MONTAGEFLOW: PHYSICALLY VALID HIGH FIDELITY EEG GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative modeling offers a promising solution to the scarcity of labeled clinical electroencephalography (EEG) data under strict privacy constraints. Although recent generative models report substantial improvements in generation fidelity, we show that these gains conceal a categorical failure. EEG is not an arbitrary multichannel signal: each channel records the voltage difference between two scalp electrodes. So, real recordings satisfy exact linear identities among their channels. Current generators invariably violate these identities, yet this has gone unnoticed. We therefore propose **MontageFlow**, a flow-matching model that enforces these identities by restricting its output to signals that satisfy them. To retain the EEG characteristics that oversmoothing washes out, we also condition the model on a physiological state vector encoded from the whole EEG window. For generation, this vector is drawn from a prior fitted in latent space. We find that ensuring validity enhances fidelity too. Across 3 large-scale benchmarks, our model improves TS-FID by 1.7–2.5× over the closest rival while ensuring physical validity.

1 INTRODUCTION

Clinical electroencephalography (EEG) data are difficult to collect at scale. Recordings with clinically meaningful labels require expert review, and privacy constraints limit data sharing. Generative models offer a way to relieve this bottleneck, synthesizing data where labeled recordings are scarce or cannot be shared. Yet EEG generation remains relatively underexplored despite recent advances in diffusion and flow-based modeling (Ho et al., 2020; Lipman et al., 2023; Peebles & Xie, 2023). The few existing approaches adapt generative adversarial networks (Hartmann et al., 2018), diffusion models (Shu et al., 2025) and flow matching models (Wang et al., 2026), typically treating EEG as a generic multichannel time series. This treatment overlooks a defining property of clinical EEG: its channels are structurally constrained by how they are recorded.

Clinical EEG is acquired from scalp electrodes, but read as potential differences between pairs of them. The chosen set of pairs, called a *montage*, links electrodes into chains. Summing the differences along a chain cancels the intermediate electrodes and leaves only the difference between its endpoints. Two chains with the same endpoints must therefore sum to the same value at every timepoint. The montage used across our benchmarks contains sixteen electrode pairs and two independent loops. These impose two independent linear constraints, restricting the sixteen channel values at each timepoint to a fourteen-dimensional subspace. Real recordings satisfy these identities, but samples from current generators do not.

This failure has remained hidden because standard evaluation metrics do not explicitly measure montage violations. More importantly, improved distributional fidelity does not guarantee physical validity. We therefore enforce these identities by construction rather than by penalizing violations. Specifically, we parameterize the model output within the subspace permitted by the montage. Invalid signals are thus excluded without introducing an additional tuning hyperparameter.

Enforcing the montage identities guarantees physical validity, but does not by itself ensure fidelity to real EEG. Capturing dominant low-frequency activity while preserving low-energy transients and heavy-tailed amplitudes remains challenging. JET (Wang et al., 2026) addresses waveform fidelity through explicitly specified auxiliary temporal, spectral, and statistical objectives. We instead use learned window-level conditioning to help preserve these characteristics. Specifically, we encode

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.

each multichannel EEG window into a single physiological state vector and condition the flow-matching decoder on it. This vector provides the decoder with window-specific context to guide generation. At sampling time, the vector is drawn from a prior fitted in the encoder's latent space.

In summary, we propose **MontageFlow**, a flow-matching framework for physically valid, high-fidelity EEG generation that combines montage-constrained outputs with learned window-level conditioning. Across TUSZ, TUEV, and TUAB, MontageFlow improves TS-FID by 1.7–$2.5\times$ over the strongest baseline while satisfying the montage identities. Our contributions are summarized below:

- **Physically valid EEG generation.** We show that state-of-the-art generators synthesize physically invalid EEG signals and introduce the *montage index* to measure this (0 = perfectly valid). MontageFlow restricts its output to the subspace of valid signals (Section 2.2).
- **Conditioning mitigates oversmoothing and validity improves fidelity.** We observe that latent conditioning mitigates oversmoothing of EEG. Additionally, we find that enforcing validity improves fidelity, lowering TS-FID on all six splits by 8 to 55 points (Section 5).
- **State-of-the-art results.** Across TUSZ, TUEV and TUAB, MontageFlow improves TS-FID by 1.7–$2.5\times$ over the strongest baseline, with a montage index of exactly 0 (Section 4).

2 METHOD

2.1 PRELIMINARIES

We first derive the exact linear constraints imposed by the bipolar EEG montage, then show that standard flow matching does not guarantee that generated outputs satisfy them.

Bipolar montage geometry. Let $X = [\mathbf{x}(\tau_1), \ldots, \mathbf{x}(\tau_L)] \in \mathbb{R}^{C\times L}$ denote an EEG window with $C = 16$ channels and L physical time points. A bipolar montage represents channel i as the potential difference between an ordered pair of electrodes:

$$x_i(\tau) = v_{a_i}(\tau) - v_{b_i}(\tau), \qquad i = 1, \ldots, C. \tag{1}$$

As illustrated in Figure 1(a), summing channels along a directed path cancels the potentials at intermediate electrodes, leaving only the difference between the path endpoints. The montage contains two pairs of paths with common endpoints. On the left hemisphere, channels 1–4 and 9–12 both connect $\mathrm{Fp1}$ to $\mathrm{O1}$. On the right hemisphere, channels 5–8 and 13–16 both connect $\mathrm{Fp2}$ to $\mathrm{O2}$. Consequently, the following equalities hold at every physical time point τ:

$$x_1(\tau) + x_2(\tau) + x_3(\tau) + x_4(\tau) = x_9(\tau) + x_{10}(\tau) + x_{11}(\tau) + x_{12}(\tau), \tag{2}$$

$$x_5(\tau) + x_6(\tau) + x_7(\tau) + x_8(\tau) = x_{13}(\tau) + x_{14}(\tau) + x_{15}(\tau) + x_{16}(\tau). \tag{3}$$

Figure 1(b) verifies Eq. 2 on a representative TUSZ window.

Montage constraint. Let $\mathbf{1}_4$ and $\mathbf{0}_4$ denote four-dimensional vectors of ones and zeros, respectively. The two path equalities can be encoded by the constraint matrix

$$A = \begin{bmatrix} \mathbf{1}_4^\top & \mathbf{0}_4^\top & -\mathbf{1}_4^\top & \mathbf{0}_4^\top \\ \mathbf{0}_4^\top & \mathbf{1}_4^\top & \mathbf{0}_4^\top & -\mathbf{1}_4^\top \end{bmatrix} \in \mathbb{R}^{2\times 16}, \tag{4}$$

where the four column groups correspond to channels 1–4, 5–8, 9–12, and 13–16, respectively. Thus, $A\mathbf{x}(\tau) = 0$ at every physical time point. Since A acts along the channel dimension, the corresponding window-level constraint is

$$AX = 0_{2\times L}. \tag{5}$$

The set of montage-valid channel vectors is therefore $\ker A = \{\mathbf{u} \in \mathbb{R}^{16} : A\mathbf{u} = 0\}$. Since the two rows of A are linearly independent, $\mathrm{rank}(A) = 2$, and $\ker A$ is a 14-dimensional subspace of $\mathbb{R}^{16}$.

