ArtifactGen: Benchmarking WGAN-GP vs Diffusion for Label-Aware EEG Artifact Synthesis

Hritik Arasu
Department of Behavior and Brain Sciences
University of Texas at Dallas
Richardson, TX 75080
hritik.arasu@UTDallas.edu

Faisal R. Jahangiri
Department of Behavior and Brain Sciences
University of Texas at Dallas
Richardson, TX 75080
faisal.jahangiri@utdallas.edu

Abstract

Artifacts in electroencephalography (EEG)—muscle, eye movement, electrode, chewing, and shiver—confound automated analysis yet are costly to label at scale. We study whether modern generative models can synthesize realistic, label-aware artifact segments suitable for augmentation and stress-testing. Using the TUH EEG Artifact (TUAR) corpus, we curate subject-wise splits and fixed-length multichannel windows (e.g., 250 samples) with preprocessing tailored to each model (per-window min–max for adversarial training; per-recording/channel z-score for diffusion). We compare a conditional WGAN-GP with a projection discriminator to a 1D denoising diffusion model with classifier-free guidance, and evaluate along three axes: (i) fidelity via Welch band-power deltas ($\Delta\delta$, $\Delta\theta$, $\Delta\alpha$, $\Delta\beta$, $\Delta\gamma$), channel-covariance Frobenius distance, autocorrelation L_2, and distributional metrics (MMD/PRD); (ii) specificity via class-conditional recovery with lightweight kNN/classifiers; and (iii) utility via augmentation effects on artifact recognition. In our setting, WGAN-GP achieves closer spectral alignment and lower MMD to real data, while both models exhibit weak class-conditional recovery, limiting immediate augmentation gains and revealing opportunities for stronger conditioning and coverage. All analyses in this version are post hoc from fixed checkpoints (no retraining). We release a reproducible pipeline—data manifests, training configurations, and evaluation scripts—to establish a baseline for EEG artifact synthesis and to surface actionable failure modes for future work.

1 Introduction

Artifacts in electroencephalography (EEG)—including muscle activity, eye movements, electrode noise, chewing, and shivering—routinely confound automated analysis and downstream clinical applications by distorting morphology, spectra, and cross-channel correlations. While artifact removal is well studied [Urigüen and García-Zapirain, 2015, Jiang et al., 2019], realistic *synthesis* of artifact segments can complement curation efforts by enabling data augmentation, algorithm stress testing, and robustness benchmarking without additional human labeling. The challenge is to synthesize multichannel windows that remain label-aware while respecting signal morphology, spectral structure, and channel covariance.

We introduce ArtifactGen, a practical and *reproducible* framework for artifact-conditioned EEG synthesis built on subject-wise splits from the Temple University Hospital EEG (TUH EEG) corpus and its artifact-annotated subset, the Temple University Artifact Corpus (TUAR) [Hamid et al., 2020]. ArtifactGen marries two complementary generative paradigms: (i) a conditional Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) with a projection discriminator

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

for stable, label-aware synthesis [Gulrajani et al., 2017, Miyato and Koyama, 2018], and (ii) a denoising Diffusion Probabilistic Model (DDPM) using a 1D U-Net with Feature-wise Linear Modulation (FiLM) conditioning [Perez et al., 2018] and classifier-free guidance (CFG) for controllability and sample quality [Ho et al., 2020, Ho and Salimans, 2022]. The pipeline standardizes preprocessing for fixed-length windows with configurable normalization, exposes training/evaluation via YAML configs, and ships analysis notebooks to facilitate faithful ablations and apples-to-apples comparisons.

Beyond single-number heuristics, ARTIFACTGEN emphasizes a time-series-appropriate evaluation suite: (i) signal-level descriptors (e.g., Welch band-power deltas and covariance/autocorrelation function (ACF) distances) to test morphology and spectra [Welch, 1967]; (ii) feature-space metrics (Fréchet Inception Distance (FID) / Kernel Inception Distance (KID) / Precision–Recall for Distributions (PRD)) to quantify fidelity–coverage trade-offs [Heusel et al., 2017, Binkowski et al., 2018, Sajjadi et al., 2018]; and (iii) functional tests—train-real/test-synth, train-synth/test-real, and AugMix-style augmentation—to probe utility and robustness [Hendrycks et al., 2020]. We release code, configuration files, and notebooks to support rigorous baselining and community progress on EEG artifact generation and augmentation.

2 Background and Related Work

Electroencephalography (EEG) is indispensable in clinical neurophysiology, yet real-world recordings are rife with non-neural artifacts—ocular movements, muscle activity, chewing, shivering, and electrode noise—that degrade downstream analysis and confound learning systems. Decades of signal-processing work have characterized these artifacts and proposed removal strategies, underscoring their broad spectral footprint and nonstationary morphology [Urigüen and García-Zapirain, 2015]. Large public corpora such as the Temple University Hospital EEG (TUH EEG) dataset [Obeid and Picone, 2016] and its artifact-focused subset, the TUH EEG Artifact Corpus (TUAR) [Hamid et al., 2020], enable supervised benchmarking but remain label- and condition-limited for models that must generalize across subjects, montages, and acquisition conditions.

Generative modeling provides a complementary route to synthesize realistic artifact segments for (i) augmenting scarce classes, (ii) stress-testing detector robustness, and (iii) studying controlled perturbations. Two families dominate recent progress: Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs). GANs are sample-efficient but historically unstable; Wasserstein GANs with gradient penalty (WGAN-GP) improved convergence via a soft Lipschitz constraint on the critic [Gulrajani et al., 2017], while projection discriminators inject label embeddings to enforce class-conditional realism [Miyato and Koyama, 2018]. In 1D biosignals, fully convolutional architectures such as WaveGAN preserve local stationarity and long-range context [Donahue et al., 2019].

