SHARED STRUCTURE, INDIVIDUAL DYNAMICS: FLOW MATCHING FOR EEG GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical electroencephalography (EEG) generation requires both patient-shared signal structure and sample-specific temporal detail. We study how the source distribution and training objective can address these requirements in Flow Matching. We propose GEMIA, a conditional Flow Matching method with a Global Structured Endpoint (GSE) and Multi-Scale Increment Alignment. GSE estimates spectral–spatial structure with equal weight for each training patient and models waveform shape and amplitude separately. The network predicts the clean endpoint while the transport loss is in velocity space; to limit endpoint error amplification, we cap the time-dependent weight in the velocity-L_1 loss. Auxiliary time-domain losses complement the structured source by aligning per-channel scale, target temporal increments at several lags, and waveform correlation. We evaluate seven generators on four patient-wise TUH protocols, assessing signal fidelity, downstream utility, and memorization. GEMIA achieves the lowest mean TS-FID on all four protocols, with reductions of 4.7%–63.1% relative to JET. It also achieves lower mean TS-FID with five Heun steps than the Flow Matching baselines with fifty steps. Source-objective ablations show that the benefit of GSE depends on the accompanying loss. These results support source-objective co-design for EEG generation.

1 INTRODUCTION

Clinical EEG data are expensive to collect and annotate, difficult to share, and highly heterogeneous across patients. Synthetic EEG could support low-resource modeling, but a generator is useful only when it preserves established temporal, spectral, and cross-channel structure and transfers to held-out patients. At the population level, generated data should reproduce the spectral distribution and cross-channel relations shared across patients. At the sample level, each generated window should preserve its channel amplitudes, temporal changes, and transient events. A low distance in one feature space does not guarantee that both levels are modeled correctly.

Recent work has made important progress. JET established conditional Flow Matching as a strong generator for multichannel EEG and introduced EEG-oriented signal constraints and TS-FID (Wang et al., 2026). EEGDM showed that a generative objective can also support EEG representation learning (Wang et al., 2025b). These studies demonstrate the value of continuous generation and EEG-specific training objectives. This raises a complementary question: can the source distribution itself use patient-shared EEG statistics, and what training objective is suitable for such a source?

The standard isotropic Gaussian is a valid and successful source for Flow Matching. Gaussian-source Flow Matching can generate EEG. However, our focus is whether a source informed by patient-shared EEG statistics can further improve generation. An isotropic Gaussian is general, but it does not contain the spectral shape or frequency-dependent cross-channel relations that can be estimated from training data. We therefore ask whether placing this label-neutral, patient-balanced information in the source can improve generation under the same model capacity and training budget.

To study this question, we introduce the Global Structured Endpoint (GSE), a source for EEG Flow Matching that models spectral-spatial shape and window amplitude separately. Pooling all training windows equally gives more weight to patients who contribute more windows. GSE instead estimates the cross-spectral structure and amplitude distribution with equal total weight for each training patient, consistent with inverse-cluster-size weighting in clustered-data analysis (Williamson et al., 2003). These statistics define a stochastic, label-neutral source.

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.

Following JiT (Li & He, 2025), the network directly predicts the clean EEG endpoint. We use a capped velocity-L_1 transport loss, together with auxiliary time-domain losses that complement the structured source. These auxiliary losses align per-channel scale, target temporal increments at several lags, and global waveform correlation. A global correlation term further measures the centered waveform similarity. A frequency-structured source and an explicit frequency-domain loss can constrain overlapping spectral statistics on both sides of the flow. Because the source already encodes spectral-spatial statistics, we use time-domain supervision to complement this structure rather than add an explicit frequency-domain loss.

The source and objective therefore have different but complementary roles. GSE provides patient-shared spectral-spatial structure before transport, whereas the objective supervises sample-specific scale and temporal dynamics during transport. We call the complete method GEMIA, combining the Global Structured Endpoint with Multi-Scale Increment Alignment.

Our contributions are:

- We propose GEMIA, a role-separated source-objective design for conditional EEG Flow Matching. GSE places patient-balanced, label-neutral spectral-spatial population structure in the source. Class semantics enter only through the AdaLN-Zero conditional network, while the auxiliary objective supervises per-channel scale and target-matched temporal increments in the time domain. This design avoids imposing overlapping label or frequency-domain information in multiple components.
- We compare a fixed GEMIA configuration with six baselines on four strict patient-wise TUH protocols, evaluating multi-seed fidelity, low-patient utility, condition agreement, structural and memorization diagnostics, and sampling efficiency. Fixed-seed source-objective and progressive loss ablations examine the structured source and auxiliary losses while keeping $\mathcal{L}_{\mathrm{cap}}$ fixed.

2 Related Work

EEG and related time-series generation. Early EEG generation methods mainly used adversarial learning. EEG-GAN generates EEG waveforms using a Wasserstein-GAN-style framework (Hartmann et al., 2018). Conditional Wasserstein GANs have also been used to generate EEG features for downstream augmentation (Luo & Lu, 2018). For general time series, TimeGAN combines adversarial learning with temporal supervision in a learned embedding space (Yoon et al., 2019), while Diffusion-TS uses clean-data prediction and a Fourier-domain loss (Yuan & Qiao, 2024). Autoregressive MEG models provide another approach to multichannel electrophysiological generation (Csaky et al., 2024; Huang et al., 2025), although they model MEG rather than EEG. JET applies conditional Flow Matching to multichannel EEG, using a Gaussian source and signal-level constraints on amplitude statistics, temporal variation, and waveform correlation (Wang et al., 2026).

Flow Matching. Flow Matching learns a velocity field along a prescribed probability path (Lipman et al., 2023). Its source distribution need not be isotropic Gaussian. Stochastic interpolants allow general endpoint distributions (Albergo et al., 2025), and TSFlow uses Gaussian-process priors to encode temporal structure before transport (Kollovieh et al., 2025). For multichannel signals, Fourier surrogate methods preserve within-channel autocorrelation and cross-channel correlation (Prichard & Theiler, 1994). When estimating an EEG source from training statistics, pooling all windows equally gives more weight to patients with more windows. GSE instead uses equal-patient weighting for its cross-spectral and amplitude statistics, consistent with inverse-cluster-size weighting for clustered data (Williamson et al., 2003). These patient-balanced statistics define a label-neutral structured source. We complement this source with time-domain alignment of per-channel scale, multi-lag increments, and waveform correlation between generated and real EEG windows, without an additional explicit frequency-domain loss. Our architecture and prediction head build on DiT and JiT (Peebles & Xie, 2023; Li & He, 2025). An extended discussion is provided in Appendix A.

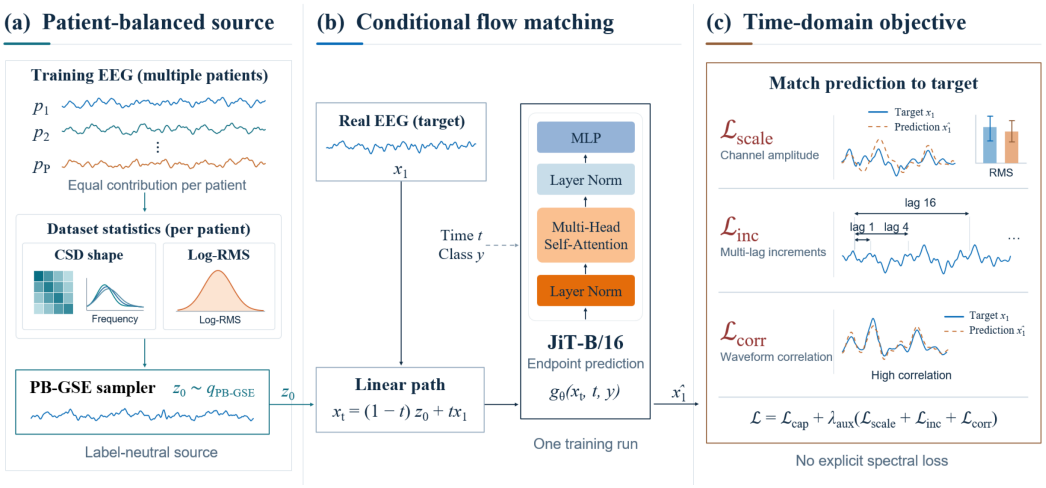

Figure 1: Overview of the proposed GEMIA framework.

3 METHOD

3.1 PROBLEM SETTING

Our goal is to generate realistic EEG. A generated EEG window should preserve the key characteristics of real data: channel-wise temporal dynamics, spectral content, and event-related structure. We formulate this as a conditional generation task: given a conditioning variable y, such as normal or abnormal; we want to learn a mapping from a random source signal to an EEG window that resembles real data from that condition. The overall pipeline of our method is shown in Figure 1.

Formally, an EEG window is denoted as $\boldsymbol{X} \in \mathbb{R}^{C\times T}$, where C is the number of channels and T is the number of time points. Entries $x_{c,s} = [\boldsymbol{X}]_{c,s}$ are scalars, and $\|\cdot\|_1$ on signal matrices denotes the entrywise norm. To perform the generation, we adopt the conditional flow matching (CFM) framework. We define a straight-line path from a source $\boldsymbol{Z}_0$ to a real target EEG window $\boldsymbol{X}_1$:

$$\boldsymbol{X}_t = (1-t)\boldsymbol{Z}_0 + t\boldsymbol{X}_1, \qquad t \in [0,1]. \tag{1}$$

Instead of predicting the velocity directly, the network predicts the clean target EEG window, denoted by $\widehat{\boldsymbol{X}}_{1,\boldsymbol{\theta}} = g_{\boldsymbol{\theta}}(\boldsymbol{X}_t, t, y)$. This follows the clean-data prediction design of JiT. The main idea is that real data occupy a structured, low-dimensional region of the full signal space, whereas noisy data do not. For our linear path, the velocity target also contains the random source $\boldsymbol{Z}_0$, and is therefore not confined to the data manifold as the clean EEG target $\boldsymbol{X}_1$ is. Directly predicting $\boldsymbol{X}_1$ gives the network a simpler and more data-structured prediction target. The velocity required by the ODE solver is then derived from $\widehat{\boldsymbol{X}}_{1,\boldsymbol{\theta}}$, rather than being predicted directly by the network.

During training, the flow time t is sampled from a logit-normal distribution rather than a uniform distribution. Specifically, we first sample a real-valued Gaussian variable $\eta \sim \mathcal{N}(\mu_t, \sigma_t^2)$ and then map it into the interval $(0,1)$ through the sigmoid function as $t = \mathrm{sigmoid}(\eta)$. The parameters μ_t and σ_t determine which parts of the transport path are sampled more often.

3.2 GLOBAL STRUCTURED ENDPOINT

Most conditional flow matching frameworks sample the source endpoint $\boldsymbol{Z}_0$ from a standard Gaussian distribution. For EEG generation, however, this choice is suboptimal. EEG signals follow a characteristic spectral pattern: power decreases as frequency increases, with peaks at physiologically meaningful frequencies. If the source endpoint already carries a rough EEG-like spectral shape, the model only needs to refine it rather than synthesize it from scratch. This leads to better fidelity with fewer inference steps. We therefore design a source endpoint that encodes the average spectral structure of training EEG, while still introducing stochasticity through the sampling process.

To separate waveform shape from overall amplitude, we subtract each channel's temporal mean from the ith EEG window $\boldsymbol{X}_{p,i}$ of patient p, obtaining $\boldsymbol{X}^{\circ}_{p,i}$. We denote its global standard deviation across channels and time points by $a_{p,i}$, computed with divisor CT and thus equal to its global RMS. The normalized shape is $\bar{\boldsymbol{X}}_{p,i} = \boldsymbol{X}^{\circ}_{p,i}/(a_{p,i} + \epsilon)$, sharing the same amplitude scale across all channels.

We apply the real fast Fourier transform (RFFT) along the time axis of each normalized window, retaining only the non-negative frequencies. Let $\widehat{\bar{\boldsymbol{x}}}_{p,i}(f) \in \mathbb{C}^C$ denote the resulting cross-channel coefficient vector at frequency f.

We then compute the cross-spectral matrix, averaged in a patient-balanced manner:

$$\widehat{\boldsymbol{S}}^{\mathrm{shape}}_{\mathrm{PB}}(f) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{1}{n_p} \sum_{i=1}^{n_p} \widehat{\bar{\boldsymbol{x}}}_{p,i}(f) \widehat{\bar{\boldsymbol{x}}}_{p,i}(f)^H. \tag{2}$$

The inner sum averages windows within each patient; the outer sum averages across patients. This ensures that each patient contributes equally to the estimated spectral structure, regardless of how many windows they have. The result $\widehat{\boldsymbol{S}}^{\mathrm{shape}}_{\mathrm{PB}}(f)$ is a Hermitian matrix at each frequency bin, capturing the cross-channel spectral correlations of the population.

At each frequency f, $\widehat{\boldsymbol{S}}^{\mathrm{shape}}_{\mathrm{PB}}(f)$ is a $C \times C$ Hermitian matrix. Its diagonal entries describe the spectral power of individual channels, while its off-diagonal entries describe how pairs of channels vary together, including their relative phase. Here, H denotes the conjugate transpose.

To generate a random shape with this spectral structure, we decompose the cross-spectral matrix as $\widehat{\boldsymbol{S}}^{\mathrm{shape}}_{\mathrm{PB}}(f) = \boldsymbol{U}_f \boldsymbol{\Lambda}_f \boldsymbol{U}_f^H$. The columns of $\boldsymbol{U}_f$ represent the main cross-channel spectral patterns, and the diagonal entries of $\boldsymbol{\Lambda}_f = \mathrm{diag}(\lambda_{f,1}, \ldots, \lambda_{f,C})$ describe the strength of these patterns. Because the matrix is estimated from a finite number of training windows, some eigenvalues can be extremely small or slightly negative because of numerical error. We therefore apply a small non-negative eigenvalue floor. Specifically, each eigenvalue is clamped as $\widetilde{\lambda}_{f,j} = \max(\lambda_{f,j}, \delta_f)$, and the corresponding diagonal matrix is formed as $\widetilde{\boldsymbol{\Lambda}}_f = \mathrm{diag}(\widetilde{\lambda}_{f,1}, \ldots, \widetilde{\lambda}_{f,C})$.

where $\delta_f > 0$ is a small numerical threshold. The resulting matrix $\widetilde{\boldsymbol{\Lambda}}_f$ is the stabilized version of $\boldsymbol{\Lambda}_f$. We then sample a standard complex Gaussian vector and transform it using the spectral patterns and their stabilized strengths:

$$\widehat{\boldsymbol{z}}^{\mathrm{shape}}(f) = \boldsymbol{U}_f \widetilde{\boldsymbol{\Lambda}}_f^{1/2} \boldsymbol{\xi}_f, \qquad \boldsymbol{\xi}_f \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}). \tag{3}$$

The random vector $\boldsymbol{\xi}_f$ provides variation between source samples. Multiplication by $\widetilde{\boldsymbol{\Lambda}}_f^{1/2}$ gives each spectral pattern the required strength, while multiplication by $\boldsymbol{U}_f$ maps these patterns to the EEG channels. Therefore, the sampled frequency-domain vector follows the stabilized patient-balanced cross-spectral structure.

After repeating this sampling operation at every non-negative frequency bin, we convert the sampled coefficients back into a real-valued waveform using the inverse RFFT as $\boldsymbol{Z}^{\mathrm{shape}} = \mathrm{IRFFT}\Big(\widehat{\boldsymbol{Z}}^{\mathrm{shape}}\Big)$, where $\widehat{\boldsymbol{Z}}^{\mathrm{shape}}$ collects the vectors $\widehat{\boldsymbol{z}}^{\mathrm{shape}}(f)$ over frequency bins.

The inverse RFFT automatically reconstructs the conjugate negative-frequency coefficients required for a real-valued signal. The resulting $\boldsymbol{Z}^{\mathrm{shape}}$ provides the spectral–spatial shape of the PB-GSE source. Its amplitude is assigned separately in the next step.

Finally, we assign an amplitude to the sampled spectral-spatial shape. For training window i from patient p, we use its global RMS $a_{p,i}$, to compute the log-RMS $r_{p,i} = \log a_{p,i}$. These log-RMS values form a patient-balanced empirical distribution, in which each patient has the same total sampling weight. We sample one value from this distribution and denote it by r.

Directly using the sampled amplitude $\exp(r)$ would retain the complete empirical amplitude variation, including rare very large or very small values. We therefore shrink the sampled log-amplitude toward a fixed reference RMS $a_0 > 0$ as $a_\alpha = \exp[(1 - \alpha_{\mathrm{amp}}) \log a_0 + \alpha_{\mathrm{amp}} r]$, where $\alpha_{\mathrm{amp}} \in [0, 1]$.

The coefficient α_{amp} controls how much empirical amplitude variation is introduced into the source. When $\alpha_{\mathrm{amp}} = 0$, every source sample uses the reference RMS a_0. When $\alpha_{\mathrm{amp}} = 1$, the source

uses the complete sampled amplitude $\exp(r)$. An intermediate value preserves the relative amplitude variation while shrinking extreme values toward a_0.

The sampled spectral-spatial shape has a random RMS of its own. To control shape and amplitude separately, we first normalize the shape and then apply the amplitude a_α:

$$\boldsymbol{Z}_0 = a_\alpha \frac{\boldsymbol{Z}^{\text{shape}}}{\text{RMS}(\boldsymbol{Z}^{\text{shape}}) + \epsilon}. \tag{4}$$

This normalization prevents the random scale of $\boldsymbol{Z}^{\text{shape}}$ from changing the independently sampled amplitude. The patient-balanced spectral statistics and amplitude distribution are computed from training patients only. The reference a_0 is fixed before training and is not estimated from validation or test data. All source statistics are computed exclusively from training patients, so no test information leaks into the endpoint prior.

3.3 Capped Clean-Endpoint Transport

Standard flow matching objectives use an L2 reconstruction loss. While this works well for many modalities, it is less suited for EEG. EEG often contains impulsive, non-Gaussian artifacts from muscle activity, electrode movement, or external noise. Moreover, EEG amplitudes exhibit heavy-tailed distributions and non-stationary dynamics. Under these conditions, an L2 loss is overly sensitive to large deviations and tends to produce over-smoothed outputs that miss sharp neural events.

