PHYSIOLOGY-GUIDED AGE–STAGE-CONSISTENT DIFFUSION FOR SLEEP EEG GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional diffusion models can generate physiological time series that appear realistic in spectral and feature statistics; however, marginal realism does not establish whether age-related changes remain compatible with an assigned sleep stage. We study this *globally realistic yet conditionally contradictory* behavior as age–stage physiological inconsistency in sleep EEG. To address it, we propose Phy-DiffKAN, which represents condition-dependent physiological structure as a factorized prior field over an interpretable 12-dimensional coordinate space. The prior field decomposes age, recorded sex, sleep stage, and their pairwise interactions into six branches, smoothly approximates the finite-anchor states of each branch, and projects them into physiology-structured tokens for a Transformer diffusion denoiser. Branch-drop and branch-mismatch constraints further enforce branch necessity and value specificity. We evaluate Phy-DiffKAN on a subject-disjoint Sleep-EDF cohort with controlled demographic allocation and complete per-subject stage coverage, using overall generation quality, fixed-stage preservation, and age-edit responsiveness. Phy-DiffKAN achieves an FID of 0.53, a feature-manifold F1 score of 0.78, a LogPSD distance of 0.06, and a five-stage balanced accuracy of 0.87. Under young→old REM reassignment with fixed diffusion noise, Phy-DiffKAN preserves REM-stage identity, suppresses leakage into N2, and retains near-real age decodability; complementary N1 age sweeps show the same stage-preservation pattern beyond REM.

1 INTRODUCTION

Recent advances in diffusion models have substantially improved the quality of EEG generation and increasingly enabled conditioning on physiological attributes, age, and sex (Yuan & Qiao, 2024; Li et al., 2025b; Torma & Szegletes, 2025). However, *whether generated samples resemble real data* and *whether age-conditioned samples preserve their assigned sleep stage* are distinct questions: the former concerns global realism, whereas the latter requires age-related modulation to remain compatible with stage physiology. A generated signal may appear realistic in the spectral domain, feature space, and downstream classification, yet remain physiologically contradictory under an age–stage assignment.

Sleep EEG provides a representative setting in which to study this problem. Sleep stages determine the dominant morphological structure of EEG, whereas age and recorded sex modulate spectral composition, amplitude characteristics, and transient activity within each stage (Sanchez & Tsaftaris, 2022; Rahman et al., 2024). When latent noise and rapid eye movement (REM) stage are fixed and only age is changed, a consistent generator should express age-related differences while preserving REM identity. Because age and stage are correlated in observational data, a model may instead respond to age by altering the stage or retain only stage-related characteristics. The target edit then changes a non-target physiological factor, producing globally realistic signals with an incoherent age–stage assignment. We refer to this failure as *age–stage physiological inconsistency*: **A model may approximate the marginal distribution of real physiological signals without learning how age and stage should jointly hold within the same sample**.

Existing approaches struggle to identify such condition substitution. Shared conditioning embeddings may achieve favorable generation metrics by reproducing population-level correlations, yet fail to learn how attributes modulate signals within each stage (Li et al., 2025b; Geirhos et al.,

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.

2020). Rigorous causal generative methods, in contrast, typically rely on causal graphs, interventional distributions, or identifiability assumptions that are difficult to obtain (Schölkopf et al., 2021; Rahman et al., 2024). Meanwhile, FID, feature-manifold coverage, and power spectral distance primarily measure overall distributional similarity, but cannot determine whether changes occur along the intended physiological dimension (Heusel et al., 2017; Kynkäänniemi et al., 2019). This raises a central question: **How can we learn and verify age–stage physiological consistency within the same sample when globally realistic signals may still carry a contradictory age–stage assignment?**

To address this question, we propose Phy-DiffKAN and investigate whether implicitly entangled conditional effects can be transformed into a structured, reassignable, and testable generative representation. The proposed method constructs a factorized physiological prior field comprising 12 interpretable coordinates and organizes age, recorded sex, sleep stage, and their pairwise interactions into six condition branches. We define the target field over 90 finite anchors formed by combinations of age, sex, and sleep stage. It continuously parameterizes this field, while each branch is encoded as an independent structured condition token that is accessed by the diffusion denoiser through cross-attention. During training, we further introduce branch-drop and branch-mismatch constraints to enforce branch necessity and value specificity, respectively.

Our contributions are:

- We formulate age–stage physiological inconsistency in sleep EEG, showing that marginal realism can coexist with age edits that shift stage identity. We identify this condition substitution as a measurable failure mode in age-conditioned sleep EEG generation.
- We propose Phy-DiffKAN, which combines a factorized physiological prior field, branch-specific condition tokens, and branch-drop/branch-mismatch constraints to transform implicit conditional control into a reassignable, testable, and auditable generative structure.
- We establish an evaluation framework centered on fixed-stage preservation and age-edit responsiveness. Our results reveal that similar overall generation quality can correspond to substantially different age–stage behaviors, motivating consistency evaluation beyond marginal realism.

All models are evaluated on the same subject-disjoint Sleep-EDF cohort with equal subject allocation across age strata, near-balanced recorded sex, and complete per-subject stage coverage, restricting simple shortcuts arising from unequal demographic frequencies. Ours attains near-best overall signal quality and a five-stage balanced accuracy of 0.87. Under young→old REM reassignment, it achieves the highest old-REM recall and the lowest hard/soft REM→N2 leakage, indicating preserved stage identity. Its age-probe balanced accuracy reaches 0.72, close to 0.73 on held-out real data, showing that age information remains detectable. Together, within the tested cohort and assignments, these results support model-level age sensitivity with stage preservation.

2 Related Work

2.1 Time-Series and Physiological Generation

TimeGAN learns temporal embeddings (Yoon et al., 2019); diffusion supports forecasting and synthesis (Rasul et al., 2021; Shen & Kwok, 2023; Kollovieh et al., 2023), with trend/seasonality (Yuan & Qiao, 2024) or distributional objectives (Li et al., 2025a). Physiological work spans conditional ECG/EEG (Chen et al., 2022; Klein et al., 2024) and spectrally constrained sleep EEG (Aristimunha et al., 2023).

N2 spindles/K-complexes, REM mixed-frequency activity, and N3 slow waves (American Academy of Sleep Medicine, 2023) undergo stage-dependent age modulation (Sun et al., 2023). Label and spectral controls (Wickramaratne & Parekh, 2023; Lou et al., 2026) leave joint condition effects implicit, motivating explicit age–stage consistency evaluation beyond marginal realism.

