ONE-STEP LATENT-FREE EEG GENERATION WITH MEAN-FIELD FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG datasets remain limited in size and diversity, leaving substantial waveform variation across subjects, sessions, and recording conditions underrepresented in training. Generative modeling offers a way to address this limitation, but the standard approaches for image or language generation may not transfer directly to EEG. Latent compression may discard amplitude structure, phase relationships, and transient events. Standard self-attention provides no explicit prior for shared cortical sources underlying cross-electrode correlations. Multi-step samplers require repeated inference, increasing the generation cost. In this work, we propose *Mean-Field Flow* (**MFF**), a one-step, latent-free method for raw EEG generation. MFF operates directly on raw EEG and uses mean-field token coupling to model channel and temporal interactions through separate shared low-dimensional fields. We train MFF with an average-velocity objective via ℓ_1 regression, reducing sampling to a single forward pass. Across 14 public EEG benchmark datasets, MFF achieves the lowest average TS-FID, reducing it by 27.8% relative to the strongest baseline. It also improves downstream brain-computer interface (BCI) classification accuracy by 2.80 percentage points on average, and its generated samples more closely match real EEG in Hjorth parameters, skewness, and scalp topographies. We hope that this study will motivate broader exploration of generative modeling for electrophysiological signals.

1 INTRODUCTION

Electroencephalography (EEG) modeling for brain-computer interface (BCI) and clinical tasks is constrained by the scale and diversity of labeled recordings (Jiang et al., 2024b; Wang et al., 2024a;b; Yang et al., 2023; Zhou et al., 2025). EEG recordings are distributed across acquisition devices, electrode montages, cohorts, sessions, experimental paradigms, and annotation schemes. Large clinical EEG corpora emphasize abnormal activity such as epileptiform discharges and seizures under hospital-specific recording conditions (Obeid & Picone, 2016; Shoeb, 2009), whereas motor imagery and affective datasets are typically collected under narrower but more controlled protocols (Brunner et al., 2008; Chen et al., 2023; Schalk et al., 2004; Zheng & Lu, 2015). Labels are expensive to obtain, inter-subject variation is substantial, and clinically meaningful events are rare. The bottleneck in downstream EEG modeling is therefore not only network architecture or learning objective, but also the availability of labeled recordings that span the waveform diversity within each task.

Generative modeling offers a way to address this limitation, but the standard approaches for image or language generation may not transfer directly to EEG. In EEG, the waveform itself is often the signal of interest, and amplitude, phase relationships, and transient waveforms can all matter for downstream analysis. Latent autoencoders and tokenizers, which are standard in image and sequence generation (Chang et al., 2022; Rombach et al., 2021; van den Oord et al., 2018), may discard this signal structure. EEG channels are not interchangeable tokens: signals at different electrodes are spatial mixtures of shared neural sources, and cross-electrode correlations are shaped in part by volume conduction (Khanna et al., 2015; Nunez & Srinivasan, 2006). Standard self-attention (Vaswani et al., 2017) provides limited inductive bias for this structure. Multi-step samplers (Ho et al., 2020; Lipman et al., 2022; Liu et al., 2022) generate samples through repeated network evaluations, increasing sampling cost. These observations suggest that EEG generation may benefit from *operating directly on raw EEG waveforms, modeling cross-electrode structure explicitly, and avoiding multi-step sampling*.

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.

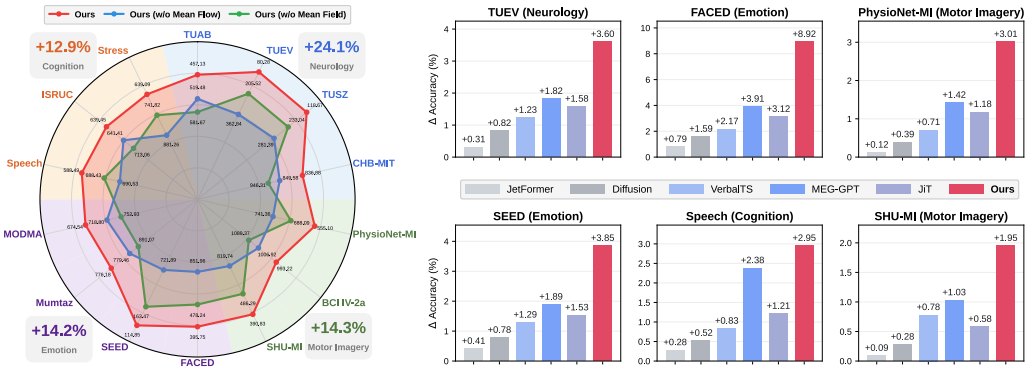


Figure 1: **Mean-Field Flow enables one-step EEG generation across heterogeneous tasks.** Left: TS-FID across 14 datasets; percentages show reductions in task-family means versus w/o Mean Field. Right: downstream classification gains from augmentation on six representative tasks.

In this work, we propose *Mean-Field Flow* (**MFF**), a one-step, latent-free method for raw EEG generation. MFF is trained separately on each dataset and conditioned on its task labels. It operates directly in raw signal space and models channel and temporal interactions through mean-field token coupling. Channel and temporal tokens interact through separate shared low-dimensional fields, instead of pairwise self-attention. Standard flow matching learns instantaneous velocity, whereas MFF uses an average-velocity objective over a finite time interval (Lipman et al., 2022; Geng et al., 2025a). This enables generation in a single forward pass from noise to an EEG sample.

We evaluate our MFF on 14 public EEG datasets spanning neurology, motor imagery, emotion, and cognition, as shown in Figure 1. MFF achieves the lowest TS-FID on all 14 datasets, with a 27.8% reduction in mean TS-FID relative to the strongest external baseline. Synthetic EEG from MFF also improves downstream BCI classification accuracy by 2.80 percentage points on average. Beyond TS-FID, MFF more closely matches real EEG in Hjorth parameters, per-channel skewness, and class-conditional scalp topographies. Its one-step sampler also reduces inference cost relative to multi-step generation. These results show that mean-field token coupling improves both signal fidelity and downstream utility in one-step raw EEG generation. Our contributions are threefold:

- We propose *Mean-Field Flow*, a one-step, latent-free method for raw EEG generation. MFF combines separate shared fields for channel and temporal interactions with an ℓ_1 average-velocity objective for one-step generation directly in EEG signal space.
- We evaluate MFF on 14 public EEG datasets spanning neurology, motor imagery, emotion, and cognition. Against seven baselines and two controlled variants, MFF achieves the lowest TS-FID on all 14 datasets and yields the largest average gain in downstream classification accuracy.
- We assess generated EEG beyond TS-FID. Hjorth statistics, skewness, scalp maps, phase locking, and motor-imagery responses show MFF better matches real EEG signal structure.

2 Related Work

Generative Modeling. Generative models for high-dimensional signals differ along three dimensions: whether generation occurs in raw signal space or compressed latent space, whether variables are continuous or discrete, and how many sampling steps they use. Early one-step models, including generative adversarial networks and non-adversarial moment-matching methods, generate raw continuous samples in a single pass (Brock et al., 2019; Goodfellow et al., 2014; Gulrajani et al., 2017; Karras et al., 2019; Li et al., 2015). Diffusion and score-based models later established stable multi-step generation in raw signal space (Bansal et al., 2023; Daras et al., 2023; Ho et al., 2020; Karras et al., 2022; Kingma et al., 2021; Nichol & Dhariwal, 2021; Song et al., 2021a;b). Flow matching, stochastic interpolants, and related continuous-time transport formulations extended this line with more flexible paths between source and target distributions (Albergo et al., 2023; Davis et al., 2024; Lipman et al., 2022; Liu et al., 2022; Ma et al., 2024; Papamakarios et al., 2021).

To scale these models to higher resolutions, recent work has moved generation to a compressed latent space (Brooks et al., 2024; Peebles & Xie, 2023; Ramesh et al., 2022; Rombach et al., 2021; Zhang

et al., 2023b; Zheng et al., 2025). This reduces the dimensionality of the generative problem, but places more burden on the quality of the tokenizer or autoencoder. A parallel line of work discretizes the signal into tokens and models generation autoregressively or with masked decoding (Chang et al., 2022; van den Oord et al., 2018). These methods retain a multi-step generation process, while replacing continuous regression with discrete prediction. Recent work has begun to move away from these defaults. Consistency models and related one-step methods reduce sampling to one or a few steps (Geng et al., 2025a; Liu et al., 2024; Luo et al., 2023; Song et al., 2023a). Latent-free transformers remove the tokenizer and generate directly in pixel space (Li & He, 2025). For continuous multi-channel signals such as EEG, natural discretizations are unclear and strong domain-specific tokenizers are not readily available. We therefore focus on raw-space generation with continuous variables and one-step sampling.

EEG Modeling. Recent work on EEG modeling has focused mainly on representation learning rather than generation. EEG models pretrain large encoders on heterogeneous corpora of EEG recordings and transfer them to downstream brain–computer interface (BCI) tasks (Chen et al., 2024b; 2025; Cui et al., 2024; Jiang et al., 2024a;b; Klein et al., 2025; Kostas et al., 2021; Wang et al., 2024a;b; 2025; Wu et al., 2022; Xiao et al., 2025; Yang et al., 2023; Zhou et al., 2025); related work on intracranial recordings applies the same strategy at finer spatial scales (Wang et al., 2023; Zhang et al., 2023a). Generative EEG modeling uses GANs (Hartmann et al., 2018), VAEs, and diffusion models (Fuhrmeister et al., 2025; Liang et al., 2021; Puah et al., 2025; Shu et al., 2024). JET generates EEG using flow matching with signal-structure constraints (Wang et al., 2026). MFF combines mean-field token coupling with an average-velocity objective for one-step raw EEG generation, evaluating both signal fidelity and downstream utility across 14 public EEG datasets. A separate line of work uses EEG as a condition to generate other modalities, including images (Bai et al., 2023; Chen et al., 2024a; Deng et al., 2025; Kavasidis et al., 2017; Lan et al., 2023; Liu et al., 2025a;b; Song et al., 2023b), text (Duan et al., 2023), and speech (Lee et al., 2023). These methods do not address the generative modeling of EEG itself.

3 METHOD

We present Mean-Field Flow, a one-step, latent-free method for raw EEG generation, following flow matching and mean-flow extensions (Geng et al., 2025a; Lipman et al., 2022; Liu et al., 2022; Lu et al., 2026). MFF uses mean-field token coupling in place of self-attention (Vaswani et al., 2017) along the channel and temporal axes, so that all tokens interact through separate shared low-dimensional fields. MFF is trained to predict an average velocity, defined as the displacement over a finite time interval divided by the interval length. This formulation reduces multi-step inference to a single forward pass. We first review flow matching in Sec. 3.1, then introduce the mean-field approximation and the token coupling in Secs. 3.2 and 3.3, and derive the Mean-Field Flow objective in Sec. 3.4.

3.1 BACKGROUND: FLOW MATCHING

We represent each EEG segment as a real-valued tensor $x \in \mathbb{R}^{C\times P\times S}$, where C is the number of electrode channels, P is the number of temporal patches, and S is the number of raw samples per patch. Given a class label y, we aim to learn a transport from Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to the conditional data distribution $p_{\text{data}}(x \mid y)$. Following standard flow matching with a linear interpolant (Lipman et al., 2022; Liu et al., 2022), we define:

$$z_t = (1-t)\,x + t\,\epsilon, \qquad t \in [0,1], \tag{1}$$

so that $z_0 = x$ and $z_1 = \epsilon$. The corresponding conditional velocity is:

$$v_t := \frac{dz_t}{dt} = \epsilon - x. \tag{2}$$

For this interpolant, the conditional velocity is constant along the path. A flow-matching network $v_\theta(z_t, t, y)$ is trained by minimizing the conditional regression loss:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{x,y,\epsilon,t}\,\|v_\theta(z_t,t,y) - v_t\|_2^2. \tag{3}$$

As in standard flow matching, minimizing this loss is equivalent to fitting the marginal velocity field induced by the path (Lipman et al., 2022). At inference time, samples are generated by integrating

the learned velocity field from $t = 1$ to $t = 0$, which requires multiple sampling steps. MFF keeps the interpolant in Eq. 1. The change is in the prediction target. Instead of the instantaneous velocity, MFF models the average velocity over a finite time interval, which is derived in Sec. 3.4.

3.2 Mean-Field Approximation

EEG channels are coupled rather than independent. At a fixed time, scalp electrodes record spatial mixtures of shared neural sources through volume conduction (Nunez & Srinivasan, 2006). EEG also exhibits recurring large-scale states, as studied in neural-mass models and microstate analysis (Breakspear, 2017; Friston et al., 2003; Jansen & Rit, 1995; Kaplan et al., 2005; Khanna et al., 2015; Wilson & Cowan, 1972). The channel vector can therefore be viewed as a readout of a smaller set of latent population states. This motivates a token interaction that is mediated by a shared field rather than by dense pairwise interactions. Here we formalize this intuition with a classical mean-field approximation. Consider M interacting scalar units $\sigma_1, \ldots, \sigma_M$ with Hamiltonian:

$$H(\sigma) = -\frac{1}{2}\sum_{i\neq j} J_{ij}\,\sigma_i\sigma_j \;-\; \sum_{i=1}^{M} h_i\,\sigma_i, \qquad J_{ij} = J_{ji}, \quad J_{ii} = 0, \tag{4}$$

where J_{ij} is the pairwise coupling and h_i is an external field on unit i. Direct evaluation of H requires all $\binom{M}{2}$ pairwise interactions.

Definition 1 (Mean-field)**.** Let $m_j = \mathbb{E}[\sigma_j]$ denote the mean of unit j. The mean field at i is

$$\eta_i(m) \;:=\; h_i \;+\; \sum_{j\neq i} J_{ij}\,m_j. \tag{5}$$

We decouple Eq. 4 using the first-order mean-field approximation (Parisi & Shankar, 1988),

$$\sigma_i\sigma_j \;\approx\; \sigma_i m_j + m_i\sigma_j - m_i m_j. \tag{6}$$

Substituting Eq. 6 into Eq. 4 gives the following decoupled form.

Lemma 1 (Mean-field factorization)**.** Under the substitution Eq. 6, the Hamiltonian becomes

$$H_{\mathrm{MF}}(\sigma; m) = -\sum_{i=1}^{M} \eta_i(m)\,\sigma_i \;+\; \frac{1}{2}\sum_{i\neq j} J_{ij}\,m_i\,m_j. \tag{7}$$

The first term in Eq. 7 is linear in each σ_i and decoupled across units. The second term is constant with respect to σ. The Gibbs partition function therefore factorizes as:

$$Z_{\mathrm{MF}}(m) = \exp\left(-\frac{\beta}{2}\sum_{i\neq j} J_{ij}\,m_i\,m_j\right)\cdot\prod_{i=1}^{M} Z_i(m), \quad Z_i(m) = \int \exp(\beta\,\eta_i(m)\,\sigma_i)\,d\mu(\sigma_i), \tag{8}$$

where $\beta > 0$ is the inverse temperature and μ is the single-unit base measure. The pairwise system is thus reduced to M single-unit responses coupled only through the shared field $\eta(m)$. This factorization motivates shared-field architectures that combine global cross-token information with token-specific local responses:

Theorem 1 (Mean-field decoupling)**.** Let $f : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ be a token-mixing layer given by

$$f(\mathbf{X})_n \;=\; g(\mathbf{X}_n,\; c(\mathbf{X}))\,, \qquad n = 1, \ldots, N, \tag{9}$$

where $c : \mathbb{R}^{N\times D} \to \mathbb{R}^{D_c}$ is permutation-invariant over the token dimension, and $g : \mathbb{R}^{D} \times \mathbb{R}^{D_c} \to \mathbb{R}^{D}$ is shared across tokens. Then, conditioned on the shared field $c(\mathbf{X})$, the layer is token-wise decoupled: each output $f(\mathbf{X})_n$ depends on other tokens only through $c(\mathbf{X})$.