We therefore adopt an L1 reconstruction loss, which implicitly assumes a Laplacian prior and is more robust to transient outliers. However, adopting L1 alone does not fully address the training instability in flow matching. In the conditional flow matching framework, the model predicts the clean endpoint $\widehat{\boldsymbol{X}}_{1,\boldsymbol{\theta}}$. Given the linear interpolation path $\boldsymbol{X}_t = (1-t)\boldsymbol{Z}_0 + t\boldsymbol{X}_1$, the induced velocity is given by $\boldsymbol{V}_{\boldsymbol{\theta}}(\boldsymbol{X}_t, t, y) = (\widehat{\boldsymbol{X}}_{1,\boldsymbol{\theta}} - \boldsymbol{X}_t)/(1-t)$. As t approaches 1, the denominator $1-t$ tends to zero. This causes the velocity gradient to be heavily amplified near the end of the path, making the training process unstable. To mitigate this, we introduce a simple cap on the training weight:

$$\mathcal{L}_{\text{cap}} = \mathbb{E}\left[\frac{\|\boldsymbol{X}_1 - \widehat{\boldsymbol{X}}_{1,\boldsymbol{\theta}}\|_1}{CT \cdot \max(1-t, \tau)}\right], \tag{5}$$

where τ is a small constant. This cap ensures that when $1-t$ becomes smaller than τ, the training weight is clipped at $1/\tau$ and does not grow further. This cap is applied only to the training loss. During inference, the solver uses the uncapped velocity field to preserve generation quality.

3.4 Multi-Scale Increment Alignment

The capped L1 loss in Eq. (5) improves robustness and training stability, but it only constrains the raw waveform values. To better preserve the temporal structure of EEG signals, we introduce three auxiliary losses that operate on different aspects of the generated waveform.

Scale loss. EEG amplitude varies across channels. To ensure the model reproduces the correct per-channel amplitude, we match the log RMS deviation of each channel:

$$\mathcal{L}_{\text{scale}} = \frac{1}{C}\sum_{c=1}^{C}\left|\log \sigma_c(\widehat{\boldsymbol{X}}_1) - \log \sigma_c(\boldsymbol{X}_1)\right|, \tag{6}$$

where $\sigma_c(\cdot)$ is the RMS deviation computed over the time dimension for channel c. The log domain ensures that both large and small amplitude scales are treated with equal relative importance.

Increment loss. EEG classification often relies on transient events such as spikes or sharp waves. A standard total variation (TV) loss would penalize all changes uniformly, which discourages the model from generating sharp transitions even when they are present in the target. Instead, we directly supervise the temporal increments at multiple lags:

$$\mathcal{L}_{\text{inc}} = \frac{1}{|\mathcal{L}_{\text{lag}}|}\sum_{\ell \in \mathcal{L}_{\text{lag}}} \frac{\|\Delta_\ell \widetilde{\widehat{\boldsymbol{X}}}_1 - \Delta_\ell \widetilde{\boldsymbol{X}}_1\|_1}{C(T-\ell)}, \qquad \mathcal{L}_{\text{lag}} = \{1, 4, 16\}, \tag{7}$$

where $\Delta_\ell x_{c,t} = x_{c,t+\ell} - x_{c,t}$, and $\widetilde{\boldsymbol{X}}$ denotes channel-wise RMS. The lags $\ell = 1, 4, 16$ correspond to adjacent, medium, and coarse temporal scales. Unlike TV loss, this formulation does not penalize sharp changes; it only penalizes discrepancies between the predicted and target increments.

Correlation loss. Finally, we add a global correlation term to align the overall waveform shape:

$$\mathcal{L}_{\text{corr}} = 1 - \rho\Big(\text{vec}(\boldsymbol{X}_1), \text{vec}(\widehat{\boldsymbol{X}}_1)\Big), \tag{8}$$

where ρ is the Pearson correlation coefficient. This encourages the model to capture the correct global trend and phase alignment.

To enforce the intended time-domain constraints, we formulate a composite structure loss: $\mathcal{L}_{\text{struct}} = \mathcal{L}_{\text{scale}} + \mathcal{L}_{\text{inc}} + \mathcal{L}_{\text{corr}}$. The full training objective integrates this with the capped velocity loss:

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \lambda_{\text{aux}} \mathcal{L}_{\text{struct}}, \tag{9}$$

where the hyperparameter λ_{aux} balances the contribution of the auxiliary structure regularization.

4 Experiments

We compare GEMIA with six baseline generators, evaluating signal fidelity, low-patient downstream utility and condition agreement, structural diagnostics, and sampling efficiency. Source-objective and progressive loss ablations further examine the contributions of the structured source and auxiliary losses. Detailed experimental settings are provided in the appendix, with generator configurations in Table B1 and complete numerical results in Tables B2-B5. Table 1 summarizes the four TUH protocols and their patient-disjoint training, validation, and test splits.

4.1 Signal Fidelity

We assess whether generated EEG preserves the spectral, temporal, and amplitude of real data. Since one feature-space distance does not capture all properties, we compare generated windows with real windows from held-out patients using complementary fidelity metrics.

TS-FID is primary distributional metric and compares standardized log-FFT feature distributions. log-PSD measures spectral mismatch, while BP-FD compares bandpower feature distributions. ACF measures temporal autocorrelation error. Amp-W1 and Env-W1 measure differences in signal-amplitude and envelope distributions, respectively. RFD compares feature distributions from a frozen CBraMod encoder. Lower values indicate better agreement with real data for all metrics. Table 2 shows that GEMIA achieves the lowest mean TS-FID on all protocols. Relative to JET, the reductions are 63.1%, 55.5%, 26.1%, and 4.7% on TUSZ v1.5.2, TUSZ v2.0.6, TUAB, and TUEV.

Complementary properties. Figure 2 summarizes the six complementary metrics using within-corpus ranks, with rank 1 indicating the lowest error. GEMIA has lower mean log-PSD error than JET on every protocol. It also achieves the lowest mean BP-FD on TUAB and TUEV.

These results show that the TS-FID gains are accompanied by improvements in selected spectral statistics, but do not imply uniformly better fidelity across all measured properties. The complementary metrics therefore help distinguish which aspects of real EEG each generator preserves. Full values and sample standard deviations are provided in Appendix Table B2.

4.2 Downstream Utility and Condition Agreement

We assess whether synthetic EEG supports classification on unseen patients when real training data are limited, and whether generated windows reflect their requested class labels. These questions are evaluated separately through downstream augmentation and bandpower-teacher agreement.

For downstream classification, all downstream classifiers use real EEG windows from the same fixed 10% of training patients. Real-only uses these windows alone, while each augmentation method adds synthetic windows equal to 50% of the real window count in each class. We train a linear classifier on a frozen CBraMod encoder, and evaluate it on real held-out test data. Balanced Accuracy averages

Table 1: Processed patient and window counts for the four EEG protocols.

Corpus	Task	Train pat.	Val pat.	Test pat.	Train win.	Val win.	Test win.	Window
TUSZ v1.5.2	background/seizure	468	119	50	35,471	20,057	20,160	5 s
TUSZ v2.0.6	background/seizure	578	53	43	41,697	10,844	9,389	5 s
TUAB v3.0.1	normal/abnormal	1,660	416	253	298,652	73,858	36,945	10 s
TUEV v2.0.1	six-class event	232	58	80	63,832	20,100	29,421	5 s

Table 2: TS-FID on held-out patients. Entries are the reported mean ± sample standard deviation over three generator runs; lower is better.

Generator	TUSZ v1.5.2	TUSZ v2.0.6	TUAB v3.0.1	TUEV v2.0.1
Conditional WGAN-GP	1419.034 ± 109.905	1652.511 ± 213.423	635.258 ± 107.370	734.579 ± 36.120
EEG-GAN	941.843 ± 306.924	1015.670 ± 388.009	754.947 ± 157.113	954.151 ± 179.739
1D U-Net DDPM	741.896 ± 59.838	764.188 ± 63.577	554.608 ± 16.203	615.079 ± 2.379
MEG-GPT-style EEG	539.810 ± 85.409	631.417 ± 67.513	439.196 ± 14.115	382.686 ± 23.985
JET	472.236 ± 18.919	380.850 ± 12.418	377.518 ± 2.614	332.134 ± 2.498
DiT-FM	519.203 ± 59.841	511.537 ± 11.586	424.968 ± 19.036	1053.420 ± 48.906
GEMIA	$\mathbf{174.139 \pm 30.792}$	$\mathbf{169.460 \pm 14.407}$	$\mathbf{278.993 \pm 4.021}$	$\mathbf{316.612 \pm 4.501}$

recall across classes; AUPRC summarizes precision-recall performance; and AUROC measures class discrimination across decision thresholds. Higher values indicate better performance for all metrics.

Figure 3(a)-(f) shows that GEMIA achieves competitive downstream performance among the compared generators. Among augmentation methods, it attains the highest mean AUPRC on TUSZ v2.0.6, TUAB, and TUEV, and the highest mean AUROC on TUSZ v2.0.6 and TUEV. On TUAB, it also achieves the highest mean Balanced Accuracy and outperforms JET and DiT-FM on all three metrics. On TUSZ v1.5.2, its AUPRC and AUROC are comparable to those of the leading augmentation methods, although its Balanced Accuracy is lower. Improvements over Real-only vary by protocol, with gains on all three metrics on TUEV. Overall, these results support GEMIA as a competitive generator for low-patient augmentation, with strong AUPRC and AUROC performance.

We also check whether the generated EEG matches the labels used for generation. We compare these labels with predictions from a classifier based on bandpower features (the teacher), using the scale-aligned evaluation protocol. For each class, BP precision is the fraction of windows predicted as that class that were also generated for it. BP recall is the fraction of windows generated for that class that the teacher assigns to it. Both scores are averaged across classes.

Figure 3(g)-(h) shows that GEMIA achieves the highest mean BP precision and recall on both TUSZ protocols. On TUAB, it also has the highest mean recall, while JET has slightly higher precision. On TUEV, it has the highest mean precision, with recall close to the best result.

These results show strong label agreement in bandpower features. However, agreement with a bandpower teacher does not ensure useful augmentation: generated samples may still lack fine signal details or introduce biases that limit a classifier's generalization to unseen patients. Condition agreement and augmentation utility are therefore complementary, not interchangeable. Full values and sample standard deviations are provided in Appendix Tables B3 and B4.

4.3 Structural Diagnostics

We further examine whether generated EEG preserves the spectral-spatial structure of real data. CSD error measures differences in channel power and cross-channel relationships across frequencies, while coherence error compares normalized cross-channel dependence. The effective-rank gap measures differences in effective spatial dimensionality. Domain gap provides a complementary check of how easily a bandpower classifier distinguishes real from generated EEG. Lower values indicate closer agreement with real data. Figure 4 shows that PB-GSE-MSI achieves the lowest mean CSD error on TUSZ v2.0.6 and TUAB, and the lowest mean effective-rank gap on TUSZ v2.0.6 and TUEV.

The results on TUEV illustrate why these diagnostics are complementary: a smaller effective-rank gap can coexist with a larger CSD error. Matching the effective spatial dimensionality does not ensure agreement in the full cross-spectral matrix. These diagnostics therefore distinguish overall structural dimensionality from detailed frequency-dependent channel relationships. Full values together with separate membership-inference diagnostics are provided in Appendix Table B4.

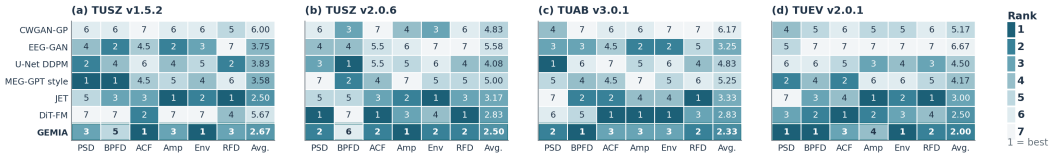

Figure 2: Complementary signal fidelity beyond TS-FID. Generators are ranked by three-run means within each corpus and metric; lower values receive better ranks (1 = best).



Figure 3: Low-patient downstream utility and condition-label agreement. (a)-(c): absolute real-test classification scores. (d)-(f): changes relative to Real-only in percentage points. (g)-(h): scale-aligned BP precision and recall. Error bars in (a)-(c) and (g)-(h) show sample standard deviations over three runs. Red stars mark the highest mean among generators within each corpus.

4.4 Generation Efficiency

We examine whether generation quality can be maintained with fewer sampling steps. We compare JET, DiT-FM, and GEMIA using the Heun solver with $K \in \{5, 10, 20, 50\}$ integration steps. For each method and training run, the checkpoint and source distribution remain fixed while K changes. We measure fidelity using TS-FID, where lower values indicate closer agreement with real data. Figure 5 shows that GEMIA achieves the lowest mean TS-FID at every evaluated step count on all four protocols. Its five-step results are also better than the fifty-step results of both baselines. For example, on TUSZ v2.0.6, GEMIA achieves 179.546 ± 12.636 with five steps, compared with 380.850 ± 12.418 for JET and 511.537 ± 11.586 for DiT-FM with fifty steps. These results show that GEMIA maintains low TS-FID with a small number of network evaluations. Complete step-wise results are provided in Appendix Table B5.

4.5 Ablation Studies

We examine whether the benefit of GSE depends on the training objective and how the auxiliary losses contribute to generation quality. We compare Gaussian and GSE sources under two objectives: $\mathcal{L}_{\text{cap}}$ alone and the full objective in Eq. (9), with $\lambda_{\text{aux}} = 0.1$. With GSE fixed, we also compare progressively larger sets of auxiliary terms: scale alignment, scale plus increment alignment, and all three terms. We evaluate each configuration using held-out TS-FID, where lower values indicate better fidelity. Within each corpus, all configurations share the backbone, patient split, training budget, and evaluation protocol. All results use seed 0 and Heun steps $K = 50$. The full GEMIA configuration reuses its seed-0 checkpoint and generated pool from the main evaluation; its score is a single-run result.

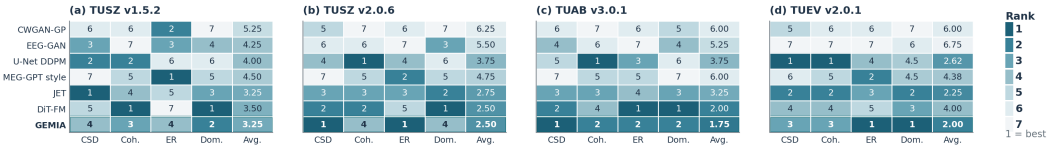

Figure 4: Structural diagnostics. Generators are ranked by three-run means within each corpus and metric; lower values receive better ranks (1 = best).

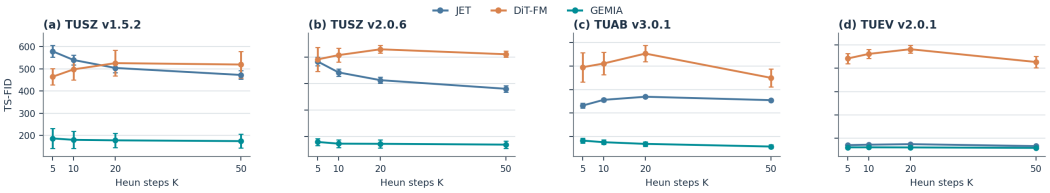

Figure 5: Solver-step sensitivity. TS-FID at different Heun step counts K; lower is better. Points and error bars show means and sample standard deviations over three generator runs.

Table 3: Source-objective and progressive loss ablations. All configurations share the capped velocity-L_1 loss. Checkmarks indicate auxiliary terms. Bold marks the lowest TS-FID within each corpus.

Source	Auxiliary terms			TS-FID ↓			
	$\mathcal{L}_{\text{scale}}$	$\mathcal{L}_{\text{inc}}$	$\mathcal{L}_{\text{corr}}$	TUSZ v1.5.2	TUSZ v2.0.6	TUAB v3.0.1	TUEV v2.0.1
Gaussian	–	–	–	1087.978	749.351	361.338	1082.464
Gaussian	✓	✓	✓	431.812	291.941	333.901	519.722
GSE	–	–	–	1018.885	821.349	289.980	673.761
GSE	✓	–	–	1237.217	1204.054	1584.089	379.405
GSE	✓	✓	–	456.460	325.229	312.908	463.822
GSE (full)	✓	✓	✓	**150.924**	**161.491**	**279.619**	**317.869**

Source-objective comparison. Table 3 shows that the full GEMIA configuration achieves the lowest TS-FID on all four protocols. With the full objective fixed, replacing the Gaussian source with GSE reduces TS-FID by 65.0%, 44.7%, 16.3%, and 38.8% on TUSZ v1.5.2, TUSZ v2.0.6, TUAB, and TUEV, respectively. The source replacement is less consistent with $\mathcal{L}_{\text{cap}}$ alone: it improves three protocols but increases TUSZ v2.0.6 TS-FID from 749.351 to 821.349. Adding the full auxiliary objective lowers TS-FID under either source on every protocol.

Contribution of the auxiliary losses. With GSE fixed, adding scale alignment alone reduces TUEV TS-FID from 673.761 to 379.405, but increases it on TUSZ protocols and TUAB. Adding increment alignment improves these three protocols, although TUEV TS-FID rises from 379.405 to 463.822. Adding waveform correlation reduces TS-FID on all four protocols, by 10.6%-66.9%. The complete objective therefore performs best among the tested combinations. These results support that combining the structured source with suitable supervision can consistently improve fidelity.

5 Conclusion

We presented GEMIA, a conditional EEG Flow Matching method combining shared spectral-spatial source structure with sample-specific time-domain supervision. Across four patient-wise TUH protocols, GEMIA achieves the lowest mean TS-FID among the compared generators, and its five-step means remain below the fifty-step means of both Flow Matching baselines. Fixed-seed ablations support the combined source-objective design. Beyond fidelity, GEMIA shows strong label agreement in bandpower features, and these findings highlight condition agreement and augmentation utility as complementary, rather than interchangeable, measures of synthetic EEG quality.