Diffusion models take an alternative route, learning to reverse a progressive noising process [Ho et al., 2020]. Subsequent refinements—learned variance, hybrid objectives, and efficient samplers—improved fidelity and speed [Nichol and Dhariwal, 2021, Dhariwal and Nichol, 2021]. Classifier-free guidance (CFG) provides a practical mechanism for label adherence without explicit classifiers [Ho and Salimans, 2022]. Adaptations to time series leverage 1D U-Nets (e.g., DiffWave [Kong et al., 2020]) and score-based SDE frameworks [Song et al., 2020], while recent neurophysiology works demonstrate high realism and controllability in EEG and ECoG generation [Vetter et al., 2024, Tosato et al., 2023]. Our work follows this line, employing a FiLM-conditioned [Perez et al., 2018] 1D U-Net [Ronneberger et al., 2015] for artifact-aware EEG synthesis.

Evaluating synthetic EEG demands neurophysiology-aligned metrics. Power spectral density (PSD) comparisons via Welch's method [Welch, 1967] assess band-power differences in $\delta/\theta/\alpha/\beta$ bands, while autocorrelation and cross-channel covariance capture temporal and spatial dependencies. Distributional fidelity and coverage are assessed using precision–recall for distributions (PRD) [Sajjadi et al., 2018], kernel maximum mean discrepancy (MMD) [Binkowski et al., 2018, Gretton et al., 2012], and classifier two-sample tests (C2ST) [Lopez-Paz and Oquab, 2017]. In practice, features from compact discriminative backbones such as EEGNet [Lawhern et al., 2018] allow EEG-specific analogs of image metrics like FID [Heusel et al., 2017] and KID [Binkowski et al., 2018].

Beyond proxy measures, *functional* evaluation—training downstream models on synthetic data—offers the most meaningful validation. Train-on-synthetic, test-on-real (TSTR) protocols

[Yoon et al., 2019] and augmentation-style robustness tests [Hendrycks et al., 2020] directly measure utility. Recent studies suggest that diffusion models often match or surpass GANs in both fidelity and coverage while being more stable to train [Dhariwal and Nichol, 2021, Nichol and Dhariwal, 2021]. Together, these insights motivate our label-aware comparison between conditional WGAN-GP and conditional diffusion on TUAR, under subject-wise splits and a comprehensive evaluation spanning spectral, temporal, multichannel, and distributional criteria.

3 Dataset and Preprocessing

We curate EEG artifact segments from the Temple University Hospital EEG resources [Obeid and Picone, 2016]. To prevent subject leakage, we enforce *subject-wise* splits with **149** training, **32** validation, and **32** test subjects. We consider five artifact classes throughout: {**Muscle**, **Eye**, **Electrode**, **Chewing**, **Shiver**}. All scripts are configuration-driven and reproducible.

Channels and sampling. We adopt a canonical eight-channel montage $\{\mathrm{Fp1}, \mathrm{Fp2}, \mathrm{C3}, \mathrm{C4}, \mathrm{O1}, \mathrm{O2}, \mathrm{T3}, \mathrm{T4}\}$ at $f_s = 250$ Hz. Only recordings with all required channels are admitted.

Windowing and overlap. Let $x \in \mathbb{R}^{C \times T}$ denote a multi-channel clip (C=8). For a target window duration S seconds, the window length (in samples) is

$$L \;=\; \lfloor S\, f_s \rfloor. \tag{1}$$

Windows are extracted with fractional overlap $\rho \in [0, 1)$ (default ρ=0.5), giving stride

$$s \;=\; \lfloor (1-\rho)\, L \rfloor. \tag{2}$$

For an annotated interval of length T_i samples, the number of windows produced is

$$N_i \;=\; \max\Bigl(0,\; \Bigl\lfloor \frac{T_i - L}{s} \Bigr\rfloor + 1\Bigr). \tag{3}$$

Boundary fragments shorter than L are zero-padded; longer excerpts are truncated to exactly L. We use S=1.0 s (L=250) for the adversarial path and S=2.0 s (L=500) for the diffusion path.

Normalization (model-specific). Two normalization schemes are implemented and selected per run:

1. **Per-window min–max to $[-1, 1]$ (adversarial path).** For window $x \in \mathbb{R}^{C \times L}$ with global per-window extrema $m = \min_{c,t} x_{c,t}$ and $M = \max_{c,t} x_{c,t}$, we map

   $$\hat{x}_{c,t} \;=\; 2\,\frac{x_{c,t} - m}{\max(M - m, \epsilon)} \;-\; 1, \qquad \epsilon = 10^{-8}. \tag{4}$$

   If configured, the pair (m, M) is persisted with the window metadata to enable consistent inverse-rescaling at load time.

2. **Per-recording, per-channel z-score (diffusion path).** For channel c with mean μ_c and standard deviation σ_c computed over the recording,

   $$\tilde{x}_{c,t} \;=\; \frac{x_{c,t} - \mu_c}{\sigma_c + \epsilon}, \qquad \epsilon = 10^{-8}. \tag{5}$$

Filtering. Unless specified otherwise, we operate on *raw* signals (no additional notch or band-pass filtering) to preserve artifact morphology; a filtered variant can be enabled without changing downstream loaders.

Channels and sampling. We adopt a canonical eight-channel montage $\{\mathrm{Fp1}, \mathrm{Fp2}, \mathrm{C3}, \mathrm{C4}, \mathrm{O1}, \mathrm{O2}, \mathrm{T3}, \mathrm{T4}\}$ at $f_s = 250$ Hz. Only recordings with all required channels are admitted.

Table 1: Subject counts per split

	Train	Val	Test
Subjects	149	32	32

Windowing and overlap. Let $x \in \mathbb{R}^{C \times T}$ denote a multi-channel clip (C=8). For a target window duration S seconds, the window length (in samples) is

$$L = \lfloor S\, f_s \rfloor. \tag{6}$$

Windows are extracted with fractional overlap $\rho \in [0, 1)$ (default ρ=0.5), giving stride

$$s = \lfloor (1 - \rho)\, L \rfloor. \tag{7}$$

For an annotated interval of length T_i samples, the number of windows produced is

$$N_i = \max\Big(0, \Big\lfloor \frac{T_i - L}{s} \Big\rfloor + 1\Big). \tag{8}$$

Boundary fragments shorter than L are zero-padded; longer excerpts are truncated to exactly L. We use S=1.0 s (L=250) for the adversarial path and S=2.0 s (L=500) for the diffusion path.