2.2 Causal and Mechanism-Based Generation

Invariant predictors (Arjovsky et al., 2019) and counterfactual structure (Sauer & Geiger, 2021) address observational shortcuts. CausalGAN uses causal graphs (Kocaoglu et al., 2018), Causal-

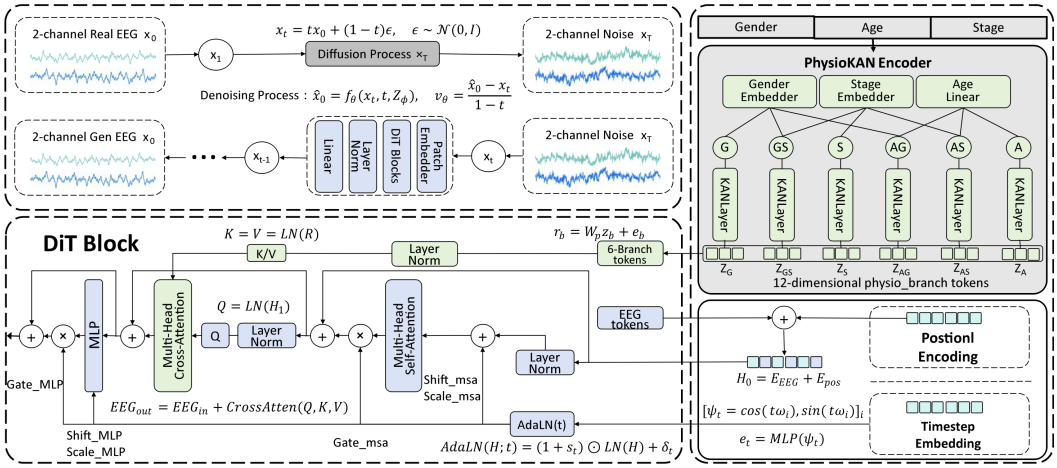

Figure 1: The overall framework of Phy-DiffKAN.

VAE/DEAR use supervised causal factors (Yang et al., 2021; Shen et al., 2022), and Diff-SCM/DCM embed diffusion in structural mechanisms (Sanchez & Tsaftaris, 2022; Chao et al., 2023). Sleep EEG provides waveforms and labels with unobserved physiological functions; our prior field supports model-level control, while causal identification requires further assumptions.

2.3 KAN Encoding and Transformer Diffusion Conditioning

KAN edge functions (Liu et al., 2024b) and scientific modularity (Liu et al., 2024a) support inspectable physiological anchors. Separate branch tokens preserve factor identities in Transformer diffusion conditioning (Peebles & Xie, 2023).

3 Method

3.1 Problem Formulation and Mechanism Interface

Figure 1 summarizes Phy-DiffKAN. Let $x_0 \in \mathbb{R}^{C\times T}$ denote a C-channel, T-sample sleep EEG epoch indexed by age a, sex g, and stage s. Conventional conditioning learns $p_\theta(x_0 \mid a, g, s)$ through a shared representation, which may conflate demographic modulation with stage morphology.

Sex is taken from the dataset's binary male/female subject metadata and encoded as $g = 0$ for male and $g = 1$ for female; no gender-identity variable is measured. Any legacy "gender" label retained in the architecture figures denotes this same recorded sex variable.

We instead condition generation on six branch-wise states from a factorized physiological prior field:

$$Z_\phi(a, g, s) = \{z_b\}_{b\in\mathcal{B}}, \qquad \mathcal{B} = \{a, g, s, ag, as, gs\}, \tag{1}$$

where each $z_b \in \mathbb{R}^{12}$ lies in a shared physiological coordinate system. The generator uses the structured set:

$$x_0 \sim p_\theta(x_0 \mid Z_\phi(a, g, s)). \tag{2}$$

Because the dataset contains waveforms and labels but not the physiological functions linking age, sex, and stage to morphology, the field is an explicit modeling prior rather than observed physiology. Paired edits reassign its inputs and use the resulting tokens as the generation interface. Three main-effect branches retain an interpretable decomposition, while three pairwise branches capture dependencies such as age-sensitive stage morphology. Higher-order terms are excluded because joint demographic–stage cells are sparse and would be weakly supported and difficult to evaluate on the finite anchor set.

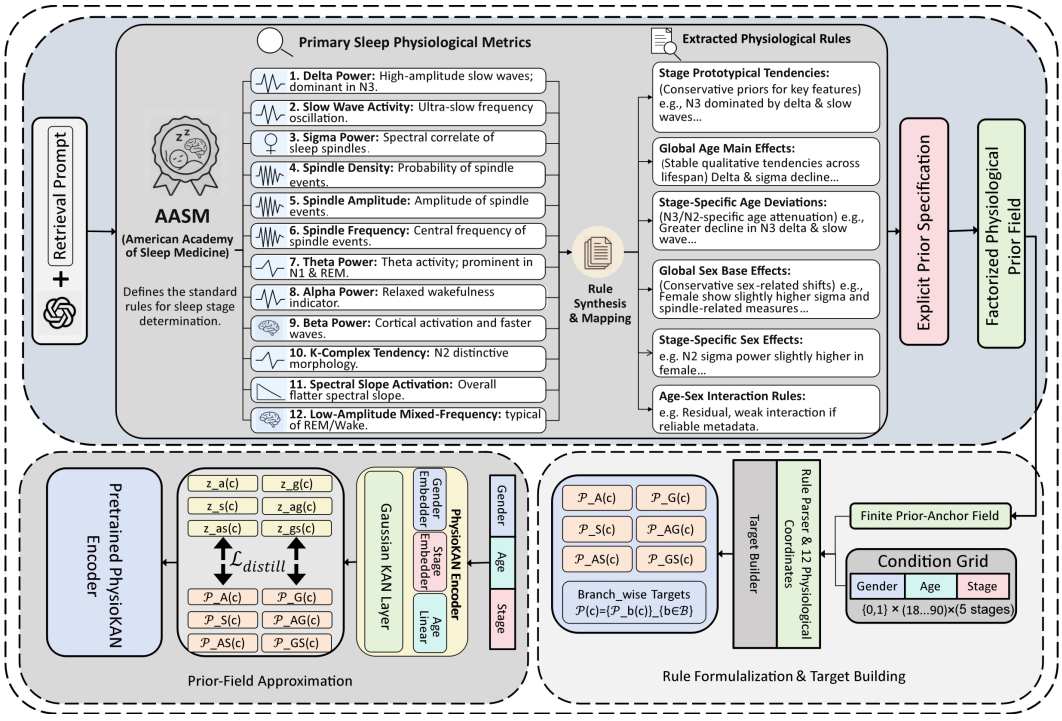

Figure 2: Factorized physiological prior-field specification and PhysioKAN approximation. The age component of the finite prior-anchor set spans 18–90 years and covers the full 25–90-year empirical EEG range.

3.2 Factorized Physiological Prior Field

Figure 2 and Table 1 define six modulation functions over 12 shared coordinates. The coordinates synthesize sleep staging, circuit physiology, age–sex spectral effects, NREM topography, and spindle phenotypes and mechanisms (Sankari & Slowik, 2026; Adamantidis et al., 2019; Sprecher et al., 2016; Carrier et al., 2001; Purcell et al., 2017; Markovic et al., 2020; Fernandez & Lüthi, 2020). They encode qualitative directions on a shared modeling scale, so the field remains an explicit modeling assumption. A machine-readable grid with branch weights and applicability masks generates targets $\tau_b(a, g, s)$ for all valid tuples; these anchors supervise PhysioKAN, which is then frozen. Appendix Section A.4 gives the complete numerical coordinate-to-anchor mapping, branch weights, and applicability rules.

Table 1: Physiological coordinates shared by all PhysioKAN branches.