Theorem 1 shows that once the shared field $c(\mathbf{X})$ is computed, each token is updated independently by the same function g. The shared field $c(\mathbf{X})$ summarizes cross-token information, and g combines this summary with each token's local features. The implementation in Sec. 3.3 scales linearly with token count for fixed feature and field dimensions.

3.3 Mean-Field Token Coupling

We now instantiate the mean-field form of Theorem 1 as a token-mixing layer, termed *mean-field token coupling*. Consider the tokens along one axis of the EEG tensor, collected as $\mathbf{X} \in \mathbb{R}^{N \times D}$. After layer normalization (Ba et al., 2016), each token is projected to a D_c-dimensional field coordinate:

$$\mathbf{Z}_n \;=\; W_2\, \phi(W_1 \widetilde{\mathbf{X}}_n), \qquad n = 1, \ldots, N, \tag{10}$$

where ϕ is the GELU activation (Hendrycks & Gimpel, 2016). The shared field is obtained by aggregating these coordinates across tokens. For each field dimension k, we define:

$$\pi_{n,k} = \frac{\exp\left(\mathbf{Z}_{n,k}\right)}{\sum_{n'=1}^{N} \exp\left(\mathbf{Z}_{n',k}\right)}, \quad \mathbf{c}_k \;=\; \sum_{n=1}^{N} \pi_{n,k}\, \mathbf{Z}_{n,k}, \qquad k = 1, \ldots, D_c, \tag{11}$$

which gives a field vector $\mathbf{c} \in \mathbb{R}^{D_c}$ with each coordinate being a softmax-weighted summary across tokens. This aggregation is permutation-invariant in the token index. Each token is then updated using its own feature together with the shared field:

$$\mathbf{Y}_n \;=\; W_4\, \phi\Big(W_3[\widetilde{\mathbf{X}}_n, \mathbf{c}]\Big), \qquad n = 1, \ldots, N, \tag{12}$$

where $[\cdot, \cdot]$ denotes feature-wise concatenation. This layer has the form in Theorem 1: once $\mathbf{c}$ is computed, each output $\mathbf{Y}_n$ depends on other tokens only through the shared field. We apply the same construction along both axes of the EEG tensor, following dual-axis designs for multivariate time-series modeling (Liu et al., 2023; Yu et al., 2026). Along channels, the C montage-ordered electrodes with 1-D channel embeddings are tokens within each patch; along time, the P position-embedded patches are tokens within each channel. The two branches share the same form but use different parameters, and their outputs are summed before the prediction head.

3.4 Mean-Field Flow Objective

Average velocity. Following mean-flow models and their raw-space extensions (Geng et al., 2025a;b; Lu et al., 2026), we replace the instantaneous velocity target in Eq. 3 with an average-velocity target. For $0 \le r < t \le 1$, we define the average velocity from t to r as:

$$\bar{u}^{\star}(z_t, t, r) \;:=\; \frac{1}{t-r} \int_r^t v(z_s, s)\, ds, \tag{13}$$

which gives:

$$z_r \;=\; z_t - (t-r)\, \bar{u}^{\star}(z_t, t, r). \tag{14}$$

At $(t, r) = (1, 0)$, this yields one-step sampling from Gaussian noise to raw EEG:

$$z_0 \;=\; \epsilon - \hat{u}_\theta(\epsilon, 1, y). \tag{15}$$

Material Identity. Differentiating Eq. 13 with respect to t, while keeping r fixed, gives:

$$v(z_t, t) \;=\; \bar{u}^{\star}(z_t, t, r) \;+\; (t-r)\, D_t\, \bar{u}^{\star}(z_t, t, r), \tag{16}$$

where $D_t := \partial_t + v(z_t, t)^\top \nabla_z$ is the material derivative along the path. This identity relates the instantaneous velocity to the average velocity and its derivative along the flow.

Training objective. We parameterize the average velocity with a network $\hat{u}_\theta(z, t, y)$. For a sampled pair (t, r), we query the same predictor twice at z_t, using t or r as the scalar time input:

$$u_t := \hat{u}_\theta(z_t, t, y), \qquad u_r := \hat{u}_\theta(z_t, r, y).$$

For the mean-flow target, we replace the material derivative term in Eq. 16 with a fixed-state bootstrap, subtracting the detached difference $u_t - u_r$ from the instantaneous velocity v_t at z_t:

$$u_{\text{tgt}} \;=\; v_t \;-\; \text{sg}[u_t - u_r]\,, \tag{17}$$

Algorithm 1 MFF training.

```
# fn(z, t, y): MFF network
# x: training batch; y: class label

t, r = sample_t_r()
e = randn_like(x)

z = (1 - t) * x + t * e
v = e - x

u_t = fn(z, t, y)
u_r = fn(z, r, y) if t > r else u_t

u_tgt = v - stopgrad(u_t - u_r)
loss = mean_abs(u_t - u_tgt)
```

Algorithm 2 MFF one-step sampling.

```
e = randn(x_shape)
x = e - fn(e, t=1, y)
```

where $\mathrm{sg}[\cdot]$ denotes stop-gradient and $v_t = \epsilon - x$. We then minimize

$$\mathcal{L}_{\mathrm{MFF}}(\theta) \;=\; \mathbb{E}_{x,y,\epsilon,t,r} \left\| \hat{u}_\theta(z_t, t, y) - u_{\mathrm{tgt}} \right\|_1. \tag{18}$$

Loss metric and sampling. We use ℓ_1 rather than ℓ_2. EEG amplitudes are often heavy-tailed, with sparse large excursions over a lower-amplitude background (Buzsáki & Mizuseki, 2014; Roberts, 2000). Compared with squared ℓ_2, the ℓ_1 loss is less dominated by large residuals. Appendix C compares exact ℓ_1, squared ℓ_2, and Huber objectives under matched settings; ℓ_1 gives the best reported TS-FID, skewness KS, and Hjorth-complexity distance. During training, we sample (t, r) with $r \leq t$ and set $r = t$ for a fixed fraction of steps. In that case, the bootstrap term vanishes and the target reduces to the instantaneous velocity v_t. Algorithm 1 summarizes training and Algorithm 2 summarizes one-step sampling. Details of the time-pair sampler, the dual-head training scheme, and the sampled-guidance variant are deferred to Appendix A. The resulting formulation keeps the flow-matching interpolant while requiring only one network evaluation at inference.

4 Experiments

We evaluate our MFF on 14 public EEG datasets spanning neurology, motor imagery, emotion, and cognition. We describe the benchmark, baselines, and evaluation protocol (Sec. 4.1). We then report distributional fidelity via TS-FID (Sec. 4.2), downstream classification gains from augmentation with synthetic EEG (Sec. 4.3), and EEG-specific statistics and scalp topographies (Sec. 4.4). Finally, we ablate the mean-field token coupling and the mean-flow objective (Sec. 4.5).

Dataset	C	L (s)	N	K
TUAB (Obeid & Picone, 2016)	16	10	409 K	2
TUEV (Obeid & Picone, 2016)	16	5	112 K	6
TUSZ (Obeid & Picone, 2016)	16	5	92 K	2
CHB-MIT (Shoeb, 2009)	16	10	127 K	2
PhysioNet-MI (Schalk et al., 2004)	64	4	9.8 K	4
BCIC-IV-2a (Brunner et al., 2008)	22	4	2.5 K	4
SHU-MI (Ma et al., 2022)	32	4	12 K	2
FACED (Chen et al., 2023)	32	10	10 K	9
SEED (Zheng & Lu, 2015)	62	1	153 K	3
Mumtaz (Mumtaz, 2016)	19	5	7.1 K	2
MODMA (Cai et al., 2022)	128	5	4.0 K	2
Speech (Jeong et al., 2022)	64	3	5.3 K	5
ISRUC (Khalighi et al., 2016)	6	30	78 K	5
Stress (Zyma et al., 2019)	20	5	1.7 K	2

Table 1: Summary of benchmarks. C: channels, L: segment length, N: sample size, K: classes.

4.1 Experimental Setup

Benchmarks. We evaluate on 14 public EEG datasets across four task domains: neurology (4), motor imagery (3), emotion (4), and cognition (3). As shown in Table 1, the datasets vary widely in channel, length, sample size, and number of classes.

Baselines. We compare MFF against seven baselines: GAN and autoregressive models (Hartmann et al., 2018; Tschannen et al., 2024), diffusion-based models (Gu et al., 2025; Ho et al., 2020), autoregressive models developed for MEG (Csaky et al., 2024; Huang et al., 2025), and a latent-free pixel-space model (Li & He, 2025). All baselines are adapted to class-conditional EEG generation, retrained on the official train split, and evaluated with the same preprocessing and generated-sample budget for each dataset. We also include two controlled variants, Ours (w/o Mean Flow) and Ours (w/o Mean Field), to isolate the contributions of *average-velocity objective* and *mean-field token coupling*. See Appendix A.6 for more details.

Evaluation. We evaluate generated EEG quality in three ways. We report TS-FID for distributional fidelity by measuring how close the synthetic distribution is to real EEG. To validate whether the synthetic data improves downstream classification, we augment CBraMod (Wang et al., 2024b)'s training set with synthetic samples and report the accuracy change ΔAcc. We also compare EEG-specific statistics and scalp topographies against real recordings. Metric definitions are in Appendix A.4.

4.2 Distributional Fidelity

We first validate the generation performance of MFF with TS-FID against seven baselines on all benchmarks, as illustrated in Table 2. Across all 14 datasets, MFF achieves the lowest TS-FID, reducing the average TS-FID from 717.97 with MEG-GPT (Huang et al., 2025), the strongest external baseline, to 518.60, a 27.8% relative reduction.

The pattern of the table suggests that the gain grows when the dataset requires more than broad spectral matching. On several coarse binary or low-variability settings, baselines remain closer, suggesting that broad spectral statistics already explain a substantial part of the distribution. On settings whose labels reflect nonstationary events or global brain states, such as epileptiform discharges, motor imagery, and emotion regulation, the gap widens substantially. The hard cases are those where the

Task	Dataset	*GAN & AR*		*Diffusion*		*MEG Transfer*		*Flow Matching*			
		EEG-GAN	JetFormer	Diffusion	VerbalTS	MEG-GPT	GPT2MEG	JiT	Ours (w/o Mean Flow)	Ours (w/o Mean Field)	**Ours**
Neurology	TUAB	670.82	658.74	647.19	544.36	538.41	513.26	562.91	519.48	581.67	**457.13**
	TUEV	486.73	477.81	454.12	425.06	386.34	346.09	399.71	362.84	205.52	**80.28**
	TUSZ	376.14	366.93	339.72	297.08	294.51	301.17	304.86	281.39	233.04	**118.67**
	CHB-MIT	1007.18	993.42	966.87	903.71	846.19	889.72	921.04	849.58	946.31	**836.88**
Motor Imagery	PhysioNet-MI	954.83	946.72	906.24	921.58	768.91	756.43	798.67	741.36	688.09	**555.10**
	BCI IV-2a	1127.05	1074.87	1101.74	1086.49	1023.81	1041.29	1058.64	1006.92	1089.37	**993.22**
	SHU-MI	1051.27	1038.45	1006.91	854.32	841.56	809.48	882.13	819.74	486.29	**390.83**
Emotion	FACED	1083.42	1081.16	1037.63	1026.03	894.71	866.25	917.48	851.96	478.24	**395.75**
	SEED	931.28	913.56	881.36	856.14	754.03	746.89	776.23	721.89	163.47	**114.85**
	Mumtaz	996.38	989.17	944.61	818.73	802.39	806.18	841.94	779.46	891.07	**776.18**
	MODMA	924.73	916.94	879.52	871.63	686.08	714.87	774.17	718.80	752.93	**674.54**
Cognition	Speech	890.52	878.83	845.29	771.43	683.47	696.31	745.09	690.53	688.43	**588.49**
	ISRUC	846.14	824.79	791.06	669.28	671.82	656.47	694.37	641.41	713.06	**639.45**
	Stress	1131.26	1114.08	1074.92	1082.37	859.31	912.41	944.63	881.26	741.82	**639.09**
Avg. TS-FID		891.27	876.82	848.37	794.87	717.97	718.34	758.71	704.76	618.52	**518.60**

Table 2: **Distributional fidelity across 14 EEG datasets.** We report TS-FID (↓) between generated and real EEG feature distributions. Methods are grouped by generative family; the final three columns are controlled variants of MFF. Lower is better. **Best**, second, and third are highlighted.

generator needs to preserve sparse high-amplitude excursions, class-dependent spatial organization, and temporal dynamics with one network evaluation. Whether this feature-space fidelity corresponds to signal-level fidelity is the question we take up in Sec. 4.4. Appendix J further compares generators under matched filtering and rereferencing pipelines on TUEV, ISRUC, and MODMA.

Takeaway: MFF achieves the lowest TS-FID on all 14 datasets, reducing average TS-FID by 27.8% over the strongest external baseline while sampling in one step.

4.3 Synthetic EEG for Downstream Classification

We next ask whether this gain in TS-FID translates into downstream classification performance. We adopt CBraMod (Wang et al., 2024b) as the BCI classifier backbone and fine-tune it under two conditions on each dataset: on the real training set, and on the real training set augmented with 20% synthetic samples drawn from each generative model, then evaluate both on the official held-out real test set and report ΔAcc with five seeds. Table 3 reports representative downstream augmentation results, while the full 14-dataset comparison is provided in Appendix B. MFF achieves the largest mean accuracy gain on all 14 datasets. Its average improvement over real-only training is 2.80 percentage points, exceeding the strongest external augmentation baseline, GPT2MEG (Csaky et al., 2024), by 1.18 percentage points. Equal-size controls with a fixed downstream update budget further show that MFF's gains are not explained by nominal data quantity alone and depend on correctly paired labels, as shown in Appendix B.3.

MFF's gain varies systematically with the structure of each dataset (Table 1). The largest gain appears on FACED (+8.92%), whose 10 K samples span 9 emotion classes: classification accuracy increases from 54.62% to 63.54%. Compared with GPT2MEG's +4.48-point improvement, MFF provides greater augmentation benefits; permuting generated labels instead decreases accuracy by 2.21 points, demonstrating the importance of class-consistent generated signals under a fixed training budget. The neurology corpora (TUAB +2.38%, TUEV +3.60%, TUSZ +2.42%) gain less because large real training sets reduce the marginal value of augmentation; here MFF's advantage comes from preserving sparse high-amplitude transients like epileptiform discharges that the labels depend on. The smallest gain appears on CHB-MIT (+0.85%), where proportional augmentation does little to correct severe class imbalance and patient-specific seizure morphology.

Takeaway: MFF gains 2.80 percentage points on average across 14 datasets; matched-budget controls show gains depend on correctly paired labels.

4.4 Statistical Analysis

In this section, we next examine the statistical structure of the generated EEG. We probe three statistics that capture temporal dynamics through Hjorth parameters, per-channel amplitude shape (skewness), and cross-channel spatial organization (scalp topographies).

Task	Dataset	*GAN & AR*		*Diffusion*		*MEG Transfer*		*Flow Matching*			
		EEG-GAN	JetFormer	Diffusion	VerbalTS	MEG-GPT	GPT2MEG	JiT	Ours (w/o Mean Flow)	Ours (w/o Mean Field)	**Ours**
Neurology	TUAB	−0.28±0.43	−0.08±0.47	+0.21±0.39	+0.61±0.38	+1.04±0.41	+1.52±0.36	+0.82±0.40	+1.73±0.37	+1.38±0.35	**+2.38±0.36**
Motor Imagery	BCI IV-2a	−0.73±0.71	−0.02±0.66	−0.18±0.63	+0.12±0.59	+0.93±0.54	+0.61±0.57	+0.42±0.53	+1.12±0.49	+0.38±0.56	**+1.55±0.52**
Emotion	FACED	+0.41±1.03	+0.79±0.97	+1.59±0.94	+2.17±0.88	+3.91±0.86	+4.48±0.91	+3.12±0.84	+5.62±0.79	+6.83±0.86	**+8.92±0.83**
Cognition	ISRUC	−0.13±0.63	+0.07±0.56	+0.31±0.54	+1.01±0.51	+1.57±0.52	+1.34±0.49	+0.82±0.48	+1.71±0.46	+1.18±0.45	**+2.28±0.48**

Table 3: **Downstream utility via ΔAcc (%, ↑) on representative tasks.** We augment real training data with additional 20% synthetic samples. Values report mean ± standard deviation across 5 random seeds, in percentage points. **Best**, second and third are highlighted.