Standard flow matching. Let $E \in \mathbb{R}^{C\times L}$ have i.i.d. standard Gaussian entries and $t \in [0, 1]$ denote flow time. Linear flow matching interpolates $Z_t = tX + (1-t)E$ and trains a velocity network $u_\theta(Z_t, t)$ to regress the target velocity $X - E$. Generation starts from $\widehat{Z}_0 = E$ and integrates

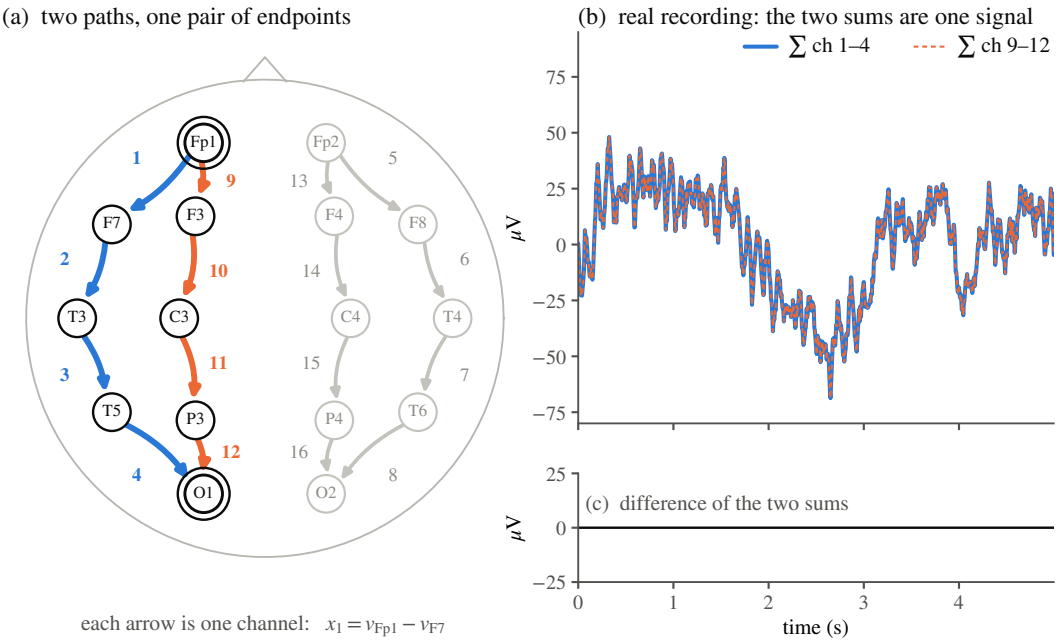

Figure 1: **The bipolar montage induces two exact linear constraints. (a)** Each channel is an oriented difference between two electrode potentials. Channels 1–4 and 9–12 form two paths from Fp1 to O1, and channels 5–8 and 13–16 do the same from Fp2 to O2. Either path sums to the potential difference between its endpoints. **(b)** The two left path sums overlap on a real TUSZ window, and **(c)** their difference is numerically zero.

the learned velocity along its own trajectory $\widehat{Z}_t$: $\widehat{X} = E + \int_0^1 u_\theta(\widehat{Z}_t, t)\, dt$. Applying the montage constraint of Eq. 5 to this sample gives

$$A\widehat{X} = AE + \int_0^1 A\, u_\theta(\widehat{Z}_t, t)\, dt \tag{6}$$

where $AE \neq 0$ almost surely and u_θ is unconstrained. Nothing forces the two terms to cancel, so standard flow matching does not guarantee montage-valid EEG.

2.2 MONTAGEFLOW

Overview. The overview of MontageFlow is shown in Figure 2. MontageFlow removes both sources of montage violation (Eq. 6) by construction: generation starts from noise projected onto $\ker A$, and the velocity network's output is always projected onto $\ker A$. For fidelity, the velocity network is additionally conditioned on a global latent encoding of the target window.

Montage projector. Projecting the noise and the velocity onto $\ker A$ requires the same operation: given any channel vector $\mathbf{u} \in \mathbb{R}^{16}$, return a vector in $\ker A$. The vector $A\mathbf{u} \in \mathbb{R}^2$ holds the amounts by which $\mathbf{u}$ violates Eqs. 2 and 3 respectively. To make $\mathbf{u}$ valid, we cancel each violation by spreading it equally over the eight channels involved. Each of the four channels on the left-hand side of the equation is decreased by one eighth of the violation, and each of the four on the right-hand side is increased by the same amount:

$$P\mathbf{u} = \mathbf{u} - \tfrac{1}{8}A^\top(A\mathbf{u}), \qquad P = I - \tfrac{1}{8}A^\top A. \tag{7}$$

Since each row of A has eight entries equal to ± 1 and the two rows share no channel, $AA^\top = 8I_2$. Hence, $AP = A - \frac{1}{8}(AA^\top)A = 0$. Applied to a window X, P acts on each column, so $A(PX) = (AP)X = 0$. Therefore, P maps every window to a montage-valid window.

Projected velocity. With P in hand, the two projections of the overview are immediate. For the velocity, let $v_\theta(X_t, t, y, h) \in \mathbb{R}^{C\times L}$ be the raw output of the velocity network, conditioned on flow

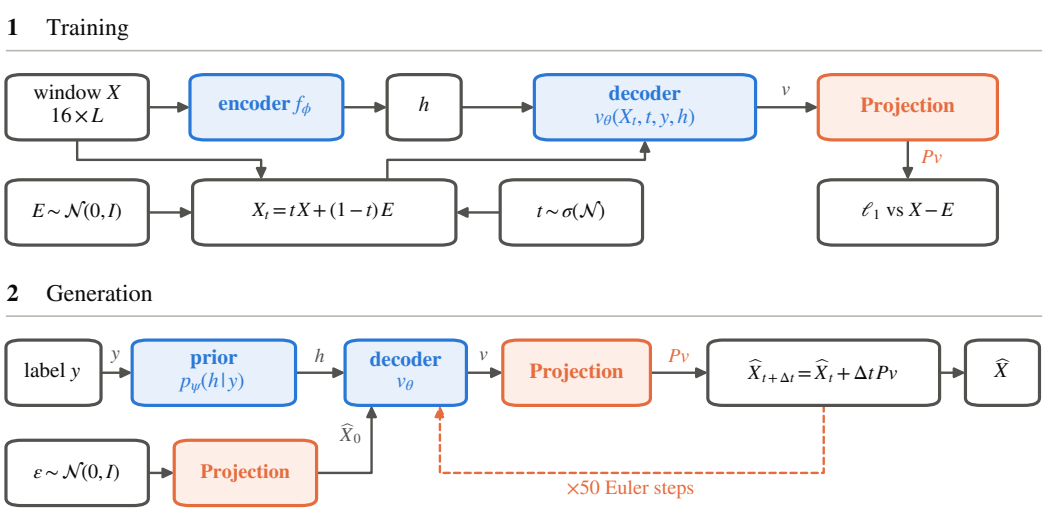

Figure 2: **Overview of MontageFlow.** During training, an encoder f_ϕ maps an EEG window X to a global latent h. The decoder receives the current flow state X_t, flow time t, class label y, and perturbed latent $\widetilde{h}$, and predicts a velocity field. After encoder-decoder training, a class-conditional rectified-flow prior $p_\psi(h \mid y)$ is fitted to the encoded training latents and replaces the encoder during generation. The predicted velocity is projected onto $\ker A$. At inference, the Gaussian source is also projected to guarantee valid EEG.

time t, class label y, and a latent h introduced below. MontageFlow uses the projected velocity

$$u_\theta(X_t, t, y, h) = P\, v_\theta(X_t, t, y, h), \tag{8}$$

which satisfies $Au_\theta = 0$. We sample $t = \mathrm{sigmoid}(z)$, with $z \sim \mathcal{N}(-0.8, 0.8^2)$, and set $X_t = tX + (1-t)E$. We optimize the following ℓ_1 velocity-matching objective:

$$\mathcal{L}_{\mathrm{MF}}(\theta, \phi) = \mathbb{E}_{X,y,E,t,\eta,\sigma}\left[\left\|(X - E) - P\, v_\theta(X_t, t, y, \widetilde{h})\right\|_1\right]. \tag{9}$$

Here, $\widetilde{h} = f_\phi(X) + \sigma\eta$ is the perturbed conditioning latent, defined in Eq. 12.

Validity by construction. The remaining source of violation is the start state, which we also project. Generation draws $\varepsilon \sim \mathcal{N}(0, I)$, starts from $\widehat{X}_0 = P\varepsilon$, and integrates the projected velocity:

$$\widehat{X} = P\varepsilon + \int_0^1 P\, v_\theta(\widehat{X}_t, t, y, h)\, dt. \tag{10}$$

Applying A gives $A\widehat{X} = AP\varepsilon + \int_0^1 AP\, v_\theta(\widehat{X}_t, t, y, h)\, dt = 0$ since $AP = 0$. The generated sample now satisfies Eq. 5.