Id	Coordinate	Main physiological role
0	delta power	Slow-wave sleep tendency
1	theta power	N1 and REM-like activity
2	alpha power	Wake-related rhythm
3	sigma power	Spindle-related band
4	beta power	Cortical activation
5	slow-wave activity	Slow oscillation tendency
6	spindle density	N2 spindle occurrence
7	spindle amplitude	Spindle strength
8	spindle frequency	Spindle peak tendency
9	K-complex tendency	N2 morphology
10	low-amplitude mixed frequency	REM/Wake-like activation
11	spectral-slope activation	High-frequency activation

Anchors are dimensionless modeling values, not calibrated powers, amplitudes, event rates, or effect sizes: sign encodes promotion or suppression and magnitude encodes prior strength. Stage branches define coordinate profiles; age, sex, and interaction branches provide offsets. Although power, event, and morphology coordinates may covary, Eq. 9 counts each branch–coordinate–condition value once, avoiding overlap-based reweighting.

Formally, let the finite physiological prior-anchor set be

$$\begin{aligned} \mathcal{A} &= \mathcal{A}_{\text{age}} \times \mathcal{A}_{\text{sex}} \times \mathcal{A}_{\text{stage}}, \\ \mathcal{A}_{\text{age}} &= \{18, 25, 35, 45, 55, 65, 75, 85, 90\}, \quad \mathcal{A}_{\text{sex}} = \{0, 1\}, \quad |\mathcal{A}_{\text{stage}}| = 5, \quad |\mathcal{A}| = 90. \end{aligned} \tag{3}$$

For any anchor condition $c \in \mathcal{A}$, the factorized field specifies

$$\mathcal{P}(c) = \{\mathcal{P}_b(c)\}_{b \in \mathcal{B}}, \qquad \mathcal{P}_b(c) = \tau_b(c) \in \mathbb{R}^{12}. \tag{4}$$

3.3 PhysioKAN Prior-Field Encoder

PhysioKAN continuously approximates the finite prior field. For $c = (a, g, s)$, it outputs

$$\widehat{\mathcal{P}}_\phi(c) = \{z_a(c), z_g(c), z_s(c), z_{ag}(c), z_{as}(c), z_{gs}(c)\}. \tag{5}$$

where $z_b \in \mathbb{R}^{12}$ is branch-specific and $z(a, g, s) = \sum_{b \in \mathcal{B}} z_b(a, g, s)$. PhysioKAN is distilled without waveform supervision from all 90 anchors using the four objectives below and then frozen.

Age is normalized by 90 for both prior-field evaluation and the encoding below. Learned univariate edge functions expose coordinate-wise input–response curves for comparison with the anchors, unlike latent MLP coordinates.

Age uses a linear encoding, while sex and sleep stage use embedding layers; all three condition encodings have hidden dimension 64:

$$f_a = W_a(a/90) + b_a, \qquad f_g = \text{Emb}_g(g), \qquad f_s = \text{Emb}_s(s). \tag{6}$$

The main-effect branches receive their own encodings, and the interaction branches receive pairwise composed inputs:

$$\begin{aligned} \eta_a &= f_a, & \eta_g &= f_g, & \eta_s &= f_s, \\ \eta_{ag} &= f_a \odot f_g, & \eta_{as} &= f_a \odot f_s, & \eta_{gs} &= f_g \odot f_s. \end{aligned} \tag{7}$$

Each branch satisfies $z_b(c) = h_b(\eta_b(c))$ and uses a KAN-style Gaussian basis layer with $K = 8$:

$$\phi_k(u_i) = \exp\left[-\frac{(u_i - c_k)^2}{2\sigma^2}\right], \qquad [h_b(u)]_o = (W_b u + b_b)_o + \sum_{i,k} \theta_{iko}^{(b)} \phi_k(u_i). \tag{8}$$

For $o \in \{1, \ldots, 12\}$, the centers $c_k = -2 + 4(k-1)/7$ and bandwidth $\sigma = 4/7$ are fixed; only W_b, b_b, and $\theta_{iko}^{(b)}$ are learned.

The anchor-matching loss uses the fixed branch weights:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{A}|} \sum_{c \in \mathcal{A}} \frac{\sum_{b \in \mathcal{B}} w_b \, \ell_{\text{SL1}}\left(\widehat{\mathcal{P}}_{\phi,b}(c), \mathcal{P}_b(c)\right)}{\sum_{b \in \mathcal{B}} w_b}. \tag{9}$$

The SmoothL1 term is averaged over the 12 coordinates. In branch order (a, g, s, ag, as, gs), the fixed weights are $(0.85, 0.50, 1.00, 0.40, 1.00, 0.80)$.

At age 45, let $z = \sum_b z_b$ and $\Delta_c(g) = z_j(45, g, s^+) - z_j(45, g, s^-)$ for $c = (j, s^+, s^-)$. Averaging over active constraints and both sex values gives

$$\mathcal{L}_{\text{rank}} = \frac{1}{2|\mathcal{C}_{\text{rank}}|} \sum_{\substack{c \in \mathcal{C}_{\text{rank}} \\ g \in \{0,1\}}} [m_{\text{rank}} - \Delta_c(g)]_+ . \tag{10}$$

With $m_{\text{rank}} = 0.2$, N3 delta/slow-wave activity and N2 sigma/spindle density must exceed the corresponding coordinates in other stages; Eq. 9 supervises the remaining prototypes.

For age monotonicity between 25 and 65 years, define $D_c(g) = q[z_j(65, g, s) - z_j(25, g, s)]$ for $c = (j, s, q)$:

$$\mathcal{L}_{\text{mono}} = \frac{1}{2|\mathcal{C}_{\text{mono}}|} \sum_{\substack{c \in \mathcal{C}_{\text{mono}} \\ g \in \{0,1\}}} [m_{\text{mono}} - D_c(g)]_+ . \tag{11}$$

All active constraints use $q = -1$ and $m_{\text{mono}} = 0.05$ for N3 delta/slow-wave activity and N2 sigma/spindle density; q fixes the expected direction of change.

Branch orthogonality limits redundant directions among the six outputs:

$$\mathcal{L}_{\text{orth}} = \frac{2}{|\mathcal{B}|(|\mathcal{B}| - 1)} \sum_{b_i < b_j} \mathbb{E} \left| \cos(z_{b_i}, z_{b_j}) \right| . \tag{12}$$

The complete prior-distillation objective is

$$\mathcal{L}_{\text{distill}} = w_p \mathcal{L}_{\text{anchor}} + w_k \mathcal{L}_{\text{rank}} + w_m \mathcal{L}_{\text{mono}} + w_o \mathcal{L}_{\text{orth}}. \tag{13}$$

We set $w_p = 3.0$, $w_k = w_m = 0.5$, and $w_o = 10^{-3}$. An age edit $Z_\phi(a, g, s) \to Z_\phi(a', g, s)$ recomputes z_a, z_{ag}, z_{as} but leaves z_g, z_s, z_{gs} unchanged.

