progressively incorporate structural and reconstruction constraints, leading to more stable convergence and better representation learning.

The loss functions used or referenced throughout this work are summarized and formalized below.

Cross-entropy (CE) loss. This classification loss compares the predicted softmax output $\hat{y} = \text{Softmax}(FC(z))$ with the one-hot encoded label $y \in \{0, 1\}^N$, and is defined as:

$$\mathcal{L}_{CE} = - \sum_{c=1}^N y_c \log \hat{y}_c$$

It trains the classifier head based on the latent representation z , enforcing clear decision boundaries in the embedding space.

Mean squared error (MSE). In ablation experiments, the CE loss is replaced by an MSE loss that uses the same classifier output $\hat{y} = FC(z)$, omitting the softmax activation:

$$\mathcal{L}_{MSE} = \|\hat{y} - y\|_2^2$$

This formulation treats classification as a regression task over class logits. In addition, we test a scaled variant denoted as MSE2. In this variant, softmax is applied to the output logits before computing the MSE:

$$\mathcal{L}_{MSE2} = \|\text{Softmax}(\hat{y}) - y\|_2^2$$

Since the softmax operation significantly reduces the magnitude of the predictions, the scale of the MSE loss also drops. To compensate for this effect, the weights for the SupCon and L1 loss terms are reduced by a factor of 10, ensuring that MSE2 plays a more dominant role in the total loss and remains comparable in scale to CE.

Supervised contrastive (SupCon) loss. To structure the latent space, the latent vector z is locally normalized and projected:

$$z_{proj} = \text{ProjHead}(\text{Norm}(z))$$

The SupCon loss is then defined as:

$$\mathcal{L}_{SupCon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}$$

where τ is the temperature parameter, $P(i)$ is the set of positives for anchor i , and $A(i)$ includes all other anchors in the batch. This term improves generalization by enforcing class-wise clustering.

L1 reconstruction loss. This loss ensures that the decoder output x_{dec} resembles the denoised signal $\hat{x}$ produced by the DDPM:

$$\mathcal{L}_{L1} = \|x_{dec} - \hat{x}\|_1$$

An additional loss $\|\hat{x} - x\|_1$ is used internally to train the DDPM but is excluded from the total loss to avoid interfering with decoder learning.

Table 6: Ablation study of loss function combinations on the SSVEP task using *DualDiff-Latent v2*. Accuracy is reported for both seen-subject and unseen-subject test splits.

Task Paradigm	Model	Loss Combination	Seen Acc. (%)	Unseen Acc. (%)
SSVEP	<i>DualDiff-Latent v2</i> (attention)	$\alpha\text{CE} + \beta\text{L1} + \gamma\text{SupCon}$ (default)	84.27	87.82
		αCE only	79.72	84.62
		αMSE only	81.82	86.86
		$\alpha\text{CE} + \beta\text{L1}$	80.77	84.29
		$\alpha\text{CE} + \gamma\text{SupCon}$	80.07	83.33
		$\alpha\text{MSE} + \beta\text{L1} + \gamma\text{SupCon}$	82.63	88.14
		$\alpha\text{MSE2} + \beta\text{L1} + \gamma\text{SupCon}$	81.70	87.50

The total loss consists of a classification term (α), a reconstruction term (β), and a contrastive term (γ). For all configurations, $\alpha = 1.0$; β and γ follow the progressive schedule in Section 3.3, except for the last (MSE2), where both are down-weighted by a factor of 10.

To evaluate the impact of each loss term, we conducted an ablation study with various configurations on the SSVEP task using the *DualDiff-Latent* model, under the default decoder input setting. In addition to the default configuration that combines CE, L1, and SupCon, we tested replacements where CE was substituted with MSE or its softmax-based variant MSE2. The results are shown in Table 6.

The MSE-based settings achieved slightly better performance on unseen-subject data but consistently degraded performance on seen-subjects. In particular, MSE2 applied softmax to the output logits and scaled down the weights of the L1 and SupCon losses by a factor of 10 to compensate for the reduced gradient scale. This adjustment resulted in a slightly more stable loss curve compared to MSE, but final accuracy was lower. For MSE alone, we observed unstable training dynamics, with loss values exhibiting an abnormal spike followed by a sudden drop, suggesting potential incompatibility with the auxiliary losses. Overall, both versions failed to outperform the CE-based setting in a balanced way.

Other partial combinations such as CE+L1 or CE+SupCon also failed to match the robustness of the full setup, typically performing worse on either seen or unseen sets. These findings reinforce our conclusion that the CE+L1+SupCon configuration remains the most reliable and well-rounded design across subject conditions.