MFF best matches real EEG temporal dynamics across task families. Table 4 reports the Wasserstein-1 distance between generated and real EEG on the three Hjorth parameters (Hjorth, 1970): activity, mobility, and complexity. We compute these statistics channel-wise on bandpass-filtered waveforms after independently normalizing real and generated trials by their median channel-wise temporal standard deviation. The empirical distributions are compared between real and generated EEG; Appendix A.4 gives details. We evaluate on one dataset per task family (TUEV, BCI IV-2a, Mumtaz, Stress). MFF obtains the lowest distance in all reported comparisons. The margin is largest on TUEV, where the complexity distance drops to 0.806 versus 3.390 for the strongest baseline; this is the dataset whose signal is dominated by sparse epileptiform transients, and reproducing their bandwidth structure is where other methods remain substantially farther from real EEG. Full results across all 14 datasets are in Appendix H.

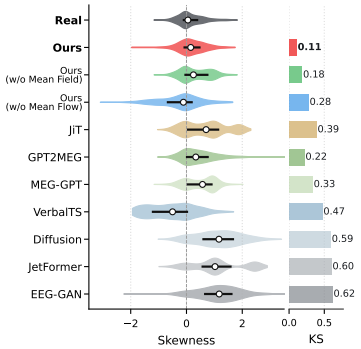


Figure 2: **Per-channel skewness on TUEV.** Left: skewness distributions. Right: KS distance.

MFF better matches the per-channel skewness of real EEG. Real EEG amplitudes are heavy-tailed, while their per-channel skewness remains concentrated near zero after preprocessing. Figure 2 compares the per-channel skewness distributions of real and generated TUEV samples; the right panel reports the Kolmogorov–Smirnov distance (Massey Jr, 1951) to real. MFF reaches KS = 0.11, 2× closer to real than the strongest baseline (GPT2MEG, 0.22) and substantially closer than the weakest baselines (EEG-GAN, 0.62). The violins reveal two distinct failure modes among the baselines: some shift the tails positive (Diffusion, JetFormer, EEG-GAN), while others compress them toward a narrower range (VerbalTS, MEG-GPT); MFF comes closest to recovering the full real distribution.

MFF reproduces class-conditional scalp topography. EEG carries diagnostic information not only in per-channel statistics but also in the spatial patterns across channels. Figure 3 shows class-averaged topographies for the six TUEV event classes, comparing real (top row) and generated (bottom row). At the class-average level, the generated samples reproduce coarse class-distinguishing spatial patterns without explicit 2-D electrode coordinates: the central focus of sporadic sharp waves (SPSW), the symmetric frontal arc of eye-movement segments (EYEM), the right-posterior focus of artifact segments (ARTF), the frontal positivity of background segments (BCKG), and the broad anterior-posterior gradient and frontal involvement that distinguish the two periodic discharge classes (GPED, PLED). The results suggest that MFF learns montage-specific organization from fixed channel order and mean-field coupling, without physical coordinate input. We further complement these qualitative topographies with correlation/PSD fidelity, waveform visualization, and measurements of band power, phase locking, and motor-imagery ERD in Appendices E and F.

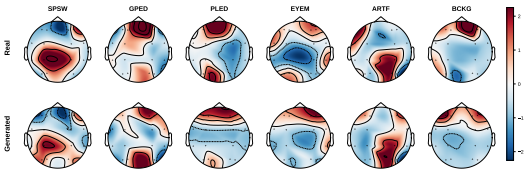


Figure 3: **Class-conditional scalp topography on TUEV.** Real (top) and MFF-generated (bottom) class-averaged topographies for the TUEV events.

Takeaway: Under the statistics evaluation, MFF best matches Hjorth dynamics, better aligns per-channel skewness distributions, and reproduces coarse class-conditional scalp topographies.

Task	Dataset	Metric	*GAN & AR*		*Diffusion*		*MEG Transfer*		*Flow Matching*			
			EEG-GAN	JetFormer	Diffusion	VerbalTS	MEG-GPT	GPT2MEG	JiT	Ours (w/o Mean Flow)	Ours (w/o Mean Field)	**Ours**
Neurology	TUEV	Act	1.113	1.175	1.169	1.091	0.950	0.935	0.888	0.923	0.443	**0.134**
		Mob	0.161	0.168	0.158	0.127	0.118	0.108	0.129	0.135	0.068	**0.026**
		Com	4.832	4.928	4.350	4.008	3.757	3.390	3.960	3.671	2.112	**0.806**
Motor Imagery	BCI IV-2a	Act	0.217	0.219	0.236	0.216	0.218	0.212	0.224	0.222	0.214	**0.202**
		Mob	0.359	0.326	0.331	0.330	0.322	0.313	0.315	0.328	0.335	**0.293**
		Com	0.640	0.626	0.646	0.656	0.622	0.615	0.620	0.630	0.608	**0.579**
Emotion	Mumtaz	Act	0.449	0.422	0.418	0.327	0.339	0.322	0.355	0.319	0.359	**0.304**
		Mob	0.403	0.400	0.350	0.301	0.289	0.279	0.311	0.275	0.335	**0.267**
		Com	1.060	0.998	0.994	0.825	0.818	0.786	0.815	0.780	0.943	**0.752**
Cognition	Stress	Act	0.453	0.436	0.460	0.456	0.321	0.362	0.343	0.347	0.262	**0.214**
		Mob	0.643	0.647	0.652	0.562	0.410	0.464	0.513	0.479	0.334	**0.275**
		Com	1.434	1.426	1.322	1.392	1.080	1.119	1.177	1.160	0.921	**0.724**

Table 4: **Hjorth-parameter distance to real EEG (↓).** We report Wasserstein-1 distances for activity (Act), mobility (Mob), and complexity (Com). **Best**, second and third are highlighted.

4.5 Ablation Study

We study two design choices in MFF: *mean-field token coupling* and *average-velocity objective*. Ours (w/o Mean Field) replaces the mean-field token coupling with self-attention (Vaswani et al., 2017). Ours (w/o Mean Flow) replaces the average-velocity objective with standard flow matching and 50-step sampling (Lipman et al., 2022). We compare variants using these metrics: TS-FID (Table 2), downstream classification gain (Table 3), Hjorth statistics (Table 4), and skewness KS (Figure 2).

Average-velocity objective. Removing the average-velocity objective increases average TS-FID to 704.76 and reduces average ΔAcc to 1.76%. Its class-conditional scalp topographies remain coherent, which suggests that mean-field token coupling helps preserve cross-electrode structure. Its amplitude-related statistics are weaker. On TUEV, the Hjorth complexity distance is 3.67, compared with 3.39 for GPT2MEG, and the skewness KS is 0.28, compared with 0.22.

Mean-field token coupling. Removing mean-field token coupling increases average TS-FID to 618.52 and reduces average ΔAcc to 1.98%. This variant remains strong on amplitude-related statistics. On TUEV, its skewness KS is 0.18 and its Hjorth complexity distance is 2.11, both better than GPT2MEG. In addition, full MFF improves average TS-FID from 618.52 to 518.60 and average ΔAcc from 1.98 to 2.80, and it better preserves class-conditional scalp topographies. These comparisons demonstrate that mean-field token coupling and the average-velocity objective improve distributional fidelity and downstream classification performance across evaluated datasets. We ablate coupling directions, token-mixing alternatives and field dimension in Appendix I.

Using both. MFF's average TS-FID is 518.60, average ΔAcc is 2.80%, and TUEV Hjorth complexity distance is 0.806. These ablations suggest that the two components contribute differently. Mean-field token coupling improves cross-electrode structure. The average-velocity objective improves one-step transport and amplitude-related fidelity. The full model performs best when both are used together.

More results. Appendix I.3 compares MFF with one-step and 50-step standard FM and a consistency model. A controlled perturbation study examines error retention in iterative sampling (Appendix I.4), without attributing the entire clean-performance gap to that mechanism. Sampling latency and loss-function comparisons are reported in Appendices A.5 and C, respectively.

5 Conclusion

We presented Mean-Field Flow, a one-step, latent-free method for raw EEG generation. MFF operates on raw EEG and uses mean-field token coupling to model interactions across channels and time. It is trained with an average-velocity objective enabling one-step sampling. Across 14 public EEG datasets, MFF achieves the lowest TS-FID, yields the largest average improvement in downstream classification and better matches real EEG in Hjorth parameters, skewness, and scalp topographies.

Limitations. MFF is trained separately on each dataset and conditioned only on class labels, so it does not yet address generation across montages, subjects, or recording sessions. Future work includes richer conditioning, privacy-preserving synthesis, and controllable generation of rare events. We hope this work encourages further progress in generative modeling for EEG and related electrophysiological signals through architectures and objectives that reflect their underlying signal structure.

AI USE STATEMENT

In accordance with ICLR policy, we only used generative AI tools for language editing, LaTeX formatting, and presentation refinement of figures and visualization code based on author-provided content. The research question and methodology were developed by the authors, who conducted the experiments and produced the reported results. The authors reviewed all AI-assisted revisions for accuracy and consistency with the underlying work, and take full responsibility for the final manuscript and accompanying artifacts.

REFERENCES

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=XH3ArccntI.

Michael Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3): 340–352, 2017.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. OpenAI Technical Report.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16:1–6, 2008.

György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature reviews neuroscience*, 15(4):264–278, 2014.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, et al. A multi-modal open dataset for mental-disorder analysis. *Scientific data*, 9(1):178, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024a.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. *arXiv preprint arXiv:2401.10278*, 2024b.

Zhige Chen, Chengxuan Qin, Wenlong You, Rui Liu, Congying Chu, Rui Yang, Kay Chen Tan, and Jibin Wu. Hear: An eeg foundation model with heterogeneous electrode adaptive representation. *arXiv preprint arXiv:2510.12515*, 2025.

Richard Csaky, Mats WJ van Es, Oiwi Parker Jones, and Mark Woolrich. Gpt2meg: Quantizing meg for autoregressive generation. *arXiv preprint arXiv:2404.09256*, 2024.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alex Dimakis, and Peyman Milanfar. Soft diffusion: Score matching with general corruptions. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=W98rebBxlQ.

Oscar Davis, Samuel Kessler, Mircea Petrache, Ismail Ilkan Ceylan, Michael M. Bronstein, and Joey Bose. Fisher flow matching for generative modeling over discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6jOScqwdHU.

Xia Deng, Shen Chen, Jiale Zhou, and Lei Li. Mind2matter: Creating 3d models from eeg signals. *arXiv preprint arXiv:2504.11936*, 2025.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*, 2023.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4): 1273–1302, 2003.

Kay Fuhrmeister, Arne Pelzer, Fabian Radke, Julia Lechinger, Mahzad Gharleghi, Thomas Köllmer, and Insa Wolf. Bridging privacy and utility: Synthesizing anonymized eeg with constraining utility functions. *arXiv preprint arXiv:2509.20454*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling, 2025a. URL https://arxiv.org/abs/2505.13447.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Shuqi Gu, Chuyue Li, Baoyu Jing, and Kan Ren. Verbalts: Generating time series from texts. *International Conference on Machine Learning*, 2025.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017. URL https://arxiv.org/abs/1704.00028.

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.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306–310, 1970. ISSN 0013-4694. doi: https://doi.org/10.1016/0013-4694(70)90143-4. URL https://www.sciencedirect.com/science/article/pii/0013469470901434.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Rukuang Huang, Sungjun Cho, Chetan Gohil, Oiwi Parker Jones, and Mark Woolrich. Meggpt: A transformer-based foundation model for magnetoencephalography data. *arXiv preprint arXiv:2510.18080*, 2025.

Ben H Jansen and Vincent G Rit. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics*, 73(4):357–366, 1995.

Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in human neuroscience*, 16:898300, 2022.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multitask foundation model for bridging the gap between language and eeg signals. *arXiv preprint arXiv:2409.00101*, 2024a.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024b.

Alexander Ya Kaplan, Andrew A Fingelkurts, Alexander A Fingelkurts, Sergei V Borisov, and Boris S Darkhovsky. Nonstationary nature of the brain activity as revealed by eeg/meg: methodological, practical and conceptual challenges. *Signal processing*, 85(11):2190–2212, 2005.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1809–1817, 2017.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124: 180–192, 2016.

Arjun Khanna, Alvaro Pascual-Leone, Christoph M Michel, and Faranak Farzan. Microstates in resting-state eeg: current status and future directions. *Neuroscience & Biobehavioral Reviews*, 49: 105–113, 2015.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. On density estimation with diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=2LdBqxc1Yv.

Timon Klein, Piotr Minakowski, and Sebastian Sager. Flexible patched brain transformer model for eeg decoding. *Scientific Reports*, 15(1):10935, 2025.

Demetres Kostas, Stephane 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.

Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the brain: image reconstruction of visual perception from human brain signals. *arXiv preprint arXiv:2308.02510*, 2023.

Young-Eun Lee, Seo-Hyun Lee, Sang-Ho Kim, and Seong-Whan Lee. Towards voice reconstruction from eeg during imagined speech. In *Proceedings of the AAAI conference on artificial intelligence*, 2023.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015.

Zhen Liang, Rushuang Zhou, Li Zhang, Linling Li, Gan Huang, Zhiguo Zhang, and Shin Ishii. Eegfusenet: Hybrid unsupervised deep feature characterization and fusion for high-dimensional eeg with an application to emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1913–1925, 2021.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Junxiang Liu, Junming Lin, Jiangtong Li, and Jie Li. Dynamind: Reconstructing dynamic visual scenes from eeg by aligning temporal dynamics and multimodal semantics to guided diffusion. *arXiv preprint arXiv:2509.01177*, 2025a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and qiang liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1k4yZbbDqX.

Xuan-Hao Liu, Yan-Kai Liu, Tianyi Zhou, Bao-Liang Lu, and Wei-Long Zheng. Mindcross: Fast new subject adaptation with limited data for cross-subject video reconstruction from brain signals. *arXiv preprint arXiv:2511.14196*, 2025b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large eeg dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.

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, 2024. URL https://arxiv.org/abs/2401.08740.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Wajid Mumtaz. MDD Patients and Healthy Controls EEG Data (New). figshare, 11 2016. URL https://figshare.com/articles/dataset/EEG_Data_New/4244171. Dataset.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://openreview.net/forum?id=-NEXDKk8gZ.

Paul L Nunez and Ramesh Srinivasan. *Electric fields of the brain: the neurophysics of EEG*. Oxford university press, 2006.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, Volume 10 - 2016, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00196.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Giorgio Parisi and Ramamurti Shankar. *Statistical Field Theory*. Addison-Wesley, 1988.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Jia Hong Puah, Sim Kuan Goh, Ziwei Zhang, Zixuan Ye, Chow Khuen Chan, Kheng Seang Lim, Si Lei Fong, Kok Sin Woon, and Cuntai Guan. Eegdm: Eeg representation learning via generative diffusion model. *arXiv preprint arXiv:2508.14086*, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Stephen J Roberts. Extreme value statistics for novelty detection in biomedical data processing. *IEE Proceedings-Science, Measurement and Technology*, 147(6):363–367, 2000.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

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*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021b. URL https://arxiv.org/abs/2011.13456.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023a. URL https://arxiv.org/abs/2303.01469.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023b.