3.4 Physiology-Structured Branch-Token Injection into the Diffusion Transformer

The denoiser retains the six PhysioKAN branches as separate condition tokens. For each $[B, 12]$ branch, the BranchConditioner computes

$$r_b = W_p z_b + e_b, \qquad b \in \mathcal{B}, \quad R = [r_a, r_g, r_s, r_{ag}, r_{as}, r_{gs}] \in \mathbb{R}^{6 \times H}. \tag{14}$$

Here e_b is a learned branch identity, H is the hidden width, and batched R has shape $[B, 6, H]$; optional inspection gates do not aggregate the tokens.

The noisy waveform is converted into temporally ordered patch tokens with positional encodings. Following the AdaLN-conditioned DiT design (Peebles & Xie, 2023), each block applies timestep-conditioned self-attention, lets the normalized waveform tokens cross-attend to the six normalized branch tokens without pooling them, and adds the result through a timestep-dependent gate. An AdaLN-conditioned MLP residual completes the block, while the prediction head remains timestep-conditioned.

3.5 Training with Branch-Sensitive Objectives

With linear forward noising, the denoiser predicts the clean waveform (Li & He, 2026):

$$x_t = t x_0 + (1 - t)\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \qquad \mathcal{L}_{\text{rec}} = \| f_\theta(x_t, t, Z_\phi(a, g, s)) - x_0 \|_2^2. \tag{15}$$

Because reconstruction permits branch-token shortcuts, we perturb one branch: Z is matched, $Z_{b \leftarrow \text{shuffle}(b)}$ substitutes b from another sample, and $Z_{b \leftarrow 0}$ drops it:

$$\begin{aligned} d_+ &= \| f_\theta(x_t, t, Z) - x_0 \|_2^2, \\ d_{\text{mis}} &= \| f_\theta(x_t, t, Z_{b \leftarrow \text{shuffle}(b)}) - x_0 \|_2^2, \quad d_{\text{drop}} = \| f_\theta(x_t, t, Z_{b \leftarrow 0}) - x_0 \|_2^2. \end{aligned} \tag{16}$$

The hinge losses require matched tokens to outperform mismatched or dropped tokens:

$$\mathcal{L}_{\text{mis}} = \mathbb{E}\Big[(1 - t)^\gamma \big[m + d_+ - d_{\text{mis}} \big]_+ \Big] , \qquad \mathcal{L}_{\text{drop}} = \mathbb{E}\Big[(1 - t)^\gamma \big[m + d_+ - d_{\text{drop}} \big]_+ \Big] . \tag{17}$$

Mismatch tests branch-value specificity, whereas drop tests branch necessity. The final diffusion objective is

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{mis}} \mathcal{L}_{\text{mis}} + \lambda_{\text{drop}} \mathcal{L}_{\text{drop}}. \tag{18}$$

Classifier-free guidance uses null conditions during training and the standard conditional/null combination at sampling without altering the branch-token interface.

3.6 Paired Token-Reassignment Sensitivity

At sampling, paired edits measure responses to reassigned mechanisms:

$$Z = Z_\phi(a, g, s), \qquad Z' = Z_\phi(a', g, s), \qquad \Delta_G(\epsilon; Z, Z') := \mathcal{G}_\theta(\epsilon; Z') - \mathcal{G}_\theta(\epsilon; Z). \tag{19}$$

Sex and stage are reassigned analogously. Fixing ϵ removes independent-noise variation, while the unedited inputs remain fixed in the paired token sets. Thus, Δ_G estimates model-level sensitivity, not an individual counterfactual or causal effect; the latter requires further structural assumptions or an identifiable interventional estimand (Sanchez & Tsaftaris, 2022; Rahman et al., 2024).

4 Experiments

4.1 Dataset and Protocol

We construct a subject-disjoint, demographically controlled cohort from Sleep-EDF Expanded (Kemp et al., 2018) to reduce shortcuts arising from unequal subgroup frequencies. The training/test splits contain 21/9 subjects, with equal subject allocation across three predefined age strata (25–39, 40–60, and 61–90): each stratum contains 7/3 training/test subjects. Female/male counts are 3/4, 4/3, and 3/4 in training and 1/2, 1/2, and 2/1 in testing, providing near-balanced recorded-sex allocation within each age stratum. Each subject contributes two nights and contains all five stages (Wake, N1, N2, N3, and REM). The resulting splits contain 42/18 nights and 35,503/6,805 epochs. After resampling and normalization, each epoch is $x \in \mathbb{R}^{2\times 3000}$ at 100 Hz. The same fixed split is used for all methods, and test subjects are excluded from model and hyperparameter selection.

4.2 Implementation Details

Prior-field distillation uses 750 epochs, one 90-condition batch per epoch, AdamW (learning rate 3×10^{-4}, weight decay 10^{-5}), and gradient clipping at 1.0; PhysioKAN is then frozen. The denoiser uses 350 epochs of AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$, weight decay 0.02), a cosine schedule (five-epoch warm-up; base, peak, and minimum learning rates 5×10^{-4}, 10^{-3}, and 10^{-5}), per-device batch size 128, four-step accumulation (effective batch size 512), and clipping at 1.0. Classifier-free dropout is 0.1. Mismatch and branch-drop weights increase linearly from zero at epoch 20 to 0.01 and 0.005 at epoch 80, with margin 0.02 and $(1-t)^\gamma$, $\gamma = 1$. EMA decays are 0.999 and 0.9996.

Sampling uses the 0.9996 EMA, 50-step Heun, and CFG scale 1.0. All evaluations use length-3000 sequences; for the proposed model and internal variants, each of five seeds generates 500 samples.

4.3 Evaluation Protocol

Marginal quality uses FID, feature-manifold F1, LogPSD error, and band errors (Table 2). Precision P, recall R, and F1 use the frozen sleep-stage classifier's 256-dimensional penultimate features; $F1 = 2PR/(P+R)$. Appendix Table 8 provides complementary, classifier-independent waveform-marker evidence. Matched-noise age reassignment tests functional age–stage separation (Tables 3 and 4), adding evidence beyond marginal metrics. Hard leakage is the fraction classified as N2, soft leakage is the mean N2 probability, and Y→O Drop is the decrease in REM recall after age reassignment.

4.4 Signal Quality

Comparison with external baselines. Under the same conditions and length-3000 protocol, Table 2 compares DiffuSETS, ImagenTime, and EEGDiffuser (Lai et al., 2025; Naiman et al., 2024; Wang et al., 2026). Phy-DiffKAN achieves the strongest overall FID/F1/LogPSD profile and leads most bands; EEGDiffuser leads only delta. This establishes competitive marginal quality, not mechanism attribution.

Ablation study. Phy_MLP retains the same six-branch diffusion architecture, physiological targets, branch objectives, and a similar parameter count as Ours, replacing only KAN with MLP; its degraded FID/F1 and spectral errors highlight the KAN advantage under this controlled comparison.

Table 2: Signal quality and spectral comparison with external generators and internal ablations.