Normalization (model-specific). Two normalization schemes are implemented and selected per run:

1. **Per-window min–max to $[-1, 1]$ (adversarial path).** For window $x \in \mathbb{R}^{C \times L}$ with global per-window extrema $m = \min_{c,t} x_{c,t}$ and $M = \max_{c,t} x_{c,t}$, we map

$$\hat{x}_{c,t} = 2\, \frac{x_{c,t} - m}{\max(M - m, \epsilon)} - 1, \qquad \epsilon = 10^{-8}. \tag{9}$$

   If configured, the pair (m, M) is persisted with the window metadata to enable consistent inverse-rescaling at load time.

2. **Per-recording, per-channel z-score (diffusion path).** For channel c with mean μ_c and standard deviation σ_c computed over the recording,

$$\tilde{x}_{c,t} = \frac{x_{c,t} - \mu_c}{\sigma_c + \epsilon}, \qquad \epsilon = 10^{-8}. \tag{10}$$

Filtering. Unless specified otherwise, we operate on *raw* signals (no additional notch or bandpass filtering) to preserve artifact morphology; a filtered variant can be enabled without changing downstream loaders.

Manifests, class maps, and splits. We supply (i) a subject-wise split CSV ensuring disjoint identities across train/val/test; (ii) a stable class map for the five artifact labels; and (iii) a consolidated manifest (JSON) that records per-window paths, labels, subject IDs, normalization statistics, and the effective L. These files fully reproduce dataset composition and preprocessing decisions.

Configuration (exact defaults). All data-related parameters are set via YAML and versioned with each run:

- `channels`: [Fp1, Fp2, C3, C4, O1, O2, T3, T4], `sample_rate`: 250 Hz, `overlap`: 0.5, `filtering`: `raw`.
- **Adversarial path (WGAN-GP):** `window_seconds` = 1.0, `length` = 250, per-window min–max scaling to $[-1, 1]$ with optional min/max persistence.
- **Diffusion path (DDPM):** `window_seconds` = 2.0, `length` = 500, per-recording, per-channel z-score normalization.
- `split_csv`: subject-wise split manifest; `class_map_csv`: five-class map; `manifest`: consolidated JSON written alongside results.

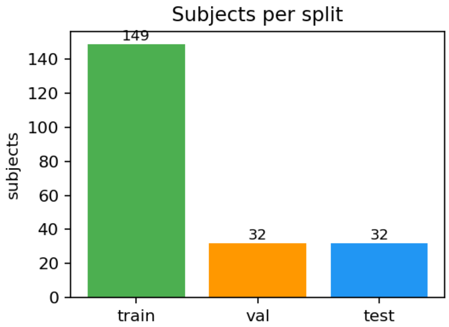


Figure 1: Bar chart of subjects per split from the manifest summary (no retraining).

Table 2: Methods Summary

Model	Win (s)	Length	Normalization	Ch	Classes	Sampler/Steps	CFG
WGAN-GP	1.0	250	`minmax_per_window`	8	5	—	—
DDPM	2.0	500	`zscore_per_recording`	8	5	ddim / 50	3.5

Manifests, class maps, and splits. We supply (i) a subject-wise split CSV ensuring disjoint identities across train/val/test; (ii) a stable class map for the five artifact labels; and (iii) a consolidated manifest (JSON) that records per-window paths, labels, subject IDs, normalization statistics, and the effective L. These files fully reproduce dataset composition and preprocessing decisions.

Configuration (exact defaults). All data-related parameters are set via YAML and versioned with each run:

- `channels`: $[\mathrm{Fp1}, \mathrm{Fp2}, \mathrm{C3}, \mathrm{C4}, \mathrm{O1}, \mathrm{O2}, \mathrm{T3}, \mathrm{T4}]$, `sample_rate`: 250 Hz, `overlap`: 0.5, `filtering`: `raw`.
- **Adversarial path (WGAN-GP):** `window_seconds` $= 1.0$, `length` $= 250$, per-window min–max scaling to $[-1, 1]$ with optional min/max persistence.
- **Diffusion path (DDPM):** `window_seconds` $= 2.0$, `length` $= 500$, per-recording, per-channel z-score normalization.
- `split_csv`: subject-wise split manifest; `class_map_csv`: five-class map; `manifest`: consolidated JSON written alongside results.

4 Methods

At-a-glance configuration. Table **??** summarizes the fixed settings per model used in this comparison.

4.1 Scope and Constraints

All updates here are post hoc using previously trained checkpoints. We do not retrain or alter model weights; new statistics and visualizations are computed from fixed checkpoints and manifests.

4.2 Conditional WGAN-GP with Projection Discriminator

We model artifact-conditioned synthesis as $G : \mathbb{R}^{d_z} \times \{1, \ldots, K\} \to \mathbb{R}^{C \times T}$, where $z \sim \mathcal{N}(0, I)$ and K is the number of artifact classes. For adversarial training we apply per-window min–max normalization to $[-1, 1]$, concatenate z with a one-hot label y, and upsample via a 1D transposed-convolutional generator to produce multi-channel windows $\tilde{x}$.

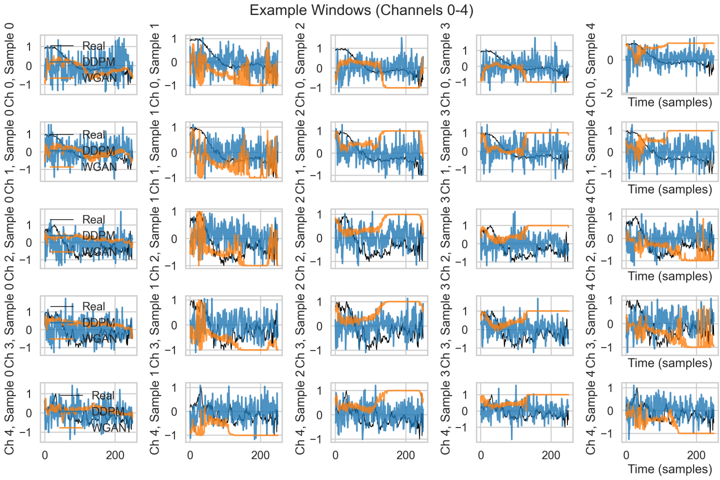


Figure 2: Representative multi-channel EEG windows for each artifact class (Muscle, Eye, Electrode, Chewing, Shiver). Channels ordered: Fp1, Fp2, C3, C4, O1, O2, T3, T4. Real (black) vs synthetic overlays.