Latent conditioning. Projection guarantees validity but distributional fidelity is also important. Hence, to give the velocity network a global view of the window it is generating, we condition it on a latent physiological state vector:

$$h = f_\phi(X) \in \mathbb{R}^{256}, \tag{11}$$

computed by an encoder f_ϕ from the clean target window X during training. At generation time, the conditioning vector is sampled from a learned latent prior rather than obtained by encoding a real window. To make the decoder robust to deviations between prior samples and encoded latents, we perturb the latent during training:

$$\widetilde{h} = f_\phi(X) + \sigma\eta, \qquad \eta \sim \mathcal{N}(0, I_{256}), \quad \sigma \sim \mathcal{U}(0, 1). \tag{12}$$

The velocity network receives $\widetilde{h}$ as conditioning in Eq. 9, and the encoder is trained jointly with it.

Latent prior. The encoder needs a target window, which is unavailable at generation time. We therefore replace it with a learned prior. After training, the training set is encoded into a latent bank $\mathcal{H} = \{(f_\phi(X_i), y_i)\}_{i=1}^n$, and a class-conditional rectified-flow model $p_\psi(h \mid y)$ is fitted to these pairs. Conditioning both the prior and the velocity network on y preserves the dependence between latent and label seen in training.

Sampling. To generate a window, we draw a class y, a latent $h \sim p_\psi(h \mid y)$, and a start state $\widehat{X}_0 = P\varepsilon$, then integrate Eq. 10 with K Euler steps.

3 EXPERIMENTAL SETUP

Datasets. We evaluate on the three Temple University Hospital (TUH) EEG corpora used by prior work: TUAB (abnormal/normal, 2 classes), TUEV (6 event types, heavily imbalanced), and TUSZ (seizure/background, 2 classes). All three are preprocessed with the released pipeline of Wang et al. (2026) into 16 channels $\times$ L length windows.

Table 1: Corpora. Windows lie on the 16-channel bipolar montage at 200 Hz.

	classes	window	train windows	test windows
TUSZ	2	16×1000	25,542	9,538
TUEV	6	16×1000	68,445	29,421
TUAB	2	16×2000	297,103	36,945

Metrics. Our primary fidelity metric is **(a) TS-FID** (lower is better): the Fréchet distance between representations of real and generated multivariate signals. It is computed from FFT magnitudes, so it sees spectral structure but is blind to physical validity. Hence, we propose the **(b) montage index**, which measures physical validity: 0 represents perfect validity, 1 is typical of independent channels. We also use **(c) shape correlation**, which compares the generated cross-channel correlation structure against the real one (higher is better, Appendix A).

Baselines. EEG-GAN (Hartmann et al., 2018), vanilla diffusion (Ho et al., 2020), vanilla flow matching (Lipman et al., 2023), JET (Wang et al., 2026), and its cross-channel variant JET$^\times$.

Implementation. The encoder of Section 2.2 is a causal Transformer (8 blocks, width 512, 8 heads). The encoder output is projected to a single 256-dimensional latent h. The decoder is a 1-D DiT-B transformer (12 blocks, width 768, 12 heads) with adaLN conditioning on the timestep, the class label and h. Each channel is split into patches of size 100, and attention runs over the patches of all channels. Together they account for 156M parameters, of which the decoder holds 130M. Training runs for $\sim$ 160k steps at an effective batch of 256. We use AdamW with a learning rate of 5×10^{-5} held constant after five epochs of warmup. Once the encoder/decoder is trained we freeze the encoder, encode the training split, and fit the prior on those latents for 12k steps. Generating a window then follows Figure 2: we draw a class label, sample h from the prior conditioned on it, and integrate the decoder ODE from projected Gaussian noise for 50 steps of Euler's method.

4 RESULTS

Table 2: Generation quality and physical validity (**bold**: best, underline: second best).

	TS-FID (train) ↓			TS-FID (test) ↓			montage index (test) ↓		
Method	TUSZ	TUEV	TUAB	TUSZ	TUEV	TUAB	TUSZ	TUEV	TUAB
Real recordings (floor)	42.7	32.0	22.2	39.1	25.3	23.3	0.000	0.000	0.000
EEG-GAN	368.2	336.5	432.6	356.7	417.2	491.5	0.615	0.516	0.629
Vanilla diffusion	164.3	<u>127.7</u>	<u>140.6</u>	162.8	<u>188.5</u>	<u>166.2</u>	<u>0.075</u>	<u>0.096</u>	<u>0.079</u>
Vanilla flow matching	<u>156.6</u>	150.2	176.5	<u>161.8</u>	233.9	201.2	0.286	0.437	0.298
JET	227.9	228.3	260.7	217.5	310.3	299.0	0.990	0.988	0.986
JET$^\times$	222.9	158.1	225.3	215.6	236.6	261.9	0.151	0.219	0.147
MontageFlow	**64.4**	**52.1**	**59.3**	**71.2**	**113.4**	**77.0**	**0.000**	**0.000**	**0.000**

Generation quality and physical validity. MontageFlow achieves the lowest TS-FID on both evaluation splits of all three datasets, with scores 1.7–2.5$\times$ lower than the strongest baseline in each setting (Table 2). In every setting, the strongest baseline is a vanilla diffusion or flow-matching model sharing our backbone. Vanilla flow matching differs from MontageFlow only in latent conditioning and montage projection, providing a controlled comparison of their combined contribution.

Their individual effects are examined in Table 4. Only real recordings and MontageFlow achieve a montage index of exactly 0. JET has an index near 0.99, while adding cross-channel attention (JET$^{\times}$) reduces it to 0.15–0.22. Vanilla diffusion achieves the lowest index among the baselines, at 0.075–0.096, but still violates the montage identities. These results show that learned cross-channel dependencies can reduce violations, whereas our construction guarantees validity.

Channel relationship. Montage validity alone does not guarantee realistic cross-channel relationships: any projected output satisfies the montage identities, even if its remaining channel dependencies differ from those of real EEG. We therefore evaluate shape correlation (Table 3). MontageFlow is closest to the real-data reference in all six settings, achieving 0.989–0.996 on train and 0.952–0.989 on test. The vanilla models, which also use cross-channel attention, achieve 0.93–0.97 but consistently remain below MontageFlow, with larger gaps on test. JET scores near 0 across all settings, indicating poor agreement with the cross-channel structure captured by this metric. Adding cross-channel attention in JET$^{\times}$ raises the scores to 0.91–0.96, supporting the importance of explicitly modeling channel dependencies. EEG-GAN achieves substantially lower scores of 0.39–0.67.

Table 3: Shape correlation across three datasets (**bold**: best, underline: second best).

	shape correlation (train) ↑			shape correlation (test) ↑		
Method	TUSZ	TUEV	TUAB	TUSZ	TUEV	TUAB
Real recordings	0.998	0.999	0.999	0.999	0.998	0.999
EEG-GAN	0.426	0.541	0.655	0.395	0.505	0.669
Vanilla diffusion	0.963	0.959	0.965	0.926	0.929	0.958
Vanilla flow matching	0.973	0.966	0.967	0.927	0.936	0.961
JET	−0.004	−0.016	0.001	−0.029	0.009	0.015
JET$^{\times}$	0.956	0.939	0.949	0.914	0.906	0.945
MontageFlow	**0.992**	**0.996**	**0.989**	**0.952**	**0.977**	**0.989**

5 Ablations and Analyses

Table 4: Incremental ablation from vanilla flow matching to MontageFlow (TS-FID ↓).