Michael Tschannen, André Susano Pinto, and Alexander Kolesnikov. Jetformer: An autoregressive generative model of raw images and text. *arXiv preprint arXiv:2411.19722*, 2024.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024a.

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. *arXiv preprint arXiv:2412.07236*, 2024b.

Xujia Wang, Xuhui Liu, Xi Liu, Qian Si, Zhaoliang Xu, Yang Li, and Xiantong Zhen. Eeg-dino: Learning eeg foundation models via hierarchical self-distillation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 196–205. Springer, 2025.

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.

Hugh R Wilson and Jack D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.

Di Wu, Siyuan Li, Jie Yang, and Mohamad Sawan. Neuro-bert: Rethinking masked autoencoding for self-supervised neurological pretraining. *arXiv preprint arXiv:2204.12440*, 2022.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *arXiv preprint arXiv:2505.18185*, 2025.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Biot: Cross-data biosignal learning in the wild. *arXiv preprint arXiv:2305.10351*, 2023.

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. *arXiv preprint arXiv:2602.18473*, 2026.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36:26304–26321, 2023a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023b.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *arXiv preprint arXiv:2506.23075*, 2025.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

APPENDICES

A Implementation Details **17**

A.1 Classifier-Free Guidance 17

A.2 Dual-Head Training 17

A.3 Hyperparameters 18

A.4 Evaluation Metrics 19

A.5 One-Step Sampling Efficiency 20

A.6 Input Preprocessing and Model Capacity 21

B Downstream Augmentation Results **22**

B.1 Absolute Downstream Accuracy 23

B.2 Generality across Downstream Architectures 23

B.3 Controlling for Added Data and Label Consistency 24

B.4 Sensitivity to the Synthetic-Data Ratio 24

C Loss Function Ablation **26**

D Class-Conditional Scalp Topographies Across 14 Datasets **27**

E Quantitative Topographic and Spectral Fidelity **29**

E.1 Physiologically Relevant Signal Structure 30

F Waveform Visualization on TUEV **31**

G Per-Channel Skewness Across 14 Datasets **32**

H Hjorth Parameters Across 14 Datasets **36**

I Additional Architecture and Objective Controls **37**

I.1 Token Interaction and Spatial–Temporal Coupling 37

I.2 Field Dimensionality on High-Density EEG 38

I.3 Average-Velocity Training and Other Sampling Formulations 38

I.4 Controlled Perturbation Accumulation 38

J Sensitivity to Preprocessing **40**

A Implementation Details

This appendix provides implementation details and model design details introduced in Secs. 3 and 4. We describe training-time classifier-free guidance (Appendix A.1) and the dual-head training design (Appendix A.2); provide the full set of training and sampling hyperparameters (Appendix A.3); give the definitions of evaluation metrics (Appendix A.4); and report one-step sampling efficiency (Appendix A.5). Input preprocessing and approximate model capacities are specified in Appendix A.6.

A.1 Classifier-Free Guidance

Standard classifier-free guidance (CFG) (Ho & Salimans, 2021) mixes conditional and unconditional predictions at sampling time, which doubles the number of forward passes per step. To retain the one-step sampler of MFF, as defined in Eq. 15, we follow previous works (Geng et al., 2025b; Lu et al., 2026) and absorb guidance into the training target instead. Sampling then remains a single forward pass.

Sampled guidance scale and interval. For each training example, we sample a guidance scale $\omega \in [1, \omega_{\max}]$ and an active interval $[t_{\min}, t_{\max}] \subseteq [0, 1]$ together with (t, r, y). The scale follows a power-law distribution

$$p(\omega) \propto \omega^{-\beta}, \qquad \omega \in [1, \omega_{\max}], \tag{19}$$

with $\omega_{\max} = 7.0$ and $\beta = 1.0$, which is uniform in $\log \omega$. We sample $t_{\min} \sim \mathcal{U}(0, 0.5)$ and $t_{\max} \sim \mathcal{U}(0.5, 1)$. When $r = t$, corresponding to the standard flow-matching branch, we use the full interval $[0, 1]$ so that guidance is applied uniformly. The effective guidance scale at (z_t, t) is

$$\omega_{\text{eff}} = \begin{cases} \omega & t \in [t_{\min}, t_{\max}], \\ 1 & \text{otherwise}, \end{cases} \tag{20}$$

clamped to $\omega_{\text{eff}} \geq 1$.

Guided velocity target. Given $z_t = (1 - t)x + t\epsilon$ and ω_{eff}, we evaluate the instantaneous-velocity head twice: once with the true class label y, yielding $\hat{v}_\theta(z_t, t, y)$, and once with the null label $\varnothing$, yielding $\hat{v}_\theta(z_t, t, \varnothing)$. We then form the guided velocity target

$$v_g = v_t + \left(1 - \tfrac{1}{\omega_{\text{eff}}}\right) \text{sg}[\hat{v}_\theta(z_t, t, y) - \hat{v}_\theta(z_t, t, \varnothing)], \tag{21}$$

where $v_t = \epsilon - x$. The guidance residual is stop-gradient detached when constructing v_g, so v_g is treated as a fixed regression target for both the average-velocity and instantaneous-velocity heads. The guided target reduces to v_t when $\omega_{\text{eff}} = 1$ and approaches $v_t + (\hat{v}_\theta - \hat{v}_\theta^{\varnothing})$ as $\omega_{\text{eff}} \to \infty$. With probability 0.1 we drop the class label, train the network on the null condition, and revert v_g to v_t (Ho & Salimans, 2021). The mean-flow regression target Eq. 17 is then computed with v_g in place of v_t.

Inference. Because guidance is folded into training, sampling remains exactly as in Algorithm 2: one forward pass of $\hat{u}_\theta$ at $t = 1$ with the desired class label and no auxiliary unconditional evaluation.

A.2 Dual-Head Training

The mean-flow target in Eq. 17 uses an identity-inspired bootstrap correction, but only the average-velocity prediction is supervised. We add a parallel head that predicts the instantaneous velocity directly. The auxiliary head provides a second regression signal at every (z_t, t), while the average-velocity head retains its bootstrap structure.

Architecture. Motivated by the parameterization of pMF (Lu et al., 2026), the network uses a shared backbone of mean-field token coupling (Sec. 3.3) followed by two prediction heads on the final hidden representation: an average-velocity head $\hat{u}_\theta$ and an instantaneous-velocity head $\hat{v}_\theta$. Both heads are linear projections of identical shape applied to the same shared features. The backbone is conditioned on (z, t, y). Inputs are arranged in the fixed channel order of each dataset and the fixed temporal patch order of each segment. Before the channel and temporal mean-field branches, we add fixed one-dimensional sinusoidal embeddings for channel indices and temporal patch indices, respectively. These embeddings provide montage-index and temporal-position information, while no physical 2-D or 3-D electrode coordinates are used. The variable r enters only through the bootstrap target in Eq. 17: it is implemented as a second query of the same network with the scalar time input set to r, rather than as an additional conditioning variable. No r is used at inference.

Loss. Both heads are trained with guided ℓ_1 norm:

$$\mathcal{L}(\theta) \;=\; \mathbb{E}\left\|\hat{u}_\theta(z_t, t, y) - u_{\text{tgt}}\right\|_1 \;+\; \mathbb{E}\left\|\hat{v}_\theta(z_t, t, y) - v_g\right\|_1, \tag{22}$$

where u_{tgt} is the mean-flow target from Eq. 17 computed with v_g in place of v_t, and v_g is the guided velocity target in Eq. 21. Gradients from both heads update the shared backbone. The $\hat{u}$-head uses the bootstrap target and requires an additional prediction with the time input set to r. The $\hat{v}$-head directly regresses the instantaneous velocity at (z_t, t), providing an additional supervision signal that does not depend on the bootstrap.

Adaptive loss balancing. The two regression terms live on different scales. Rather than tune their weights by hand, we rescale each per-sample loss by $(\ell + \varepsilon)^{-p}$ with $p = 1$ and $\varepsilon = 10^{-2}$, following (Geng et al., 2025a). The weights are detached from the computation graph, so they reweight the losses without changing the underlying regression targets.

Inference. At sampling time we keep only $\hat{u}_\theta$ (Algorithm 2). The $\hat{v}$-head and its extra forward passes are used only during training.

A.3 Hyperparameters

Table 5 lists the training and sampling hyperparameters. The same configuration is used across all 14 datasets. Only quantities fixed by dataset shape vary, namely the channel count C, the patch grid (P, S) that determines segment length, and the number of classes K. All experiments are run on a single NVIDIA A100 GPU.

Optimization		**Time and Noise Sampling**	
optimizer	AdamW	time sampler	sigmoid-normal
(β_1, β_2)	$(0.9, 0.95)$	$(P_{\text{mean}}, P_{\text{std}})$	$(-0.8,\ 0.8)$
base learning rate	5×10^{-5}	flow-matching fraction	0.5
schedule	constant	numerical guard t_ε	5×10^{-2}
warmup epochs	5	noise distribution	$\mathcal{N}(0, I)$
weight decay	0	label drop probability	0.1
gradient clipping	none		
batch size (per GPU)	32		
training epochs	200		
numerical precision	bfloat16		
Classifier-Free Guidance		**Inference**	
$\omega_{\max}$	7.0	sampling steps	1
β (power-law exponent)	1.0	sampler	one-step prediction
$t_{\min}$	$\mathcal{U}(0, 0.5)$	inference ω	1.0
$t_{\max}$	$\mathcal{U}(0.5, 1)$	inference interval	$[0, 1]$
adaptive p	1.0	EMA decay (slow)	0.9999
adaptive ε	1×10^{-2}	EMA decay (fast)	0.9996
loss norm	ℓ_1	EMA used at inference	slow

Table 5: MFF training and sampling hyperparameters. The same configuration is used on all 14 datasets; only quantities determined by dataset shape vary.

Backbone configuration. MFF uses the configuration with hidden size $H = 768$, six temporal mean-field blocks and six channel mean-field blocks. The shared-field/core dimension is $D_c = H/4 = 192$, and each block uses a two-layer feed-forward network with expansion ratio 2. EEG segments are patchified with patch size $S = 200$, so the number of temporal patches is $P = T/S$ where T is the number of samples. The temporal and channel branches use separate patch projections and fixed one-dimensional sinusoidal positional embeddings. Their outputs are summed and decoded by two patchwise heads, one for the average velocity $\hat{u}_\theta$ and one for the instantaneous velocity $\hat{v}_\theta$; only $\hat{u}_\theta$ is used during sampling.

A.4 Evaluation Metrics

Time-Series Fréchet Inception Distance (TS-FID). We compute TS-FID with a deterministic EEG feature map rather than an image-domain classifier. For an EEG segment $x \in \mathbb{R}^{C\times T}$, we first apply a channel-wise Fast Fourier Transform (FFT) and retain the magnitude of the one-sided spectrum. This emphasizes spectral content while reducing sensitivity to arbitrary temporal phase shifts. When the dataset-specific preprocessing includes a high-frequency cutoff f_{cut}, we truncate the spectrum accordingly to reduce the influence of measurement noise. The remaining magnitudes are pooled into $K_{\text{freq}} = 256$ frequency bins and $K_{\text{spat}} = 4$ coarse spatial groups, followed by a log transform and per-sample standardization:

$$\phi(x) = \mathrm{Std}\big(\log\big(1 + \mathrm{Pool}_{\text{spat}}\big(\mathrm{Pool}_{\text{freq}}(|\mathcal{F}(x)|_{0:f_{\text{cut}}})\big)\big)\big)\,, \tag{23}$$

where $\mathcal{F}(\cdot)$ denotes the channel-wise temporal FFT and $\mathrm{Std}(\cdot)$ denotes per-sample standardization. For every dataset and method, we use the same preprocessing, feature map, sample budget, and class-conditioning protocol. Let (μ_r, Σ_r) and (μ_g, Σ_g) be the empirical mean and covariance of the real and generated feature sets. TS-FID is then

$$\text{TS-FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big)\,. \tag{24}$$

Lower TS-FID indicates that the generated EEG better matches the real distribution under this frequency–spatial feature representation.

Downstream accuracy change (ΔAcc). Distributional fidelity alone does not guarantee that synthetic EEG is useful for training. We therefore measure augmentation improvement by comparing the same downstream model trained with and without generated data. Let $\mathcal{D}_{\text{train}}$ be the real training set, $\mathcal{D}_{\text{gen}}$ the synthetic set produced by a given generator, and $\mathcal{D}_{\text{test}}$ an official held-out real test set. We report

$$\Delta\text{Acc} = \mathrm{Acc}\big(h_{\mathcal{D}_{\text{train}}\cup\mathcal{D}_{\text{gen}}}, \mathcal{D}_{\text{test}}\big) - \mathrm{Acc}\big(h_{\mathcal{D}_{\text{train}}}, \mathcal{D}_{\text{test}}\big)\,, \tag{25}$$

where $h_{\mathcal{D}}$ denotes CBraMod (Wang et al., 2024b) fine-tuned on dataset $\mathcal{D}$. In the downstream experiments, $\mathcal{D}_{\text{gen}}$ contains 20% as many samples as the real training set, and all reported values in Tables 3 and 7 are mean $\pm$ standard deviation over five random seeds. Positive ΔAcc means that the synthetic samples provide useful training signal for a classifier evaluated only on real EEG.

Hjorth parameter distance. We evaluate signal-level dynamics with the Hjorth parameters (Hjorth, 1970), which summarize variance, characteristic temporal variation, and departure from a single dominant oscillatory scale. Before computing these statistics, we apply the same bandpass filter used elsewhere in evaluation. To make the comparison insensitive to global amplitude scale, real and generated samples are normalized independently. For a trial $x \in \mathbb{R}^{C\times T}$, we first compute the temporal standard deviation of each channel,

$$\sigma_c(x) \;:=\; \mathrm{Std}\big(x_{c,1}, \ldots, x_{c,T}\big), \qquad c = 1, \ldots, C, \tag{26}$$

and then define the trial-level normalization factor

$$s(x) \;:=\; \mathrm{median}\big\{\sigma_c(x) : c = 1, \ldots, C\big\} \;+\; \epsilon, \tag{27}$$

where ϵ is a small numerical constant. We normalize all channels by

$$\widetilde{x}_{c,t} \;:=\; \frac{x_{c,t}}{s(x)}. \tag{28}$$

For a normalized channel waveform $\widetilde{x}_c = (\widetilde{x}_{c,1}, \ldots, \widetilde{x}_{c,T})$, we define the first- and second-order temporal differences as

$$\Delta\widetilde{x}_{c,t} = \widetilde{x}_{c,t+1} - \widetilde{x}_{c,t}, \qquad \Delta^2\widetilde{x}_{c,t} = \Delta\widetilde{x}_{c,t+1} - \Delta\widetilde{x}_{c,t}. \tag{29}$$

Let

$$v_0 = \mathrm{Var}(\widetilde{x}_c), \qquad v_1 = \mathrm{Var}(\Delta\widetilde{x}_c), \qquad v_2 = \mathrm{Var}(\Delta^2\widetilde{x}_c), \tag{30}$$

where the variance is computed over time. The Hjorth parameters are

$$\mathrm{Activity}(\widetilde{x}_c) = v_0, \tag{31}$$

$$\mathrm{Mobility}(\widetilde{x}_c) = \sqrt{\frac{v_1 + \epsilon}{v_0 + \epsilon}}, \tag{32}$$

$$\mathrm{Complexity}(\widetilde{x}_c) = \frac{\sqrt{(v_2 + \epsilon)/(v_1 + \epsilon)}}{\sqrt{(v_1 + \epsilon)/(v_0 + \epsilon)}}. \tag{33}$$

For each statistic $q \in \{\mathrm{Activity}, \mathrm{Mobility}, \mathrm{Complexity}\}$, we compute $q(\widetilde{x}_c)$ for every trial and channel, pool the resulting values over trials and channels, and compare the real and generated empirical distributions using Wasserstein-1 distance:

$$d^{(q)}_{\mathrm{Hjorth}} = W_1\Big(\{q(\widetilde{x}_c) : x \in \mathcal{D}_{\mathrm{real}},\, c = 1, \ldots, C\}, \{q(\widetilde{\hat{x}}_c) : \hat{x} \in \mathcal{D}_{\mathrm{gen}},\, c = 1, \ldots, C\}\Big). \tag{34}$$

Lower values indicate that the generated EEG better matches the corresponding real EEG signal statistic.