ETHICS AND REPRODUCIBILITY

The experiments use de-identified research EEG datasets. Generated signals should be identified as synthetic and should not be treated as clinical evidence without independent validation. Any release should respect the access conditions of the source datasets.

All GSE statistics are computed from training patients. We plan to release source-construction, training, generation, and evaluation code with versioned configurations and result records. These records should identify splits, source statistics, checkpoints and selected EMA weights, generated pools, and evaluator versions through configurations. They should specify the centering and RMS operators, a_0, α_{amp}, τ, λ_{aux}, $\mathcal{L}_{\mathrm{lag}}$, training budgets, and training and sampling seeds.

Evaluation records should include signal units and scale conversions, TS-FID pooling and feature dimensions, reference identities, diagnostic sample selection, Heun steps, network-evaluation counts, and timing scope. Both BP precision conventions remain available with their underlying predictions, and raw MIA AUC should accompany the transformed advantage. This distinguishes reporting changes from changes to models or generated data.

AI USE STATEMENT

We used ChatGPT and Codex to assist with literature search and organization, methodological discussion and experimental design, mathematical derivations and proof drafting, including Lean proof scripts, code development and debugging, preprocessing checks, analysis and interpretation of experimental outputs, and manuscript drafting and revision. The authors take full responsibility for the methodological choices, mathematical claims, code, references, figures, reported results, and final manuscript. The authors reviewed the AI-assisted material and checked the reported numerical values against the experimental records.

REFERENCES

Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 290–306. PMLR, 2022. URL https://proceedings.mlr.press/v162/alaa22a.html.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://jmlr.org/papers/v26/23-1605.html.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Richard Csaky, Mats W. J. van Es, Oiwi Parker Jones, and Mark Woolrich. Foundational GPT Model for MEG. *arXiv preprint arXiv:2404.09256v1*, 2024. doi: 10.48550/arXiv.2404.09256. URL https://arxiv.org/abs/2404.09256v1. Version 1; later versions use the title GPT2MEG: Quantizing MEG for Autoregressive Generation.

Shuai Cui, Wenxuan Chen, Wenjie Du, Jian Lou, Dan Li, and Wenjie Feng. AdaPCLA: Curriculum prior internalization for long-tailed longitudinal EHR generation. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=jdm8WK08ec.

Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation. *arXiv preprint arXiv:2111.08095*, 2021. doi: 10.48550/arXiv.2111.08095. URL https://arxiv.org/abs/2111.08095.

Thomas Donoghue, Matar Haller, Erik J. Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Antonio H. Lara, Joni D. Wallis, Robert T. Knight, Avgusta Shestyuk, and Bradley Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, 23(12):1655–1665, 2020. doi: 10.1038/s41593-020-00744-x. URL https://doi.org/10.1038/s41593-020-00744-x.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017. doi: 10.48550/arXiv.1706.02633. URL https://arxiv.org/abs/1706.02633. NIPS 2017 Machine Learning for Health Workshop; arXiv version.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5767–5777, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv preprint arXiv:1806.01875*, 2018. doi: 10.48550/arXiv.1806.01875. URL https://arxiv.org/abs/1806.01875.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019. doi: 10.2478/popets-2019-0008. URL https://petsymposium.org/popets/2019/popets-2019-0008.php.

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*, volume 30, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Elaine B. Hoffman, Pranab K. Sen, and Clarice R. Weinberg. Within-cluster resampling. *Biometrika*, 88(4):1121–1134, 2001. doi: 10.1093/biomet/88.4.1121. URL https://academic.oup.com/biomet/article-abstract/88/4/1121/226041.

Chubin Huang, Gaoqiang Li, Sheng-hua Zhong, Rongrong Lu, and Tianhao Gao. DiffEEGBooth: A diffusion-based EEG generation framework for motor imagery with temporal consistency and neurophysiological constraint. *Journal of Visual Communication and Image Representation*, 117: 104758, 2026. doi: 10.1016/j.jvcir.2026.104758. URL https://www.sciencedirect.com/science/article/abs/pii/S1047320326000532.

Rukuang Huang, Sungjun Cho, Chetan Gohil, Oiwi Parker Jones, and Mark Woolrich. MEG-GPT: A transformer-based foundation model for magnetoencephalography data. *arXiv preprint arXiv:2510.18080*, 2025. doi: 10.48550/arXiv.2510.18080. URL https://arxiv.org/abs/2510.18080.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Timon Klein, Piotr Minakowski, and Sebastian Sager. Flexible Patched Brain Transformer model for EEG decoding. *Scientific Reports*, 15(1):10935, 2025. doi: 10.1038/s41598-025-86294-3. URL https://doi.org/10.1038/s41598-025-86294-3.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow Matching with Gaussian Process Priors for Probabilistic Time Series Forecasting. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ee1a1ecc92f35702b5c29dad3dc909ea-Abstract-Conference.html.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2021.653659/full.

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*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Tianhong Li and Kaiming He. Back to Basics: Let Denoising Generative Models Denoise. *arXiv preprint arXiv:2511.13720*, 2025. doi: 10.48550/arXiv.2511.13720. URL https://arxiv.org/abs/2511.13720.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Yun Luo and Bao-Liang Lu. EEG Data Augmentation for Emotion Recognition Using a Conditional Wasserstein GAN. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2535–2538. IEEE, 2018. doi: 10.1109/EMBC.2018.8512865. URL https://doi.org/10.1109/EMBC.2018.8512865.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers. In *Computer Vision – ECCV 2024*, volume 15135 of *Lecture Notes in Computer Science*, pp. 23–40. Springer, 2024. doi: 10.1007/978-3-031-72980-5_2. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/9828_ECCV_2024_paper.php.

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://www.sciencedirect.com/science/article/pii/S1388245704001993.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205. IEEE, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Dean Prichard and James Theiler. Generating surrogate data for time series with several simultaneously measured variables. *Physical Review Letters*, 73(7):951–954, 1994. doi: 10.1103/PhysRevLett.73.951. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.73.951.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868. PMLR, 2021. URL https://proceedings.mlr.press/v139/rasul21a.html.

Suman Ravuri and Oriol Vinyals. Classification Accuracy Score for Conditional Generative Models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fcf55a303b71b84d326fb1d06e332a26-Abstract.html.

Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5):1–9, 2017. doi: 10.1093/gigascience/gix019. URL https://academic.oup.com/gigascience/article/6/5/gix019/3071704.

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*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f7696a9b362ac5a51c3dc8f098b73923-Abstract.html.

Thomas Schreiber and Andreas Schmitz. Improved Surrogate Data for Nonlinearity Tests. *Physical Review Letters*, 77(4):635–638, 1996. doi: 10.1103/PhysRevLett.77.635. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.77.635.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250. URL https://ieeexplore.ieee.org/document/9991178/.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0165. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract.html.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1–4):77–94, 1992. doi: 10.1016/0167-2789(92)90102-S. URL https://doi.org/10.1016/0167-2789(92)90102-S.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

Shaocong Wang, Tong Liu, Yihan Li, Ming Li, Kairui Wen, Pei Yang, Wenqi Ji, Minjing Yu, and Yong-Jin Liu. EEGDM: Learning EEG Representation with Latent Diffusion Model. *arXiv preprint arXiv:2508.20705v4*, 2025b. doi: 10.48550/arXiv.2508.20705. URL https://arxiv.org/abs/2508.20705. First submitted in 2025; version 4 revised in 2026.

Yifan Wang, Yijia Ma, Wen Li, and Chenyu You. Let EEG Models Learn EEG. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.21280. Accepted at ICML 2026; arXiv version.

Yihe Wang, Taida Li, Yujun Yan, Wenzhan Song, and Xiang Zhang. How to evaluate your medical time series classification? *arXiv preprint arXiv:2410.03057v1*, 2024b. doi: 10.48550/arXiv.2410.03057. URL https://arxiv.org/abs/2410.03057v1. Version 1; later versions use a revised title.

John M. Williamson, Somnath Datta, and Glen A. Satten. Marginal Analyses of Clustered Data When Cluster Size Is Informative. *Biometrics*, 59(1):36–42, 2003. doi: 10.1111/1541-0420.00005. URL https://doi.org/10.1111/1541-0420.00005.

Andrew Wilson and Ryan Adams. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1067–1075. PMLR, 2013. URL https://proceedings.mlr.press/v28/wilson13.html.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

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://proceedings.neurips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html.

Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica I. Aviles-Rivero, and Shujun Wang. Decentralized Attention Fails Centralized Signals: Rethinking Transformers for Medical Time Series. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.18473. ICLR 2026; arXiv version.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable Diffusion for General Time Series Generation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b5b66077d016c037576cc56a82f97f66-Abstract-Conference.html.

A Extended Related Work

EEG generation and representation learning. Early EEG generation methods mainly used adversarial learning, building on the GAN framework (Goodfellow et al., 2014). EEG-GAN generates EEG waveforms using a Wasserstein-GAN-style framework (Hartmann et al., 2018). Conditional Wasserstein GANs have also been used to generate EEG feature representations for downstream data augmentation (Luo & Lu, 2018). These studies show that synthetic EEG can support neuroscience and classification tasks, but adversarial training can be unstable (Gulrajani et al., 2017) and does not provide an explicit continuous transport process. Building on denoising diffusion (Ho et al., 2020), recent work also generates neurophysiological and motor-imagery EEG signals (Vetter et al., 2024; Huang et al., 2026).

JET is the closest prior work to our generation setting. It applies conditional Flow Matching to multichannel EEG and introduces signal-level constraints for amplitude statistics, temporal variation, and waveform correlation (Wang et al., 2026). Our work uses JET as a strong external baseline, but studies a different question. JET starts from Gaussian noise and places its EEG-specific information mainly in the training objective, whereas we study a patient-balanced structured source together with a complementary time-domain objective. Generative modeling can also be used to learn EEG representations. Other EEG encoders use self-supervised or cross-data learning (Kostas et al., 2021; Wang et al., 2024a; Jiang et al., 2024; Yang et al., 2023). EEGDM uses latent diffusion as a self-supervised training objective for an EEG encoder, whose representations are used in downstream tasks (Wang et al., 2025b). While EEGDM emphasizes learning representations through generation, our work focuses on generating EEG waveforms and evaluating their usefulness for downstream classification.

General time-series generators provide several related modeling choices, including recurrent conditional GANs and variational autoencoders (Esteban et al., 2017; Desai et al., 2021). TimeGAN combines adversarial learning with a supervised temporal objective in a learned embedding space (Yoon et al., 2019). Diffusion-TS directly predicts clean time-series samples and adds a Fourier-domain loss to preserve temporal structure (Yuan & Qiao, 2024). TimeGrad and CSDI apply diffusion to forecasting and imputation, respectively (Rasul et al., 2021; Tashiro et al., 2021). For longitudinal electronic health records (EHRs), AdaPCLA (Cui et al., 2026) uses curriculum-based prior internalization to improve the modeling of dependencies involving rare medical codes. It focuses on discrete clinical-event sequences rather than continuous EEG waveforms. These studies address temporal dependence and data-specific structure, both of which are relevant to EEG generation.

Flow Matching, source design, and architecture. Flow Matching trains a continuous vector field by regressing the velocity of a prescribed probability path (Lipman et al., 2023). Related continuous-time models use neural ODEs, score-based SDEs, or rectified flows (Chen et al., 2018; Song et al., 2021; Liu et al., 2023). Although an isotropic Gaussian is a common and successful source, continuous transport frameworks are not limited to one source distribution. Stochastic interpolants can connect general endpoint distributions (Albergo et al., 2025), and TSFlow gives a time-series example in which Gaussian-process priors encode temporal structure before transport (Kollovieh et al., 2025). Spectral Gaussian-process kernels also encode temporal patterns (Wilson & Adams, 2013). GSE follows this source-design idea, but uses a different construction: It estimates cross-spectral statistics with equal weight for each training patient and uses these statistics to construct a label-neutral EEG source. This weighting is related to inverse-cluster-size estimation (Williamson et al., 2003; Hoffman et al., 2001).

Our network architecture uses existing Transformer designs (Vaswani et al., 2017; Dosovitskiy et al., 2021; Ma et al., 2024). DiT uses Transformer blocks with adaptive layer normalization for generative modeling (Peebles & Xie, 2023). JiT argues that directly predicting clean data can be easier than predicting noise or velocity in a high-dimensional space (Li & He, 2025). Patch-based Transformers are also established in EEG decoding, including PBT and CBraMod (Klein et al., 2025; Wang et al., 2025a). EEG Conformer combines convolution with attention (Song et al., 2023). We therefore treat patch tokenization, AdaLN-Zero, and the clean-data prediction head as inherited components.

Spectral–spatial structure of EEG. Neural power spectra include an aperiodic background and may contain periodic oscillatory peaks (Donoghue et al., 2020). For multichannel EEG, the power spectrum of each channel describes its frequency content but does not describe how different channels

vary together (Nolte et al., 2004). Fourier surrogate methods preserve selected signal statistics (Theiler et al., 1992; Schreiber & Schmitz, 1996). Prichard and Theiler extend Fourier surrogate generation to multivariate time series while preserving both within-channel autocorrelation and cross-channel correlation (Prichard & Theiler, 1994). They demonstrate the method on multichannel EEG.

GSE uses spectral and cross-channel statistics to construct its source. Its cross-spectral matrix is estimated from amplitude-normalized training windows and describes their channel-wise power and frequency-dependent cross-channel relations. A separate amplitude distribution controls the overall scale of each source sample. We estimate both the cross-spectral statistics and the amplitude distribution with equal weight for each training patient, without using class labels. GSE provides a statistical source distribution for Flow Matching, while the conditional flow learns to transform samples from this source into class-specific EEG waveforms.

Training objectives and synthetic-data evaluation. Signal structure can guide the training objective as well as the source distribution. Diffusion-TS includes a Fourier-domain loss (Yuan & Qiao, 2024). JET combines endpoint reconstruction with statistical, total-variation, and correlation constraints (Wang et al., 2026). We study how such supervision can complement a structured source. GSE supplies patient-shared spectral–spatial structure. Our auxiliary losses compare predicted and real EEG windows through per-channel scale, temporal increments at several lags, and waveform correlation. Since GSE already injects spectral-spatial structure into the source, we deliberately formulate these comparisons in the time domain to decouple the objective from the source and prevent redundant supervision.

Generated data also need to be evaluated beyond the training objective (Stein et al., 2023; Alaa et al., 2022). FID compares the means and covariances of Inception features from real and generated images (Heusel et al., 2017). Precision–recall measures further distinguish fidelity from coverage (Sajjadi et al., 2018; Kynkäänniemi et al., 2019). TimeGAN uses discriminative tests to assess whether real and synthetic sequences can be distinguished. It also uses predictive tests to assess whether a model trained on synthetic data can make accurate predictions on real data (Yoon et al., 2019). Related classification-based tests assess conditional utility (Ravuri & Vinyals, 2019). For EEG, JET uses TS-FID to compare spectral feature distributions and evaluates synthetic-data augmentation through downstream classification (Wang et al., 2026). We additionally compare real and generated EEG in the feature space of a frozen CBraMod encoder (Wang et al., 2025a). These features are not used in the generator's training objective. Patient-disjoint evaluation tests transfer to unseen subjects (Saeb et al., 2017; Wang et al., 2024b; Yu et al., 2026), while membership-inference attacks assess memorization risk (Hayes et al., 2019). Our evaluation reports signal fidelity, downstream utility, mechanism and safety diagnostics, and generation efficiency separately.

B Complete Numerical Results and Plotting Conventions

B.1 Datasets and Patient-Wise Splits

Table 1 in the main text summarizes the four protocols, their tasks, and processed patient and window counts. The task labels define the conditioning variable y. TUSZ and TUEV windows have shape $\boldsymbol{X} \in \mathbb{R}^{16\times1000}$, while TUAB windows have shape $\boldsymbol{X} \in \mathbb{R}^{16\times2000}$.

Patient-wise assignment. The split audits report zero patient overlap between any pair of training, validation, and test sets within each protocol. For TUSZ v1.5.2, the patient-wise partition is constructed from existing preprocessed windows using a test-priority rule: patients present in both the original training and test partitions are assigned entirely to the test partition. The processed split therefore differs from the original allocation of preprocessed windows.

GEMIA estimates both the spectral source statistics and the amplitude distribution from training patients only. Equal-patient weighting in source estimation is distinct from patient-disjoint evaluation and from class-balanced sampling during training.

Input representation and scaling. The collected Flow Matching data loaders reshape each TUSZ or TUEV window to $16 \times 5 \times 200$ and each TUAB window to $16 \times 10 \times 200$, then divide the stored signal values by 100. The loaded samples are converted to single-precision tensors. The final two

Table B1: Generator families and training objectives. The four non-flow baselines are budget-limited adapted implementations. JET and DiT-FM are the main Flow Matching baselines.

Method	Model family	Training objective
Conditional WGAN-GP	GAN	Wasserstein objective with gradient penalty
EEG-GAN	GAN	Adversarial objective
1D U-Net DDPM	Diffusion	Denoising objective
MEG-GPT-style EEG	Adapted sequence model	Not established in collected run records[†]
JET	Flow Matching	JET signal-constrained objective
DiT-FM	Flow Matching	Velocity-L_2
GEMIA	Flow Matching	Capped velocity-L_1 and time-domain alignment

[†]The historical training objective of the adapted MEG-GPT-style EEG run family is not established by the collected run-specific source records. It is therefore not characterized here as autoregressive likelihood training.

axes partition the temporal dimension into 200-sample blocks without changing the window duration. The TUEV loader converts stored labels $1, \ldots, 6$ to $0, \ldots, 5$; the binary protocols use labels 0 and 1. This fixed input scaling is distinct from the centering and RMS normalization used to estimate GSE shape statistics, and from feature standardization used by individual evaluators.