Method	Signal		Spectral						
	FID ↓	F1 ↑	LogPSD ↓	Delta ↓	Theta ↓	Alpha ↓	Sigma ↓	Beta ↓	Gamma ↓
External generators									
DiffuSETS	1.93±0.44	0.61±0.02	1.71±0.65	16.87±6.80	2.83±1.12	0.70±0.12	0.33±0.03	0.13±0.00	0.01±0.00
ImagenTime	0.87±0.15	0.67±0.01	0.15±0.02	0.36±0.15	1.02±0.06	0.23±0.04	0.10±0.02	0.06±0.01	0.01±0.00
EEGDiffuser	0.64±0.05	0.73±0.02	0.11±0.01	**0.07±0.07**	0.83±0.08	0.21±0.04	0.02±0.01	0.02±0.02	0.01±0.00
Internal ablations									
Recon Only	0.72±0.07	0.69±0.03	0.81±0.05	0.40±0.19	7.94±0.30	0.62±0.07	0.10±0.04	0.02±0.01	**0.00±0.00**
w/o Drop Loss	0.88±0.06	0.74±0.01	0.34±0.04	0.63±0.20	2.34±0.18	0.65±0.06	0.07±0.04	0.03±0.02	0.01±0.01
w/o Mismatch Loss	1.05±0.04	0.76±0.01	0.35±0.03	0.63±0.24	3.09±0.17	0.25±0.03	0.02±0.01	0.01±0.01	0.01±0.01
Random KAN	1.65±0.26	0.71±0.05	0.48±0.08	**0.07±0.01**	1.83±0.26	0.84±0.11	0.52±0.10	0.47±0.12	0.11±0.04
MLP Cond.	1.02±0.10	0.75±0.00	0.50±0.04	1.79±0.37	2.98±0.17	0.51±0.06	0.11±0.03	0.09±0.03	0.03±0.01
Phy_MLP	1.72±0.13	0.62±0.00	1.61±0.34	0.19±0.03	9.83±0.31	2.75±0.16	1.31±0.09	0.94±0.05	0.17±0.06
Ours	**0.53±0.03**	**0.78±0.01**	**0.06±0.01**	0.33±0.07	**0.06±0.03**	**0.08±0.02**	**0.01±0.01**	**0.01±0.00**	**0.00±0.00**

Recon Only shows that pointwise reconstruction does not make condition branches indispensable. Removing drop or mismatch degrades the joint profile, supporting branch necessity and value specificity. Random KAN matches delta but has the worst FID and larger errors elsewhere. Thus, balanced fidelity requires the factorized field and both branch-sensitive objectives.

4.5 Age–Stage Conditional Stress Test

Age and sleep stage are correlated in the observational corpus, so a label-conditioned generator may use age as a shortcut for stage morphology. We therefore fix the requested stage as REM or N2, change only the age condition, and reuse the initial diffusion noise. A frozen EEGNet measures whether the generated waveform remains recognizable as the requested stage. On held-out real EEG, this readout reaches 0.809 accuracy, 0.759 macro-F1, 0.904 REM recall, and 0.767 N2 recall.

Table 3: Young-to-old age–stage stress test.

Method	5-stage Bal. Acc. ↑	Young REM ↑	Y→O Drop ↓	Old REM ↑	Old N2 ↑	R→N2 H ↓	R→N2 S ↓
External generators							
DiffuSETS	0.69±0.01	0.90±0.02	0.56±0.03	0.34±0.01	0.66±0.02	0.60±0.03	0.59±0.01
ImagenTime	0.68±0.01	0.92±0.03	0.72±0.03	0.20±0.03	0.60±0.01	0.68±0.02	0.68±0.03
EEGDiffuser	0.75±0.01	0.91±0.01	0.72±0.03	0.19±0.02	**0.88±0.00**	0.79±0.01	0.79±0.02
Internal ablations							
Recon Only	0.68±0.01	0.92±0.03	0.50±0.02	0.42±0.01	0.69±0.01	0.42±0.02	0.34±0.03
w/o Drop Loss	0.69±0.01	0.94±0.00	0.52±0.03	0.42±0.03	0.63±0.02	0.25±0.01	0.18±0.03
w/o Mismatch Loss	0.72±0.01	0.94±0.01	0.53±0.01	0.42±0.00	0.71±0.03	0.39±0.02	0.26±0.01
Random KAN	0.75±0.01	0.91±0.01	0.50±0.00	0.41±0.02	0.63±0.01	0.38±0.03	0.35±0.02
MLP Cond.	0.78±0.01	0.93±0.02	0.82±0.03	0.11±0.01	**0.85±0.02**	0.88±0.00	0.83±0.03
Random $A \times S$	0.79±0.01	0.89±0.02	0.48±0.03	0.41±0.03	0.69±0.02	0.36±0.01	0.34±0.00
Phy_MLP	0.75±0.01	0.87±0.03	0.25±0.03	0.62±0.04	0.76±0.02	0.22±0.01	0.21±0.01
Ours	**0.87±0.01**	**0.95±0.01**	**0.20±0.03**	**0.75±0.03**	0.84±0.01	**0.18±0.02**	**0.17±0.01**

Tables 3 and 4 form a two-step test: Table 3 checks preservation of stage identity after age editing, whereas Table 4 checks that age remains decodable. The frozen age probe is trained on subject-disjoint real REM epochs from young (≤ 35) and old (> 60) subjects. Passing both supports model-level age–stage separation within the tested assignments. Appendix Table 9 extends the stress test to N1 across young and old requests, showing the stage-preservation pattern beyond REM.

Table 4: Stage-specific age decodability in REM EEG.

Method	Old Rec. ↑	Young Rec. ↑	Bal. Acc. ↑	Macro F1 ↑
Held-out real REM				
Real reference	0.71±0.01	0.74±0.00	0.73±0.02	0.73±0.01
External generators				
ImagenTime	0.39±0.02	0.53±0.01	0.46±0.00	0.46±0.02
EEGDiffuser	0.42±0.00	0.55±0.02	0.49±0.01	0.49±0.01
DiffuSETS	0.41±0.01	0.54±0.02	0.47±0.02	0.47±0.00
Internal variants				
MLP Cond.	0.24±0.02	0.48±0.01	0.36±0.00	0.35±0.01
Random $A \times S$	0.62±0.01	0.62±0.00	0.62±0.02	0.62±0.01
w/o Drop Loss	0.60±0.00	0.70±0.02	0.65±0.01	0.65±0.02
Random KAN	0.60±0.02	0.66±0.01	0.63±0.00	0.63±0.01
Recon Only	0.59±0.01	0.67±0.02	0.63±0.01	0.63±0.00
w/o Mismatch Loss	0.66±0.02	0.71±0.00	0.69±0.01	0.69±0.02
Phy_Cond.	0.64±0.03	0.70±0.02	0.67±0.01	0.67±0.00
Ours	**0.69±0.01**	**0.74±0.02**	**0.72±0.00**	**0.72±0.01**

Comparison with external baselines. Table 3 shows that, unlike external generators, Phy-DiffKAN preserves fixed REM instead of shifting toward N2 after age editing; Table 4 shows that its near-real age decodability is retained. The retained age information distinguishes this result from age-agnostic stage preservation.

Ablation study. MLP Cond. evaluates an alternative shared-conditioning design: competitive marginal and stage scores accompany poor REM preservation, high leakage, and weak age decodability, showing that a shared embedding can mix age and stage. Phy_MLP provides the controlled encoder comparison, retaining the six branches, physiological targets, branch objectives, and a similar parameter count while replacing KAN with MLP. It trails Ours in five-stage balanced accuracy (0.75 vs. 0.87) and old-REM recall (0.62 vs. 0.75), with higher REM-to-N2 leakage (0.22/0.21 vs. 0.18/0.17), supporting the advantage of KAN.