	TUSZ		TUEV		TUAB	
	train	test	train	test	train	test
Vanilla flow matching (x-parameterisation)	210.3	204.7	217.3	290.8	236.8	262.0
+ velocity parameterisation	162.0	159.2	172.2	250.2	198.3	229.2
+ cross-channel attention	156.6	161.8	150.2	233.9	176.5	201.2
+ latent conditioning	95.0	103.6	77.0	144.4	106.6	130.7
+ latent perturbation	87.1	92.8	60.0	125.3	105.1	131.8
+ montage projection (inference only)	90.1	94.5	59.7	126.4	128.9	160.1
+ montage projection (inference + training)	**64.4**	**71.2**	**52.1**	**113.4**	**59.3**	**77.0**

Incremental ablation with respect to vanilla flow matching. Table 4 starts from vanilla flow matching without cross-channel attention and adds one change per row until it reaches MontageFlow. Switching from x- to velocity parameterisation lowers TS-FID by 33–48 on every split. Adding cross-channel attention gives the vanilla flow matching baseline of Table 2. It barely moves TS-FID on TUSZ and lowers it by 16–28 on TUEV and TUAB. Latent conditioning (row 4) lowers TS-FID by a further 58–90, the largest gain in the ladder. Perturbing the latent during training (row 5) lowers it by another 8–19 on TUSZ and TUEV and leaves TUAB unchanged. Finally, the projection buys physical validity outright for every generated EEG window, and it could be applied at inference time alone. Rows 6 and 7 compare inference-only projection with training the velocity network through the projector. Inference-only projection barely changes train TS-FID on TUSZ and TUEV but worsens it on TUAB by 24. Training with projected velocities instead lowers train TS-FID relative to row 5 by 22.7 on TUSZ, 7.9 on TUEV, and 45.8 on TUAB. We hypothesize that the unconstrained model spends capacity representing signals no electrode montage can produce.

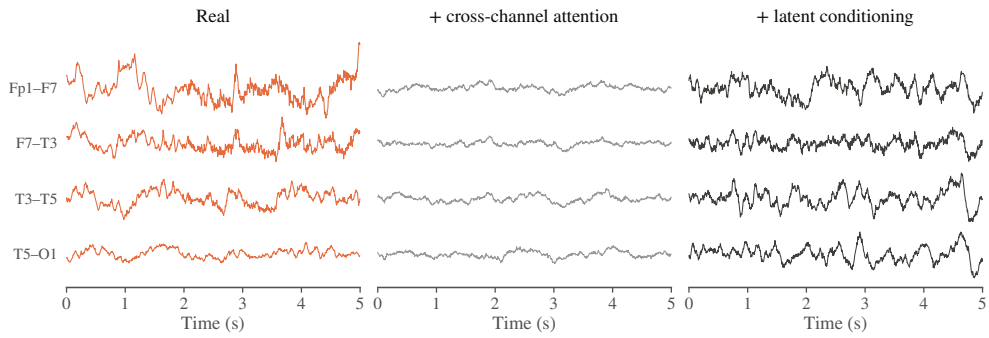

Figure 3: **Latent conditioning mitigates oversmoothing.** Representative samples from TUAB are shown.

Projected training noise. MontageFlow uses unprojected Gaussian noise during training but projects the initial noise at sampling time. We evaluate an alternative that also projects the training noise. This change worsens test TS-FID by 26.2 on TUSZ, 8.6 on TUEV, and 35.7 on TUAB. We hypothesize that unprojected noise adds regularization making the velocity network more robust.

Latent conditioning mitigates oversmoothing. Clinical EEG varies widely from window to window in amplitude and rhythm. Generators trained with a regression loss can average this variety away. As shown in Figure 3, latent conditioning mitigates oversmoothing. Without the latent (row 3 of Table 4), generated windows collapse to a flat average and keep 42–52% of the real median amplitude. With it (row 4), they keep 78–84%, although fast rhythms stay weaker than in real EEG.

Conditioning the prior on the label. The decoder is conditioned on both a window-level latent h and a class label y. Since the latent can encode class-related characteristics, sampling it from a marginal prior $p_\psi(h)$ independently of y can produce inconsistent latent-label pairs. We instead fit $p_\psi(h \mid y)$ and sample it with the same label that is passed to the decoder. On the train split, this improves TS-FID by 15 points on TUEV, which has 6 classes. The gain is small on TUSZ (1.6) and TUAB (0.2), which have 2 classes each. On the test split, the TUEV gain is 8.9 points.

The regression loss (L_1 versus L_2). Flow matching conventionally uses squared-error velocity regression. Compared with L_2, L_1 places less emphasis on large residuals. We evaluate this choice by training MontageFlow under each loss. L_1 yields lower TS-FID on all three datasets and both splits. Using L_2 instead raises TS-FID by approximately 5–9 points on TUSZ and TUEV and 32–36 points on TUAB. On the test split, TUEV rises from 113.4 to 122.8 and TUAB from 77.0 to 113.4.

Computational cost. Table 5 reports the decoder patch, parameters, peak training memory, seconds per optimiser step at batch 256, estimated training GPU-hours, and per-window sampling latency at 50 solver steps. All rows are measured in one run on the same H100. The encoder of MontageFlow adds 26M parameters. MontageFlow's latency includes sampling the latent from the prior, which adds only 2 ms per window. At matched decoder patch, MontageFlow therefore generates about as fast as vanilla flow matching (40 vs. 37 ms). Halving the decoder patch from 200 to 100 raises step time and latency by about 1.7× and peak memory by about 40%.

Table 5: Measured cost on one H100. Latency is per window at 50 steps.

Method	Patch	Params	VRAM	s/step	GPU-h	ms/window
EEG-GAN	–	117M	4.4 GB	4.28	64	2
Vanilla diffusion	100	130M	6.8 GB	0.65	29	37
Vanilla flow matching	100	130M	6.8 GB	0.64	29	37
JET	200	130M	4.9 GB	0.42	19	25
JET$^\times$	200	130M	4.8 GB	0.37	17	21
MontageFlow	100	156M	9.2 GB	0.87	39	40

6 Physiological Diagnostics

We qualitatively examine whether the generated signals behave like EEG by studying Figure 4. **(1) Power spectrum:** The top row shows the power at each frequency. Real EEG has most of

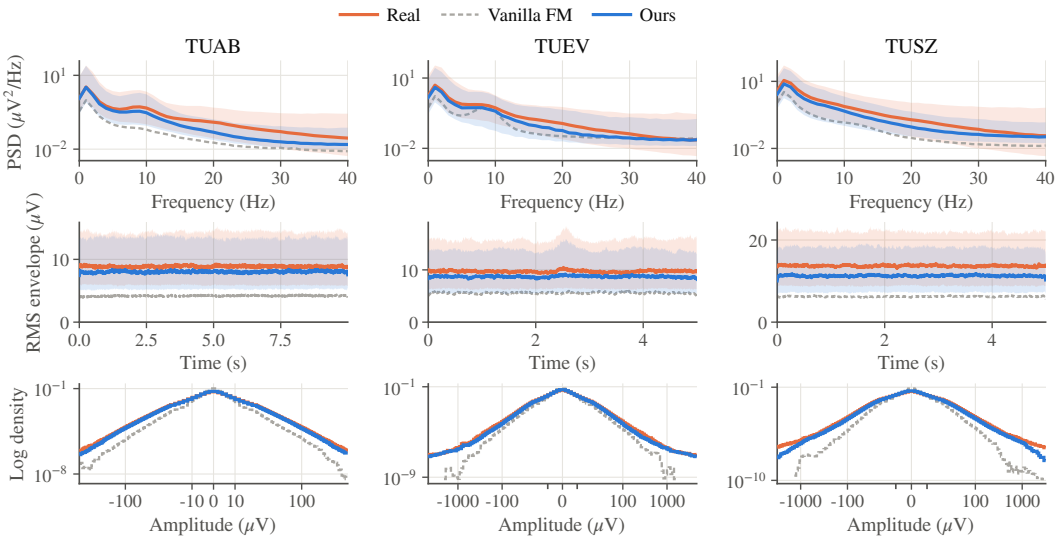

Figure 4: **Physiological diagnostics. Top:** power spectrum, median over windows, with shading from the 10th to the 90th percentile. **Middle:** amplitude over time (root-mean-square across channels), median over windows, with shading over the middle half of windows. **Bottom:** distribution of sample amplitudes (log density). Windows are 10 s on TUAB and 5 s on TUEV and TUSZ.