A.5 One-Step Sampling Efficiency

A practical advantage of MFF is that generation requires only one forward pass. To quantify this benefit, we compare the wall-clock latency of the one-step endpoint sampler with that of a 50-step sampler under the same model, batch size, and hardware setting. Timing excludes data loading and is measured with synchronized GPU execution. Table 6 reports batch latency, per-sample latency, and throughput at batch size 64.

Sampling	Batch latency ($B = 64$)	Per-sample latency	Throughput
1 step	49.68 ± 0.11 **ms**	0.78 **ms**	1,288 **samples/s**
50 step	4730.33 ± 1.12 ms	73.91 ms	13.5 samples/s
Speedup	–	–	$\times 95.2$

Table 6: **Sampling efficiency.** One-step sampling substantially reduces latency relative to a 50-step sampler. Timing is measured on a single NVIDIA A100 GPU with batch size 64 under the same model setting. Colors compare sampling settings within each metric.

A.6 Input Preprocessing and Model Capacity

Generator inputs. We use the official or subject-disjoint split associated with each benchmark and train each generator only on its training partition. Inputs follow the selected montage and fixed channel order of each dataset. We resample EEG to 200 Hz, divide recordings into fixed-length segments, and multiply the signals by 0.01 for numerical stability. Channel counts and segment durations are listed in Table 1; the durations range from 1 to 30 seconds. Beyond channel selection and reordering, resampling, segmentation, and numerical scaling, the default generator-input pipeline adds no filtering, rereferencing, artifact rejection, or cleaning to the processing already present in each public release. This input construction is distinct from the transformations used to compute individual evaluation metrics in Appendix A.4.

Downstream inputs. For each downstream backbone, real and generated signals undergo the same native input transformation, including channel ordering, sampling rate, segment length, signal scale, and model-specific normalization. The numerical scaling used for generation does not replace the backbone-specific input transformation. Appendix B.2 evaluates multiple downstream architectures, while Appendix J examines explicitly varied generator-input pipelines.

Model capacity. MFF and all external baselines except JetFormer are configured at approximately the 100M-parameter scale; JetFormer uses approximately 300M parameters. Small differences arise from architecture- and dataset-specific input and output layers. The controlled variants retain the hidden dimension and backbone depth of MFF and remain in the same approximate capacity regime. These comparisons therefore use similar model scales rather than exactly identical parameter counts.

B Downstream Augmentation Results

Table 7 reports the downstream augmentation results on all 14 EEG datasets. In the main paper, we show one representative dataset from each task domain for readability; here we include every dataset and baseline under the same protocol. For each generator, we augment the real training set with synthetic data equal to 20% of the original training size, fine-tune CBraMod (Wang et al., 2024b), and measure the resulting change in classification accuracy on the official held-out real test set. Positive ΔAcc indicates that the synthetic samples improve downstream classification. Across the 14 datasets, MFF yields the largest average gain (+2.80%), exceeding both the strongest external baseline (+1.62%) and the two controlled ablations.

Task	Dataset	*GAN & AR*		*Diffusion*		*MEG Transfer*		*Flow Matching*			
		EEG-GAN	JetFormer	Diffusion	VerbalTS	MEG-GPT	GPT2MEG	JiT	Ours (w/o Mean Flow)	Ours (w/o Mean Field)	**Ours**
Neurology	TUAB	−0.28±0.43	−0.08±0.47	+0.21±0.39	+0.61±0.38	+1.04±0.41	+1.52±0.36	+0.82±0.40	+1.73±0.37	+1.38±0.35	**+2.38±0.36**
	TUEV	+0.17±0.64	+0.31±0.59	+0.82±0.57	+1.23±0.53	+1.82±0.61	+2.41±0.54	+1.58±0.51	+2.13±0.49	+2.87±0.53	**+3.60±0.52**
	TUSZ	+0.08±0.52	+0.19±0.47	+0.43±0.42	+1.09±0.45	+1.31±0.43	+0.93±0.41	+0.71±0.38	+1.53±0.41	+1.81±0.39	**+2.42±0.40**
	CHB-MIT	−0.57±0.37	−0.21±0.31	−0.03±0.30	+0.29±0.33	+0.73±0.29	+0.42±0.28	+0.22±0.29	+0.51±0.26	+0.33±0.27	**+0.85±0.28**
Motor Imagery	PhysioNet-MI	−0.04±0.61	+0.12±0.54	+0.39±0.50	+0.71±0.52	+1.42±0.48	+2.08±0.51	+1.18±0.47	+1.83±0.45	+2.51±0.44	**+3.01±0.47**
	BCI IV-2a	−0.73±0.71	−0.02±0.66	−0.18±0.63	+0.12±0.59	+0.93±0.54	+0.61±0.57	+0.42±0.53	+1.12±0.49	+0.38±0.56	**+1.55±0.52**
	SHU-MI	−0.19±0.46	+0.09±0.43	+0.28±0.41	+0.78±0.39	+1.03±0.42	+1.12±0.38	+0.58±0.35	+1.32±0.36	+1.58±0.39	**+1.95±0.38**
Emotion	FACED	+0.41±1.03	+0.79±0.97	+1.59±0.94	+2.17±0.88	+3.91±0.86	+4.48±0.91	+3.12±0.84	+5.62±0.79	+6.83±0.86	**+8.92±0.83**
	SEED	+0.22±0.94	+0.41±0.85	+0.78±0.81	+1.29±0.78	+1.89±0.75	+2.79±0.77	+1.53±0.73	+2.41±0.70	+3.14±0.74	**+3.85±0.72**
	Mumtaz	−0.62±0.51	−0.23±0.43	+0.02±0.47	+0.52±0.40	+0.89±0.39	+0.97±0.41	+0.43±0.38	+1.03±0.36	+0.47±0.37	**+1.35±0.38**
	MODMA	−0.41±0.55	−0.13±0.50	+0.11±0.48	+0.42±0.44	+1.21±0.43	+1.09±0.42	+0.51±0.45	+0.83±0.39	+0.92±0.40	**+1.62±0.42**
Cognition	Speech	+0.11±0.73	+0.28±0.70	+0.52±0.65	+0.83±0.61	+2.38±0.58	+1.73±0.63	+1.21±0.59	+1.52±0.55	+2.23±0.60	**+2.95±0.58**
	ISRUC	−0.13±0.63	+0.07±0.56	+0.31±0.54	+1.01±0.51	+1.57±0.52	+1.34±0.49	+0.82±0.48	+1.71±0.46	+1.18±0.45	**+2.28±0.48**
	Stress	−0.26±0.83	+0.03±0.75	+0.19±0.77	+0.51±0.70	+1.48±0.66	+1.18±0.69	+0.92±0.68	+1.31±0.63	+2.12±0.65	**+2.45±0.65**
Avg. ΔAcc		−0.16	+0.12	+0.39	+0.83	+1.54	+1.62	+1.00	+1.76	+1.98	**+2.80**

Table 7: **Downstream utility via ΔAcc (%, ↑).** We augment real training data with additional 20% synthetic samples. Values report mean ± standard deviation across 5 random seeds, in percentage points. **Best**, second and third are highlighted.

B.1 Absolute Downstream Accuracy

Table 8 reports the absolute CBraMod accuracies underlying the MFF gains in Table 7. The real-only and augmented conditions use the same five seeds and official held-out real test sets, and accuracy changes are computed from paired runs. MFF improves the mean accuracy on all 14 datasets, including an increase from 65.24% to 68.84% on TUEV and from 54.62% to 63.54% on FACED. These absolute values complement, rather than replace, the mean and standard deviation of the paired gains reported in Table 7.

Task	**Dataset**	**Real only** (%) ↑	**+20% MFF** (%) ↑	**ΔAcc** (pp) ↑
Neurology	TUAB	81.31	**83.69**	+2.38
	TUEV	65.24	**68.84**	+3.60
	TUSZ	74.36	**76.78**	+2.42
	CHB-MIT	72.62	**73.47**	+0.85
Motor imagery	PhysioNet-MI	63.55	**66.56**	+3.01
	BCI IV-2a	52.07	**53.62**	+1.55
	SHU-MI	61.28	**63.23**	+1.95
Emotion	FACED	54.62	**63.54**	+8.92
	SEED	44.73	**48.58**	+3.85
	Mumtaz	94.83	**96.18**	+1.35
	MODMA	76.84	**78.46**	+1.62
Cognition	Speech	54.21	**57.16**	+2.95
	ISRUC	75.96	**78.24**	+2.28
	Stress	71.92	**74.37**	+2.45

Table 8: **Absolute downstream accuracy across 14 datasets.** Values are mean accuracies from the paired five-seed CBraMod evaluation. MFF adds synthetic samples equal to 20% of the real training set. Accuracy is measured in percent; ΔAcc is measured in percentage points (pp). Standard deviations of the paired gains are reported in Table 7. Colors compare real-only and augmented accuracy within each row.

B.2 Generality across Downstream Architectures

Pretrained and from-scratch classifiers. We repeat 20% MFF augmentation on the official TUEV split with LaBraM and TFM-Token, in addition to CBraMod. We also evaluate EEGNet, DeepConvNet, and ShallowFBCSPNet with random initialization. For the from-scratch comparisons, each fixed-seed real-only/augmented pair shares its initialization, optimizer-update budget, and model-selection protocol. Table 9 reports positive gains for all six backbones. The improvement therefore does not require CBraMod or a pretrained encoder.

Initialization	**Backbone**	**Real only** (%) ↑	**+20% MFF** (%) ↑	**ΔAcc** (pp) ↑
Pretrained	CBraMod	65.24	**68.84**	+3.60
	LaBraM	54.52	**57.73**	+3.21
	TFM-Token	60.86	**62.80**	+1.94
From scratch	EEGNet	54.48	**57.61**	+3.13
	DeepConvNet	52.16	**54.64**	+2.48
	ShallowFBCSPNet	53.77	**55.09**	+1.32

Table 9: **Downstream augmentation across six backbones on TUEV.** All augmented conditions add 20% MFF-generated data and are evaluated on held-out real EEG. Colors compare real-only and augmented accuracy within each row. The additional backbone comparisons report point estimates.

Cross-dataset evaluation. We extend the from-scratch comparison to PhysioNet-MI, MODMA, Speech, and BCI IV-2a under the same fixed-seed and fixed-update protocol. Table 10 shows improvements in all 15 dataset–classifier combinations. These results extend the downstream evidence beyond TUEV, while remaining evaluations of task-specific augmentation.

Dataset	EEGNet	DeepConvNet	ShallowFBCSPNet
TUEV	+3.13	+2.48	+1.32
PhysioNet-MI	+2.14	+1.92	+2.06
MODMA	+1.37	+1.13	+1.09
Speech	+2.16	+2.05	+1.71
BCI IV-2a	+1.13	+0.92	+1.01

Table 10: **Cross-dataset augmentation for classifiers trained from scratch.** Values are ΔAcc in percentage points (↑), relative to the corresponding real-only classifier. All conditions use a 20% synthetic-data ratio. Results are fixed-seed paired comparisons.

B.3 Controlling for Added Data and Label Consistency

We compare correctly labeled MFF samples with three equal-sized controls: duplicated real samples, conventional temporal-shift/amplitude-jitter augmentation, and the same MFF samples with permuted labels. Each condition adds 20% as many samples as the real training set, and the number of downstream optimization updates is fixed. We evaluate CBraMod on TUEV, BCI IV-2a, FACED, and ISRUC.

Added data	TUEV	BCI IV-2a	FACED	ISRUC	Avg.
Duplicated real samples	+0.11	+0.06	+0.28	+0.09	+0.14
Temporal shift + amplitude jitter	+0.61	+0.37	+1.53	+0.42	+0.73
MFF with permuted labels	-1.08	-0.76	-2.21	-0.69	-1.19
MFF with correct labels	**+3.60**	**+1.55**	**+8.92**	**+2.28**	**+4.09**

Table 11: **Equal-size augmentation controls.** Values are CBraMod ΔAcc in percentage points (↑). Dataset size and downstream optimizer-update count are matched across the four augmented conditions. The average covers only the four datasets shown, not the full 14-dataset benchmark. The top three values in each column are highlighted.

Correctly labeled MFF samples yield an average gain of $+4.09$ percentage points, compared with $+0.14$ from duplication and $+0.73$ from conventional augmentation. Permuting the labels of the same generated signals instead decreases accuracy by 1.19 points on average. Thus, increasing the nominal training-set size or optimization exposure alone does not explain the observed gains; their benefit depends on class-consistent generated signals. This control assesses downstream utility and is not, by itself, a test of physiological equivalence.

B.4 Sensitivity to the Synthetic-Data Ratio

We define the augmentation ratio as $|\mathcal{D}_{\text{gen}}|/|\mathcal{D}_{\text{train}}|$. The default 20% ratio was selected a priori as a uniform moderate budget, rather than tuned for each dataset. We first evaluate ratios from 0% to 100% on TUEV, PhysioNet-MI, MODMA, and Speech, then examine the 20%–50% interval in 5% increments on five datasets with CBraMod and EEGNet.

Results. In the four-dataset sweep, the average gain is $+2.80$ points at 20%, $+2.77$ at 50%, and $+2.61$ at 100% (Table 12). The finer CBraMod sweep shows a plateau: the five-dataset average is $+2.55$ at 20% and approximately $+2.51$ from 25% to 35% (Table 13). Dataset-level behavior is not identical: PhysioNet-MI is numerically highest at 50%, whereas BCI IV-2a decreases from $+1.55$ at 20% to $+0.95$ at 50%. With EEGNet, every tested ratio remains beneficial on every dataset, and 20% gives the largest reported gain in each of the five columns. We retain 20% as an efficiency-oriented default: moving to 50% requires $2.5\times$ as many generated samples without a consistent utility gain. These sweeps do not establish a statistically optimal or universally transferable augmentation ratio.

Ratio	TUEV	PhysioNet-MI	MODMA	Speech	Avg.
0%	+0.00	+0.00	+0.00	+0.00	+0.00
10%	+2.45	+1.98	+1.12	+1.89	+1.86
20%	**+3.60**	+3.01	**+1.62**	**+2.95**	**+2.80**
50%	+3.55	**+3.12**	+1.59	+2.82	+2.77
100%	+3.21	+2.84	+1.58	+2.81	+2.61

Table 12: **Broad augmentation-ratio sweep with CBraMod.** Values are ΔAcc in percentage points (↑); the 0% condition is the real-only reference. The average covers the four datasets shown. The top three values in each column are highlighted.

(a) CBraMod

Ratio	TUEV	PhysioNet-MI	MODMA	Speech	BCI IV-2a
20%	**+3.60**	+3.01	**+1.62**	**+2.95**	**+1.55**
25%	+3.51	+2.99	+1.61	+2.89	+1.53
30%	+3.49	+3.03	+1.58	+2.91	+1.54
35%	+3.57	+3.00	+1.58	+2.91	+1.50
40%	+3.52	+3.05	+1.60	+2.84	+1.08
45%	+3.55	+3.11	+1.61	+2.85	+1.11
50%	+3.55	**+3.12**	+1.59	+2.82	+0.95

(b) EEGNet

Ratio	TUEV	PhysioNet-MI	MODMA	Speech	BCI IV-2a
20%	**+3.13**	**+2.14**	**+1.37**	**+2.16**	**+1.13**
25%	+2.94	+2.01	+1.31	+2.12	+1.11
30%	+3.01	+2.07	+1.22	+2.04	+1.06
35%	+2.91	+2.10	+1.09	+2.14	+0.93
40%	+2.73	+1.98	+1.28	+2.08	+0.72
45%	+2.88	+2.11	+1.33	+2.11	+0.76
50%	+2.52	+2.03	+1.29	+2.05	+0.59

Table 13: **Fine-grained augmentation-ratio sweeps.** Values are ΔAcc in percentage points (↑) from fixed-seed, fixed-update comparisons. Both backbones are evaluated on the same five datasets. The top three distinct reported values within each dataset and backbone are highlighted; ties share a color.