5 Conclusion

Phy-DiffKAN addresses age–stage inconsistency in conditional sleep EEG through a factorized prior field, its PhysioKAN approximation, and branch-token diffusion conditioning. By separating conditions and pairwise interactions, the design makes their modulation inspectable, while the field remains a modeling assumption. Signal metrics, controlled ablations, REM/N2 tests, and the supplementary N1 sweep indicate preserved realism and reduced shortcut-driven hybrid generation. Fixed-noise token reassignments isolate targeted responses from sampling variation. Together, the evidence supports model-level age sensitivity with stage preservation within the tested cohort and condition range.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT) for language polishing, LaTeX editing, formatting equations and tables, and checking presentation issues. We have not used generative AI tools for research ideation, methodological design, data generation, experiment execution, result analysis, or scientific conclusions, and the remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

The main paper specifies the factorized prior field, six branch interface, branch-sensitive objectives, fixed-noise reassignment protocol, and evaluation metrics. Appendix Sections A.2–A.10 provide the complete published-rule mapping, coordinate anchors, branch weights and masks, training and sampling settings, subject-disjoint Sleep-EDF split, preprocessing, metric definitions, and supplementary stress-test and waveform diagnostics (Tables 5–11). These details, together with the fixed seeds and sample counts reported in the main paper, identify the data processing, model configuration, and evaluation steps needed to reproduce the reported comparisons.

References

Antoine R. Adamantidis, Carolina Gutierrez Herrera, and Thomas C. Gent. Oscillating circuitries in the sleeping brain. *Nature Reviews Neuroscience*, 20(12):746–762, 2019. doi: 10.1038/s41583-019-0223-4. URL https://doi.org/10.1038/s41583-019-0223-4.

American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, Darien, IL, version 3 edition, 2023. URL https://aasm.org/clinical-resources/scoring-manual/.

Bruno Aristimunha, Raphael Yokoingawa de Camargo, Sylvain Chevallier, Oeslle Lucena, Adam Thomas, M. Jorge Cardoso, Walter Lopez Pinaya, and Jessica Dafflon. Synthetic sleep EEG signal generation using latent diffusion models. In *Deep Generative Models for Health Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=mDwURmlapW.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. doi: 10.48550/arXiv.1907.02893. URL https://arxiv.org/abs/1907.02893.

Julie Carrier, Sarah Land, Daniel J. Buysse, David J. Kupfer, and Timothy H. Monk. The effects of age and gender on sleep EEG power spectral density in the middle years of life (ages 20–60 years old). *Psychophysiology*, 38(2):232–242, 2001. doi: 10.1111/1469-8986.3820232. URL https://doi.org/10.1111/1469-8986.3820232.

Patrick Chao, Patrick Blöbaum, Sapan Patel, and Shiva Prasad Kasiviswanathan. Modeling causal mechanisms with diffusion models for interventional and counterfactual queries. *arXiv preprint arXiv:2302.00860*, 2023. doi: 10.48550/arXiv.2302.00860. URL https://arxiv.org/abs/2302.00860.

Jintai Chen, Kuanlun Liao, Kun Wei, Haochao Ying, Danny Z. Chen, and Jian Wu. ME-GAN: Learning panoptic electrocardio representations for multi-view ECG synthesis conditioned on heart diseases. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3360–3370. PMLR, 2022. URL https://proceedings.mlr.press/v162/chen22n.html.

Laura M. J. Fernandez and Anita Lüthi. Sleep spindles: Mechanisms and functions. *Physiological Reviews*, 100(2):805–868, 2020. doi: 10.1152/physrev.00042.2018. URL https://doi.org/10.1152/physrev.00042.2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Bob Kemp, A. H. Zwinderman, Bert Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé. Sleep-EDF Database Expanded. PhysioNet, 2018. URL https://physionet.org/content/sleep-edfx/1.0.0/. Version 1.0.0.

Guido Klein, Pierre Guetschel, Gianluigi Silvestri, and Michael Tangermann. Synthesizing EEG signals from event-related potential paradigms with conditional diffusion models. In *Proceedings of the 9th Graz Brain-Computer Interface Conference*, pp. 438–443, 2024. doi: 10.3217/978-3-99161-014-4-077. URL https://arxiv.org/abs/2403.18486.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5a1a10c2c2c9b9af1514687bc24b8f3d-Abstract-Conference.html.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Yongfan Lai, Jiabo Chen, Qinghao Zhao, Deyun Zhang, Yue Wang, Shijia Geng, Hongyan Li, and Shenda Hong. DiffuSETS: 12-lead ECG generation conditioned on clinical text reports and patient-specific information. *Patterns*, 6(10):101291, 2025. doi: 10.1016/j.patter.2025.101291. URL https://doi.org/10.1016/j.patter.2025.101291.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. doi: 10.48550/arXiv.2511.13720. URL https://arxiv.org/abs/2511.13720.

Yang Li, Han Meng, Zhenyu Bi, Ingolv T. Urnes, and Haipeng Chen. Population aware diffusion for time series generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (17):18520–18529, 2025a. doi: 10.1609/aaai.v39i17.34038. URL https://ojs.aaai.org/index.php/AAAI/article/view/34038.

Yang Li, Han Meng, Zhenyu Bi, Ingolv T. Urnes, and Haipeng Chen. Population aware diffusion for time series generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (17):18520–18529, 2025b. doi: 10.1609/aaai.v39i17.34038. URL https://ojs.aaai.org/index.php/AAAI/article/view/34038.

Ziming Liu, Pingchuan Ma, Yixuan Wang, Wojciech Matusik, and Max Tegmark. KAN 2.0: Kolmogorov–arnold networks meet science. *arXiv preprint arXiv:2408.10205*, 2024a. doi: 10.48550/arXiv.2408.10205. URL https://arxiv.org/abs/2408.10205.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. *arXiv preprint arXiv:2404.19756*, 2024b. doi: 10.48550/arXiv.2404.19756. URL https://arxiv.org/abs/2404.19756.

Y. Lou, Y. Zhang, Z. Shan, F. Gao, and Z. Qian. An EEG-based framework for sleep quality assessment and modulation with conditional convolutional diffusion modeling. *IEEE Journal of Biomedical and Health Informatics*, 2026. doi: 10.1109/JBHI.2026.3710928. URL https://doi.org/10.1109/JBHI.2026.3710928. Early Access.

Andjela Markovic, Michael Kaess, and Leila Tarokh. Gender differences in adolescent sleep neurophysiology: A high-density sleep EEG study. *Scientific Reports*, 10(1):15935, 2020. doi: 10.1038/s41598-020-72802-0. URL https://doi.org/10.1038/s41598-020-72802-0.

Ilan Naiman, Nimrod Berman, Itai Pemper, Idan Arbiv, Gal Fadlon, and Omri Azencot. Utilizing image transforms and diffusion models for generative modeling of short and long time series. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3868. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/dc6748383752138af7f00b3185a0a404-Abstract-Conference.html.