its power at the lowest frequencies and falls off with frequency, with an alpha peak near 8–9 Hz on TUAB and TUEV. Below about 12 Hz, our spectrum follows the real shape and reproduces the alpha peak at the same frequency. Vanilla flow matching has less power at almost every frequency, and on TUEV it exaggerates the alpha peak. Above 12 Hz, our spectrum falls below the real one, most visibly on TUAB. However, real windows have much higher variance at these frequencies, and our median stays within the real shading throughout. **(2) Amplitude over time:** The middle row tracks signal amplitude across the window. Real EEG keeps a flat median envelope with no drift, and so does MontageFlow. Our envelope sits slightly below the real one, at 82–91% of its level, and its spread across windows is similar. Vanilla flow matching has about half the real amplitude. On TUEV, the real envelope rises near 2.5 s. MontageFlow reproduces this rise, while vanilla flow matching barely shows it. **(3) Heavy-tailed distributions:** The amplitude distribution of clinical EEG is heavy-tailed. A generator that collapses towards average behaviour loses the extremes, where transient events such as spikes live (bottom row). Our density follows the real curve from the sharp central peak down through the tails. Vanilla flow matching is too narrow everywhere, and its tails vanish by 1000 μV. Above 100 μV, our samples keep 60–70% of the real probability mass, against 3–12% for vanilla flow matching.

We also report five more diagnostic metrics in Table 6 (Appendix A), which show that MontageFlow is closest to real EEG on three of the five. JET$^{\times}$ is slightly closer on the envelope and Hjorth activity, which both measure amplitude.

Table 6: Physiological diagnostics on TUSZ test split (**bold**: best, underline: second best).

Method	TS-FID ↓	PSD slope ↓	Envelope ↓	Hj. act. ↓	Hj. mob. ↓	Drift ↓
Real vs. real floor	39.1	0.024	0.052	0.048	0.030	0.052
EEG-GAN	356.7	1.845	1.185	1.353	0.832	0.698
Vanilla diffusion	162.8	0.760	1.795	1.804	0.626	0.328
Vanilla flow matching	161.8	0.196	1.657	1.599	0.403	0.124
JET	217.5	0.861	0.596	0.488	0.678	0.414
JET$^{\times}$	215.6	0.899	**0.502**	**0.449**	0.699	0.280
MontageFlow	**71.2**	**0.172**	0.521	0.530	**0.250**	**0.105**

7 Related works

EEG models. EEG is routinely used to monitor brain activity and is essential for epilepsy diagnosis (Tatum et al., 2018). Deep EEG models are predominantly discriminative and task-specific. For instance, EEGNet uses a compact CNN for EEG classification (Lawhern et al., 2018), while EEG Conformer combines convolution with self-attention to capture long-range temporal dependencies (Song et al., 2023). Gemein et al. (2020) compare deep and feature-based decoders for distinguishing pathological from normal recordings on TUAB, while Tang et al. (2022) and REST (Afzal et al., 2024) use graph neural networks for seizure detection and classification on TUSZ. Even pretrained EEG foundation models (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025) typically require fine-tuning or linear probing on labelled data for downstream tasks (Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025). However, achieving expert-level detection requires large expert-annotated datasets (Jing et al., 2023). Annotation typically relies on certified neurologists, whose expertise is in short supply (Shah et al., 2018). Moreover, clinically significant target events are rare: seizures account for less than 0.1% of the TUH EEG archive (Shah et al., 2018). Data access is also constrained by patient privacy: even TUH, the largest public EEG corpus, required HIPAA de-identification before release (Obeid & Picone, 2016). Hand-crafted augmentation can expand training data, but its effectiveness is task-dependent, and some transformations may alter the underlying semantics (Rommel et al., 2022). Per-channel augmentation may also violate montage identities (Eq. 5). Alternatively, synthetic seizure EEG can substitute for real seizures in detector training, yielding comparable performance with lower re-identification risk (Pascual et al., 2021). Despite extensive research on EEG-based predictive modelling, EEG generation remains comparatively underexplored.

Generative modeling. Generative modeling has advanced rapidly, driven largely by vision. GANs enabled high-fidelity, high-resolution image synthesis (Goodfellow et al., 2014; Karras et al., 2019), but adversarial training can be unstable and prone to mode collapse (Arjovsky et al., 2017; Gulrajani et al., 2017). Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) use a stable, non-adversarial training objective and have surpassed GANs in image quality (Dhariwal & Nichol, 2021). Flow matching learns a velocity field along a prescribed path from noise to data (Lipman et al., 2023; Liu et al., 2023b). Under Gaussian paths, it is closely related to diffusion, with differences largely attributable to noise-level weighting (Kingma & Gao, 2023). Architectures have evolved from U-Nets (Ronneberger et al., 2015; Ho et al., 2020) to transformers operating on latent patches (Peebles & Xie, 2023) and, recently, raw pixels with clean-signal prediction (Li & He, 2026). For time series, GAN- and diffusion-based generators (Yoon et al., 2019; Yuan & Qiao, 2024) employ generic inductive biases without encoding domain-specific physical constraints. Constrained generative models address general or inequality-defined feasible sets through sampling-time interventions (Christopher et al., 2024) or modified diffusion processes (Lou & Ermon, 2023; Liu et al., 2023a). In contrast, the montage constraint is a linear equality with a fixed, closed-form projector that we apply during both training and sampling. EEG generation remains relatively underexplored: EEG-GAN (Hartmann et al., 2018) uses progressive training with a modified WGAN-GP objective, Shu et al. (2025) synthesize preictal EEG with diffusion for seizure prediction, and JET (Wang et al., 2026) uses a flow-matching transformer with auxiliary temporal and statistical losses. However, EEG-GAN shows low fidelity (Table 2). JET needs dataset-specific tuning of auxiliary-loss weights. None of them generates physically valid EEG.

8 Conclusion

We show that current EEG generators synthesize physically invalid signals. MontageFlow rules out such signals by restricting its output to the subspace of valid signals. MontageFlow further conditions the flow matching decoder on a physiological state vector. With latent conditioning, enforcing validity during training also improves fidelity. Across three publicly available large-scale EEG corpora, MontageFlow improves TS-FID by 1.7–$2.5\times$ over the strongest baseline with perfect physical validity. MontageFlow is applicable to other electrical recordings such as ECG though we restrict the scope of our study to EEG. More broadly, our results suggest that validity and fidelity need not trade off when known physical laws are built into the generator.

AI USE STATEMENT

In this work, we used generative AI tools for AI-assisted coding, drawing the diagrams and polishing the writing. Additionally, we used generative AI tools for finding related works and writing the latex code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Arshia Afzal, Grigorios Chrysos, Volkan Cevher, and Mahsa Shoaran. REST: Efficient and accelerated EEG seizure analysis through residual state updates. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 271–290. PMLR, 2024. URL https://proceedings.mlr.press/v235/afzal24a.html.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 89307–89333, 2024. doi: 10.52202/079017-2834. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a28af221f2f70be183afc16797a56b91-Abstract-Conference.html.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG - adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ZeFMtRBy4Z.