C Loss Function Ablation

On TUEV, we compare MeanFlow with squared ℓ_2 regression (Geng et al., 2025a) against MFF with its exact ℓ_1 average-velocity objective and a Huber variant with $\delta = 1.0$, under matched training conditions. The Huber threshold is defined in the same scaled training space as the other objectives. Architecture, training schedule, adaptive loss balancing, and evaluation protocol are held fixed. We report TS-FID, skewness KS distance, and the Wasserstein-1 distance of Hjorth complexity, following the metrics used in the main paper.

Loss	**TS-FID** ↓	**Skewness KS** ↓	**Hjorth Com.** ↓
Squared ℓ_2 (MSE)	358.49	0.412	2.918
Huber ($\delta = 1.0$)	<u>223.24</u>	<u>0.248</u>	<u>1.738</u>
ℓ_1 (default)	**80.28**	**0.110**	**0.806**

Table 14: **Loss-function ablation on TUEV.** MeanFlow uses squared ℓ_2 regression; MFF uses exact ℓ_1 regression and achieves the best reported results across all three metrics. Hjorth Com. denotes Hjorth-complexity Wasserstein-1 distance. Values are point estimates.

Huber loss reduces TS-FID from 358.49 to 223.24 relative to squared ℓ_2, but remains above the 80.28 obtained with ℓ_1. The same ordering holds for skewness KS (0.412, 0.248, and 0.110) and Hjorth-complexity distance (2.918, 1.738, and 0.806). Thus, the preference for ℓ_1 is not based solely on a comparison against MSE.

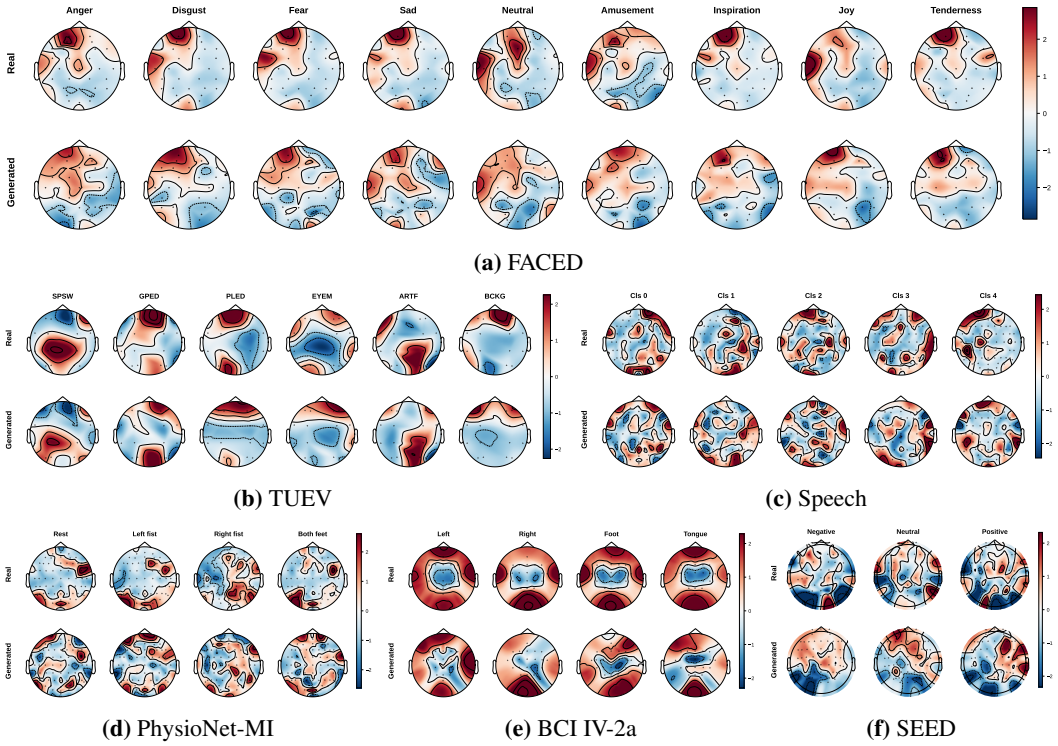
Interpretation. Reducing the influence of large residuals is a plausible explanation for the improvement over MSE, particularly when training uses a bootstrap-coupled target and sampling makes a direct endpoint prediction. However, these experiments do not isolate the full optimization mechanism or explain the entire gap between Huber and ℓ_1. They establish the empirical importance of the loss choice in the tested MFF configuration, rather than a general superiority of ℓ_1 for all EEG generators. All default MFF results use the exact ℓ_1 objective, not a Huber approximation.

D Class-Conditional Scalp Topographies Across 14 Datasets

Figures 4 and 5 show class-conditional scalp topographies across all 14 datasets. In each panel, the top row shows class-averaged topographies from real test samples, and the bottom row shows the corresponding averages from MFF-generated samples. For each dataset, we train the same MFF configuration separately, using class labels as the only conditioning signal. The model does not receive physical electrode coordinates as input; it produces signals in the fixed channel order of each dataset, with channel-index positional embeddings used by the channel branch, and the resulting channels are projected to the corresponding montage only at evaluation time.

Across datasets, MFF recovers the coarse class-distinguishing spatial structure of real EEG. Examples include the central focus of SPSW and the frontal arc of EYEM in TUEV, the anterior–posterior gradient of the abnormal class in TUAB, the lateralized focus of seizure segments in TUSZ, the bilateral frontal positivity in CHB-MIT, and the broadly distributed activity in MODMA. The consistency of these patterns indicates that MFF learns montage-specific channel organization from the training distribution through its global channel-communication mechanism, even without explicit 2-D electrode coordinates.

Two trends are visible across datasets. Datasets whose labels depend mainly on coarse spatial contrasts, such as the binary neurology datasets and the MDD-versus-healthy comparisons, are reproduced most cleanly: the dominant class signature is largely preserved with little visible distortion. Datasets whose labels depend on subtler within-region differences, such as the four-class motor imagery datasets and the multi-class emotion datasets, retain the overall spatial envelope but with lower fidelity in finer focal structure. These topographic comparisons reveal dataset-specific differences in spatial fidelity and complement the downstream utility results in Sec. 4.3. The same model configuration handles channel counts ranging from 6 (ISRUC) to 128 (MODMA), with only the dataset-specific output shape changing. Quantitative fidelity of the corresponding correlation matrices and log-PSD representations is reported in Appendix E.



(a) FACED

(b) TUEV

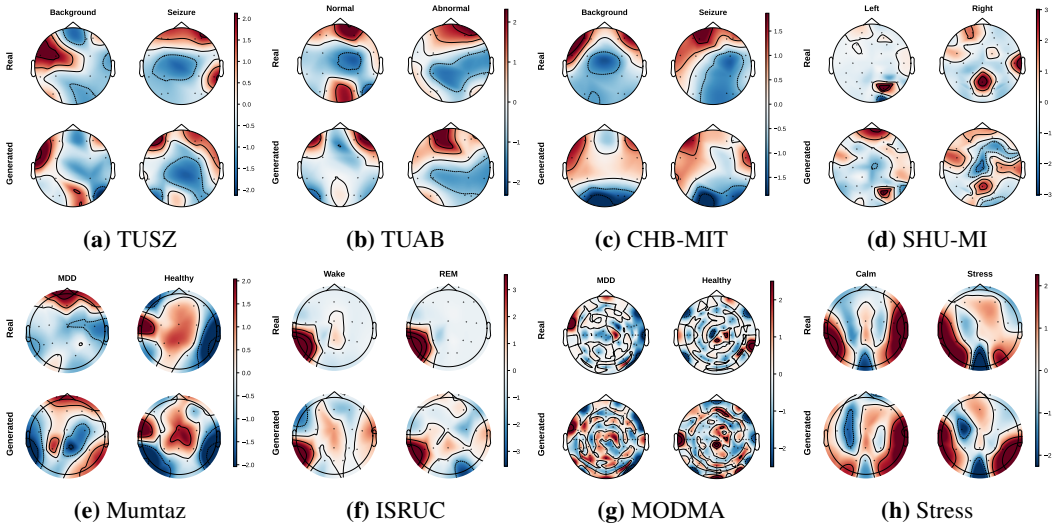
(c) Speech

(d) PhysioNet-MI

(e) BCI IV-2a

(f) SEED

Figure 4: **Real vs. generated EEG scalp topography on multi-class datasets.** Each panel shows class-averaged topographies from real test samples (top row) and generated samples (bottom row), with one column per class.



(a) TUSZ **(b)** TUAB **(c)** CHB-MIT **(d)** SHU-MI

(e) Mumtaz **(f)** ISRUC **(g)** MODMA **(h)** Stress

Figure 5: **Real vs. generated EEG scalp topography on additional datasets.** Each panel shows class-averaged topographies from real test samples (top row) and generated samples (bottom row), with one column per displayed class. ISRUC shows Wake and REM.

E Quantitative Topographic and Spectral Fidelity

We complement the class-conditional scalp topographies in Appendix D with dataset-level quantitative comparisons. Table 15 reports three metrics computed between MFF-generated and official held-out real EEG on each dataset: two metrics on the channel correlation matrix, which measure cross-channel structure, and one metric on the per-channel log-PSD matrix, which measures spectral structure.

Metrics. For each class, we compute the class-mean channel correlation matrix $R \in \mathbb{R}^{C\times C}$ and the class-mean per-channel log-PSD matrix $P \in \mathbb{R}^{C\times F}$ (Welch estimator) from real and generated samples separately. We then report: (1) **Corr-Cos** and **Corr-r**, the cosine similarity and Pearson correlation between the flattened off-diagonal entries of R^{real} and R^{gen}. Removing the diagonal isolates inter-channel coupling and avoids the trivial unit entries. (2) **PSD-Cos**, the cosine similarity between P^{real} and P^{gen}, which measures how well channel-wise spectral profiles match across frequencies. For each dataset, we average these metrics over classes using the official held-out real samples and a matched number of generated samples.

Dataset	Corr-Cos ↑	Corr-r ↑	PSD-Cos ↑
TUAB	0.880	0.894	0.986
TUEV	0.988	0.979	0.975
TUSZ	0.990	0.983	0.975
CHB-MIT	0.819	0.834	0.953
PhysioNet-MI	0.205	0.323	0.860
BCI IV-2a	0.239	0.440	0.883
SHU-MI	0.684	0.592	0.998
FACED	0.356	0.465	0.876
SEED	0.929	0.891	0.990
Mumtaz	0.882	0.515	0.902
MODMA	0.876	0.266	0.987
Speech	0.751	0.271	0.989
ISRUC	0.773	0.748	0.975
Stress	0.836	0.459	0.907
Average	0.729	0.619	0.947

Table 15: **Correlation-matrix and spectral fidelity across 14 datasets.** For each dataset, we compute class-mean channel correlation matrices and per-channel log-PSD matrices from held-out real samples and an equal number of generated samples. Corr-Cos and Corr-r are computed on the off-diagonal entries of the correlation matrix; PSD-Cos is computed on the class-mean log-PSD matrix.

Table 15 shows that spectral fidelity is uniformly strong. PSD-Cos is at least 0.86 on every dataset and 0.95 on 9 of the 14 datasets, with an average of 0.947. This indicates that MFF matches the channel-wise spectral profile of real EEG not only on clinical corpora but also on motor imagery, emotion, and cognition datasets.

Correlation-matrix fidelity depends more strongly on the granularity of spatial structure. The highest values appear on datasets whose class structure is governed by coarse spatial organization: the four neurology datasets all achieve Corr-Cos ≥ 0.819, and SEED, Mumtaz, MODMA, Stress, and ISRUC also remain high. TUEV and TUSZ are especially close to real EEG under this metric, with Corr-Cos near 0.99. By contrast, PhysioNet-MI (0.205), BCI IV-2a (0.239), and FACED (0.356) are substantially lower, while SHU-MI is intermediate at 0.684. On these datasets, the gap between spectral and correlation fidelity identifies inter-channel dependence as a remaining challenge for accurate EEG generation. Even in these cases, PSD-Cos remains high, suggesting that the remaining gap lies primarily in fine inter-channel coupling rather than in overall spectral shape.

E.1 Physiologically Relevant Signal Structure

We complement correlation-matrix and PSD fidelity with three measurements: relative band-power distance over the delta, theta, alpha, beta, and gamma bands; correlation between real and generated class-conditional phase-locking-value (PLV) matrices; and spatial correlation between class-conditional mu/beta event-related desynchronization (ERD) maps. Band-power and PLV results are averaged over TUEV, BCI IV-2a, ISRUC, and MODMA. ERD is evaluated on BCI IV-2a relative to its pre-cue baseline. Each comparison uses equally sized generated and held-out real sets. A real–real reference is obtained by averaging five random equal splits of the held-out data.

Method	Band-power distance ↓	PLV-matrix correlation ↑	Mu/beta ERD correlation ↑
EEG-GAN	0.156	0.427	0.378
GPT2MEG	<u>0.102</u>	<u>0.626</u>	<u>0.514</u>
MFF	**0.032**	**0.829**	**0.798**
Real split vs. real split	0.021	0.913	0.887

Table 16: **Physiologically relevant structure in generated EEG.** Band-power and PLV results are four-dataset averages; ERD is evaluated only on BCI IV-2a. The real–real row is a finite-sample reference, not another generator or a formal bound. Ranking highlights compare only the three evaluated generators.

MFF gives the lowest band-power distance (0.032) and the highest PLV-matrix and ERD-map correlations (0.829 and 0.798) among the evaluated generators. The real–real reference remains closer under all three measurements (0.021, 0.913, and 0.887). These results support preservation of the measured spectral, phase, and event-related spatial structure beyond the statistics in Sec. 4.4. They do not establish full physiological equivalence or preservation of neural properties not tested by these metrics.

F WAVEFORM VISUALIZATION ON TUEV

The class-averaged topographies in Appendix D evaluate coarse spatial organization, but they do not show the morphology of individual generated trials. We therefore visualize representative single-trial waveforms on TUEV in Fig. 6. We choose TUEV because it is a multi-class neurology dataset whose labels correspond to clinically interpretable event morphologies, making raw trace visualization meaningful at the single-trial level.