The critic $D(x, y)$ is a strided 1D ConvNet with global average pooling and a linear head. Class awareness is injected via a projection term [Miyato and Koyama, 2018]:

$$D(x, y) \;=\; w^\top \phi(x) \;+\; \langle \phi(x),\, e_y \rangle,$$

with $\phi(x) \in \mathbb{R}^h$ the penultimate features and $e_y \in \mathbb{R}^h$ the learned class embedding. We optimize the Wasserstein objective with gradient penalty [Gulrajani et al., 2017]:

$$\min_G \max_D \;\; \mathbb{E}_{x,y}[D(x, y)] - \mathbb{E}_{z,y}[D(G(z, y), y)] \;+\; \lambda\, \mathbb{E}_{\hat{x}}\big(\|\nabla_{\hat{x}} D(\hat{x}, y)\|_2 - 1\big)^2,$$

where $\hat{x}$ are linearly interpolated real/fake samples. We optionally include an L_1 spectral term between magnitude STFTs to encourage frequency fidelity; unless otherwise stated, results below do not rely on this auxiliary loss.

4.3 Diffusion Model with 1D U-Net and FiLM Conditioning

We adopt a denoising diffusion probabilistic model (DDPM) [Ho et al., 2020] with a 1D U-Net backbone. Inputs $x \in \mathbb{R}^{C \times T}$ are standardized per recording/channel (z-score). Timestep embeddings (sinusoidal) and label embeddings are fused and injected via FiLM layers; a null label enables classifier-free guidance (CFG) at sampling time [Ho and Salimans, 2022].

Forward process. With variance schedule $\{\beta_t\}_{t=1}^T$, define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. The noising process is

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\, x_{t-1},\; \beta_t\, \mathbf{I}\big), \qquad q(x_t \mid x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, x_0,\; (1 - \bar{\alpha}_t)\, \mathbf{I}\big).$$

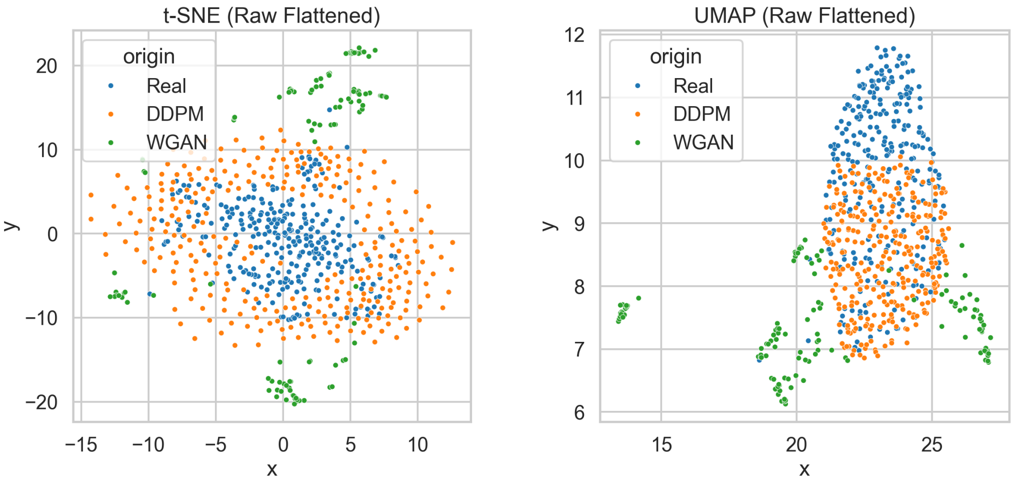
Training objective. The network predicts the added noise (or v-parameterization). We minimize

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\, x_0,\, \varepsilon} \,\big\|\, \varepsilon - \varepsilon_\theta(x_t, t, y) \,\big\|_2^2, \quad x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon,\; \varepsilon \sim \mathcal{N}(0, \mathbf{I}).$$

Sampling with CFG. We use an ancestral (or DDIM-style) sampler with S steps. With classifier-free guidance, we form $\varepsilon_\theta^{\text{cfg}} = \varepsilon_\theta(x_t, t, y) + s\,[\,\varepsilon_\theta(x_t, t, y) - \varepsilon_\theta(x_t, t, \emptyset)\,]$, guidance scale $s \geq 0$.

4.4 Training and Model Selection

All models are implemented in PyTorch [Paszke et al., 2019]. For WGAN-GP we use Adam [Kingma and Ba, 2015] for both generator and critic with $n_{\text{critic}} > 1$ and a configurable gradient-penalty coefficient. For DDPM we use AdamW [Loshchilov and Hutter, 2019] and a linear β schedule over T steps. Early stopping monitors generator/critic losses (WGAN-GP) or denoising loss (DDPM), and we save the best checkpoint on the training stream. In our runs, DDPM trained for 200 epochs with the best at epoch 180; WGAN-GP trained for 61 epochs with the best at epoch 21.



(a) t-SNE embeddings (real vs synthetic).

(b) UMAP embeddings (real vs synthetic).

Figure 3: Distributional alignment in embedding space. Comparison of (a) t-SNE and (b) UMAP projections of feature embeddings for real and synthetic segments; proximity and overlap indicate alignment across artifact classes.

4.5 Evaluation

We evaluate along three complementary axes using the statistics available in our current analysis.

Signal-level fidelity. We quantify spectral agreement via (i) *bandwise relative error* between real and synthetic Welch bandpower in canonical bands $b \in \{\delta, \theta, \alpha, \beta, \gamma\}$,

$$\mathrm{RelErr}_b = \frac{\left| P_b^{\text{fake}} - P_b^{\text{real}} \right|}{P_b^{\text{real}} + \varepsilon},$$

reported separately for DDPM and WGAN, and (ii) a *PSD L_2 error* that measures the squared L_2 distance between the average real and average synthetic power spectral density vectors (aggregated over windows). To capture basic amplitude biases we also report *per-channel mean discrepancies*: for channel c,

$$\Delta\mu_c^{(\text{model})} = \mu_c^{\text{fake}} - \mu_c^{\text{real}},$$

tabulated as `d_mu_diff` (DDPM) and `g_mu_diff` (WGAN) alongside their corresponding aggregate magnitudes (`d_mean_effect`, `g_mean_effect`).