Lukas A. W. Gemein, Robin T. Schirrmeister, Patryk Chrabaszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220:117021, October 2020. doi: 10.1016/j.neuroimage.2020.117021. URL https://doi.org/10.1016/j.neuroimage.2020.117021.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F. Struck, Aline Herlopian, Ioannis Karakis, Jonathan J. Halford, Marcus C. Ng, Emily L. Johnson, Brian L. Appavu, Rani A. Sarkis, Gamaleldin Osman, Peter W. Kaplan, Monica B. Dhakar, Lakshman Arcot Jayagopal, Zubeda Sheikh, Olga Taraschenko, Sarah Schmitt, Hiba A. Haider, Jennifer A. Kim, Christa B. Swisher, Nicolas Gaspard, Mackenzie C. Cervenka, Andres A. Rodriguez Ruiz, Jong Woo Lee, Mohammad Tabaeizadeh, Emily J. Gilmore, Kristy Nordstrom, Ji Yeoun Yoo, Manisha G. Holmes, Susan T. Herman, Jennifer A. Williams, Jay Pathmanathan, Fábio A. Nascimento, Ziwei Fan, Samaneh Nasiri, Mouhsin M. Shafi, Sydney S. Cash, Daniel B. Hoch, Andrew J. Cole, Eric S. Rosenthal, Sahar F. Zafar, Jimeng Sun, and M. Brandon Westover. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17):e1750–e1762, 2023. doi: 10.1212/WNL.0000000000207127. URL https://doi.org/10.1212/WNL.0000000000207127.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405. IEEE, 2019. doi: 10.1109/CVPR.2019.00453. URL http://dx.doi.org/10.1109/CVPR.2019.00453.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 65484–65516. Curran Associates, Inc., 2023. doi: 10.52202/075280-2858. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ce79fbf9baef726645bc2337abb0ade2-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, June 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.653659. URL http://dx.doi.org/10.3389/fnhum.2021.653659.

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

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Back_to_Basics_Let_Denoising_Generative_Models_Denoise_CVPR_2026_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou, and Molei Tao. Mirror diffusion models for constrained and watermarked generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42898–42917, 2023a. doi: 10.52202/075280-1858. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/85f5c7372625d1e0df0e3996f85062d6-Abstract-Conference.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=XVjTT1nw5z.

Aaron Lou and Stefano Ermon. Reflected diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22675–22701. PMLR, 2023. URL https://proceedings.mlr.press/v202/lou23a.html.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, May 2016. doi: 10.3389/fnins.2016.00196. URL https://doi.org/10.3389/fnins.2016.00196.

Damian Pascual, Alireza Amirshahi, Amir Aminifar, David Atienza, Philippe Ryvlin, and Roger Wattenhofer. EpilepsyGAN: Synthetic epileptic brain activities with privacy preservation. *IEEE Transactions on Biomedical Engineering*, 68(8):2435–2446, August 2021. ISSN 1558-2531. doi: 10.1109/TBME.2020.3042574. URL https://doi.org/10.1109/TBME.2020.3042574.

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, October 2023.

Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on EEG: a systematic comparison. *Journal of Neural Engineering*, 19(6):066020, November 2022. doi: 10.1088/1741-2552/aca220. URL https://doi.org/10.1088/1741-2552/aca220.

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

Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, November 2018. doi: 10.3389/fninf.2018.00083. URL https://doi.org/10.3389/fninf.2018.00083.

Kai Shu, Le Wu, Yuchang Zhao, Aiping Liu, Ruobing Qian, and Xun Chen. Data augmentation for seizure prediction with generative diffusion model. *IEEE Transactions on Cognitive and Developmental Systems*, 17(3):577–591, June 2025. ISSN 2379-8939. doi: 10.1109/tcds.2024.3489357. URL http://dx.doi.org/10.1109/TCDS.2024.3489357.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

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://openreview.net/forum?id=PxTIG12RRHS.

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://doi.org/10.1109/TNSRE.2022.3230250.

Siyi Tang, Jared Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=k9bx1EfHI_-.

W. O. Tatum, G. Rubboli, P. W. Kaplan, S. M. Mirsattari, K. Radhakrishnan, D. Gloss, L. O. Caboclo, F. W. Drislane, M. Koutroumanidis, D. L. Schomer, D. Kasteleijn-Nolst Trenite, Mark Cook, and S. Beniczky. Clinical utility of EEG in diagnosing and monitoring epilepsy in adults. *Clinical Neurophysiology*, 129(5):1056–1082, May 2018. doi: 10.1016/j.clinph.2018.01.019. URL https://doi.org/10.1016/j.clinph.2018.01.019.

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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Yifan Wang, Yijia Ma, Wen Li, and Chenyu You. Let EEG models learn EEG. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=TP8OuKKmsf.

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, pp. 78240–78260, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4h1apFjO99.

Appendix

A METRIC DEFINITIONS

TS-FID, shape correlation and the physiological metrics compare two finite sets of windows, so even two sets of real windows do not score perfectly. Tables 2, 3 and 6 therefore also report the *real-versus-real* value, computed between two disjoint blocks of real windows from the same split. The montage index needs no such reference. Real recordings score 0.

TS-FID. The Fréchet distance between Gaussians fitted to the features $\phi(X)$ of real and generated windows,

$$\text{TS-FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\big),$$

with (μ, Σ) the mean and covariance of the features over $n = 2{,}000$ windows. Following Wang et al. (2026), ϕ takes the magnitude of the discrete Fourier transform along time and keeps the lower half of the frequency bins. It average-pools frequency to at most 256 bins and channels to 4 groups, applies $\log(1 + \cdot)$, and standardises within the window.

Montage index. Each constraint in Eqs. 2–3 says that two chains of channels between the same pair of electrodes must sum to the same signal. The *residual* is the difference between the two sums. Within each window, we average its absolute value over time and divide by the root-mean-square of the $m = 8$ channels involved. We average this over windows and both constraints, then divide by $\sqrt{2m/\pi} \approx 2.257$, its value for independent Gaussian channels. So 0 is physically valid and 1 is typical of independent channels.

Shape correlation. We average the 16×16 per-window Pearson correlation matrix of the channels over windows. Shape correlation is the Pearson correlation between the off-diagonal entries of this matrix for generated and for real windows. It asks whether the same channel pairs are coupled.

Physiological metrics. Each metric in Table 6 is a per-window scalar $g(X)$. Its distributions over generated and real windows are compared by the 1-Wasserstein distance, divided by the interquartile range of the real values (lower is better),

$$d(g) = W_1\big(P_g^{\text{gen}},\, P_g^{\text{real}}\big) \,/\, \text{IQR}\big(P_g^{\text{real}}\big).$$

PSD slope. The exponent χ of a least-squares fit $\log_{10} S(f) \approx a - \chi \log_{10} f$ to each channel's Welch spectrum over 1–40 Hz, averaged over channels. A signal that is too smooth has a steeper slope.

Envelope. The $\log_{10}$ RMS over time of $e(t) = \big(\frac{1}{C}\sum_c x_c(t)^2\big)^{1/2}$. It measures how much energy a window carries.

Hjorth activity. $\frac{1}{C}\sum_c \log_{10} \text{var}(x_c)$, the per-channel counterpart of the envelope.

Hjorth mobility. $\frac{1}{C}\sum_c \sqrt{\text{var}(\dot{x}_c)/\text{var}(x_c)}$, with $\dot{x}_c$ the first difference. It is scale-free, so it reflects frequency content rather than amplitude.

Drift. The least-squares slope of $e(t)$ across the window, multiplied by L and divided by the mean of $e(t)$. It detects windows whose amplitude ramps up or decays.

B QUALITATIVE SAMPLES

Real EEG mixes an ongoing background rhythm with brief sharp events. MontageFlow reproduces the rhythm on all three corpora. On TUSZ its traces are visibly smoother than real. This is the time-domain form of the high-frequency attenuation seen in Figure 4. The baselines fail differently. Vanilla diffusion and vanilla flow matching are far too quiet. JET's traces are over-regular: they rise and fall with a fixed rhythm.