3.4 Evaluation Metrics

Downstream classification accuracy is used. Measures the proportion of correctly classified EEG samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

4 Baseline

We used EEGNet as a unified baseline model across all tasks due to its proven effectiveness in EEG classification. Its lightweight architecture, based on depthwise and separable convolutions, efficiently captures both spatial and temporal features, making it suitable for diverse paradigms such as Motor Imagery, Brain-to-Speech Decoding, SSVEP, and P300.

To ensure fair cross-task comparisons, we standardized the model architecture and relied on publicly available implementations. This consistency eliminates architectural variability and allows us to focus solely on performance differences driven by task complexity and data characteristics. Accordingly, we kept all EEGNet settings at their default values, except for task-specific adjustments to the number of output classes (`nb_classes`), input channels (`Chans`), and time samples (`Samples`).

All tasks were framed as classification problems, and evaluation was based on downstream accuracy (see Equation 1) for both seen- and unseen-subject test splits.

5 Experiments and Results

Among all model variants evaluated, the *DualDiff-Latent* series demonstrated strong performance improvements over the EEGNet baseline, particularly on the 26-class SSVEP task. The gains were most pronounced on unseen-subject data, while for harder tasks like Motor Imagery (MI) and imagined speech (FEIS), improvements were especially notable on seen-subject accuracy.

The model evolved over three versions. *DualDiff-Latent v1* introduced local normalization of latent features. *DualDiff-Latent v2* incorporated attention pooling to enhance the encoder’s discriminative power and was used for all non-SSVEP experiments. Building on both, *DualDiff-Latent v3* further optimized the decoder input by combining the latent vector z with the DDPM-refined signal $\hat{x}$, resulting in the best generalization performance (see ablation in Table 3). Specifically, *DualDiff-Latent v3* achieved the highest unseen-subject accuracy (88.14%) while maintaining strong seen-subject performance (84.27%). Training curves for this configuration (Figure 5) show fast convergence and stable validation accuracy, supporting the effectiveness of the progressive multi-loss weighting

Table 7: EEGNet performance across multiple EEG datasets. Accuracy is reported for both seen-subject (intra-subject) and unseen-subject (cross-subject) test splits, where applicable. Tasks are ranked by difficulty.

Task Paradigm	Dataset	Subjects	Classes	Seen Acc. (%)	Unseen Acc. (%)
1*SSVEP	Tsinghua BCI Benchmark	35	26	85.54	81.03
2*P300	BCI Competition III P300 Speller	2	2	74.00 (mixed)	
		2	2	77.98 (mixed)	
Motor Imagery	BCI Competition IV 2a	9	4	69.00 (mixed)	
2*Speech Production	2*FEIS	21	2	74.83 (mixed)	
		21	16	5.97	6.78
2*Heard Speech	2*FEIS	21	2	74.87 (mixed)	
		21	16	5.99	5.78
2*Imagined Speech	2*FEIS	21	2	74.84 (mixed)	
		21	16	6.02	6.36

strategy. A mild degree of overfitting can be observed in the later epochs, though it does not substantially affect overall generalization.

A mini t-SNE visualization of latent features from *DualDiff-Latent* v2 (Figure 2, center) reveals distinct label-level clustering, where samples of the same class form compact and well-separated groups. In contrast, subjects are broadly distributed within each label cluster, suggesting that the encoder effectively abstracts away subject-specific variability. This indicates that the model captures class-discriminative yet subject-invariant representations, which is critical for cross-subject EEG decoding. The latent structure learned through attention-enhanced encoding and supervised contrastive loss aligns with the strong generalization performance observed in unseen-subject settings.

Overall, *DualDiff-Latent* provides a unified framework that balances generative and discriminative objectives, scales across diverse EEG paradigms, and consistently outperforms task-specific baselines.

Table 8: Final results across tasks and models. Accuracy is reported for both seen-subject (intra-subject) and unseen-subject (cross-subject) test splits. Tasks are ranked by task difficulty.