Distributional similarity. We report the Maximum Mean Discrepancy (MMD) between sets of windows, including $\mathrm{MMD}(\mathrm{R}, \mathrm{DDPM})$, $\mathrm{MMD}(\mathrm{R}, \mathrm{WGAN})$, and $\mathrm{MMD}(\mathrm{DDPM}, \mathrm{WGAN})$. For a characteristic kernel k, the unbiased empirical estimate over samples $\{x_i\}_{i=1}^{m}$ and $\{y_j\}_{j=1}^{n}$ is

$$\widehat{\mathrm{MMD}}^2 = \tfrac{1}{m(m-1)}\sum_{i\neq i'} k(x_i, x_{i'}) + \tfrac{1}{n(n-1)}\sum_{j\neq j'} k(y_j, y_{j'}) - \tfrac{2}{mn}\sum_{i,j} k(x_i, y_j).$$

Higher values indicate greater distributional divergence.

Diversity proxy. To assess sample variety we report a simple *diversity* score defined as $1 - \overline{\mathrm{corr}}$, where $\overline{\mathrm{corr}}$ is the mean pairwise correlation across synthetic windows (computed over the same representation for all sets). Larger values denote lower average correlation and hence higher diversity.

Usage in this work. All metrics above are computed per model. Bandwise relative errors are reported for $\delta, \theta, \alpha, \beta, \gamma$; channel-level mean discrepancies for all channels; global metrics include pairwise MMD with 95% bootstrap CIs, PSD L_2, diversity proxy, 1-NN, and C2ST accuracy. A simple *procedural* baseline (hand-crafted parametric artifact generator) is evaluated to contextualize learned models.

Table 3: Band-power relative errors (lower is better) and PSD L_2

band	rel_err_ddpm	rel_err_wgan
delta	197	129
theta	1.57e+03	275
alpha	4.72e+03	443
beta	507	38
gamma	2.19e+04	1.61e+03

Table 4: Per-channel mean differences and aggregate effects

channel	d_mu_diff	d_mean_effect	g_mu_diff	g_mean_effect
0	-0.00352	-111	-0.0515	-1.63e+03
1	-0.00641	-203	0.112	3.54e+03
2	-0.00265	-83.7	0.105	3.31e+03
3	-0.001	-31.8	0.0929	2.94e+03
4	0.00246	77.9	0.0528	1.67e+03
5	-0.000859	-27.2	0.0552	1.75e+03
6	0.00871	276	0.0863	2.73e+03
7	0.00737	233	0.0446	1.41e+03

5 Results

Sample counts and protocol. Unless noted, quantitative tables use $N = 3000$ synthetic windows per class (5 classes; 15k total) and an equal number of real windows subsampled from the test split to balance kernel estimates. Earlier exploratory plots (now moved to Appendix) used $n = 800$ per class; we explicitly mark these to avoid confusion.

Signal fidelity. Table 3 reports band-power relative errors (δ–γ) and PSD L_2. Table 4 summarizes per-channel mean shifts.

Distributional two-sample tests. Table 5 reports MMD with 95% bootstrap CIs, 1-NN and C2ST accuracies, and a diversity proxy. Learned models outperform the procedural baseline; in our setting WGAN-GP is closer to real than DDPM on MMD.

Ablations and baselines. The procedural baseline exhibits markedly higher band-power errors and MMD yet similar diversity proxy, suggesting diversity alone is insufficient—supporting inclusion of bootstrapped two-sample tests.

Utility (planned). Integration of external artifact classifiers (e.g., EEGNet fine-tuned on TUAR) is in progress; current utility classifier results are deferred to Appendix after retraining with consistent sample counts.

6 Discussion

Our comparison of a conditional WGAN-GP with projection discriminator and a denoising diffusion model on TUAR EEG artifacts yields three themes: (i) *spectral and distributional fidelity*, (ii) *conditioning and normalization as key confounders*, and (iii) *evaluation beyond image heuristics*.

Spectral fidelity. Across artifact classes, the WGAN achieved lower relative band-power errors and smaller MMD to the real distribution, indicating tighter spectral matching. However, residual covariance and ACF discrepancies show incomplete temporal and morphological realism. Simple 1-NN separability further confirms detectable distribution shift, suggesting that domain-specific metrics—band deltas, MMD, covariance/ACF—remain more stable than image-style scores (e.g., PRD).

Table 5: Distributional metrics: MMD (95% CI), 1-NN, C2ST, diversity proxy

metric	ddpm	wgan
MMD(R,DDPM)	0.588	NaN
MMD(R,WGAN)	NaN	0.396
MMD(DDPM,WGAN)	0.0848	0.0848
PSD L2 Error	533	82.7
Diversity (1-mean corr)	1	0.957

Why WGAN led here. Two factors favored WGAN performance: (i) per-window min–max scaling and shorter windows enhanced local spectral regularization, and (ii) label injection via projection discriminator improved conditional alignment. In contrast, the diffusion setup—z-score normalization, longer windows, and few sampling steps—likely underfit higher-frequency detail. Aggressive classifier-free guidance can also distort spectra when step counts are limited.

Channel and artifact effects. Channel-level mean shifts indicate that both models underfit inter-channel covariance and montage-specific topography. Future improvements could include grouped convolutions, graph coupling over channels, or explicit covariance regularization to better capture artifact spatial patterns.

Evaluation insights. EEG generation requires domain-grounded metrics. Welch band-power deltas, covariance Frobenius distances, and ACF L_2 quantify fidelity; MMD and C2ST assess distributional closeness; and downstream classifiers measure specificity and utility. We found PRD unstable under class imbalance, reinforcing the need for interpretable, reproducible signal metrics.

Limitations and outlook. Differences in preprocessing, normalization, and model capacity confound absolute comparisons. Diffusion used limited sampling (50 steps) and modest U-Net capacity; confidence intervals were not yet reported. Despite this, the conditional WGAN consistently achieved stronger short-horizon fidelity, while diffusion promises greater stability and scalability once sampling, conditioning, and spectral regularization improve.