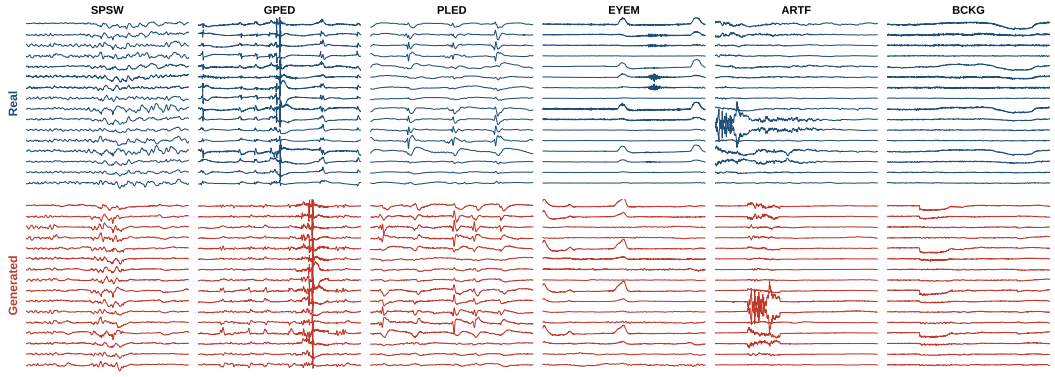
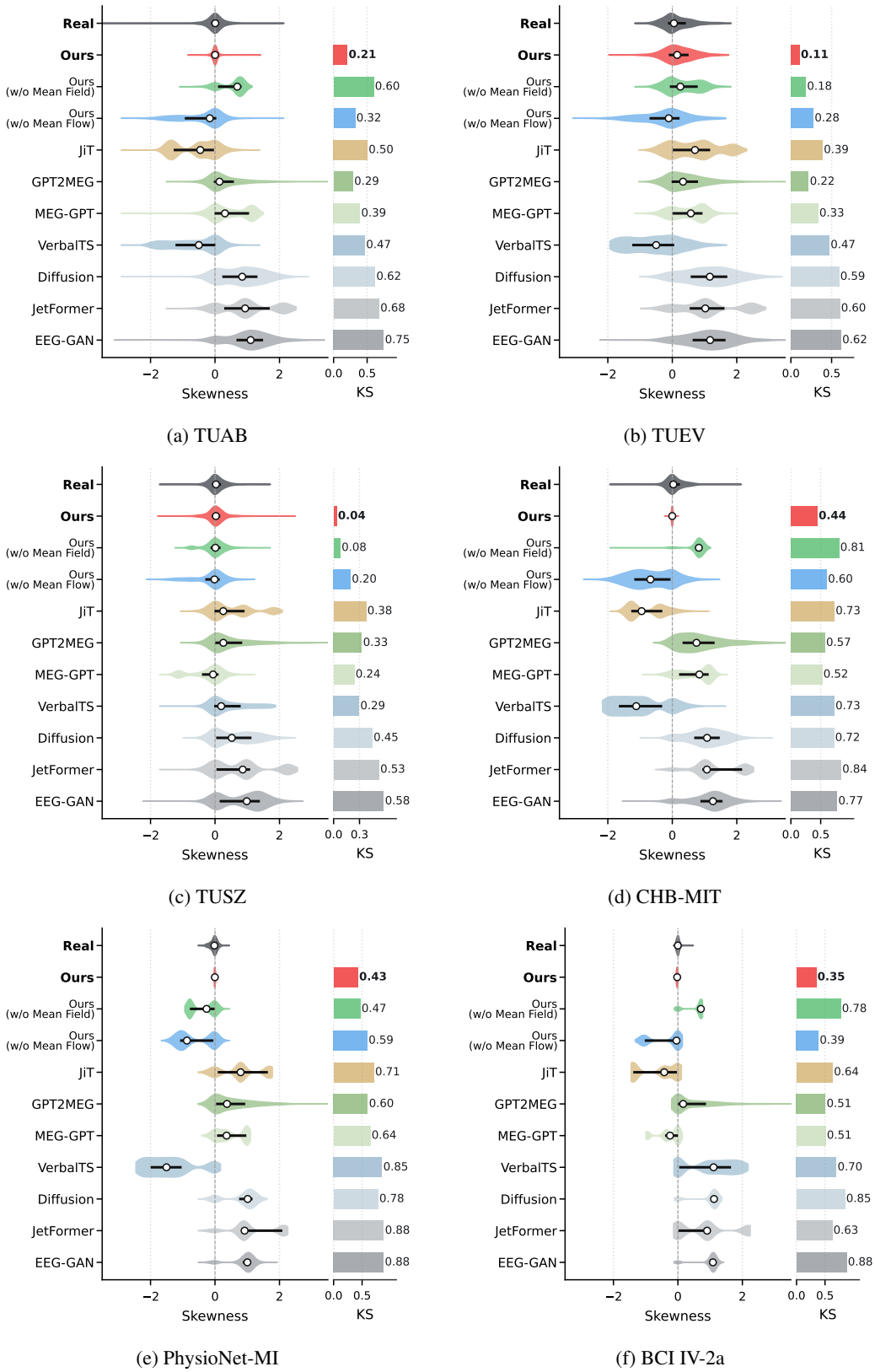


Figure 6: **Single-trial waveform examples on TUEV.** For each TUEV event class, we show stacked multi-channel waveforms from a real sample and an MFF-generated sample under the same preprocessing used in evaluation. Channels are vertically offset for visualization. The generated traces reproduce characteristic single-trial temporal morphology, including sharp-wave activity in SPSW, periodic discharge patterns in GPED and PLED, eye-movement deflections in EYEM, artifact-like transients in ARTF, and lower-amplitude background activity in BCKG. These examples complement the class-averaged scalp topographies by visualizing individual raw EEG morphology.

G PER-CHANNEL SKEWNESS ACROSS 14 DATASETS

Figures 7, 8 and 9 show per-channel skewness distributions across all 14 datasets. Each panel compares real and generated EEG with violin plots and reports, in the rightmost column, the Kolmogorov–Smirnov distance from each method to the real skewness distribution. All values are computed under the same protocol as in the main text: skewness is measured channel-wise on bandpass-filtered single trials, with real and generated samples normalized independently by the median per-trial channel standard deviation. Across the 14 datasets, MFF attains the lowest KS distance on 13 datasets, with values ranging from 0.04 on TUSZ to 0.58 on Speech. On Speech, MFF ranks second (0.58), behind MEG-GPT (0.56).

The baseline failure modes are also consistent across datasets and mirror the pattern highlighted on TUEV in the main text. Methods such as VerbalTS and MEG-GPT tend to produce overly narrow per-channel skewness distributions concentrated near zero or mildly negative values. Methods such as Diffusion, JetFormer, and EEG-GAN under-represent the negative tail and shift the distribution toward positive skew. MFF comes closest to the real distribution in support, center, and spread. The two ablations retain part of this shape, but with broader or shifted distributions, consistent with the complementary roles of mean-field token coupling and the average-velocity objective discussed in Sec. 4.5.

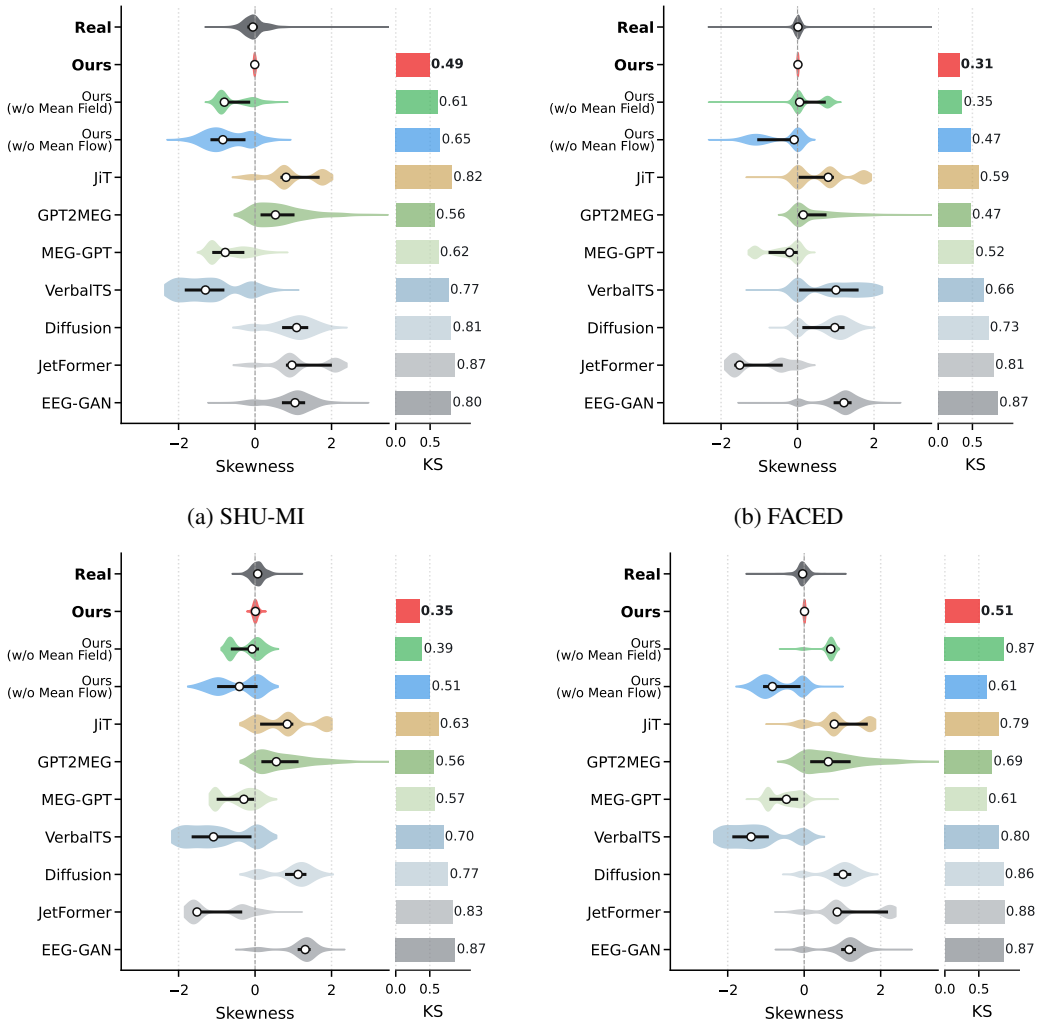


(a) TUAB (b) TUEV

(c) TUSZ (d) CHB-MIT

(e) PhysioNet-MI (f) BCI IV-2a

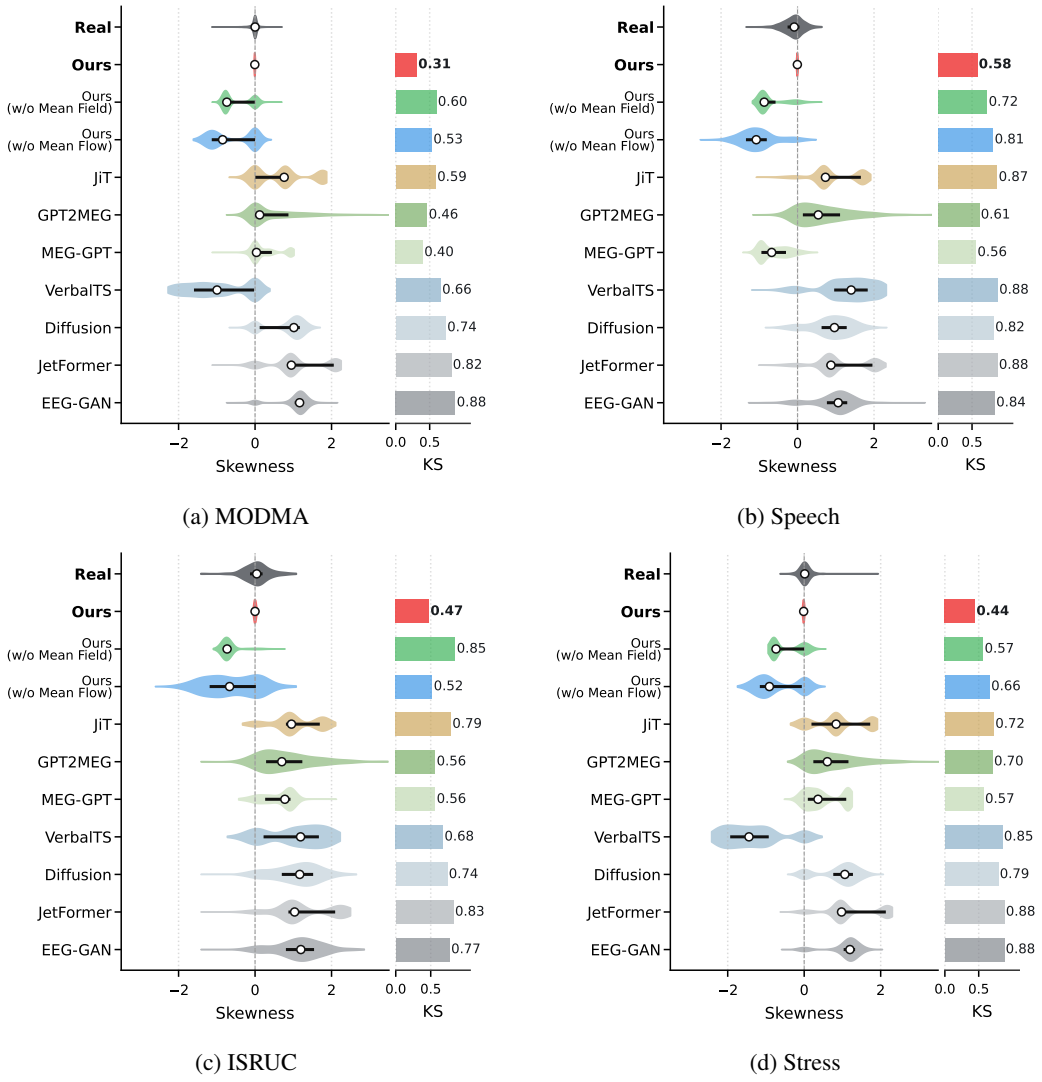
Figure 7: Per-channel skewness distributions on Neurology and Motor Imagery datasets (part 1/3). Each panel shows skewness of generated signals from all methods alongside Real data (top), with the KS distance to Real annotated on the right. Lower KS indicates a closer match to the real distribution. **Ours** achieves the lowest KS on every dataset.



(a) SHU-MI (b) FACED

(c) SEED (d) Mumtaz

Figure 8: Per-channel skewness distributions on Motor Imagery and Emotion datasets (part 2/3). Each panel shows skewness of generated signals from all methods alongside Real data (top), with the KS distance to Real annotated on the right. Lower KS indicates a closer match to the real distribution. **Ours** achieves the lowest KS on every dataset.



(a) MODMA (b) Speech

(c) ISRUC (d) Stress

Figure 9: Per-channel skewness distributions on Emotion and Cognition datasets (part 3/3). Each panel shows skewness of generated signals from all methods alongside Real data (top), with the KS distance to Real annotated on the right. Lower KS indicates a closer match to the real distribution. **Ours** achieves the lowest KS on three of the four datasets shown and ranks second on Speech.

H HJORTH PARAMETERS ACROSS 14 DATASETS

Table 17 extends the representative four-dataset comparison in Table 4 to all 14 datasets. For each method, we report Wasserstein-1 distance to real EEG for the three Hjorth parameters under the same protocol used in the main text: Hjorth statistics are computed channel-wise on bandpass-filtered signals after real and generated trials are normalized independently by the median channel-wise temporal standard deviation, and the resulting values are pooled across trials and channels.

Across all 14 datasets and all three Hjorth parameters, MFF achieves the lowest distance to real EEG in every case. The two controlled ablations account for most of the second- and third-best results, and no external baseline matches MFF on any dataset–parameter pair. The conclusion from the main text therefore holds across task families and channel counts: under these signal-level summaries of temporal dynamics, MFF is consistently the closest to real EEG.

The margin over the strongest external baseline also depends on dataset structure, echoing the TS-FID pattern in Sec. 4.2. On lower-variability settings such as TUAB, BCI IV-2a, and MODMA, the gap is modest because strong baselines already capture much of the Hjorth structure. On datasets whose labels track global brain states or nonstationary events, such as TUEV, SEED, and FACED, the gap becomes much larger. On SEED, for example, MFF reduces the complexity distance to 1.037, compared with 8.561 for MEG-GPT, the strongest external baseline.