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. doi: 10.48550/arXiv.2212.09748. URL https://arxiv.org/abs/2212.09748.

Shaun M. Purcell, Dara S. Manoach, Charmaine Demanuele, Brian E. Cade, Sara Mariani, Robert Cox, Georgia Panagiotaropoulou, Richa Saxena, John Q. Pan, Jordan W. Smoller, Susan Redline, and Robert Stickgold. Characterizing sleep spindles in 11,630 individuals from the national sleep research resource. *Nature Communications*, 8:15930, 2017. doi: 10.1038/ncomms15930. URL https://doi.org/10.1038/ncomms15930.

Md Musfiqur Rahman, Matt Jordan, and Murat Kocaoglu. Conditional generative models are sufficient to sample from any causal effect estimand. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8d6233829f97fe8eb3fa7d2ab774aedc-Abstract-Conference.html.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868. PMLR, 2021. URL https://proceedings.mlr.press/v139/rasul21a.html.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 647–668. PMLR, 2022. URL https://proceedings.mlr.press/v177/sanchez22a.html.

Abdulghani Sankari and Jennifer M. Slowik. Sleep study. In *StatPearls*. StatPearls Publishing, Treasure Island, FL, 2026. URL https://www.ncbi.nlm.nih.gov/books/NBK563147/. Updated January 10, 2026.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://doi.org/10.1109/JPROC.2021.3058954.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31016–31029. PMLR, 2023. URL https://proceedings.mlr.press/v202/shen23d.html.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022. URL https://jmlr.org/papers/v23/21-0080.html.

Kate E. Sprecher, Brady A. Riedner, Richard F. Smith, Giulio Tononi, Richard J. Davidson, and Ruth M. Benca. High resolution topography of age-related changes in non-rapid eye movement sleep electroencephalography. *PLOS ONE*, 11(2):e0149770, 2016. doi: 10.1371/journal.pone.0149770. URL https://doi.org/10.1371/journal.pone.0149770.

Haoqi Sun, Elissa Ye, Luis Paixao, Wolfgang Ganglberger, Catherine J. Chu, Can Zhang, Jonathan Rosand, Emmanuel Mignot, Sydney S. Cash, David Gozal, Robert J. Thomas, and M. Brandon Westover. The sleep and wake electroencephalogram over the lifespan. *Neurobiology of Aging*, 124:60–70, 2023. doi: 10.1016/j.neurobiolaging.2023.01.006. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9957961/.

Szabolcs Torma and Luca Szegletes. Generative modeling and augmentation of EEG signals using improved diffusion probabilistic models. *Journal of Neural Engineering*, 22(1):016001, 2025. doi: 10.1088/1741-2552/ada0e4. URL https://doi.org/10.1088/1741-2552/ada0e4.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. EEGDiffuser: Label-guided EEG signals synthesis via diffusion model for BCI applications. *Neurocomputing*, 670:132636, 2026. doi: 10.1016/j.neucom.2026.132636. URL https://doi.org/10.1016/j.neucom.2026.132636.

Sajila D. Wickramaratne and Ankit Parekh. SleepSIM: Conditional GAN-based non-REM sleep EEG signal generator. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1–4. IEEE, 2023. doi: 10.1109/EMBC40787.2023.10341043. URL https://doi.org/10.1109/EMBC40787.2023.10341043.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yang_CausalVAE_Disentangled_Representation_Learning_via_Neural_Structural_Causal_Models_CVPR_2021_paper.html.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2403.01742. URL https://arxiv.org/abs/2403.01742.

A Supplementary Material

A.1 Scope and Causal Boundary

This supplement documents the rulebook, PhysioKAN distillation, visualization diagnostics, training and sampling parameters, metric definitions, and dataset split. The contribution is operational and model-level: paired edits evaluate age–stage consistency by reassigning physiological mechanism tokens under fixed diffusion noise. The empirical scope centers on the learned generator.

A.2 Rulebook Creation and Audit Protocol

The machine-readable rulebook encodes published sleep-physiology findings as soft priors for the Physiological KAN Encoder (Sankari & Slowik, 2026; Adamantidis et al., 2019; Sprecher et al., 2016; Carrier et al., 2001; Purcell et al., 2017; Markovic et al., 2020; Fernandez & Lüthi, 2020). Every stage, age, sex, and pairwise rule listed below is traced to this published evidence; the dimensionless anchor magnitudes encode its reported direction on a shared modeling scale. Local EEG supports validation and calibration, and the authors verified each retained direction, stage scope, coordinate, and source against the original publication.

Sleep stages follow Wake, N1, N2, N3, and REM with labels 0–4. Age is normalized as

$$\mathrm{age01} = \mathrm{clamp}((\mathrm{age} - 18)/(90 - 18), 0, 1).$$

The dataset records binary biological sex, encoded as 0 for male and 1 for female. The legacy word "gender" may remain inside previously generated figure artwork or filenames; there it denotes this same binary sex variable, not self-reported gender identity. A learned null condition used for classifier-free guidance is a model token and is not a third sex category.

A.3 Physiological Coordinates

Table 5: Physiological coordinate definitions used by all PhysioKAN branches.

Id	Coordinate	Physiological reference	Rulebook unit
0	delta power	0.5–4 Hz slow-wave band tendency	dimensionless soft target
1	theta power	4–8 Hz N1/REM-like activity	dimensionless soft target
2	alpha power	8–13 Hz wake-related rhythm	dimensionless soft target
3	sigma power	11–16 Hz spindle-related band	dimensionless soft target
4	beta power	13–30 Hz cortical activation	dimensionless soft target
5	slow-wave activity	0.5–2 Hz slow-oscillation tendency	dimensionless soft target
6	spindle density	N2 spindle occurrence tendency	dimensionless soft target
7	spindle amplitude	spindle-strength tendency	dimensionless soft target
8	spindle frequency	spindle-peak tendency	dimensionless soft target
9	K-complex tendency	N2 morphology	dimensionless soft target
10	low-amplitude mixed frequency	REM/Wake-like activation	dimensionless soft target
11	spectral-slope activation	high-frequency activation	dimensionless soft target

Table 6: Rulebook stage prototypes. Abbreviations follow the coordinate order in Table 5.

Stage	d	th	a	sig	b	swa	spd	spa	spf	kc	lamf	slope
Wake	0.0	0.2	0.8	-0.5	0.6	-0.8	-0.8	-0.8	0.0	-0.8	0.7	0.5
N1	0.2	0.8	-0.4	-0.3	-0.2	-0.2	-0.5	-0.4	0.0	-0.5	0.5	0.1
N2	0.3	0.2	-0.3	1.0	-0.3	0.2	1.0	0.8	0.2	1.0	-0.2	-0.2
N3	1.0	0.5	-0.6	-0.2	-0.6	1.0	-0.2	-0.1	-0.1	0.2	-0.8	-0.8
REM	0.0	0.7	0.1	-0.7	0.2	-0.8	-0.9	-0.9	0.0	-0.9	0.8	0.4

The targets in Table 5 are directional priors on a shared modeling scale, distinct from calibrated power, microvolt amplitude, event rate, or clinical effect size. Delta power and slow-wave activity overlap statistically, as do sigma power and spindle measures, but their roles are not duplicated: band-integrated tendencies and event/morphology tendencies occupy separate coordinates. A finding is assigned once to the construct it measures, repeated citations do not create additional loss terms, and the rule loss averages each of the 12 coordinates once. Thus, related coordinates may be dependent without receiving implicit duplicate weighting.