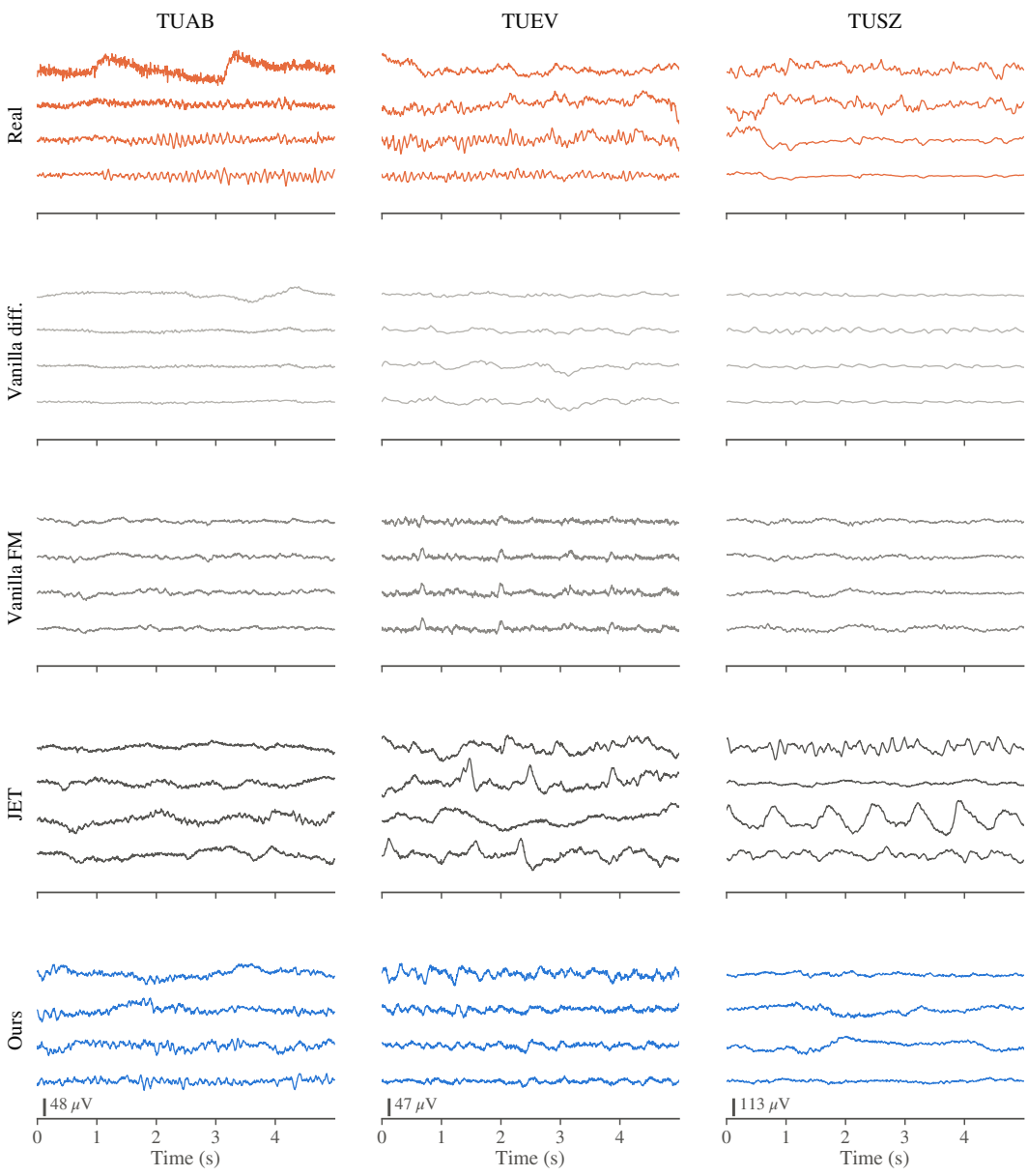

Figure 5: **Generated and real windows.** First four channels of the bipolar montage, five seconds, one window per corpus and method.

C CHANNEL-WISE ANALYSIS

Figure 4 averages over the sixteen channels. Figures 6, 7 and 8 show the same three analyses for each channel separately, so an error on a few channels cannot hide in the average. Each curve is a median over the 2,000 windows of that channel.

(1) Power spectrum: Figure 6 shows the power spectrum of each channel. Real EEG has the most alpha power on P3–O1 and P4–O2 and the weakest alpha peak on Fp1–F7, Fp2–F8, Fp1–F3 and Fp2–F4. MontageFlow shows the same pattern on all three corpora. Vanilla flow matching has less power at almost every frequency, and on TUEV it exaggerates the alpha peak on every channel. Above 12 Hz, our spectrum falls below the real one on every channel but stays within the real shading.

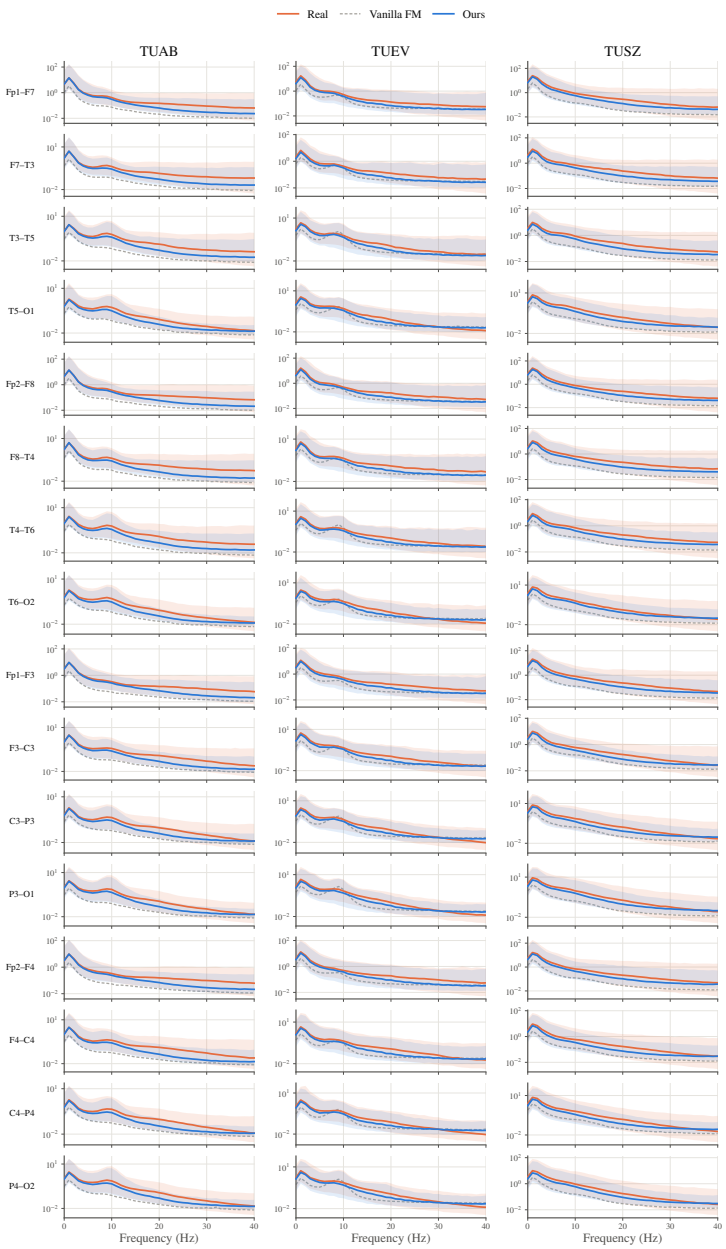

Figure 6: **Channel-wise power spectrum.** Median over windows for each channel, with shading from the 10th to the 90th percentile.

(2) Amplitude over time: Figure 7 tracks the amplitude of each channel across the window. Every envelope is flat, for real EEG and for MontageFlow, so no channel drifts. Real EEG has the largest amplitude on Fp1–F7, Fp2–F8, Fp1–F3 and Fp2–F4 and smallest on T6–O2, C3–P3 and C4–P4. MontageFlow ranks the channels the same way on all three corpora, at 81–95% of the real level. Vanilla flow matching stays at 44–72%.