Task	Dataset	Metric	*GAN & AR*		*Diffusion*		*MEG Transfer*		*Flow Matching*			
			EEG-GAN	JetFormer	Diffusion	VerbalTS	MEG-GPT	GPT2MEG	JiT	Ours (w/o Mean Flow)	Ours (w/o Mean Field)	**Ours**
Neurology	TUAB	Act	0.352	0.344	0.361	0.285	0.264	0.249	0.280	0.267	0.298	**0.217**
		Mob	0.440	0.467	0.433	0.346	0.345	0.335	0.376	0.339	0.361	**0.288**
		Com	3.259	3.343	3.278	2.630	2.520	2.396	2.766	2.435	2.739	**2.064**
	TUEV	Act	1.113	1.175	1.169	1.091	0.950	0.935	0.888	0.923	0.443	**0.134**
		Mob	0.161	0.168	0.158	0.127	0.118	0.108	0.129	0.135	0.068	**0.026**
		Com	4.832	4.928	4.350	4.008	3.757	3.390	3.960	3.671	2.112	**0.806**
	TUSZ	Act	0.780	0.800	0.693	0.627	0.554	0.699	0.543	0.559	0.441	**0.193**
		Mob	0.230	0.216	0.195	0.156	0.164	0.182	0.159	0.162	0.130	**0.053**
		Com	0.865	0.818	0.737	0.618	0.665	0.713	0.691	0.633	0.529	**0.236**
	CHB-MIT	Act	0.653	0.643	0.640	0.556	0.585	0.579	0.586	0.572	0.592	**0.536**
		Mob	0.487	0.506	0.502	0.405	0.383	0.424	0.436	0.393	0.459	**0.370**
		Com	2.194	2.435	2.173	1.960	1.898	2.012	2.022	1.907	2.286	**1.816**
Motor Imagery	PhysioNet-MI	Act	0.977	1.040	0.996	0.983	0.802	0.770	0.744	0.742	0.695	**0.502**
		Mob	0.695	0.715	0.653	0.593	0.524	0.503	0.539	0.483	0.419	**0.288**
		Com	2.707	2.626	2.677	2.415	2.042	1.999	2.145	2.041	1.879	**1.379**
	BCI IV-2a	Act	0.217	0.219	0.236	0.216	0.218	0.212	0.224	0.222	0.214	**0.202**
		Mob	0.359	0.326	0.331	0.330	0.322	0.313	0.315	0.328	0.335	**0.293**
		Com	0.640	0.626	0.646	0.656	0.622	0.615	0.620	0.630	0.608	**0.579**
	SHU-MI	Act	2.432	2.503	2.525	2.016	1.838	1.953	1.852	2.095	0.991	**0.701**
		Mob	0.953	0.940	0.856	0.679	0.750	0.702	0.737	0.695	0.358	**0.259**
		Com	7.420	6.547	6.727	5.262	5.275	6.177	5.985	5.567	3.067	**2.385**
Emotion	FACED	Act	0.992	1.030	1.035	1.015	0.801	0.933	0.779	0.824	0.430	**0.335**
		Mob	0.806	0.883	0.823	0.721	0.642	0.637	0.641	0.651	0.267	**0.188**
		Com	3.359	3.333	3.281	3.130	2.706	2.809	2.965	2.540	1.346	**1.111**
	SEED	Act	0.632	0.662	0.668	0.638	0.518	0.628	0.494	0.531	0.109	**0.068**
		Mob	1.860	1.881	1.812	1.516	1.312	1.289	1.464	1.526	0.267	**0.161**
		Com	10.826	10.577	9.441	9.721	8.561	9.378	8.653	8.857	1.678	**1.037**
	Mumtaz	Act	0.449	0.422	0.418	0.327	0.339	0.322	0.355	0.319	0.359	**0.304**
		Mob	0.403	0.400	0.350	0.301	0.289	0.279	0.311	0.275	0.335	**0.267**
		Com	1.060	0.998	0.994	0.825	0.818	0.786	0.815	0.780	0.943	**0.752**
	MODMA	Act	0.167	0.166	0.146	0.147	0.103	0.113	0.122	0.108	0.115	**0.096**
		Mob	0.398	0.440	0.394	0.405	0.303	0.305	0.328	0.312	0.323	**0.280**
		Com	2.036	2.117	1.954	1.899	1.495	1.540	1.653	1.591	1.708	**1.427**
Cognition	Speech	Act	0.784	0.837	0.812	0.768	0.641	0.665	0.618	0.633	0.564	**0.513**
		Mob	0.800	0.827	0.783	0.636	0.537	0.529	0.608	0.573	0.555	**0.430**
		Com	7.439	7.635	7.082	5.662	5.202	5.189	5.955	5.084	5.236	**4.214**
	ISRUC	Act	4.500	4.493	4.085	3.449	3.411	3.270	3.338	3.246	3.787	**3.148**
		Mob	0.462	0.446	0.411	0.339	0.336	0.333	0.353	0.346	0.360	**0.315**
		Com	1.190	1.177	1.115	0.943	0.957	0.920	0.990	0.942	1.046	**0.884**
	Stress	Act	0.453	0.436	0.460	0.456	0.321	0.362	0.343	0.347	0.262	**0.214**
		Mob	0.643	0.647	0.652	0.562	0.410	0.464	0.513	0.479	0.334	**0.275**
		Com	1.434	1.426	1.322	1.392	1.080	1.119	1.177	1.160	0.921	**0.724**

Table 17: **Hjorth-parameter Wasserstein-1 distance (↓) across 14 datasets.** All results are computed under the same evaluation protocol. **Best**, second, and third are highlighted.

I Additional Architecture and Objective Controls

We extend the component ablations in Sec. 4.5 with token-interaction replacements, directional controls, a field-width sweep, comparisons with other sampling formulations, and a controlled perturbation diagnostic. Unless a setting is explicitly varied, each comparison retains the corresponding dataset split and evaluation protocol. We report point estimates for the additional controlled experiments.

I.1 Token Interaction and Spatial–Temporal Coupling

Scope and controls. The temporal field is recomputed independently for each segment and network block. It couples temporal patches within one channel and one segment; the spatial field couples channels within one temporal patch. The temporal field is therefore shared over the current 1–30-second segment, not an entire recording. To isolate the two directions, we replace the temporal or spatial mean-field sublayer with self-attention while retaining the other mean-field branch. The dual-axis layout, hidden dimension, depth, mean-flow objective, and training schedule remain unchanged. Replacing both directions gives Ours (w/o Mean Field).

We additionally replace the mean-field sublayers with MLP-Mixer-style token mixing on TUEV, keeping the dual-axis layout, hidden size, depth, objective, and training protocol fixed. Channel-only and time-only token-MLP variants provide additional axis-restricted controls. Table 18 combines these alternatives with the directional mean-field ablations.

Variant	TS-FID ↓	Hjorth Com. ↓	ΔAcc ↑
TUEV: 5-second segments			
Self-attention (w/o Mean Field)	205.52	2.112	+2.87
MLP-Mixer replacement	246.47	2.384	+2.18
Channel-only token MLP	321.93	3.018	+1.16
Time-only token MLP	277.61	2.897	+1.29
Spatial mean field only	132.41	1.412	+3.04
Temporal mean field only	146.88	1.784	+2.91
MFF (both mean fields)	**80.28**	**0.806**	**+3.60**
ISRUC: 30-second segments			
Self-attention (w/o Mean Field)	713.06	1.046	+1.18
Spatial mean field only	701.39	1.223	+1.72
Temporal mean field only	676.85	0.968	+1.91
MFF (both mean fields)	**639.45**	**0.884**	**+2.28**

Table 18: **Token-interaction and directional controls.** A single-axis mean-field variant uses self-attention in the other branch; it does not remove that branch. Hjorth Com. is Hjorth-complexity Wasserstein-1 distance, and ΔAcc is CBraMod accuracy change in percentage points with 20% augmentation. Rankings are computed separately within each dataset.

Results. On TUEV, MFF reduces TS-FID to 80.28, compared with 205.52 for self-attention and 246.47 for the dual-axis MLP-Mixer replacement. Retaining only the spatial or temporal mean field gives intermediate TS-FID values of 132.41 and 146.88. On ISRUC, temporal-only coupling gives lower TS-FID than spatial-only coupling (676.85 versus 701.39), while both fields reduce it further to 639.45. The directional effects are dataset- and metric-dependent: on ISRUC, spatial-only coupling increases Hjorth-complexity distance from 1.046 to 1.223 relative to self-attention. Nevertheless, using both fields gives the best reported TS-FID, Hjorth-complexity distance, and downstream gain on both datasets. The result supports complementary spatial and temporal interactions, not a uniform improvement from every single-axis replacement.

I.2 Field Dimensionality on High-Density EEG

We vary the field dimension on 128-channel MODMA while holding the hidden size, depth, training schedule, and evaluation protocol fixed. The tested dimensions are $D_c \in \{96, 192, 384\}$, corresponding to $H/8$, $H/4$, and $H/2$ for $H = 768$. Corr-Cos and Corr-r use the off-diagonal class-mean channel correlations defined in Appendix E.

Field dimension	TS-FID ↓	Corr-Cos ↑	Corr-r ↑	ΔAcc ↑
96 ($H/8$)	701.82	0.754	0.202	+1.19
192 ($H/4$, default)	**674.54**	**0.876**	**0.266**	**+1.62**
384 ($H/2$)	678.91	0.868	0.261	+1.52

Table 19: **Field-width ablation on 128-channel MODMA.** The default $D_c = 192$ gives the best point estimate under all four measurements. Accuracy changes are in percentage points with 20% augmentation. Hidden size and depth are fixed; total parameter counts need not be identical when field width changes.

Reducing the field dimension to 96 degrades all four metrics, while increasing it to 384 does not improve the reported values. Within this dataset and tested width range, simply enlarging the field beyond 192 does not improve the evaluated fidelity or utility. This is not a controlled scaling study over channel count, nor does it establish that fine-grained inter-channel structure is fully captured on arbitrary high-density EEG.

I.3 Average-Velocity Training and Other Sampling Formulations

We compare MFF with standard flow matching (FM), a consistency model (Song et al., 2023a), and EEG-GAN on TUEV. The compared methods use the same split, preprocessing, class conditioning, approximate 100M-parameter scale, training budget, and evaluation. Standard FM retains mean-field coupling and is evaluated with either a single sampling step or 50-step Heun integration. The consistency model enforces agreement between adjacent noise levels under a clean-endpoint boundary and generates a sample in one evaluation.

Method	Steps	TS-FID ↓	ΔAcc ↑
EEG-GAN	1	486.73	+0.17
Standard FM (one step)	1	706.31	+0.63
Standard FM (Heun)	50	362.84	+2.13
Consistency model	1	227.48	+2.72
MFF	1	**80.28**	**+3.60**

Table 20: **Sampling-formulation comparisons on TUEV.** Standard FM retains the mean-field architecture. Steps denote sampling updates, not the number of network evaluations; a Heun update may require more than one evaluation. ΔAcc is measured in percentage points using 20% augmentation.

Fifty-step Heun sampling improves standard FM over its one-step use, reducing TS-FID from 706.31 to 362.84. The consistency model further reduces it to 227.48, but remains above MFF at 80.28; downstream gains follow the same ordering. Since both consistency and EEG-GAN also support one-step generation, one-step sampling alone does not account for MFF's performance in these comparisons. The results support the evaluated MFF configuration, not a claim that flow matching is the only viable framework for one-step EEG generation. The separate loss ablation in Appendix C also shows that the regression objective materially affects performance.

I.4 Controlled Perturbation Accumulation

We test error retention in a valid 50-step standard-FM sampler on TUEV, rather than recursively applying a model trained for one-step endpoint prediction. For paired random seeds and class

labels, we store an unperturbed trajectory and inject the same sequence of zero-mean Gaussian signal-space perturbations after the sampling updates, with a cumulative signal-to-noise ratio of 20 dB. Periodic resets to the paired clean state control the maximum number of steps over which perturbations are retained. The model, time grid, number of network evaluations, initial states, and injected perturbations remain fixed.

We measure endpoint deviation from the paired clean sample as

$$\mathrm{NRMSE} = \frac{\|x_{\mathrm{pert}} - x_{\mathrm{clean}}\|_2}{\|x_{\mathrm{clean}}\|_2}. \tag{35}$$

Condition	Retained horizon	TS-FID ↓	NRMSE ↓
Clean standard FM	0	362.84	0.000
Reset every step	1	**374.90**	**0.018**
Reset every 5 steps	5	401.37	0.043
Reset every 10 steps	10	432.85	0.067
No reset	50	487.26	0.112

Table 21: **Controlled perturbation accumulation in 50-step standard FM.** Horizon is measured in sampling steps. All perturbed conditions receive the same injected perturbation sequence; resetting changes how long perturbations remain in the evolving state. The clean condition is a reference without injected perturbations. Colors rank only the perturbed conditions within each metric.

As the maximum retained horizon increases from 1 to 50 steps, TS-FID rises from 374.90 to 487.26 and endpoint NRMSE from 0.018 to 0.112. This behavior is consistent with cumulative perturbation retention during iterative sampling. It does not establish that acquisition noise is injected during ordinary generation, that errors are amplified rather than accumulated, or that this mechanism alone explains the clean-performance gap between standard FM and MFF.

J SENSITIVITY TO PREPROCESSING

Matched-pipeline evaluation. We retrain MFF, EEG-GAN, Diffusion, and GPT2MEG independently under alternative preprocessing pipelines. On TUEV, we evaluate the default pipeline, 1–40 Hz band-pass filtering, common-average referencing (CAR), and band-pass filtering with CAR. On ISRUC and 128-channel MODMA, we evaluate the default, band-pass, and band-pass-plus-CAR pipelines. For every dataset and setting, generated and held-out real EEG are compared under identical preprocessing. The data split, conditioning, training and generated-sample budgets, and evaluation protocol are otherwise held fixed.

For TUEV, we also evaluate 20% MFF augmentation with pretrained CBraMod and randomly initialized EEGNet. Each real-only/augmented pair uses the same fixed seed, initialization, optimizer-update budget, and model-selection protocol. These are within-pipeline comparisons after retraining, not tests of a single fixed generator under an unseen preprocessing transformation.

(a) Generation fidelity: TS-FID ↓

Dataset	Preprocessing	EEG-GAN	Diffusion	GPT2MEG	MFF
TUEV	Default	486.73	454.12	346.09	**80.28**
	Band-pass	425.66	411.03	304.49	**74.32**
	CAR	519.94	481.29	391.70	**91.97**
	Band-pass + CAR	468.85	463.12	339.37	**77.84**
ISRUC	Default	846.14	791.06	656.47	**639.45**
	Band-pass	793.52	742.68	615.31	**586.72**
	Band-pass + CAR	759.84	708.53	581.94	**552.17**
MODMA	Default	924.73	879.52	714.87	**674.54**
	Band-pass	861.35	822.46	665.72	**623.81**
	Band-pass + CAR	812.67	774.15	621.08	**574.36**

(b) TUEV downstream utility: ΔAcc (pp) ↑

Preprocessing	CBraMod	EEGNet
Default	+3.60	+3.13
Band-pass	+3.66	+2.97
CAR	+3.05	+2.56
Band-pass + CAR	+3.59	+3.32

Table 22: **Sensitivity to matched preprocessing pipelines.** Band-pass denotes 1–40 Hz filtering. Panel (a) compares generators within each dataset–pipeline setting; ranking colors apply within rows. Panel (b) reports augmentation gains over the matching real-only classifier on TUEV. CAR-only and downstream comparisons are evaluated only on TUEV.

Results. MFF gives the lowest TS-FID in all ten reported dataset–pipeline settings. On TUEV, its augmentation gains remain positive across all four pipelines, ranging from $+3.05$ to $+3.66$ points for CBraMod and from $+2.56$ to $+3.32$ points for EEGNet. Absolute TS-FID values should not be compared across preprocessing pipelines as if they measured an unchanged target distribution: filtering and rereferencing alter that distribution. The supported conclusion is that MFF retains its advantage over the evaluated generators within the tested matched settings, not that MFF is invariant to arbitrary preprocessing.