Task Paradigm	Model	Subjects	Classes	Seen Acc. (%)	Unseen Acc. (%)
6*SSVEP	EEGNet	35	26	85.54	81.03
	<i>DualDiff-Latent</i> v1 (local z norm)	35	26	83.33	86.22
	<i>DualDiff-Latent</i> v2 (attention)	35	26	84.27	87.82
	<i>DualDiff-Latent</i> v3 (decoder input: $z + \hat{x}$)	35	26	84.27	88.14
	<i>DualDiff-Mixup</i> v1 (input mixup)	35	26	74.00	33.00
	<i>DualDiff-Mixup</i> v2 (output mixup)	35	26	75.00	42.00
	<i>DualDiff-Hybrid</i> v1 (2-stage + temporal)	35	26	84.15	87.50
	<i>DualDiff-Hybrid</i> v2 (cold-start + temporal)	35	26	81.59	84.94
2*P300	EEGNet	–	2	74.00–77.98 (mixed)	
	<i>DualDiff-Latent</i> v2 (attention)	–	2	–	
2*Motor Imagery (MI)	EEGNet	9	4	69.00 (mixed)	
	<i>DualDiff-Latent</i> v2 (attention)	9	4	81.12	67.36
2*Heard Speech	EEGNet	21	16	5.99	5.78
	<i>DualDiff-Latent</i> v2 (attention)	21	16	43.85	6.38
2*Speech Production	EEGNet	21	16	5.97	6.78
	<i>DualDiff-Latent</i> v2 (attention)	21	16	41.69	6.48
2*Imagined Speech	EEGNet	21	16	6.02	6.36
	<i>DualDiff-Latent</i> v2 (attention)	21	16	45.05	7.36

6 Summary and Discussion

Among all evaluated tasks, the SSVEP setting showed the most favorable results under the *DualDiff-Latent* framework. In particular, the *DualDiff-Latent* v3 variant achieved strong accuracy on both seen- and unseen-subject splits. The latent representation z was key to this success, functioning not only as a class-discriminative embedding for the classifier, but also as a generative feature that enhanced decoder reconstruction. Interestingly, the model performed slightly better on unseen subjects than on

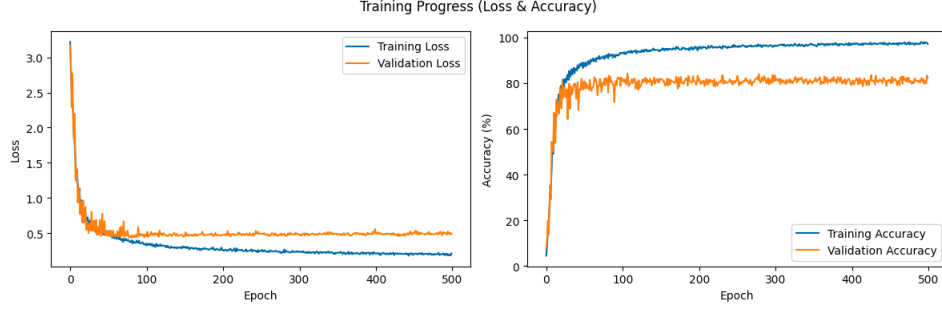


Figure 5: Training curves for the best model (*DualDiff-Latent v3*): (Left) Training and validation loss; (Right) Training and validation accuracy.

seen ones, possibly because the training emphasized class-level structure while avoiding overfitting to individual-specific EEG patterns. The training curves indicate a mild degree of overfitting in later epochs, suggesting room for improvement in regularization or data diversity.

For all the mixup strategies tested, *DualDiff-Mixup* directly used the original and generated inputs without exploiting the latent space, z , obtained by training the autoencoder. The results showed deteriorated performance and heavy overfitting. This suggested that the decoders reconstructed the between-subject variance more than expected and added additional noise. It was better to work in the embedded space. *DualDiff-Hybrid* confirmed that the latent z trained on original inputs only already contained enough between-label variance. In addition, the reconstruction might have focused more on recreating the between-subject variance (similar to what was seen in *DualDiff-Mixup*'s). The mixup injected more of this noise back to the model, and eventually to latent space, resulting in equal or slightly worse performance.

For the MI task, *DualDiff-Latent* achieved strong performance in the subject-dependent setting, suggesting that it can effectively extract discriminative motor-related features. However, similar to speech-related tasks, the model struggled to generalize under cross-subject evaluation. This reflects the inherent difficulty of motor imagery decoding, which involves more abstract and variable neural representations across subjects, and underscores the need for improved subject-invariant modeling strategies.

For the FEIS dataset, which involved the 16-class classification of phonemes during heard speech, speech production, and imagined speech, our results reveal a stark contrast in performance between subject-dependent and subject-independent scenarios. For data from subjects seen during training ('seen'), our proposed *DualDiff-Latent v2* model demonstrated a dramatic improvement over the EEGNet baseline across all three conditions (6% $\rightarrow$ 43%). This substantial gain highlights the effectiveness of the dual-objective approach in capturing complex, subject-specific neural patterns associated with phoneme processing.

However, the performance on 'unseen' subjects, tells a different story. Both the EEGNet baseline and our *DualDiff-Latent v2* model performed poorly, achieving accuracies only slightly above the chance level. This significant drop underscores the challenge posed by inter-subject variability in EEG signals for complex tasks like phoneme decoding. The current architecture and dual-objective training strategy, despite their benefits, were insufficient to learn the generalizable, subject-invariant features necessary to overcome the high variability inherent in the FEIS dataset for cross-subject classification. Bridging this generalization gap for complex EEG decoding remains a critical direction for future research.