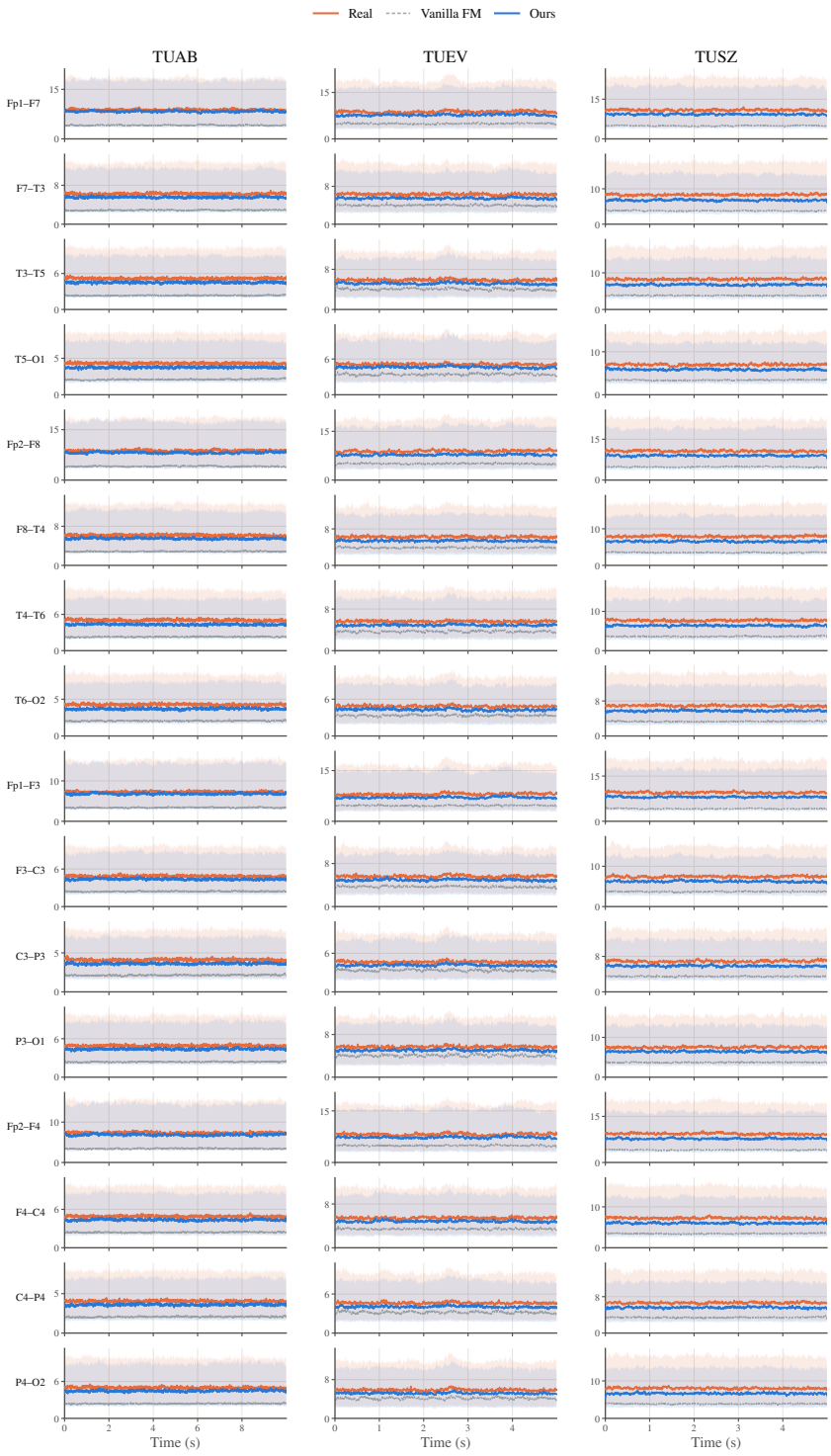

Figure 7: **Channel-wise amplitude over time.** Absolute amplitude of each channel, median over windows, with shading over the middle half of windows.

(3) Heavy-tailed distributions: Figure 8 shows the amplitude distribution of each channel. Our density follows the real curve from the sharp central peak down through the tails on every channel. On TUAB and TUEV, real EEG and MontageFlow both have their heaviest tails on Fp1–F7, Fp2–F8, Fp1–F3 and Fp2–F4. Above 100 μV, our samples keep 43–94% of the real probability mass per channel, against at most 18% for vanilla flow matching.

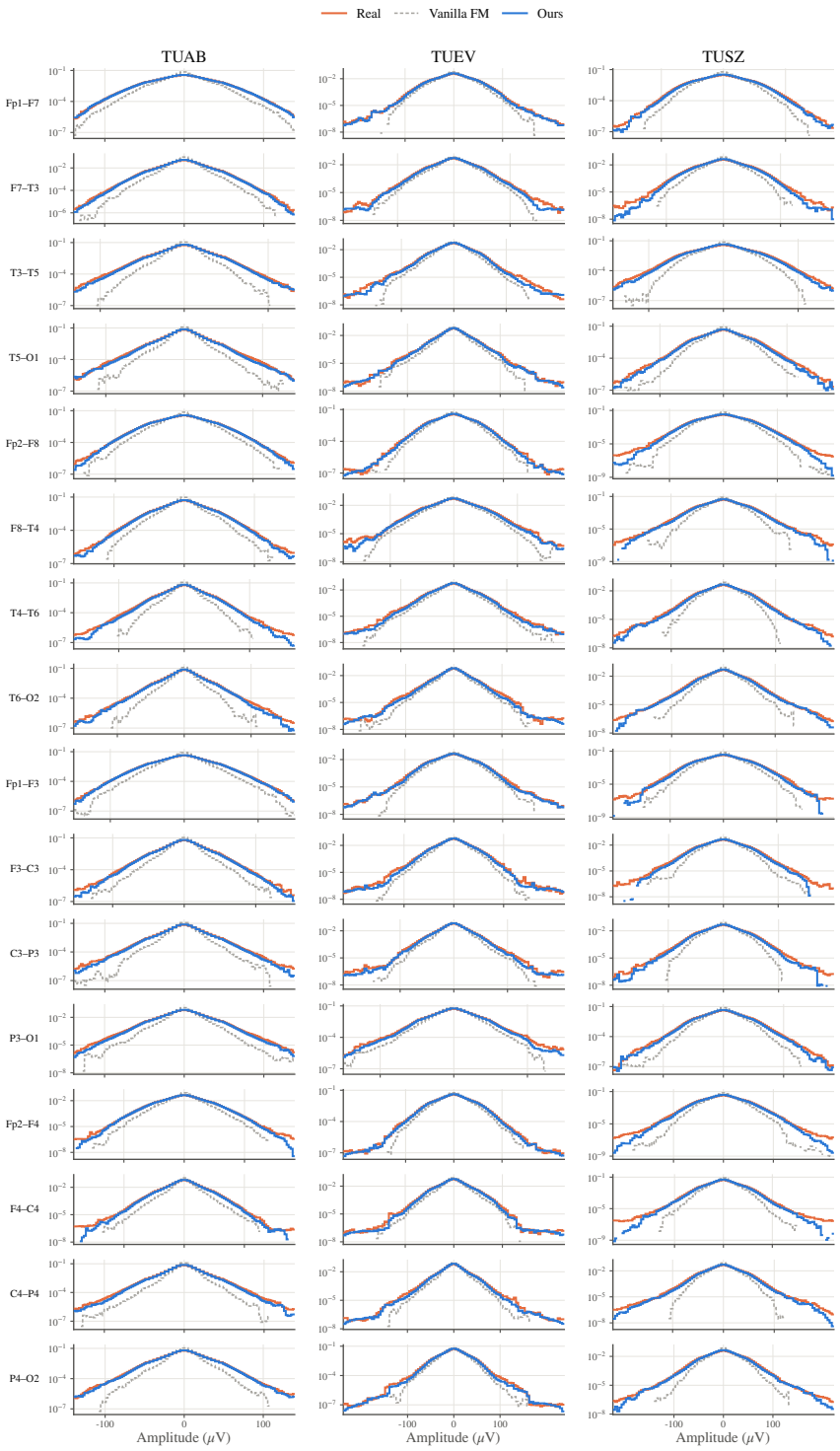

Figure 8: **Channel-wise amplitude distribution.** Distribution of sample amplitudes for each channel (log density).