Collected preprocessing implementation. The collected preprocessing scripts specify resampling to 200 Hz, a 0.3–75 Hz bandpass filter, a 60 Hz notch filter, and bipolar differences computed from signals read in microvolts. Their common ordered montage is

FP1–F7, F7–T3, T3–T5, T5–O1,
FP2–F8, F8–T4, T4–T6, T6–O2,
FP1–F3, F3–C3, C3–P3, P3–O1,
FP2–F4, F4–C4, C4–P4, P4–O2.

The manifest-based TUAB script extracts complete non-overlapping 10-second windows. The manifest-based TUSZ script extracts complete 5-second windows within contiguous seizure or background intervals. The manifest-based TUEV script instead extracts 5-second windows centered on annotated events and uses reflection at recording boundaries; non-overlap is not imposed between different event windows. These statements describe the collected implementations. The historical script version that produced each processed corpus has not been independently established by this collection; TUSZ v1.5.2 explicitly inherits previously processed windows. The input artifacts and split counts remain those listed in Table 1.

B.2 Compared Methods and Evaluation Protocol

Compared methods. Table B1 separates generator families from training objectives. JET is the closest baseline to our conditional EEG setting: it uses a Gaussian source and EEG-oriented signal constraints. DiT-FM uses the DiT backbone family with a Gaussian source and a velocity-L_2 objective, without auxiliary signal losses. GEMIA combines the patient-balanced source with the capped velocity-L_1 loss, channel-scale alignment, multi-lag increment alignment, and waveform correlation. These comparisons evaluate complete methods; source and loss controls are examined separately in Table 3.

Conditional WGAN-GP, EEG-GAN, 1D U-Net DDPM, and MEG-GPT-style EEG are budget-limited adapted baselines, not exact reproductions of every detail of the original architectures. Their training costs are not matched to the controlled Flow Matching comparison. We compare these adapted implementations under the reported evaluation protocol, without attributing all differences to the generative framework alone.

GEMIA implementation. GEMIA uses the JiT-B/16 backbone to predict clean EEG endpoints, with a temporal patch length of 200 samples. The archived seed-0 training logs report approximately 129.86 million trainable parameters: 129,858,248 for the binary tasks and 129,861,320 for the six-class TUEV task. Attention and projection dropout are set to zero, while class-label dropout is 0.1. Flow times follow the logit-normal parameterization with $\mu_t = -0.8$ and $\sigma_t = 0.8$. The

source amplitude strength is $\alpha_{\text{amp}} = 0.20$; the loss uses $\tau = 0.05$, $\lambda_{\text{aux}} = 0.1$, and increment lags $\mathcal{L}_{\text{lag}} = \{1, 4, 16\}$.

Training schedule. The recorded GEMIA schedule uses 200 epochs and batch size 256 for both TUSZ protocols, 131 epochs and batch size 256 for TUEV, and 28 epochs and batch size 128 for TUAB. The training implementation uses AdamW, with weight decay set to zero in the recorded runs. The base learning rate is 5×10^{-5} per 256 samples. The logged learning rates before warm-up scaling are 5×10^{-5} for TUSZ and TUEV and 2.5×10^{-5} for TUAB. All four protocols use five warm-up epochs followed by a constant learning-rate schedule. Two EMA tracks are maintained with decay coefficients 0.9999 and 0.9996. The archived training engine uses bfloat16 autocasting, with one optimizer update per minibatch. The recorded runs use four data-loader workers, a single process, and no early stopping. The main method is evaluated from the final-budget checkpoints. These settings describe GEMIA and are not assumed to apply to all non-flow baselines.

Training sampling. The seed-0 records enable class-weighted sampling for TUSZ and TUEV and disable it for TUAB. The collected loader implements inverse-class- frequency sampling with replacement, while the unweighted TUAB loader shuffles training windows. This sampling rule controls the class exposure of the generator during training; it does not replace patient-balanced estimation of the source statistics.

Hardware and software. GPU training and neural-network inference jobs use one NVIDIA GeForce RTX 4090 with 24 GB memory per job. An archived environment snapshot, collected on 18 September 2026, records Python 3.10.20 and PyTorch `2.11.0+cu128`, built with CUDA 12.8 and cuDNN 9.19.0. The snapshot documents the collected software environment rather than a separately verified installation history for every run. CPU-based diagnostic computations do not require GPU inference.

Sampling and repeated runs. The main Flow Matching evaluation uses Heun sampling with $K = 50$ integration steps, sampling batch size 64, and guidance scale 1.0. Generated pools match the class counts of the corresponding processed held-out test pool; source samples are drawn independently of the real test waveforms. The latter are used only as evaluation references. Main fidelity results report means and sample standard deviations over generator seeds $\{0, 2025, 2026\}$. The source–objective and progressive loss ablations use seed 0; their full-model row reuses the corresponding main-model checkpoint rather than adding another training run. The sampling study additionally evaluates $K \in \{5, 10, 20\}$ while holding the checkpoint and source distribution fixed within each run. Our implementation uses $2K - 1$ network evaluations, so $K = 5$ and $K = 50$ correspond to 9 and 99 evaluations, respectively. These comparisons assess low-step sampling rather than measured wall-clock acceleration.

Fidelity evaluation. TS-FID is computed from standardized log-FFT features with a frequency-bin cap of $B_f = 256$ on complete generated and real test pools. RFD uses a frozen CBraMod encoder and the complete test pools. The five remaining signal-level metrics in Table B2 use the fixed class-stratified audit subset. CSD error is reported only in Table B4, together with the other structural diagnostics. The detailed metric definitions specify feature maps, scaling, sampling limits, and class aggregation for each evaluation.

Low-patient downstream evaluation. All downstream classifiers use the same fixed subset of 10% of training patients, selected with patient seed 4407. Real-only uses these real windows alone; each augmentation method adds synthetic windows equal to 50% of the real window count in each class. A linear classifier is trained on a frozen CBraMod encoder for 50 epochs with batch size 256 and is evaluated on real held-out patients. Classifier checkpoints are selected using real-validation AUROC. The three generator/classifier seed pairs are $(0, 0)$, $(2025, 2025)$, and $(2026, 2026)$; they are not nine independent combinations. The low-patient restriction applies to the classifier's real training subset, not to the full generator-training datasets listed in Table 1.

Structural and memorization diagnostics. The structural and condition-agreement measurements retain the protocols and reporting conventions in Table B4. BP precision and recall measure agreement between requested labels and bandpower-teacher predictions, not population coverage

Table B2: Signal fidelity on held-out test patients. Lower is better in every metric column. Values are means and sample standard deviations across generator runs. TS-FID uses per-window standardized log-FFT features with frequency-bin cap $B_f = 256$. TS-FID and RFD use complete test pools; signal-level diagnostics use the fixed class-stratified audit subset. GEMIA uses training/sampling seeds $\{0, 2025, 2026\}$. JET now has three-run results on all four protocols.

Corpus	Method	TS-FID	log-PSD	ACF	Amp-W1	BP-FD	Env-W1	RFD
TUSZ v1.5.2	Conditional WGAN-GP	1419.034±109.905	3.630±0.235	0.152±0.014	0.398±0.394	7492.705±12836.200	0.596±0.644	1.831±2.102
	EEG-GAN	941.843±306.924	1.437±0.137	0.127±0.025	0.204±0.152	85.697±6.250	0.274±0.234	2.362±2.118
	1D U-Net DDPM	741.896±59.838	1.412±0.028	0.133±0.000	0.224±0.006	148.617±23.237	0.327±0.010	0.468±0.007
	MEG-GPT-style EEG	539.810±85.409	1.378±0.075	0.127±0.003	0.383±0.007	79.869±0.315	0.316±0.012	2.011±0.664
	JET	472.236±18.919	1.942±0.113	0.107±0.019	0.172±0.033	106.107±26.899	0.247±0.056	0.458±0.180
	DiT-FM	519.203±59.841	4.426±0.297	0.052±0.009	0.613±0.099	70339.207±6090.775	0.945±0.165	1.419±0.579
	GEMIA	174.139±30.792	1.726±0.383	0.059±0.024	0.705±0.438	1720.614±2101.833	1.100±0.721	1.804±1.928
TUSZ v2.0.6	Conditional WGAN-GP	1652.511±213.423	3.970±1.376	0.139±0.006	0.311±0.035	87.656±6.194	0.438±0.062	2.280±0.385
	EEG-GAN	1015.670±388.009	3.093±1.302	0.130±0.041	0.376±0.252	88.460±26.926	0.544±0.382	5.396±4.420
	1D U-Net DDPM	764.188±63.577	2.645±0.052	0.130±0.000	0.349±0.009	79.343±0.084	0.515±0.014	1.514±0.009
	MEG-GPT-style EEG	631.417±67.513	4.233±0.475	0.128±0.000	0.439±0.033	80.223±0.097	0.496±0.018	2.242±0.297
	JET	380.850±12.418	3.885±0.335	0.092±0.006	0.252±0.015	110.982±10.696	0.383±0.019	1.334±0.068
	DiT-FM	511.537±11.586	1.543±0.200	0.049±0.007	0.276±0.078	2460.934±1091.803	0.441±0.128	1.313±0.137
	GEMIA	169.460±14.407	3.087±0.084	0.077±0.007	0.641±0.047	137.085±25.500	0.994±0.077	1.729±0.065
TUAB v3.0.1	Conditional WGAN-GP	635.258±107.370	1.995±0.931	0.162±0.001	0.221±0.229	6112.133±8020.796	0.350±0.372	2.497±2.587
	EEG-GAN	754.947±157.113	1.841±0.232	0.161±0.023	0.097±0.020	75.804±5.999	0.112±0.015	0.789±0.255
	1D U-Net DDPM	554.608±16.203	1.430±0.046	0.177±0.000	0.197±0.005	97.168±3.859	0.283±0.008	0.440±0.014
	MEG-GPT-style EEG	439.196±14.115	2.014±0.498	0.161±0.003	0.355±0.027	79.699±0.253	0.282±0.022	2.023±0.149
	JET	377.518±2.614	3.135±0.228	0.141±0.002	0.137±0.015	73.659±0.545	0.183±0.016	0.163±0.012
	DiT-FM	424.968±19.036	2.166±0.123	0.063±0.005	0.082±0.006	82.405±22.687	0.109±0.002	0.339±0.058
	GEMIA	278.993±4.021	2.459±0.108	0.216±0.027	0.228±0.046	69.966±1.653	0.331±0.068	0.207±0.035
TUEV v2.0.1	Conditional WGAN-GP	734.579±36.120	1.819±0.246	0.130±0.010	0.132±0.060	14195.074±23771.154	0.194±0.098	1.847±1.149
	EEG-GAN	954.151±179.739	1.981±0.579	0.133±0.012	0.226±0.210	377956.076±652737.860	0.345±0.349	1.877±1.134
	1D U-Net DDPM	615.079±2.379	2.023±0.003	0.114±0.000	0.110±0.001	26660.231±2934.717	0.177±0.002	0.400±0.011
	MEG-GPT-style EEG	382.686±23.985	1.711±0.126	0.097±0.001	0.194±0.029	5397.165±2366.190	0.204±0.019	0.852±0.253
	JET	332.134±2.498	2.047±0.042	0.112±0.001	0.060±0.002	2400.592±127.641	0.095±0.004	0.166±0.007
	DiT-FM	1053.420±48.906	1.739±0.114	0.090±0.003	0.085±0.003	1704.400±470.637	0.132±0.007	0.781±0.118
	GEMIA	316.612±4.501	1.698±0.037	0.161±0.010	0.115±0.011	421.204±110.831	0.176±0.018	0.281±0.021

or augmentation benefit. The zero-precision convention for a class without teacher predictions is distinct from replacing a failed computation with zero. Membership-inference results describe the specified attack and score direction, not a general privacy guarantee. The legacy interpretation of NN gap remains as stated in that table.

Numerical records and plotting conventions. Tables 1–B5 collect the dataset, method, and numerical records. In particular, Tables B2–B5 preserve the complete results underlying the main-text visualizations, with their uncertainties and evaluation scopes. Figures are derived from reported results without rerunning evaluators. Rank maps use means within each corpus and metric, with average ranks for exact ties at the displayed precision. Rank 1 is best. The fidelity figure averages ranks over its six non-TS-FID metrics; the structural figure averages ranks over its four displayed metrics. These unweighted averages are descriptive summaries, not combined fidelity or safety scores, effect sizes, or significance tests. Utility differences are computed from the displayed means in percentage points. No paired-difference standard deviation is reconstructed from marginal standard deviations. Red stars identify the highest generator mean within each corpus and panel; Real-only is a reference and is excluded from this comparison set.

Table B3: Downstream utility on real held-out test data. Higher is better. Real-only uses a fixed 10% subset of training patients (subset seed 4407). All other rows add synthetic data to the same subset, using a per-class synthetic-to-real ratio of 0.5. We train a linear classifier on a frozen CBraMod encoder for 50 epochs with batch size 256, and select the checkpoint by validation AUROC. Values are mean $\pm$ sample standard deviation over seeds $\{0, 2025, 2026\}$. Generator and classifier seeds are paired in the augmentation runs.

Corpus	Training source	Balanced Accuracy	AUPRC	AUROC
TUSZ v1.5.2	Real-only	0.692 ± 0.001	0.817 ± 0.001	0.778 ± 0.001
	Conditional WGAN-GP	0.637 ± 0.073	0.812 ± 0.009	0.771 ± 0.004
	EEG-GAN	0.670 ± 0.007	0.808 ± 0.010	0.765 ± 0.011
	1D U-Net DDPM	0.672 ± 0.005	0.815 ± 0.001	0.772 ± 0.001
	MEG-GPT-style EEG	0.615 ± 0.073	0.818 ± 0.001	0.773 ± 0.002
	JET	0.564 ± 0.060	0.818 ± 0.004	0.772 ± 0.003
	DiT-FM	0.667 ± 0.022	0.818 ± 0.005	0.774 ± 0.003
	GEMIA	0.584 ± 0.085	0.818 ± 0.003	0.774 ± 0.004
TUSZ v2.0.6	Real-only	0.713 ± 0.001	0.738 ± 0.001	0.777 ± 0.001
	Conditional WGAN-GP	0.672 ± 0.003	0.682 ± 0.005	0.730 ± 0.009
	EEG-GAN	0.688 ± 0.037	0.700 ± 0.045	0.746 ± 0.044
	1D U-Net DDPM	0.697 ± 0.002	0.710 ± 0.003	0.759 ± 0.002
	MEG-GPT-style EEG	0.683 ± 0.012	0.704 ± 0.004	0.748 ± 0.007
	JET	0.562 ± 0.101	0.680 ± 0.010	0.727 ± 0.007
	DiT-FM	0.686 ± 0.008	0.699 ± 0.003	0.747 ± 0.006
	GEMIA	0.653 ± 0.000	0.714 ± 0.006	0.761 ± 0.006
TUAB v3.0.1	Real-only	0.749 ± 0.000	0.814 ± 0.000	0.828 ± 0.000
	Conditional WGAN-GP	0.708 ± 0.013	0.753 ± 0.021	0.784 ± 0.001
	EEG-GAN	0.710 ± 0.003	0.767 ± 0.014	0.789 ± 0.010
	1D U-Net DDPM	0.737 ± 0.003	0.800 ± 0.001	0.818 ± 0.002
	MEG-GPT-style EEG	0.702 ± 0.003	0.737 ± 0.006	0.781 ± 0.007
	JET	0.716 ± 0.001	0.780 ± 0.002	0.798 ± 0.002
	DiT-FM	0.725 ± 0.005	0.785 ± 0.006	0.805 ± 0.006
	GEMIA	0.745 ± 0.006	0.801 ± 0.005	0.809 ± 0.002
TUEV v2.0.1	Real-only	0.210 ± 0.001	0.266 ± 0.002	0.721 ± 0.002
	Conditional WGAN-GP	0.233 ± 0.007	0.267 ± 0.004	0.746 ± 0.004
	EEG-GAN	0.226 ± 0.003	0.268 ± 0.006	0.756 ± 0.003
	1D U-Net DDPM	0.222 ± 0.001	0.273 ± 0.002	0.761 ± 0.003
	MEG-GPT-style EEG	0.209 ± 0.013	0.278 ± 0.004	0.777 ± 0.005
	JET	0.218 ± 0.000	0.280 ± 0.001	0.774 ± 0.002
	DiT-FM	0.216 ± 0.006	0.277 ± 0.004	0.764 ± 0.001
	GEMIA	0.222 ± 0.006	0.282 ± 0.001	0.781 ± 0.002

Table B4: Signal-structure and memorization diagnostics on strict held-out pools. Entries show the mean (first line) and sample standard deviation (second line) over generator seeds $\{0, 2025, 2026\}$, with diagnostic seed 3407 fixed. BP and MIA use the scale-aligned rescores. BP Prec.$^{(0)}$ macro-averages all condition classes and assigns zero only to a class with no predicted samples; this reporting convention was added after inspecting undefined values, and the original strict-NA results are retained separately. CSD, coherence error, rank gap, and domain gap retain their preceding values. †NN gap retains the earlier, non-scale-rescored values and is not used as evidence for memorization. MIA reports $\max(0, 2\,\mathrm{AUC} - 1)$ under a fixed score direction; these diagnostics do not establish a formal privacy guarantee.