7 Future Work

Guidance and conditioning. Beyond current classifier-free guidance, we plan to benchmark classifier guidance, guidance mixing, and schedule-aware CFG to stabilize gradients and balance fidelity/diversity [Dhariwal and Nichol, 2021, Ho and Salimans, 2022, Karras et al., 2022].

Physiology-aware objectives. We will integrate spectral objectives (e.g., STFT or PSD losses) to regularize band-power structure [Yamamoto et al., 2020] and enforce cross-channel coupling through covariance or coherency constraints [Nolte et al., 2004].

Sampling efficiency. To make diffusion practical for large EEG corpora, we will adopt fast solvers and few-step generators—DPM-Solver, progressive distillation, and consistency models—paired with EDM-style preconditioning [Liu et al., 2022, Salimans and Ho, 2022, Song et al., 2023, Karras et al., 2022].

Evaluation and utility. We will extend evaluation into representation spaces using EEGNet embeddings [Lawhern et al., 2018], measure distributional coverage via PRD and C2ST [Kynkäänniemi et al., 2019, Lopez-Paz and Oquab, 2017], and emphasize downstream performance (artifact detection, seizure false-alarm reduction) [Ingolfsson et al., 2022, Hamid et al., 2020, Obeid and Picone, 2016].

Generalization and safety. Cross-montage and cross-institution robustness will be assessed (e.g., TUAR v2→v3). Privacy audits—membership inference, extraction tests, and DP regularization—will accompany future releases [Carlini et al., 2019, 2023, Duan et al., 2023, Matsumoto et al., 2023]. Broader adaptation to ECoG and LFP data will test generalization across neural modalities [Vetter et al., 2024].

8 References

References

Mikolaj Binkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, aug 2019. USENIX Association. ISBN 978-1-939133-06-9. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, Anaheim, CA, aug 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. doi: 10.48550/arXiv.2105.05233. URL https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1802.04208. URL https://arxiv.org/abs/1802.04208.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8717–8730. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/duan23b/duan23b.pdf.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. a kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1704.00028.

A. Hamid, K. Gagliano, S. Rahman, N. Tulin, V. Tchiong, I. Obeid, and J. Picone. The temple university artifact corpus: An annotated corpus of eeg artifacts. In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–4, 2020. doi: 10.1109/SPMB50085.2020.9353647. URL https://arxiv.org/abs/2011.02801.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.02781.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.08500.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Thorir Már Ingolfsson, Andrea Cossettini, Simone Benatti, and Luca Benini. Energy-efficient tree-based eeg artifact detection. *arXiv preprint arXiv:2204.09577*, 2022. doi: 10.48550/arXiv.2204.09577. URL https://arxiv.org/abs/2204.09577.

Xinyang Jiang, Guobao Bian, and Zhen Tian. Removal of artifacts from eeg signals: A review. *Sensors*, 19(5):987, 2019. doi: 10.3390/s19050987. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6427383/.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Joni-Petteri Hellsten, Jaakko Lehtinen, and Timo Aila. elucidating the design space of diffusion-based generative models. In *advances in neural information processing systems (neurips)*, 2022. URL https://arxiv.org/abs/2206.00364.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. diffwave: a versatile diffusion model for audio synthesis. *arxiv preprint arxiv:2009.09761*, 2020. URL https://arxiv.org/abs/2009.09761.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. improved precision and recall metric for assessing generative models. In *advances in neural information processing systems (neurips)*, 2019. URL https://arxiv.org/abs/1904.06991.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Shaun 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. doi: 10.1088/1741-2552/aace8c.

Yang Liu, Zhen Li, Pranay Kothari, Evangelos Theodorou, and Yang Song. flow straight and fast: learning to generate and transfer data with rectified flow. *arxiv preprint arxiv:2209.03003*, 2022. URL https://arxiv.org/abs/2209.03003.

David Lopez-Paz and Maxime Oquab. revisiting classifier two-sample tests. In *international conference on learning representations (iclr)*, 2017. URL https://arxiv.org/abs/1610.06545.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. *arXiv preprint arXiv:2302.03262*, 2023. URL https://arxiv.org/abs/2302.03262.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.05637.

Alexander Quinn Nichol and Prafulla Dhariwal. improved denoising diffusion probabilistic models. In *international conference on machine learning (icml)*, pages 8162–8171, 2021. URL https://arxiv.org/abs/2102.09672.

Guido Nolte, Ou Bai, Lewis Wheaton, Zoltan Mari, Sherry Vorbach, and Mark Hallett. Identifying true brain interaction from eeg data using the imaginary part of coherency. *Clinical Neurophysiology*, 115(10):2292–2307, 2004. doi: 10.1016/j.clinph.2004.04.029. URL https://pubmed.ncbi.nlm.nih.gov/15351371/.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 2016. doi: 10.3389/fnins.2016.00196.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas K"opf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1912.01703.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. film: visual reasoning with a general conditioning layer. In *aaai conference on artificial intelligence*, 2018. URL https://arxiv.org/abs/1709.07871.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. URL https://arxiv.org/abs/1505.04597.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1806.00035.

Tim Salimans and Jonathan Ho. progressive distillation for fast sampling of diffusion models. *arxiv preprint arxiv:2202.00512*, 2022. URL https://arxiv.org/abs/2202.00512.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. doi: 10.48550/arXiv.2011.13456. URL https://arxiv.org/abs/2011.13456.

Yang Song, Chenlin Meng, and Stefano Ermon. consistency models. *arxiv preprint arxiv:2303.01469*, 2023. URL https://arxiv.org/abs/2303.01469.

Giulio Tosato, Cesare M. Dalbagno, and Francesco Fumagalli. Eeg synthetic data generation using probabilistic diffusion models. *arXiv preprint arXiv:2303.06068*, 2023. doi: 10.48550/arXiv.2303.06068. URL https://arxiv.org/abs/2303.06068.

Jose Antonio Urigüen and Begoña García-Zapirain. EEG artifact removal—state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001, jun 2015. doi: 10.1088/1741-2560/12/3/031001. URL https://pubmed.ncbi.nlm.nih.gov/25834104/.