7 Future Work

Next, we plan to run all experiments with multiple random seeds to assess statistical significance and ensure result reliability. Instead of mixing raw input signals, we aim to experiment with mixing latent embeddings within the encoder, performing ablations at different intermediate layers to determine the most effective fusion point. Also, we plan to incorporate spectrogram-based features into our pipeline to better capture the spectral structure of imagined speech and motor commands. In parallel,

we aim to fuse EEG signals from multiple stages of speech processing, such as imagined, heard, and spoken, by using separate encoders for each stage, enabling the model to learn complementary representations. To improve generalization on low-resource tasks like imagined speech decoding, we will also explore transfer learning by pretraining models on related datasets such as overt speech EEG.

8 Code

Github: <https://github.com/eddieguo-1128/DualDiff>

9 Acknowledgments

The authors would like to thank Professor Bhiksha Raj (Carnegie Mellon University) for proposing the research direction, and Kateryna Shapovalenko (Carnegie Mellon University) for mentorship and guidance throughout the project.

10 Team Contributions

Mengchun Zhang: Ran baseline experiments for both P300 datasets, developed the complete *DualDiff-Latent* architecture, and performed comprehensive ablation studies on SSVEP to evaluate the contribution of the latent space.

Eddie Guo: Ran the baseline experiments on the FEIS dataset and conducted all evaluations of the *DualDiff-Latent v2* model on FEIS.

Yucheng Shao: Ran the SSVEP baseline and implemented the full suite of *DualDiff-Mixup* models, conducting all related experiments and ablations on the SSVEP dataset.

Parusha Pradhan: Implemented the baseline model for the Motor Imagery dataset and ran *DualDiff-Latent v2* evaluations on Motor Imagery and partially on FEIS.

References

- Yang An, Yuhao Tong, Weikai Wang, and Steven W. Su. Enhancing eeg signal generation through a hybrid approach integrating reinforcement learning and diffusion models, 2024. URL <https://arxiv.org/abs/2410.00013>.
- Benjamin Blankertz, Dean Krusienski, and Gerwin Schalk. BCI Competition III - Data Set II: P300 Speller Paradigm. Wadsworth Center, NYS Department of Health, 2004. URL <http://www.bbc.de/competition/iii/>. Available online: <http://www.bbc.de/competition/iii/> (accessed March 17, 2025).
- Manuel Carzaniga and Lorenzo Gualniera. P300 speller dataset - open-source eeg-based p300 speller implementation. GitHub repository, 2019. URL <https://github.com/Manucar/p300-speller>. Available online: <https://github.com/Manucar/p300-speller> (accessed March 17, 2025).
- Benjamin Cichy, Jamie Lukos, Mohammad Alam, J Cortney Bradford, and Nicholas Wymbs. A convolutional network adaptation for cortical classification during mobile brain imaging. *arXiv preprint arXiv:2310.07941*, 2023.
- Soowon Kim, Young-Eun Lee, Seo-Hyun Lee, and Seong-Whan Lee. Diff-e: Diffusion-based learning for decoding imagined speech eeg. *arXiv preprint arXiv:2307.14389*, 2023.
- 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. *Journal of neural engineering*, 15(5):056013, 2018.

- Tian-jian Luo and Zikun Cai. Diffusion models-based motor imagery eeg sample augmentation via mixup strategy. *Expert Systems with Applications*, 235:125585, 2024. doi: 10.1016/j.eswa.2024.125585. URL <https://www.sciencedirect.com/science/article/pii/S0957417424024527>.
- Yashwanth Ravipati, Nader Pouratian, Corey Arnold, and William Speier. Evaluating deep learning performance for p300 neural signal classification. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1218, 2024.
- Szabolcs Torma and Luca Szegletes. Generative modeling and augmentation of eeg signals using improved diffusion probabilistic models. *Journal of Neural Engineering*, 22(1):016001, 2025. doi: 10.1088/1741-2552/ada0e4.
- Giulio Tosato, Cesare M. Dalbagno, and Francesco Fumagalli. Eeg synthetic data generation using probabilistic diffusion models, 2023. URL <https://arxiv.org/abs/2303.06068>.
- Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for ssvep-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2017. doi: 10.1109/TNSRE.2016.2578980. URL <https://ieeexplore.ieee.org/document/7740878>.
- S. Wellington and J. Clayton. Fourteen-channel eeg with imagined speech (feis) dataset, 2019. URL <https://doi.org/10.5281/zenodo.3369178>.