Corpus	Generator	CSD ↓	Coh. err. ↓	Rank gap ↓	BP Prec.$^{(0)}$ ↑	BP Recall ↑	Domain gap ↓	NN gap† (legacy)	MIA adv. ↓
TUSZ v1.5.2	Conditional WGAN-GP	1.6328 ±1.0963	0.6958 ±0.0751	0.8460 ±0.1040	0.4170 ±0.2893	0.5015 ±0.0025	0.4971 ±0.0006	−3.6828 ±1.1562	0.1032 ±0.0214
	EEG-GAN	1.0197 ±0.0310	0.7634 ±0.1139	0.9425 ±0.1341	0.6576 ±0.3792	0.7279 ±0.2529	0.4930 ±0.0035	−4.3161 ±0.0248	0.1127 ±0.0262
	1D U-Net DDPM	0.9991 $\pm 9.96\times10^{-6}$	0.0523 ±0.0018	1.4582 ±0.0282	0.2500 ±0.0000	0.5000 ±0.0000	0.4961 ±0.0008	−4.2970 ±0.0035	0.0809 ±0.0008
	MEG-GPT-style EEG	1.8352 ±0.1548	0.2300 ±0.0181	0.3929 ±0.0104	0.2500 ±0.0000	0.5000 ±0.0000	0.4937 ±0.0015	−4.3252 ±0.0066	0.1419 ±0.0230
	JET	0.9939 ±0.0029	0.1529 ±0.0009	1.2429 ±0.2164	0.6728 ±0.1002	0.5107 ±0.0119	0.4843 ±0.0027	−4.0401 ±0.0308	0.0000 ±0.0000
	DiT-FM	1.6164 ±0.2731	0.0445 ±0.0017	1.6505 ±0.0140	0.5709 ±0.0509	0.5651 ±0.0551	0.4718 ±0.0129	−1.6009 ±0.5660	0.0533 ±0.0131
	GEMIA	2.3151 ±1.2477	0.1433 ±0.0592	1.1543 ±0.1977	0.8059 ±0.0742	0.7699 ±0.0989	0.4740 ±0.0241	−3.5971 ±0.4170	0.0204 ±0.0354
TUSZ v2.0.6	Conditional WGAN-GP	0.9994 ±0.0004	0.7239 ±0.0790	1.5488 ±0.1211	0.4529 ±0.4788	0.5452 ±0.4315	0.4995 ±0.0002	−4.1669 ±0.2447	0.0891 ±0.0772
	EEG-GAN	1.0190 ±0.0382	0.7219 ±0.1142	1.5541 ±0.1528	0.5103 ±0.3492	0.6271 ±0.2243	0.4901 ±0.0117	−4.1123 ±0.4759	0.0208 ±0.0361
	1D U-Net DDPM	0.9986 $\pm 5.19\times10^{-5}$	0.0759 ±0.0018	0.8103 ±0.0290	0.5121 ±0.0594	0.5007 ±0.0023	0.4966 ±0.0015	−4.3711 ±0.0039	0.0000 ±0.0000
	MEG-GPT-style EEG	1.0500 ±0.0339	0.1846 ±0.0075	0.5529 ±0.0151	0.2499 ±0.0002	0.4996 ±0.0007	0.4952 ±0.0004	−4.3841 ±0.0077	0.0000 ±0.0000
	JET	0.9835 ±0.0007	0.1045 ±0.0137	0.6940 ±0.0511	0.7250 ±0.0384	0.5412 ±0.0244	0.4720 ±0.0118	−4.0316 ±0.0170	0.0000 ±0.0000
	DiT-FM	0.9605 ±0.0050	0.0765 ±0.0017	1.0010 ±0.0308	0.6655 ±0.0661	0.5899 ±0.0894	0.3395 ±0.0568	−2.9238 ±0.5200	0.0029 ±0.0049
	GEMIA	0.9412 ±0.0506	0.1775 ±0.0530	0.4602 ±0.1197	0.7743 ±0.0325	0.7653 ±0.0334	0.4917 ±0.0014	−3.7049 ±0.0815	0.0000 ±0.0000
TUAB v3.0.1	Conditional WGAN-GP	1.5282 ±0.8768	0.6514 ±0.1978	1.5997 ±0.2857	0.4269 ±0.2804	0.4995 ±0.0021	0.4942 ±0.0021	−3.7044 ±0.2871	0.0152 ±0.0136
	EEG-GAN	0.9935 ±0.0047	0.6323 ±0.0238	1.6002 ±0.0257	0.5577 ±0.2694	0.5577 ±0.0962	0.4939 ±0.0017	−3.8955 ±0.0062	0.0000 ±0.0000
	1D U-Net DDPM	0.9983 ±0.0002	0.0525 ±0.0015	0.7076 ±0.0147	0.5565 ±0.0697	0.5013 ±0.0028	0.4969 ±0.0004	−3.9214 ±0.0029	0.0203 ±0.0009
	MEG-GPT-style EEG	4.2547 ±1.2573	0.2817 ±0.0219	0.8551 ±0.1328	0.3429 ±0.3976	0.3720 ±0.3274	0.4985 ±0.0009	−3.9010 ±0.0013	0.0415 ±0.0144
	JET	0.9659 ±0.0028	0.0780 ±0.0027	0.8301 ±0.0074	0.7107 ±0.0127	0.5372 ±0.0105	0.4888 ±0.0011	−3.8197 ±0.0092	0.0228 ±0.0037
	DiT-FM	0.9635 ±0.0071	0.0804 ±0.0054	0.5123 ±0.0799	0.5744 ±0.0074	0.5611 ±0.0079	0.4590 ±0.0060	−3.8418 ±0.0114	0.0000 ±0.0000
	GEMIA	0.9152 ±0.0103	0.0724 ±0.0034	0.5536 ±0.0312	0.7059 ±0.0036	0.5652 ±0.0083	0.4832 ±0.0022	−3.7760 ±0.0216	0.0411 ±0.0024
TUEV v2.0.1	Conditional WGAN-GP	2.5348 ±1.7405	0.6208 ±0.0504	1.4736 ±0.0653	0.0374 ±0.0078	0.1669 ±0.0003	0.4993 ±0.0002	−4.4058 ±0.0196	0.0613 ±0.0103
	EEG-GAN	20.8657 ±33.4519	0.6902 ±0.0148	1.5871 ±0.0224	0.0851 ±0.0903	0.1947 ±0.0485	0.4985 ±0.0019	−4.3355 ±0.1343	0.0242 ±0.0102
	1D U-Net DDPM	0.9837 ±0.0008	0.1154 ±0.0005	0.9303 ±0.0097	0.0328 ±0.0000	0.1667 ±0.0000	0.4984 ±0.0000	−4.4304 ±0.0003	0.0391 ±0.0015
	MEG-GPT-style EEG	6.1672 ±2.6521	0.2520 ±0.0246	0.6755 ±0.0989	0.3005 ±0.2461	0.3254 ±0.2651	0.4984 $\pm 1.06\times10^{-5}$	−4.4558 ±0.0048	0.0895 ±0.0207
	JET	1.0236 ±0.0345	0.1298 ±0.0008	0.9289 ±0.0093	0.2774 ±0.0145	0.2669 ±0.0124	0.4771 ±0.0026	−4.4330 ±0.0010	0.0016 ±0.0015
	DiT-FM	1.0978 ±0.0931	0.2170 ±0.0004	1.0161 ±0.0057	0.1985 ±0.0464	0.2144 ±0.0706	0.4913 ±0.0021	−4.4008 ±0.0004	0.0000 ±0.0000
	GEMIA	2.2711 ±1.0413	0.1893 ±0.0304	0.6466 ±0.1277	0.3269 ±0.0383	0.3248 ±0.0477	0.4652 ±0.0034	−4.4358 ±0.0025	0.0001 ±0.0002

Table B5: Solver-step sensitivity for Flow Matching models. Entries report TS-FID with $B_f = 256$ as mean $\pm$ sample standard deviation over three independent generator training runs with seeds $\{0, 2025, 2026\}$; lower is better. K denotes Heun integration steps, corresponding to $2K - 1$ network evaluations in our implementation.

Corpus	Generator	$K = 5$	$K = 10$	$K = 20$	$K = 50$
TUSZ v1.5.2	JET	578.825 ± 26.898	539.641 ± 22.554	503.782 ± 21.513	472.236 ± 18.919
	DiT-FM	464.223 ± 36.952	497.335 ± 48.449	525.513 ± 57.751	519.203 ± 59.841
	GEMIA	186.197 ± 45.026	179.935 ± 38.261	177.605 ± 32.380	174.139 ± 30.792
TUSZ v2.0.6	JET	484.185 ± 16.201	442.809 ± 13.814	413.521 ± 11.307	380.850 ± 12.418
	DiT-FM	492.153 ± 45.311	508.283 ± 25.934	530.398 ± 14.992	511.537 ± 11.586
	GEMIA	179.546 ± 12.636	173.132 ± 14.513	172.534 ± 14.983	169.460 ± 14.407
TUAB v3.0.1	JET	366.008 ± 5.065	378.153 ± 3.279	384.767 ± 2.231	377.518 ± 2.614
	DiT-FM	447.499 ± 31.262	455.350 ± 23.556	476.498 ± 17.488	424.968 ± 19.036
	GEMIA	291.654 ± 4.955	288.206 ± 4.757	284.633 ± 4.757	278.993 ± 4.021
TUEV v2.0.1	JET	340.427 ± 4.816	344.101 ± 4.220	348.754 ± 3.814	332.134 ± 2.498
	DiT-FM	1083.329 ± 43.140	1122.485 ± 39.354	1163.070 ± 33.317	1053.420 ± 48.906
	GEMIA	321.423 ± 7.621	321.656 ± 5.191	320.725 ± 4.689	316.612 ± 4.501

C METRIC DEFINITIONS AND EVALUATION SCOPE

This section expands the metrics in Tables B2, B3, and B4 without changing their values, sample pools, or evaluation rules. Signal fidelity, predictive utility, condition-label agreement, and membership-inference diagnostics describe different properties. Neither a small feature distance nor a high teacher agreement score is, by itself, a statement about clinical validity or privacy.

C.1 NOTATION, AGGREGATION, AND COORDINATE SYSTEMS

Let $\mathcal{R}$ and $\mathcal{G}$ denote the real and generated windows used by one evaluation unit, and let $\mathcal{R}_y, \mathcal{G}_y$ denote their subsets with condition y. A window has $C = 16$ channels and sampling rate $f_s = 200$ Hz, with $T = 1000$ for the two TUSZ protocols and TUEV, and $T = 2000$ for TUAB. Let K_{cls} be the number of condition classes; this differs from the Heun step count K. Superscript s identifies a generator run, not an individual EEG window.

For three seed-level scores $m^{(0)}, m^{(2025)}, m^{(2026)}$, the reported summary is

$$\bar{m} = \frac{1}{3} \sum_{s \in \{0,2025,2026\}} m^{(s)}, \qquad s_m = \sqrt{\frac{1}{2} \sum_{s} (m^{(s)} - \bar{m})^2}. \tag{C.1}$$

Thus s_m describes variation between the three runs, not the standard error of the mean, a confidence interval, or variation across individual patients. Seeds are not combined into one generated pool before evaluating a nonlinear metric. In particular, averaging Fréchet distances, AUROCs, or attack advantages differs from evaluating those metrics on pooled samples.

When a metric is class-macro averaged, its aggregation is $K_{\text{cls}}^{-1} \sum_y m(\mathcal{G}_y, \mathcal{R}_y)$, rather than a class-count-weighted mean. This is also different from calculating $m(\mathcal{G}, \mathcal{R})$ on the full mixture. The current Table B2 establishes the full-pool versus audit subset distinction, but does not establish one shared class reducer for every column. The column-specific frozen evaluation rule must be retained.

Numerical coordinates. All signal operators must receive the coordinates used in their recorded protocol. If a waveform is multiplied by b, its Fourier coefficients are multiplied by b and its power by b^2. Consequently, changing coordinates before taking logarithms, computing distances, or standardizing features is not generally equivalent to changing a reported score afterward. The factor 100 used in the completed BP/MIA rescore is specific to that recorded coordinate conversion; it must not be applied retrospectively to the other columns of Tables B2 and B4.

Specification boundary. The current manuscript and supplied Table B4 definition file are the primary specification. Earlier evaluator files available with the project clarify some mathematical operations, but their identity with every final run is not established by a table of rounded scores. Where the current materials omit an operator or reduction setting, the descriptions below state that boundary explicitly instead of assigning an unrecorded default. This is a documentation issue, not a request to rerun any experiment.

C.2 SIGNAL-FIDELITY METRICS IN TABLE B2

Evaluation objects and direction. All seven metric columns of Table B2 are lower-is-better comparisons of generated and held-out real EEG. TS-FID and RFD use the complete strict-test pools. The other five metrics use the fixed, class-stratified signal-audit subset. These are comparisons between signal populations or their summaries, not a claim that each generated window is the reconstruction of a particular real test waveform. Stored reference windows remain evaluation references; the generator is not given them at inference.

A shared Fréchet functional. For feature vectors $\boldsymbol{\phi}(\boldsymbol{X}) \in \mathbb{R}^d$ extracted from a sample set $\mathcal{A}$ with $n_{\mathcal{A}} \geq 2$, define

$$\boldsymbol{\mu}_{\mathcal{A}} = \frac{1}{n_{\mathcal{A}}} \sum_{\boldsymbol{X} \in \mathcal{A}} \boldsymbol{\phi}(\boldsymbol{X}), \qquad \boldsymbol{\Sigma}_{\mathcal{A}} = \frac{1}{n_{\mathcal{A}} - 1} \sum_{\boldsymbol{X} \in \mathcal{A}} (\boldsymbol{\phi}(\boldsymbol{X}) - \boldsymbol{\mu}_{\mathcal{A}})(\boldsymbol{\phi}(\boldsymbol{X}) - \boldsymbol{\mu}_{\mathcal{A}})^{\top}, \tag{C.2}$$

$$\mathcal{F}_{\boldsymbol{\phi}}(\mathcal{G}, \mathcal{R}) = \|\boldsymbol{\mu}_{\mathcal{G}} - \boldsymbol{\mu}_{\mathcal{R}}\|_2^2 + \operatorname{tr}\Big(\boldsymbol{\Sigma}_{\mathcal{G}} + \boldsymbol{\Sigma}_{\mathcal{R}} - 2(\boldsymbol{\Sigma}_{\mathcal{R}}^{1/2} \boldsymbol{\Sigma}_{\mathcal{G}} \boldsymbol{\Sigma}_{\mathcal{R}}^{1/2})^{1/2}\Big). \tag{C.3}$$

The symmetric positive-semidefinite square-root form states the ideal real-arithmetic functional. A numerical implementation may use $\mathrm{tr}(\boldsymbol{\Sigma}_{\mathcal{G}}\boldsymbol{\Sigma}_{\mathcal{R}})^{1/2}$ together with its recorded stabilization rule. Numerical jitter, feature scaling, and any normalization by d are implementation choices and cannot be inferred from the metric name. TS-FID, BP-FD, and RFD use different feature maps; they are not interchangeable. The functional compares first and second feature moments, not every aspect of the underlying EEG distribution (Heusel et al., 2017).

TS-FID: standardized log-FFT distribution distance. The manuscript specifies a log-FFT feature map with per-window feature standardization and a frequency-bin cap $B_f = 256$. Let $\boldsymbol{F}_{\boldsymbol{X}} = |\,\mathrm{RFFT}(\boldsymbol{X})|$ and write $\mathcal{P}$ for the frozen frequency selection and pooling operations. The feature construction is

$$\boldsymbol{q}(\boldsymbol{X}) = \log(1 + \mathrm{vec}(\mathcal{P}(\boldsymbol{F}_{\boldsymbol{X}}))), \quad \boldsymbol{\phi}_{\mathrm{TS}}(\boldsymbol{X}) = \frac{\boldsymbol{q}(\boldsymbol{X}) - \overline{q}(\boldsymbol{X})\mathbf{1}}{\max\{s(\boldsymbol{q}(\boldsymbol{X})), \epsilon_{\mathrm{TS}}\}}, \quad \mathrm{TSFID} = \mathcal{F}_{\boldsymbol{\phi}_{\mathrm{TS}}}(\mathcal{G}, \mathcal{R}). \tag{C.4}$$

The bar and s in this equation operate within a window's feature vector; they are not statistics fitted over test patients. The main reported score is the full-pool distribution comparison, rather than an arithmetic mean of per-window distances. Standardization reduces sensitivity to an overall feature offset and scale, but does not imply exact waveform-amplitude invariance, especially with the $\log(1 + \cdot)$ transform and finite floors.

The supplied earlier `metrics.py` makes $\mathcal{P}$ explicit: it retains the first $\lfloor F/2 \rfloor$ of the $F = \lfloor T/2 \rfloor + 1$ RFFT bins, average-pools frequency to $\min(B_f, \lfloor F/2 \rfloor)$ bins, average-pools channels to four spatial bins, then applies $\log(1 + \cdot)$ and within-vector sample-standard-deviation normalization with floor 10^{-6}. Under *that implementation*, $d = 1000$ for $T = 1000$ and $d = 1024$ for $T = 2000$; 256 is a cap, not the final feature dimension. The current table confirms $B_f = 256$, but does not provide the final extractor hash. These extra pooling settings describe the supplied implementation and must not be represented as a newly verified hash match to every final run. Absolute TS-FID values depend on feature dimension and sample budget, so the principal comparison is between methods evaluated on the same protocol.

Log-PSD error: spectral shape and power mismatch. A power spectral density estimate assigns power to frequency bins. A standard Welch form for channel c is

$$P_{\boldsymbol{X}}(c, f_k) = \frac{d_k}{J f_s \sum_{u=0}^{L-1} w_u^2} \sum_{j=1}^{J} \left| \sum_{u=0}^{L-1} w_u x_{c,j,u}^{\mathrm{detrend}} e^{-2\pi i k u / L} \right|^2, \tag{C.5}$$

where $\boldsymbol{w}$ is the taper, J the number of segments, and d_k the one-sided power multiplier (one at self-conjugate bins and two at interior bins). For a frozen population log-spectrum operator Ψ_{PSD}, a log-PSD mean-absolute-error comparison has the form

$$E_{\mathrm{logPSD}}(\mathcal{G}_y, \mathcal{R}_y) = \mathrm{mean}_{c,k} \left| \Psi_{\mathrm{PSD}}(\mathcal{G}_y)_{c,k} - \Psi_{\mathrm{PSD}}(\mathcal{R}_y)_{c,k} \right|. \tag{C.6}$$

Writing $\boldsymbol{P}_{\boldsymbol{X}} = [P_{\boldsymbol{X}}(c, f_k)]_{c,k}$, for example, $\Psi(\mathcal{A}) = \log(\mathrm{mean}_{\boldsymbol{X} \in \mathcal{A}} \boldsymbol{P}_{\boldsymbol{X}} + \epsilon)$ and $\Psi(\mathcal{A}) = \mathrm{mean}_{\boldsymbol{X} \in \mathcal{A}} \log(\boldsymbol{P}_{\boldsymbol{X}} + \epsilon)$ are distinct operators. The final audit's segment settings, frequency range, logarithm, stabilizer, and order of averaging and taking logarithms are not specified in the current table. Equation (C.6) therefore identifies the mathematical comparison without choosing between those implementation variants. A smaller error means closer spectra under the recorded operator; it does not prove that high-frequency signal content is physiological rather than artifact, or that all transient events are retained.