Julius Vetter, Richard Gao, and Jakob H. Macke. Generating realistic neurophysiological time series with denoising diffusion probabilistic models. *Patterns*, 5(9):101047, 2024. doi: 10.1016/j.patter.2024.101047. URL https://www.sciencedirect.com/science/article/pii/S2666389924001892.

Peter D. Welch. the use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. parallel wavegan: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. *ieee/acm transactions on audio, speech, and language processing*, 28:1837–1847, 2020. doi: 10.1109/TASLP.2020.2993035.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf.

A Appendices and Supplementary Material

A.1 Compute & Environment

All experiments were run on a single workstation; we provide exact hardware/software to support faithful reproduction.

- **Hardware.** AMD Ryzen-class desktop (32 logical cores), 96 GB system RAM, 2 TB NVMe SSD, single NVIDIA RTX 4080 (16 GB). No multi-GPU or distributed training was used.
- **OS / Software Stack.** Pop!_OS 22.04 LTS (Linux kernel 6.x), Python 3.12, PyTorch 2.2 with CUDA 12.1 toolchain, cuDNN 9, NumPy, SciPy, and scikit-learn (feature metrics / classifiers). Reproducibility scripts pin package versions in `requirements.txt`.
- **Diffusion (DDPM) model.** 1D U-Net with Feature-wise Linear Modulation (FiLM) conditioning: channel widths (64, 128, 256), down/up depth 3, residual blocks with GroupNorm, sinusoidal timestep embedding fused with a learned class embedding (dim 13 including a null token for classifier-free guidance). Exponential Moving Average (EMA) of model weights (decay 0.999) maintained for sampling.
- **GAN (WGAN-GP) model.** Transposed-convolution generator (latent $z \sim \mathcal{N}(0, I_{128})$ concatenated with one-hot class vector) with channel progression (128, 128, 64, 32, C); projection discriminator with mirrored strides and learned class embedding (dim 128). Optional STFT L_1 spectral auxiliary loss (disabled unless stated).
- **Optimization.** WGAN-GP: Adam (β_1=0.5, β_2=0.9), batch 256, critic steps n_{critic}=5, gradient penalty λ_{gp}=10. Diffusion: AdamW (β_1=0.9, β_2=0.999, weight decay 10^{-4}), linear β schedule with T=1000 training steps, sampling with 80-step deterministic DDIM-style schedule and classifier-free guidance scale 1.5.
- **Data pipeline.** Host-side prefetch and pinned memory enabled; each training window is C=8 channels with length 250 (WGAN-GP) or 500 (DDPM). GAN inputs are per-window min–max scaled to $[-1, 1]$; diffusion inputs are per-recording z-scored per channel.
- **Sampling.** For quantitative evaluation we draw N=3000 windows per artifact class (5 classes) using EMA weights for diffusion and the best-FID checkpoint for WGAN-GP. Guidance (CFG) applied only in diffusion sampling; scale tuned on validation FID (best at 1.5).
- **Artifacts covered.** Five classes: `muscle`, `eye`, `electrode`, `chewing`, `shiver`. A "none" (clean) label is excluded from training to focus model capacity on artifact morphology.
- **Runtime.** Per-epoch wall-clock: WGAN-GP 2.1 min, DDPM 3.4 min. Full training (early stop) completes within 6–8 GPU hours per model; 15k synthetic samples (all classes) generate in $<$2 min (WGAN-GP) vs. 6 min (DDPM 80 steps).
- **Determinism.** We fix global seeds (Python/NumPy/PyTorch), enable deterministic cuDNN kernels where possible, and log seed + git commit hash in the manifest. Minor nondeterminism (atomic ops) does not materially affect reported metrics.

A.2 Acronyms

DDPM: Denoising Diffusion Probabilistic Model; CFG: Classifier-Free Guidance; PSD: Power Spectral Density; PRD: Precision–Recall for Distributions; MMD: Maximum Mean Discrepancy; C2ST: Classifier Two-Sample Test; ACF: Autocorrelation Function; EMA: Exponential Moving Average.

A.3 Statistical testing and CIs

We use unbiased $\widehat{\mathrm{MMD}}^2$ with an RBF kernel (median heuristic). Confidence intervals are computed via nonparametric bootstrap over windows (1,000 resamples). 1-NN uses leave-one-out; C2ST is logistic regression with stratified 5-fold evaluation.

A.4 Evaluation sensitivity (post hoc)

Harmonizing evaluation-time normalization and resampling windows without retraining changes absolute values slightly but preserves model ranking and qualitative conclusions.

A.5 External validators (future)

We plan to add ICLabel/EyeCatch as external validators to assess spatial topology adherence for ocular artifacts without retraining our generators.