ACF error: normalized temporal dependence. Let $x_c^{\circ}(t) = x_c(t) - T^{-1} \sum_u x_c(u)$. An unadjusted normalized finite-window autocorrelation is

$$A_{\boldsymbol{X}}(c, \ell) = \frac{\sum_{t=0}^{T-\ell-1} x_c^{\circ}(t) x_c^{\circ}(t + \ell)}{\sum_{t=0}^{T-1} (x_c^{\circ}(t))^2 + \epsilon_A}. \tag{C.7}$$

If the audit averages these curves before comparing them, its ACF error is $\mathrm{mean}_{c,\ell \in \mathcal{L}_A} |\overline{A}_{\mathcal{G}}(c, \ell) - \overline{A}_{\mathcal{R}}(c, \ell)|$. Lag selection $\mathcal{L}_A$, an optional $T/(T - \ell)$ finite-length adjustment, and the order of averaging are part of the frozen implementation. The current materials identify a normalized-ACF error but do not fix these options; Equation (C.7) is the unadjusted reference definition, not a retrospective change to a stored score. ACF captures lag-dependent second-order temporal dependence. Signals with different higher-order or phase organization can still have similar autocorrelation curves.

Amp-W1: marginal amplitude distance. For two one-dimensional distributions with cumulative distribution functions $F_{\mathcal{G}}, F_{\mathcal{R}}$ and generalized quantiles $Q_{\mathcal{G}}, Q_{\mathcal{R}}$, their Wasserstein-1 distance is

$$W_1(\mathcal{G}, \mathcal{R}) = \int_{\mathbb{R}} |F_{\mathcal{G}}(u) - F_{\mathcal{R}}(u)|\, du = \int_0^1 |Q_{\mathcal{G}}(p) - Q_{\mathcal{R}}(p)|\, dp. \tag{C.8}$$

For equally sized empirical samples this is the mean absolute difference of the sorted samples. A fixed quantile-grid approximation replaces the integral by its recorded quadrature. Amp-W1 applies this comparison to waveform amplitudes, using the channel and class aggregation of the frozen audit. It is not the distance between window RMS values. A per-channel mean of W_1 and a single W_1 after pooling all channels need not agree; the current Table B2 does not specify which implementation produced its entries. The metric assesses amplitude marginals and is insensitive to permutations that preserve those marginals, so it cannot replace ACF or CSD.

BP-FD: bandpower-distribution distance. A bandpower feature is obtained by integrating a PSD over a fixed band, applying the specified compression, and concatenating channel-band entries:

$$b_{c,j}(\boldsymbol{X}) = \int_{f \in \mathcal{B}_j} P_{\boldsymbol{X}}(c, f)\, df, \quad \boldsymbol{\phi}_{\mathrm{BP}}(\boldsymbol{X}) = \mathrm{vec}_{c,j} \log(1 + b_{c,j}(\boldsymbol{X})), \quad \mathrm{BPFD} = \mathcal{F}_{\boldsymbol{\phi}_{\mathrm{BP}}}(\mathcal{G}, \mathcal{R}). \tag{C.9}$$

The project records describe a class-conditional log-Welch-bandpower Fréchet family. Its final band boundaries, standardization, class reducer, numerical square-root handling, and possible dimension normalization must remain those of the final evaluator. They are not established merely by the shared prefix "BP" in BP-FD and BP precision. In particular, the fully specified bands and factor-100 rescore in Section C.4 do not authorize modifying Table B2. BP-FD compares the feature distribution; BP precision and recall compare condition labels under a classifier.

Env-W1: temporal-envelope amplitude distance. Let $\mathcal{E}$ denote the fixed envelope extractor. For a Hilbert envelope, $\mathcal{E}(\boldsymbol{x}_c)(t) = |x_c(t) + \mathrm{i}\mathcal{H}(\boldsymbol{x}_c)(t)|$; for a local RMS envelope, $\mathcal{E}(\boldsymbol{x}_c)(t) = \sqrt{\mathrm{mean}_{u \in \mathcal{W}_t}\, x_c(u)^2}$. These are different signal representations. Env-W1 applies Equation (C.8) to the envelope values, retaining the frozen audit's operator, boundary handling, pooling, and quantile settings. The current files do not identify which envelope operator produced Table B2, so neither example is asserted to be its implementation. A smaller Env-W1 means closer envelope-value distributions; it need not imply that envelope events occur at the same temporal locations.

RFD: independent representation-space distance. Let $\boldsymbol{h}_{\mathrm{CB}}(\boldsymbol{X})$ be the pooled output of the frozen pretrained CBraMod encoder. Representation Fréchet distance evaluates $\mathcal{F}_{\boldsymbol{h}_{\mathrm{CB}}}(\mathcal{G}, \mathcal{R})$ with the recorded pooling and scaling. Its encoder features are not part of the generator's training objective. RFD therefore provides a different representation-space diagnostic from TS-FID, although it is not independent of all dataset or pretraining biases. The current table establishes full-pool use. Whether the exported field is the pooled-mixture distance or a macro mean of full per-class distances, as well as any normalization by representation dimension, must be read from the final evaluator rather than inferred from its numeric range.

Required evaluator metadata for exact reproduction. The remaining documentation fields are: final evaluator/configuration hashes; class aggregation per column; Welch taper, overlap, detrending and frequency range; ACF lag grid and reduction; channel and quantile handling for Amp-W1; envelope extractor and boundary handling; BP feature normalization and Fréchet implementation; and RFD pooling/class/dimension reduction. These fields can be recovered from the existing evaluator and configuration; no new generated samples or metric recomputation is required to document them.

C.3 Downstream-utility metrics in Table B3

Training, checkpoint selection, and test scoring. Table B3 compares Real-only with seven real-plus-synthetic augmentation conditions. The real subset contains 10% of training patients, selected with patient seed 4407. Each augmentation condition uses the same real subset and a per-class synthetic-to-real window ratio of 0.5. A linear classifier is fitted on a frozen CBraMod encoder for 50 epochs with batch size 256. Real-validation AUROC selects the checkpoint. Final metrics use only real held-out test windows. The three paired generator/classifier seeds are $\{0, 2025, 2026\}$;

these are not three new patient-subset draws. Real-only has three classifier runs but no generator seed. Consequently, the 84 augmented fits and 12 Real-only fits form 96 classifier results, without multiplying the number of independent generator models.

For n real test windows, write y_i for the true class, $p_{i,c}$ for the classifier's probability of class c, and $\widehat{y}_i = \arg\max_c p_{i,c}$ for its predicted label. Define the confusion matrix $\boldsymbol{M}$ by $m_{ab} = [\boldsymbol{M}]_{ab} = \sum_{i=1}^{n} \mathbf{1}\{y_i = a, \widehat{y}_i = b\}$. All metrics below are calculated per run before the mean and sample SD in Equation (C.1) are formed. Patient-disjoint splits prevent a patient from occurring in both classifier training and final test, but the reported scores aggregate test *windows*, not patient-level averages.

Balanced Accuracy. With $N_c = \sum_b m_{cb} > 0$, Balanced Accuracy is macro recall:

$$\mathrm{BA} = \frac{1}{K_{\mathrm{cls}}} \sum_{c=0}^{K_{\mathrm{cls}}-1} \frac{m_{cc}}{N_c}. \tag{C.10}$$

For binary classification it equals $(\mathrm{sensitivity} + \mathrm{specificity})/2$. For TUEV it gives each of the six true classes equal weight, irrespective of their different test-window counts. BA lies in $[0, 1]$ and depends on the final decision rule. Always predicting a single class gives BA $1/K_{\mathrm{cls}}$ when every true class is present; this is a useful diagnostic reference, not a guarantee about a trained model.

AUPRC: average precision from ranked probability scores. For a binary target $z_i \in \{0, 1\}$ and score p_i, group identical scores and sweep thresholds from highest to lowest. Let P_j and R_j be precision and recall after including score group j, with $R_0 = 0$. The average- precision convention used by the supplied utility evaluator is

$$\mathrm{AP} = \sum_j (R_j - R_{j-1}) P_j. \tag{C.11}$$

This is the stepwise average-precision estimator, not trapezoidal interpolation between precision–recall points. For binary protocols the positive class is class 1 (seizure or abnormal), and the score is $p_{i,1}$. For TUEV, form six one-versus-rest targets $z_i^{(c)} = \mathbf{1}\{y_i = c\}$ and report

$$\mathrm{AUPRC}_{\mathrm{macro}} = K_{\mathrm{cls}}^{-1} \sum_c \mathrm{AP}(\boldsymbol{z}^{(c)}, \boldsymbol{p}_{\cdot,c}). \tag{C.12}$$

AP emphasizes ranking positives above negatives, and its interpretation changes with positive-class prevalence. It is therefore inappropriate to compare raw AUPRC values across different protocols as though they measured one common difficulty scale. TUEV macro AP is not a prevalence-weighted or micro-averaged score.

AUROC: pairwise ranking discrimination. For binary scores, let $\mathcal{I}_+ = \{i : y_i = 1\}$ and $\mathcal{I}_- = \{i : y_i = 0\}$, with both sets nonempty. The empirical ROC AUC is

$$\mathrm{AUC} = \frac{1}{|\mathcal{I}_+||\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} \left[\mathbf{1}\{p_i > p_j\} + \tfrac{1}{2}\mathbf{1}\{p_i = p_j\} \right]. \tag{C.13}$$

It is the fraction of correctly ordered positive–negative pairs, counting a tie as one half. The binary protocols use $p_{i,1}$. TUEV uses the unweighted mean of six one-versus-rest AUROCs, not one-versus-one AUC. AUROC lies in $[0, 1]$; 0.5 is the non-discriminative population reference for an uninformative continuous score. It assesses ranking, not probability calibration or the quality of one fixed decision threshold. Thus AUROC and BA can change in different directions without a mathematical inconsistency.

Effect relative to Real-only and uncertainty. For a fixed corpus, metric, and seed s, define $d_s = m_{\mathrm{aug}}^{(s)} - m_{\mathrm{real}}^{(s)}$. A figure built from the published means displays $100(\bar{m}_{\mathrm{aug}} - \bar{m}_{\mathrm{real}})$ percentage points. With the same three pairs, this equals $100\bar{d}$. However,

$$s_d^2 = s_{\mathrm{aug}}^2 + s_{\mathrm{real}}^2 - 2\widehat{\mathrm{Cov}}(m_{\mathrm{aug}}, m_{\mathrm{real}}) \tag{C.14}$$

requires the paired seed-level values. The two marginal SDs alone do not identify s_d. The current difference panels therefore do not invent such an error bar. Three-run SDs are not significance tests. The fixed low-patient subset also means that they do not measure variation over alternative patient selections.

Why these are utility metrics. Every Table B3 score belongs to a classifier trained under one data condition and evaluated on real test data. It is not a classifier score obtained by presenting generated EEG directly to the final test classifier. Better condition-label agreement under the bandpower teacher in Table B4 does not mathematically guarantee improvement in these real-test utility scores.

C.4 Signal-structure and memorization diagnostics in Table B4

Inputs, sampling, and repeated runs. The diagnostics use the existing strict-test generated pools from seven methods on four corpora, with generator seeds $\{0, 2025, 2026\}$. These 84 units are not enlarged by repeated execution or by duplicate result archives. Diagnostic seed 3407 is fixed. The class-stratified sampler retains a pool in full when its size is at most 5000. Otherwise, it selects up to $\lfloor 5000/K_{\mathrm{cls}} \rfloor$ windows from each observed condition class, without redistributing unfilled quotas. Here K_{cls} is the number of conditions. Consequently, the completed three-column rescores use 5000 generated windows per unit for both TUSZ protocols and TUAB, and 4228 for TUEV. Each corpus uses one fixed real reference container and the same diagnostic sampling rule across methods and generator seeds.

For a corpus–method pair, the reported mean is the average of its three seed-level values and the sample standard deviation has denominator $3 - 1$. Predictions from the seeds are not pooled to compute a different score. All numerical primary entries have three values under the stated reporting conventions. The original strict BP precision remains undefined in 26 units and cannot form a complete three-seed summary in 12 rows; these records are retained separately rather than discarded.

CSD error: frequency-dependent cross-channel structure. For a population $\mathcal{A}$, using the RFFT coefficient vector $\widehat{\boldsymbol{x}}(f)$ of each window $\boldsymbol{X}$, the second-moment matrix at frequency f is

$$\boldsymbol{S}_{\mathcal{A}}(f) = |\mathcal{A}|^{-1} \sum_{\boldsymbol{X} \in \mathcal{A}} \widehat{\boldsymbol{x}}(f)\widehat{\boldsymbol{x}}(f)^{H}. \tag{C.15}$$

Let $\mathbf{S}_{\mathcal{A}}$ stack the matrices $\boldsymbol{S}_{\mathcal{A}}(f)$ over frequency bins. A relative discrepancy for matched generated and reference subsets is $\|\mathbf{S}_{\mathcal{G}} - \mathbf{S}_{\mathcal{R}}\|_F / (\|\mathbf{S}_{\mathcal{R}}\|_F + \epsilon_S)$, with the Frobenius norm over the frequency and channel axes. Diagonal entries describe channel power and complex off-diagonal entries describe cross-channel relations. This second-moment tensor is not a centered covariance across windows unless its frequency-wise mean is zero. CSD error is reported only in Table B4; its explicit class-macro version is stated below.

Spectral and cross-channel diagnostics. These columns retain the preceding strict-pool evaluation. For a set of windows $\mathcal{A}$, let $\widehat{\boldsymbol{x}}(f)$ denote the coefficient vector at frequency f of the unnormalized temporal RFFT of $\boldsymbol{X} \in \mathcal{A}$, and define

$$\boldsymbol{S}_{\mathcal{A}}(f) = \frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{X} \in \mathcal{A}} \widehat{\boldsymbol{x}}(f)\widehat{\boldsymbol{x}}(f)^{H}. \tag{C.16}$$

With the frozen evaluator's numerical floors understood, the magnitude-squared coherence matrix $\boldsymbol{H}_{\mathcal{A}}(f)$ has entries

$$h_{\mathcal{A},ij}(f) = [\boldsymbol{H}_{\mathcal{A}}(f)]_{ij} = \frac{|s_{\mathcal{A},ij}(f)|^2}{s_{\mathcal{A},ii}(f)\, s_{\mathcal{A},jj}(f)}. \tag{C.17}$$

Let $q_{\mathcal{A},j}(f)$ be the normalized nonnegative eigenvalues of the Hermitian cross-spectral matrix, with $s_{\mathcal{A},ij}(f) = [\boldsymbol{S}_{\mathcal{A}}(f)]_{ij}$. Its effective rank is

$$r_{\mathcal{A}}(f) = \exp\left[-\sum_{j} q_{\mathcal{A},j}(f) \log q_{\mathcal{A},j}(f) \right]. \tag{C.18}$$

For generated and real subsets $\mathcal{G}_c$ and $\mathcal{R}_c$ in condition c, the three errors are

$$E_{\mathrm{CSD}} = \frac{1}{K_{\mathrm{cls}}} \sum_c \frac{\|\mathbf{S}_{\mathcal{G}_c} - \mathbf{S}_{\mathcal{R}_c}\|_F}{\|\mathbf{S}_{\mathcal{R}_c}\|_F + \epsilon}, \tag{C.19}$$

$$E_{\mathrm{Coh}} = \frac{1}{K_{\mathrm{cls}}} \sum_c \mathrm{mean}_{f,i,j} \, |h_{\mathcal{G}_c,ij}(f) - h_{\mathcal{R}_c,ij}(f)|, \tag{C.20}$$

$$E_{\mathrm{rank}} = \frac{1}{K_{\mathrm{cls}}} \sum_c \mathrm{mean}_f \, |\log(r_{\mathcal{G}_c}(f) + \epsilon) - \log(r_{\mathcal{R}_c}(f) + \epsilon)|. \tag{C.21}$$

The CSD norm includes frequency and channel axes. The coherence average includes the full matrix, including its diagonal. The released metric core specifies the exact floors and eigenvalue clipping. The field named `coherence` is an error, not the absolute coherence of a generated signal.

Bandpower features and scale alignment. Each window has 16 channels sampled at 200 Hz. Welch estimation uses segment length $\min(256, T)$. Trapezoidal integration over the bands $[0.5, 4)$, $[4, 8)$, $[8, 13)$, $[13, 30)$, and $[30, 45)$ Hz is followed by $\log(1 + \cdot)$. Concatenating the channel-band features gives an 80-dimensional feature vector. The three-column rescore converts stored generated waveforms to the numerical scale of the raw PKL diagnostic windows by multiplying by a fixed factor of 100 *before* Welch estimation. Raw PKL inputs are unchanged. The conversion follows the recorded contract that generated and reference waveforms share the stored training coordinates. Matched reference-to-PKL windows verify this conversion; this check is not an independent verification of every historical generator export implementation.

Bandpower teacher and precision reporting. The teacher is the saved standardized, class-balanced logistic-regression classifier fitted on 8000 fixed real training windows per corpus. It is not the downstream CBraMod classifier. The rescore reuses its parameters without fitting again. The confusion matrix $\boldsymbol{M}$ compares the generated condition labels (rows) with the teacher's predicted labels (columns). Writing $m_{ij} = [\boldsymbol{M}]_{ij}$ and $TP_c = m_{cc}$, the original per-class precision is $TP_c / \sum_i m_{ic}$ when the denominator is nonzero. Original strict macro precision is undefined if any condition receives no predictions.

For a fully numerical report, we additionally define

$$P^{(0)} = \frac{1}{K_{\mathrm{cls}}} \sum_c \begin{cases} \dfrac{m_{cc}}{\sum_i m_{ic}}, & \sum_i m_{ic} > 0, \\ 0, & \sum_i m_{ic} = 0. \end{cases} \tag{C.22}$$

This `zero_division=0` reporting convention was added after inspecting undefined precision values and is applied to every method, corpus, and seed. It sets only the undefined *per-class* precision to zero; it does not replace an entire missing experiment or an entire macro score by zero. The original strict-NA scores, missing-class counts, and confusion matrices are retained. No predictions are changed.

Recall retains the original definition,

$$R = \frac{1}{K_{\mathrm{cls}}} \sum_c \frac{m_{cc}}{\sum_j m_{cj}}, \tag{C.23}$$

with all target conditions represented in the evaluated pools. BP precision and recall measure condition-label agreement under this teacher, not precision and recall of the full generative distribution.

Retained domain-gap evaluation. Domain gap retains its preceding values rather than being recomputed in the three-column rescore. It is $|\mathrm{AUC}_{\mathrm{domain}} - 0.5|$ for the real-versus-generated logistic classifier using the recorded bandpower feature pipeline. A $70/30$ fit/score split is stratified by domain with seed 3407; the domain scaler and classifier are fitted on the diagnostic fit subset. These stored-coordinate results must not be described as newly recomputed raw-coordinate diagnostics. Values near 0.5 indicate strong separation by this classifier, not strong similarity.

Scale-aligned membership-inference diagnostic. Let ϕ_{atk} be the bandpower feature transform standardized using only the fixed attack-calibration subset. A fixed permutation with seed $3407+7919$ splits the 8000 diagnostic training windows into 5600 calibration and 2400 member windows. The 8000 real test windows, selected with seed 3408, are nonmembers. Calibration and member indices are disjoint; no additional patient-disjointness between these two subsets is asserted. The generated bank uses the scale-aligned features.

For $d(\boldsymbol{u}, \mathcal{A}) = \min_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{u} - \boldsymbol{a}\|_2$, the fixed-direction query score is

$$s(\boldsymbol{X}) = -d(\boldsymbol{\phi}_{\mathrm{atk}}(\boldsymbol{X}), \boldsymbol{\phi}_{\mathrm{atk}}(\mathcal{G}))\,. \tag{C.24}$$

Higher scores are fixed in advance to indicate membership. The reported quantity is

$$a_{\mathrm{MIA}} = \max\{0, 2\,\mathrm{AUC}_{\mathrm{raw}} - 1\}. \tag{C.25}$$

The transform is applied to each generator seed before averaging. Raw AUCs are retained, and the score direction is not reversed after observing the final scoring data. A zero can correspond to an AUC below 0.5 and does not provide a formal privacy guarantee.

Legacy nearest-neighbor gap. The earlier signed NN gap is retained separately as a historical result. It compares generated-to-training and real-test-to-training nearest-neighbor distances under the earlier real-training-standardized bandpower pipeline. That evaluation did not receive the raw-coordinate scale alignment used for the BP and MIA rescore. Its signed values are not corrected by this assembly and are not used to support conclusions about memorization or privacy. Table B4 marks this column with †; a main-text layout excluding it is supplied without discarding its data.

Implementation and provenance. The four unchanged structural/domain columns and the legacy NN column come from the original strict-pool result snapshots. The three updated columns and the additional precision report are derived from their respective versioned result snapshots. Completion provides the adopted three-column results. Resume is retained as a cross-check, not as additional seeds. Actual executor, input-pool, selected-window, teacher, and feature hashes are retained per unit. Low-memory I/O and the fixed-manifest symlink adapter are implementation records, not additional generator methods or training runs.

Interpretation and degenerate cases. CSD error, coherence error, effective-rank gap, and domain gap are discrepancies; smaller values describe closer statistics under their respective operators. BP precision and BP recall are agreement scores and are larger-is-better. MIA advantage is an attack statistic; it is not another fidelity coordinate. The signed historical NN gap has no blanket "smaller is better" ordering.

If every generated condition class is represented but the teacher predicts only one class j, then its macro recall is exactly $1/K_{\mathrm{cls}}$: recall is one for class j and zero for every other class. Its zero-convention macro precision is π_j/K_{cls}, where π_j is the fraction of evaluated generated windows requested under class j. This explains why a binary recall near 0.5, or a six-class recall near $1/6$, can accompany poor condition-label agreement. It does not, by itself, identify whether the cause lies in generated signals, class conditioning, or the teacher. The retained confusion matrices distinguish these cases more directly than a scalar score.

The fixed MIA transform satisfies $0 \le a_{\mathrm{MIA}} \le 1$ and returns exactly zero whenever $\mathrm{AUC}_{\mathrm{raw}} \le 0.5$. Thus a reported zero is compatible with a finite, nontrivial raw AUC below 0.5. The convention does not allow reversing an attack after inspecting the final scoring set. It also does not imply that every possible attack has zero advantage. For example, the two raw AUCs 0.45 and 0.55 map to 0 and 0.10; their average advantage is 0.05, whereas transforming their mean AUC gives zero. This is why the per-seed transformation is performed before aggregation.

An ideal positive-power coherence entry lies in $[0, 1]$, and a nonzero positive-semidefinite C-channel spectrum has effective rank in $[1, C]$. Numerical floors and a zero-trace spectrum require the recorded implementation convention. Coherence diagonals are ideally one, so including them in the mean reduces the relative contribution of off-diagonal discrepancies; the published full-matrix reduction is retained rather than silently replaced by an off-diagonal-only score.

D MATHEMATICAL PROPERTIES OF THE SOURCE AND OBJECTIVE

This section derives properties of the equations in Section 3. The statements concern the finite-dimensional source construction and the sample-level training objective. They do not prove superiority to Gaussian sources, global optimization convergence, exact recovery of the data population, privacy, or a universal reduction in integration steps. Those claims would require additional assumptions or empirical evidence.

D.1 NOTATION AND ASSUMPTIONS

Write $P = |\mathcal{P}| \geq 1$ and assume $n_p \geq 1$ for every training patient. All signal arrays are finite, with $C, T \geq 1$. Inner products on complex vectors use conjugation, and matrix adjoints are denoted by H. Positive reference amplitudes are required wherever logarithms appear. The time variable satisfies $0 < t < 1$ and the training cap satisfies $\tau > 0$. Source normalization is analyzed with the actual stabilizer $\epsilon > 0$; idealized $\epsilon = 0$ results are stated separately and require nonzero RMS. No eigenvalue floor or training weight is assumed to depend on a test label.

The waveform entering multi-scale increment alignment. The current Method describes $\widetilde{\boldsymbol{X}}$ as "channel-wise RMS," but $\Delta_\ell \widetilde{\boldsymbol{X}}$ requires a time-indexed waveform, not one scalar per channel. The increment propositions below hold for arbitrary waveform-valued arrays $\boldsymbol{A}, \boldsymbol{B}$ and therefore do not require silently resolving that ambiguity. When discussing channel-normalized signals, we explicitly use the convention

$$\mathcal{N}_\epsilon(\boldsymbol{X})_{c,t} = \frac{x_{c,t} - \mu_c(\boldsymbol{X})}{\sigma_c(\boldsymbol{X}) + \epsilon}, \qquad \sigma_c(\boldsymbol{X}) = \sqrt{T^{-1} \sum_t (x_{c,t} - \mu_c(\boldsymbol{X}))^2}. \tag{D.1}$$

This is an explicit convention for the statements that invoke it, not a verification that the current training implementation uses an additive ϵ rather than a clamped denominator. Either actual convention must be identified before claiming implementation-level correspondence.

Supplementary Lean 4 material. The mathematical arguments are presented in full below. Draft Lean 4 formalizations of selected statements, their explicit hypotheses, a statement-to-source correspondence, and reproduction instructions are provided in a separate supplementary archive. These sources are not claimed to be kernel-verified in the current version. No theorem about the learned neural network or the experimental scores is imported as an assumption.

D.2 THE LOGIT-NORMAL TIME SCHEDULE

Proposition D.1 (Interior times and the induced time density)**.** *For any finite $\eta \in \mathbb{R}$, $t = (1 + e^{-\eta})^{-1}$ belongs to $(0, 1)$. If $\eta \sim \mathcal{N}(\mu_t, \sigma_t^2)$ with $\sigma_t > 0$, then*

$$p_t(t) = \frac{1}{\sigma_t \sqrt{2\pi}\, t(1-t)} \exp\left[-\frac{(\log(t/(1-t)) - \mu_t)^2}{2\sigma_t^2} \right], \quad 0 < t < 1. \tag{D.2}$$

The median time is $\mathrm{sigmoid}(\mu_t)$. In particular, $\mu_t < 0$ places more than one half of the probability mass below $1/2$.

Proof. The exponential is positive, so the denominator is greater than one. The sigmoid is strictly increasing with inverse $\eta = \log(t/(1-t))$ and inverse derivative $1/[t(1-t)]$. Substitution into the normal density gives Equation (D.2). A normal variable with positive standard deviation has median μ_t, and a strictly increasing map preserves medians. For $\mu_t < 0$, $\Pr(t < 1/2) = \Pr(\eta < 0) > 1/2$. This does not make time sampling uniform, nor does it prove that the resulting finite-capacity optimization is equivalent to a uniformly weighted objective. □

D.3 PATIENT BALANCING AND SPECTRAL SOURCE CONSTRUCTION

Proposition D.2 (Centering and stabilized shape RMS)**.** *For one channel, let $\mu = T^{-1} \sum_t x_t$ and $x_t^\circ = x_t - \mu$. Then $\sum_t x_t^\circ = 0$. For the full centered window define $a = \sqrt{(CT)^{-1} \sum_{c,t} (x_{c,t}^\circ)^2}$ and*

$\bar{\boldsymbol{X}} = \boldsymbol{X}^\circ/(a+\epsilon)$, *with* $\epsilon > 0$. *Then*

$$\mathrm{RMS}(\bar{\boldsymbol{X}}) = \frac{a}{a+\epsilon} \in [0,1), \qquad 1 - \mathrm{RMS}(\bar{\boldsymbol{X}}) = \frac{\epsilon}{a+\epsilon}. \tag{D.3}$$

For $a > 0$ *and* $\epsilon = 0$*, the idealized normalized shape has RMS exactly one.*

Proof. The centered sum is $\sum_t x_t - T\mu = 0$. For any real scalar b,

$$\mathrm{RMS}(b\boldsymbol{U}) = \sqrt{b^2(CT)^{-1}\sum_{c,t} u_{c,t}^2} = |b|\,\mathrm{RMS}(\boldsymbol{U}). \tag{D.4}$$

Use $b = (a+\epsilon)^{-1} > 0$ and $\mathrm{RMS}(\boldsymbol{X}^\circ) = a$. Nonnegativity of a and positivity of ϵ give the bounds and subtraction gives the second identity. If $a = 0$, then $\boldsymbol{X}^\circ = \boldsymbol{0}$ and the stabilized shape is zero, not a unit-energy signal. □

Proposition D.3 (Equal patient mass)**.** *Assign every window of patient* p *weight* $w_{p,i} = 1/(Pn_p)$*. Then*

$$\sum_{i=1}^{n_p} w_{p,i} = \frac{1}{P}, \qquad \sum_{p=1}^{P}\sum_{i=1}^{n_p} w_{p,i} = 1. \tag{D.5}$$

For any statistic h*, the empirical expectation is* $\sum_p w_p \bar{h}_p = P^{-1}\sum_p n_p^{-1}\sum_i h(\boldsymbol{X}_{p,i})$*. Uniformly replicating all windows of any patient does not change that patient's mean or the patient-balanced expectation.*

Proof. Within a patient, summing the constant weight gives $n_p/(Pn_p) = 1/P$. Summing over P patients gives one. If every window of patient p is repeated $r \geq 1$ times, its new contribution is $(rn_p)^{-1}\sum_{i=1}^{n_p}\sum_{j=1}^{r} h(\boldsymbol{X}_{p,i}) = n_p^{-1}\sum_i h(\boldsymbol{X}_{p,i})$. Substitution in the outer average proves the invariance. Repeating only selected windows changes the within-patient empirical distribution and need not leave the statistic invariant. □

Proposition D.4 (Hermitian positive-semidefinite cross-spectrum)**.** *For any finite complex vectors* $\boldsymbol{z}_{p,i}(f) \in \mathbb{C}^C$*, the patient-balanced matrix* $\boldsymbol{S}(f) = \sum_{p,i} w_{p,i}\boldsymbol{z}_{p,i}(f)\boldsymbol{z}_{p,i}(f)^H$ *is Hermitian and positive semidefinite. Its diagonal entries equal weighted channel power,* $s_{cc}(f) = [\boldsymbol{S}(f)]_{cc} = \sum_{p,i} w_{p,i}|z_{p,i,c}(f)|^2$.

Proof. Each outer product is Hermitian and the weights are nonnegative real numbers. For any $\boldsymbol{v} \in \mathbb{C}^C$,

$$\boldsymbol{v}^H\boldsymbol{S}(f)\boldsymbol{v} = \sum_{p,i} w_{p,i}(\boldsymbol{v}^H\boldsymbol{z}_{p,i})(\boldsymbol{z}_{p,i}^H\boldsymbol{v}) = \sum_{p,i} w_{p,i}|\boldsymbol{z}_{p,i}^H\boldsymbol{v}|^2 \geq 0. \tag{D.6}$$

The formula for a diagonal entry follows by setting both channel indices to c. Equivalently, place columns $\sqrt{w_{p,i}}\boldsymbol{z}_{p,i}$ in a matrix $\boldsymbol{B}$; then $\boldsymbol{S} = \boldsymbol{B}\boldsymbol{B}^H$. This factorization also shows that strict positivity is not guaranteed: the vectors may span a proper subspace. Small negative numerical eigenvalues are not properties of the exact empirical matrix. □

Proposition D.5 (Eigenvalue flooring and colored coefficients)**.** *Let* $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^H$ *with* $\boldsymbol{U}$ *unitary, and let* $\widetilde{\lambda}_j = \max\{\lambda_j, \delta\}$ *for* $\delta > 0$*. Then* $\widetilde{\boldsymbol{S}} = \boldsymbol{U}\widetilde{\boldsymbol{\Lambda}}\boldsymbol{U}^H \succeq \delta\boldsymbol{I}$*. If* $\boldsymbol{L} = \boldsymbol{U}\widetilde{\boldsymbol{\Lambda}}^{1/2}$ *and a random complex vector* $\boldsymbol{\xi}$ *satisfies* $\mathbb{E}\boldsymbol{\xi} = \boldsymbol{0}$ *and* $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^H] = \boldsymbol{I}$*, then*

$$\mathbb{E}[\boldsymbol{L}\boldsymbol{\xi}] = \boldsymbol{0}, \qquad \mathbb{E}[(\boldsymbol{L}\boldsymbol{\xi})(\boldsymbol{L}\boldsymbol{\xi})^H] = \boldsymbol{L}\boldsymbol{L}^H = \widetilde{\boldsymbol{S}}. \tag{D.7}$$

For proper circular complex $\boldsymbol{\xi}$*, its pseudocovariance remains zero.*

Proof. For any $\boldsymbol{v}$, put $\boldsymbol{u} = \boldsymbol{U}^H\boldsymbol{v}$. Unitarity gives $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2$, so $\boldsymbol{v}^H\widetilde{\boldsymbol{S}}\boldsymbol{v} = \sum_j \widetilde{\lambda}_j|u_j|^2 \geq \delta\|\boldsymbol{v}\|_2^2$. Linearity of expectation yields $\mathbb{E}[\boldsymbol{L}\boldsymbol{\xi}] = \boldsymbol{L}\mathbb{E}\boldsymbol{\xi} = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{L}\boldsymbol{\xi}\boldsymbol{\xi}^H\boldsymbol{L}^H] = \boldsymbol{L}\boldsymbol{I}\boldsymbol{L}^H$. Multiplying the diagonal square roots gives $\boldsymbol{L}\boldsymbol{L}^H = \boldsymbol{U}\widetilde{\boldsymbol{\Lambda}}\boldsymbol{U}^H$. Properness adds $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] = \boldsymbol{0}$, hence $\mathbb{E}[\boldsymbol{L}\boldsymbol{\xi}\boldsymbol{\xi}^\top\boldsymbol{L}^\top] = \boldsymbol{0}$. The covariance calculation only needs the stated first two moments; Gaussianity is sufficient but not necessary. □

Scope. The matrix and floor identities are separated from the analytical expectation argument in the formalization. A covariance transfer identity is not a formal derivation of the moments of a Gaussian probability measure.

Remark D.6 (Real Fourier endpoints and time-domain covariance). For a real length-T waveform $\boldsymbol{x}$, $\widehat{x}(T-k) = \overline{\widehat{x}(k)}$. The DC coefficient and, for even T, the Nyquist coefficient are real. The proper-complex construction in Proposition D.5 applies to interior positive-frequency bins. A mathematically consistent real sampler must use real random vectors at self-conjugate bins and pair the remaining frequencies by conjugacy. If an IRFFT routine discards an imaginary part at a self-conjugate bin, the covariance of that projected coefficient is not the unmodified proper-complex covariance.

Under these endpoint conventions and independent frequency draws, inverse Fourier transformation is a real linear map. If $\boldsymbol{F}_{\mathbb{R}}$ is its real-coordinate matrix and $\boldsymbol{C}_{\boldsymbol{\xi}}$ the stacked coefficient covariance, then the pre-normalization time-domain covariance is $\boldsymbol{F}_{\mathbb{R}}\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{F}_{\mathbb{R}}^{\top}$. This follows by expanding the second moment of a deterministic linear transformation. No claim of exact final-source covariance is implied after sample-wise RMS normalization.