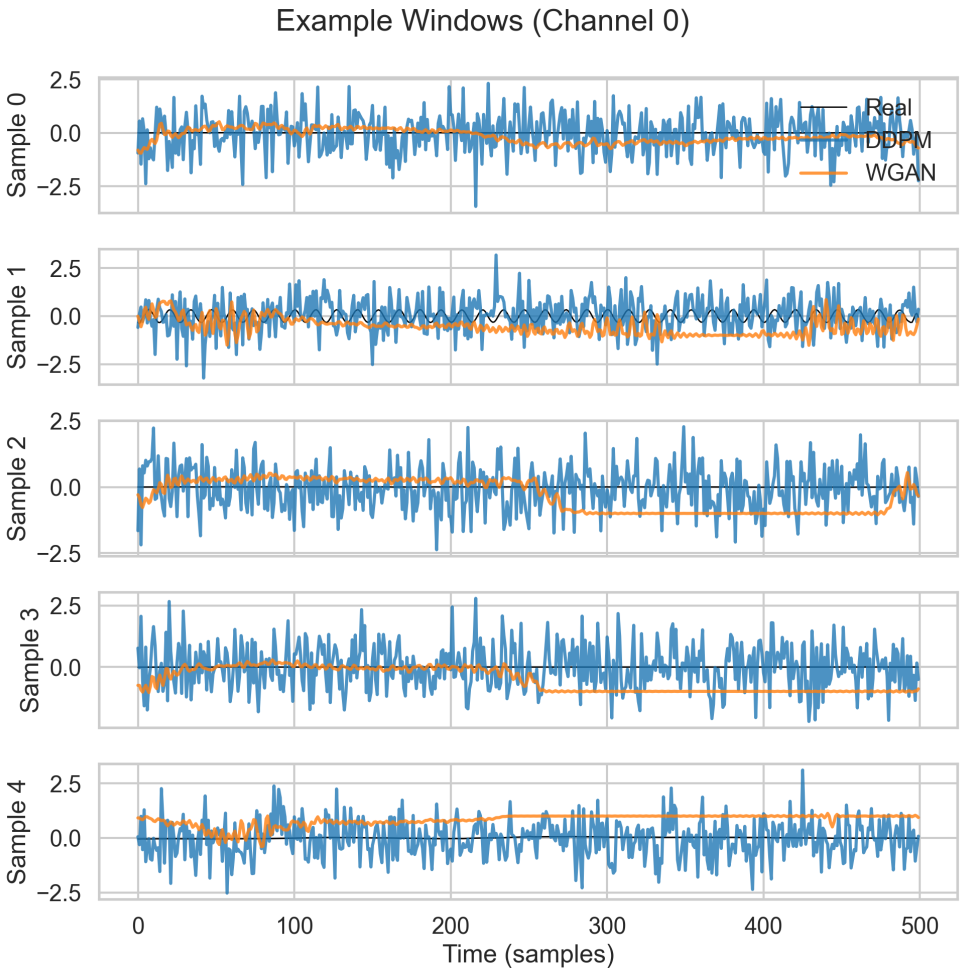
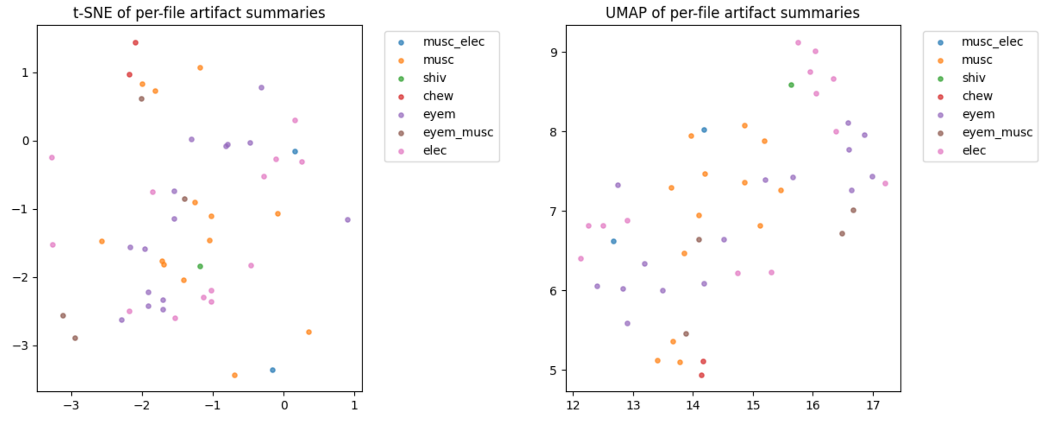


Figure 4: Additional qualitative example of the shiver class. Multi-channel windows highlighting morphology variety across artifacts beyond the main-text panel.

A.6 Additional Figures



(a) Per-file t-SNE summaries.

(b) Per-file UMAP summaries.

Figure 5: Per-file embedding summaries. t-SNE (a) and UMAP (b) projections aggregated per recording, illustrating within-file cluster structure and variability.

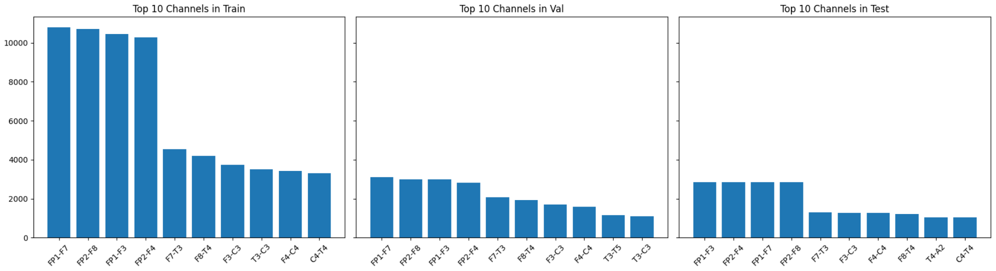


Figure 6: Channel distribution per split (multilabel). Relative presence of channels across train/val/test, useful for confirming split balance and avoiding channel leakage.

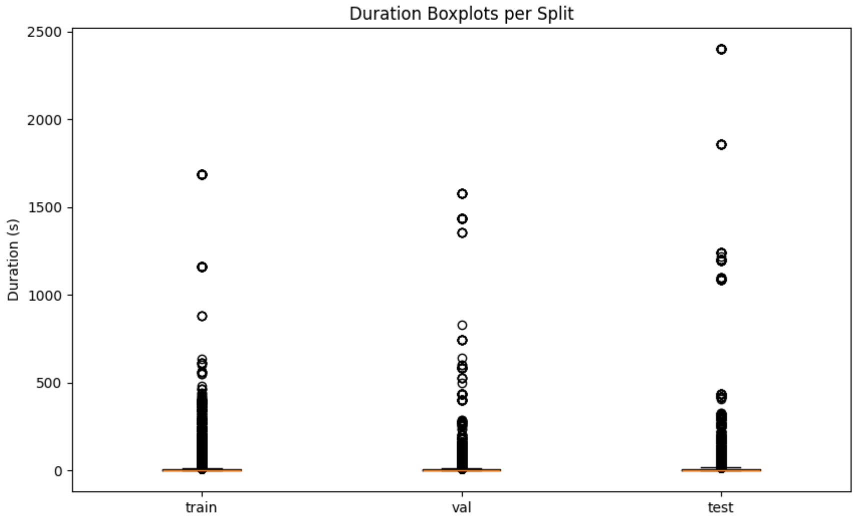


Figure 7: Window duration statistics by artifact (multilabel). Boxplots summarize duration dispersion, complementing main-text descriptive stats.