D.4 Amplitude shrinkage and label-neutral sampling

Proposition D.7 (Log-amplitude contraction). *Let $a_0 > 0$, $r \in \mathbb{R}$, $0 \le \alpha \le 1$, and $a_\alpha = \exp((1-\alpha)\log a_0 + \alpha r)$. Then*

$$a_\alpha > 0, \qquad a_{\alpha=0} = a_0, \qquad a_{\alpha=1} = e^r, \tag{D.8}$$

$$\log a_\alpha - \log a_0 = \alpha(r - \log a_0), \qquad |\log a_\alpha - \log a_0| = \alpha |r - \log a_0|. \tag{D.9}$$

The log amplitude lies between $\log a_0$ and r. Under the finite patient-balanced empirical distribution of r,

$$\mathbb{E}\log a_\alpha = (1-\alpha)\log a_0 + \alpha \mathbb{E}r, \qquad \mathrm{Var}(\log a_\alpha) = \alpha^2 \mathrm{Var}(r). \tag{D.10}$$

Proof. The exponential is positive and $\log(e^u) = u$. Subtracting $\log a_0$ gives the affine identity. Taking absolute values and using $\alpha \ge 0$ gives the contraction. The interval inclusion follows because the log amplitude is a convex combination. For the expectation, use weights summing to one. Subtract the resulting mean to obtain $\log a_\alpha - \mathbb{E}\log a_\alpha = \alpha(r - \mathbb{E}r)$, square both sides, and average. These variance identities apply to *log amplitude*, not to amplitude itself or the final epsilon-normalized source. □

Proposition D.8 (The actual RMS after source normalization). *Let $r_z = \mathrm{RMS}(\boldsymbol{Z}^{\mathrm{shape}}) \ge 0$, $a_\star = a_\alpha > 0$, and $\epsilon > 0$. For $\boldsymbol{Z}_0 = a_\star \boldsymbol{Z}^{\mathrm{shape}}/(r_z + \epsilon)$,*

$$\mathrm{RMS}(\boldsymbol{Z}_0) = \frac{a_\star r_z}{r_z + \epsilon} \le a_\star, \qquad a_\star - \mathrm{RMS}(\boldsymbol{Z}_0) = \frac{a_\star \epsilon}{r_z + \epsilon}. \tag{D.11}$$

When $r_z > 0$ the relative RMS deficit is $\epsilon/(r_z+\epsilon)$. For $\epsilon = 0$ and $r_z > 0$, the RMS equals $a_\star$.

Proof. Apply Equation (D.4) with the positive multiplier $a_\star/(r_z+\epsilon)$. Subtract the resulting fraction from $a_\star$ and simplify. The denominator is positive and at least r_z, giving the inequality. □

Scope. The denominator depends on the random shape. Therefore the exact second moment is $\mathbb{E}[{a_\star}^2 zz^H/(r_z+\epsilon)^2]$, not in general a fixed scalar times $\mathbb{E}[zz^H]$. For example, in one dimension with $\epsilon = 0$ and fixed $a_\star$, normalizing a nonzero symmetric variable yields $a_\star$ times its sign and variance ${a_\star}^2$, regardless of its original variance. Hence Proposition D.5 must not be restated as exact covariance matching for the final source.

Proposition D.9 (Conditional label neutrality). *Fix the training dataset $\mathcal{D}$ and all statistics computed from it. Let U be source randomness independent of the requested condition Y, and suppose $\boldsymbol{Z}_0 = F_{\mathcal{D}}(U)$ with no Y argument. Then $\mathcal{L}(\boldsymbol{Z}_0 \mid Y = y, \mathcal{D}) = \mathcal{L}(\boldsymbol{Z}_0 \mid \mathcal{D})$ for every conditioning value with positive probability (or under the corresponding constant conditional kernel).*

Proof. For measurable sets $\mathcal{A}, \mathcal{B}$, independence gives

$$\begin{aligned}\Pr(F_{\mathcal{D}}(U) \in \mathcal{A}, Y \in \mathcal{B} \mid \mathcal{D}) &= \Pr(U \in F_{\mathcal{D}}^{-1}(\mathcal{A}), Y \in \mathcal{B} \mid \mathcal{D}) \\ &= \Pr(U \in F_{\mathcal{D}}^{-1}(\mathcal{A}) \mid \mathcal{D}) \Pr(Y \in \mathcal{B} \mid \mathcal{D}).\end{aligned}$$

This is conditional independence of $\boldsymbol{Z}_0$ and Y given $\mathcal{D}$. Equivalently, the sampler kernel is constant in its label argument. The result does not say that $\mathcal{D}$ contains no disease-related structure, that $\boldsymbol{Z}_0$ is independent of $\mathcal{D}$, or that releasing a generator has a differential-privacy guarantee. □

D.5 Clean-endpoint transport and capped weighting

Proposition D.10 (Endpoint–velocity error identity)**.** *For $\boldsymbol{X}_t = (1-t)\boldsymbol{Z}_0 + t\boldsymbol{X}_1$ and $t < 1$, the path derivative is $\boldsymbol{U}_t = \boldsymbol{X}_1 - \boldsymbol{Z}_0$. With $\boldsymbol{V}_{\boldsymbol{\theta}} = (\widehat{\boldsymbol{X}}_1 - \boldsymbol{X}_t)/(1-t)$,*

$$\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{U}_t = \frac{\widehat{\boldsymbol{X}}_1 - \boldsymbol{X}_1}{1-t}, \qquad \|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{U}_t\|_1 = \frac{\|\widehat{\boldsymbol{X}}_1 - \boldsymbol{X}_1\|_1}{1-t}. \tag{D.12}$$

The squared L_2 identity instead has denominator $(1-t)^2$.

Proof. Differentiate the affine path coordinate by coordinate. Substituting it into $\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{U}_t$ gives the numerator $\widehat{\boldsymbol{X}}_1 - (1-t)\boldsymbol{Z}_0 - t\boldsymbol{X}_1 - (1-t)(\boldsymbol{X}_1 - \boldsymbol{Z}_0) = \widehat{\boldsymbol{X}}_1 - \boldsymbol{X}_1$. The norm identities follow from homogeneity and $1 - t > 0$. This is an algebraic identity for a sampled source–target pair; it does not assume an optimal network or exact ODE integration. □

Proposition D.11 (Cap bounds and unchanged uncapped region)**.** *Let $d = CT > 0$ and $w_\tau(t) = 1/\max\{1-t, \tau\}$ with $\tau > 0$. For $t < 1$,*

$$0 < w_\tau(t) \le \frac{1}{\tau}, \qquad w_\tau(t) \le \frac{1}{1-t}. \tag{D.13}$$

If $1 - t \ge \tau$, the capped and uncapped weights coincide. If $1 - t \le \tau$, the capped weight is $1/\tau$. The sample loss $\ell_\tau(\widehat{\boldsymbol{X}}, \boldsymbol{X}; t) = w_\tau(t)\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\|_1/d$ obeys

$$|\ell_\tau(\boldsymbol{A}, \boldsymbol{X}; t) - \ell_\tau(\boldsymbol{B}, \boldsymbol{X}; t)| \le \frac{w_\tau(t)}{d}\|\boldsymbol{A} - \boldsymbol{B}\|_1 \le \frac{1}{d\tau}\|\boldsymbol{A} - \boldsymbol{B}\|_1. \tag{D.14}$$

Every coordinate of a loss subgradient with respect to $\widehat{\boldsymbol{X}}$ has magnitude at most $1/(d\tau)$.

Proof. The maximum is positive and no smaller than either positive denominator in Equation (D.13); reciprocal inequalities give the bounds. The two regimes follow directly from the definition of the maximum. The reverse triangle inequality for L_1 gives $|\|\boldsymbol{A} - \boldsymbol{X}\|_1 - \|\boldsymbol{B} - \boldsymbol{X}\|_1| \le \|\boldsymbol{A} - \boldsymbol{B}\|_1$. Multiplication by the positive weight and its upper bound yields Equation (D.14). The scalar absolute-value subgradient belongs to $[-1, 1]$, including at zero; rescaling by w_τ/d gives the coordinate bound. A parameter gradient also contains the network Jacobian, so this result alone does not bound $\|\nabla_{\boldsymbol{\theta}} L\|$ or prove stable optimization. □

Remark D.12 (Positive time weights and the scope of L_1 regression). At a fixed input $(\boldsymbol{X}_t, t, y)$, a strictly positive weight depending only on t does not change the set of minimizers of the conditional L_1 risk. That risk is minimized coordinatewise by conditional medians, whereas squared L_2 risk is minimized by the conditional mean when its second moment exists. Restricting a network class or adding auxiliary terms can change both optima. Consequently, the standard conditional-mean Flow Matching argument cannot be transferred automatically to the complete capped-L_1 plus-auxiliary objective.

To see the distinction without regularity issues, let X be zero with probability $3/4$ and one with probability $1/4$. For $a \in [0, 1]$,

$$\mathbb{E}|a - X| = \tfrac{1}{4} + \tfrac{1}{2}a, \qquad \mathbb{E}(a - X)^2 = (a - \tfrac{1}{4})^2 + \tfrac{3}{16}. \tag{D.15}$$

The L_1 optimum is zero, but the squared-L_2 optimum is $1/4$. The example does not claim that the trained EEG model follows this particular conditional law. It shows why an exact generative-distribution theorem would need assumptions beyond the loss identities above.

D.6 What the time-domain auxiliary terms constrain

Proposition D.13 (Log-scale discrepancy)**.** *For positive channel scales $s_c, \widehat{s}_c$, each scale term satisfies*

$$|\log \widehat{s}_c - \log s_c| = |\log(\widehat{s}_c/s_c)| \geq 0. \tag{D.16}$$

It is zero exactly when $\widehat{s}_c = s_c$, and is invariant to multiplying both scales by the same positive constant. If a single channel's term is at most η, then $e^{-\eta} \leq \widehat{s}_c/s_c \leq e^{\eta}$.

Proof. The logarithm quotient rule gives the first equality. The logarithm is injective on positive reals, and adding $\log b$ to both logarithms does not change their difference. The final statement follows by exponentiating $-\eta \leq \log(\widehat{s}_c/s_c) \leq \eta$. □

Scope. The Method's unstabilized logarithms require positive channel scales. A flat channel has scale zero, so an actual additive-epsilon or clamp policy must be declared. Such stabilization generally destroys exact scale invariance. Also, an average scale loss at most η only guarantees an individual channel term at most $C\eta$, not at most η.

Proposition D.14 (Increment matching is an error seminorm)**.** *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{C\times T}$ and $1 \leq \ell < T$. With $\boldsymbol{E} = \boldsymbol{A} - \boldsymbol{B}$ and $\Delta_\ell a_{c,t} = a_{c,t+\ell} - a_{c,t}$ for $0 \leq t < T - \ell$,*

$$\Delta_\ell \boldsymbol{A} - \Delta_\ell \boldsymbol{B} = \Delta_\ell \boldsymbol{E}. \tag{D.17}$$

The multi-lag term is a nonnegative weighted sum of L_1 seminorms of this error. It is invariant to adding arbitrary constants independently to each channel of $\boldsymbol{E}$, and satisfies

$$\frac{\|\Delta_\ell \boldsymbol{E}\|_1}{C(T-\ell)} \leq \frac{2\|\boldsymbol{E}\|_1}{C(T-\ell)}. \tag{D.18}$$

Proof. Expand both sides of Equation (D.17). A constant channel offset cancels in a difference. Absolute homogeneity and the triangle inequality for each linear difference map prove the seminorm properties. Finally, $|e_{c,t+\ell} - e_{c,t}| \leq |e_{c,t+\ell}| + |e_{c,t}|$. Each of the two cropped sums is bounded by the full $\|\boldsymbol{E}\|_1$, giving Equation (D.18) after division by a positive count. □

Scope. Here $\boldsymbol{A}, \boldsymbol{B}$ may be the normalized waveforms entering multi-scale increment alignment. The result does not identify the unspecified normalization in the main text. A target with a sharp change has zero increment discrepancy when it is reproduced exactly, even when its own total variation is nonzero. Therefore target increment matching is not the same objective as penalizing the generated waveform's total variation toward zero.

Proposition D.15 (The lag-one zero set)**.** *Suppose the lag set contains one and every lag has a positive weight. If $L_{\text{increment}}(\boldsymbol{A}, \boldsymbol{B}) = 0$, then $a_{c,t} - b_{c,t} = k_c$ for constants k_c. If, in addition, $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same temporal mean in each channel, then $\boldsymbol{A} = \boldsymbol{B}$.*

Proof. All terms are nonnegative, so zero total loss forces every lag-one absolute difference to vanish. Thus $e_{c,t+1} = e_{c,t}$. Induction over t gives $e_{c,t} = e_{c,0} = k_c$. Equality of channel means gives $0 = T^{-1}\sum_t e_{c,t} = k_c$. Without lag one, a larger null space may remain; for example, lag-four differences alone do not connect the four residue classes of time indices. □

Remark D.16 (A spectral interpretation without changing the multi-scale increment loss). For a *periodic* length-T error $\boldsymbol{e}$ and circular shift $\boldsymbol{S}_\ell$, its unitary DFT obeys $\left[\widehat{(\boldsymbol{S}_\ell - \boldsymbol{I})\boldsymbol{e}}\right]_k = (e^{2\pi \mathrm{i} k\ell/T} - 1)\widehat{e}_k$. Using $|e^{\mathrm{i}u} - 1|^2 = 4\sin^2(u/2)$ and Parseval gives

$$\|(\boldsymbol{S}_\ell - \boldsymbol{I})\boldsymbol{e}\|_2^2 = \sum_k 4\sin^2(\pi k\ell/T)|\widehat{e}_k|^2. \tag{D.19}$$

This follows by substituting the DFT shift identity and expanding the squared modulus. The actual Method uses cropped differences and an L_1 norm, so Equation (D.19) is an explanatory idealization, not a claimed equality for its multi-scale increment loss. For squared L_2, the circular energy equals the cropped energy plus the wrap-around boundary terms. A time-domain loss can therefore have spectral effects without being an explicit frequency-domain loss; source and objective are not mathematically independent simply because their computations use different domains.

Proposition D.17 (Pearson correlation as centered shape agreement)**.** *Flatten a waveform and subtract its global mean, obtaining nonzero real vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Define $\rho(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/(\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2)$. Then*

$$-1 \le \rho \le 1, \qquad 0 \le L_{\mathrm{corr}} = 1 - \rho \le 2, \qquad \left\| \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} - \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} \right\|_2^2 = 2L_{\mathrm{corr}}. \tag{D.20}$$

Global translations and positive rescaling of either original waveform do not change ρ. Moreover, $L_{\mathrm{corr}} = 0$ exactly when the two centered vectors have the same direction.

Proof. Cauchy–Schwarz gives $|\boldsymbol{u}^\top \boldsymbol{v}| \le \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$; divide by a positive product. Expanding the squared distance between the unit vectors gives $1 + 1 - 2\rho$. Centering cancels a global translation, and a positive scale cancels between numerator and denominator. A squared norm is zero exactly when its argument is zero, proving the final equivalence. □

Scope. Global centering after flattening is not per-channel centering. A constant waveform makes unstabilized Pearson correlation undefined. Adding a stabilizer to the denominator changes the exact unit-vector identity and can make an identical finite-amplitude pair have nonzero correlation loss. The statement above applies to the unstabilized formula on nonconstant vectors.

Proposition D.18 (Nonnegativity and a conditional zero-loss implication)**.** *On the nondegenerate domain where all four losses are defined, suppose $\lambda_{\mathrm{aux}} \ge 0$. Then $L_{\mathrm{full}} \ge L_{\mathrm{cap}} \ge 0$. If $\lambda_{\mathrm{aux}} > 0$ and a sample's total loss is zero, each of its four nonnegative terms is zero. In particular, the positive capped weight implies $\widehat{\boldsymbol{X}}_1 = \boldsymbol{X}_1$ for that sample.*

Proof. Absolute values and L_1 norms are nonnegative; the correlation bound was proved in Proposition D.17. A sum of nonnegative terms can vanish only if every term vanishes. Since $\max(1 - t, \tau) > 0$ and $CT > 0$, zero capped loss forces $\|\widehat{\boldsymbol{X}}_1 - \boldsymbol{X}_1\|_1 = 0$. This is a pointwise implication. For an expected loss, an almost-everywhere conclusion additionally requires measurable, nonnegative integrands and the standard zero-integral argument. □

D.7 Two boundaries relevant to source and solver claims

Proposition D.19 (Independent pairing does not imply shorter paths)**.** *Let $\boldsymbol{x}, \boldsymbol{z}$ be independent real vectors with finite second moments. Then*

$$\mathbb{E}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 = \operatorname{tr}\operatorname{Cov}(\boldsymbol{x}) + \operatorname{tr}\operatorname{Cov}(\boldsymbol{z}) + \|\mathbb{E}\boldsymbol{x} - \mathbb{E}\boldsymbol{z}\|_2^2. \tag{D.21}$$

In particular, two centered sources with equal total second moment have the same expected squared displacement under independent pairing to a fixed target population, even if their spectral shapes differ.

Proof. Expand $\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{z}\|_2^2 - 2\boldsymbol{x}^\top \boldsymbol{z}$. Independence factors each cross moment, $\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{z}] = (\mathbb{E}\boldsymbol{x})^\top (\mathbb{E}\boldsymbol{z})$. Use $\mathbb{E}\|\boldsymbol{x}\|_2^2 = \operatorname{tr}\operatorname{Cov}(\boldsymbol{x}) + \|\mathbb{E}\boldsymbol{x}\|_2^2$ and the analogous identity for $\boldsymbol{z}$. The result concerns the stated independent coupling, not an optimal-transport coupling or the learned marginal ODE. □

Scope. This analytical moment argument is not included as a probability- measure theorem in the Lean project. It prevents the stronger, unsupported claim that spectral source structure alone must shorten every sampled path.

Proposition D.20 (Network-evaluation count for the recorded solver)**.** *For $K \ge 1$ intervals, if the first $K - 1$ intervals each use two network evaluations and the terminal interval uses one, the total is $2(K - 1) + 1 = 2K - 1$. Thus $K = 5, 10, 20, 50$ give $9, 19, 39, 99$ calls.*

Proof. Add the calls over the stated intervals and simplify. No order-of-accuracy or wall-clock theorem follows from counting calls. In particular, the one-call terminal rule is part of the recorded implementation, not a universal property of every Heun solver, and avoiding a call at $t = 1$ does not itself prove regularity of the uncapped endpoint-derived field there. □