A.4 Rulebook Targets

The implemented branch set is

$$\mathcal{B} = \{A, G, S, A\times G, A\times S, G\times S\},$$

where (G) denotes the binary sex variable. The six rule-loss branch weights for stage, age, sex, age–stage, sex–stage, and age–sex are 1.00, 0.85, 0.50, 1.00, 0.80, and 0.40, respectively. The evaluated model instantiates exactly these six branches; a residual note in the YAML records a future variant. Table 6 and the coefficient lists below provide the complete numerical rule-to-anchor mapping used for every branch.

Stage prototypes. The following vectors use the coordinate order in Table 5.

Age effects. Coefficients multiplying $\mathrm{age01}$ are: delta -0.80, theta -0.40, alpha -0.10, sigma -0.50, beta +0.30, slow-wave activity -1.00, spindle density -0.50, spindle amplitude -0.40, spindle frequency +0.20, K-complex tendency -0.20, low-amplitude mixed-frequency +0.20, and spectral-slope activation +0.40. Key age–stage offsets are N3 delta -0.60 and slow-wave activity -0.80; N2 sigma -0.50, spindle density -0.60, spindle amplitude -0.50, spindle frequency +0.20, and K-complex tendency -0.20; N1 theta -0.10; REM low-amplitude mixed-frequency +0.20 and theta -0.10; and Wake alpha -0.05.

Sex and pairwise rules. Female-coded base effects are deliberately weak priors: delta +0.15, theta +0.10, alpha +0.10, sigma +0.20, beta +0.10, slow-wave activity +0.10, spindle density +0.25, spindle amplitude +0.15, and spindle frequency +0.10. Sex–stage interactions include N2 sigma +0.20, spindle density +0.25, spindle amplitude +0.15, spindle frequency +0.10; N3 delta +0.10 and slow-wave activity +0.10; and REM beta +0.05. The age–sex branch is near zero by default, with sigma +0.05 and spindle density +0.05 multiplied by $\mathrm{female01}(1 - \mathrm{age01})$.

A.5 PhysioKAN Distillation and Auditability

Encoder implementation. PhysioKAN converts an assigned condition triple $c = (a, g, s)$ into a structured dictionary of rule-aligned mechanism vectors rather than a single opaque label embedding. Age is linearly encoded after division by 100, whereas binary sex and sleep stage use embedding tables. The latter tables additionally contain a classifier-free null id (sex id 2 and stage id K):

$$\begin{aligned} f_a &= W_a(a/100) + b_a, \\ f_g &= \mathrm{Emb}_g(g), \qquad f_s = \mathrm{Emb}_s(s), \end{aligned} \tag{20}$$

where $f_a, f_g, f_s \in \mathbb{R}^{64}$. The three main-effect inputs are $\eta_a = f_a$, $\eta_g = f_g$, and $\eta_s = f_s$; the pairwise inputs use element-wise composition,

$$\eta_{ag} = f_a \odot f_g, \qquad \eta_{as} = f_a \odot f_s, \qquad \eta_{gs} = f_g \odot f_s. \tag{21}$$

Six independent KAN functions then produce

$$\begin{aligned} Z_\phi(a, g, s) &= \{z_a, z_g, z_s, z_{ag}, z_{as}, z_{gs}\}, \\ z_b &= h_b(\eta_b) \in \mathbb{R}^{12}. \end{aligned} \tag{22}$$

and their sum $z = \sum_{b \in \mathcal{B}} z_b$ is the total rule-aligned physiological profile. This factorization separates demographic and stage main effects from their interactions, while retaining a named 12-coordinate output for every branch.

Each h_b is a KAN-style Gaussian basis layer with a linear residual path. For input coordinate u_i, output coordinate o, and basis k,

$$\begin{aligned} \phi_k(u_i) &= \exp\left[-\frac{(u_i - c_k)^2}{2\sigma^2}\right], \\ [h_b(u)]_o &= (W_b u + b_b)_o + \sum_{i,k} \theta^{(b)}_{iko} \phi_k(u_i). \end{aligned} \tag{23}$$

The implementation uses eight fixed centers $c_k = -2 + 4(k-1)/7$ on $[-2, 2]$ and fixed width $\sigma = 4/7$. Only the base linear map and Gaussian correction weights are learned. Thus, the layer is a Gaussian-basis KAN variant rather than the original B-spline parameterization; its coordinate-wise response curves nevertheless remain directly inspectable.

Rule-guided distillation. The rule-distillation grid uses ages $\{18, 25, 35, 45, 55, 65, 75, 85, 90\}$, both sex codes, and all five stages, giving 90 conditions. Each branch emits a 12-dimensional vector. For branch targets τ_b, the weighted rule loss is

$$\mathcal{L}_{\text{rule}} = \frac{\sum_{b \in \mathcal{B}} \omega_b \, \text{SmoothL1}(z_b, \tau_b)}{\sum_{b \in \mathcal{B}} \omega_b}, \tag{24}$$

with branch weights

$$\begin{aligned} (\omega_a, \omega_g, \omega_s) &= (0.85, 0.50, 1.00), \\ (\omega_{ag}, \omega_{as}, \omega_{gs}) &= (0.40, 1.00, 0.80). \end{aligned}$$

Stage-ranking constraints impose a margin of 0.2 on total-profile landmarks (N3 delta/slow-wave activity and N2 sigma/spindle density), while age-monotonic constraints compare ages 25 and 65 with margin 0.05 for the same landmarks. Branch redundancy is discouraged by

$$\mathcal{L}_{\text{ortho}} = \frac{2}{|\mathcal{B}|(|\mathcal{B}| - 1)} \sum_{b_i < b_j} \mathbb{E} \left|\cos(z_{b_i}, z_{b_j})\right|. \tag{25}$$

The complete objective is

$$\begin{aligned} \mathcal{L} = 3.0\mathcal{L}_{\text{rule}} &+ 0.5\mathcal{L}_{\text{stage_rank}} \\ &+ 0.5\mathcal{L}_{\text{age_mono}} + 10^{-3}\mathcal{L}_{\text{ortho}}. \end{aligned}$$

Distillation uses seed 42, 750 epochs, one full 90-condition batch per epoch, AdamW with learning rate 3×10^{-4} and weight decay 10^{-5}, and gradient clipping at 1.0. The distilled encoder is frozen for diffusion training. The 25–90-year empirical EEG cohort lies fully within the 18–90 rule-distillation support, with age 90 included explicitly as the upper grid endpoint. Intermediate observed ages use the continuous normalized age covariate, and no 85-year boundary clamping is applied.

KAN is valuable here not only as a nonlinear conditioner. Its univariate edge functions can be plotted as coordinate-level input–response curves, allowing their direction, nonlinearity, and saturation to be checked against the source rulebook. This curve-level traceability provides an auditability advantage over an MLP embedding whose hidden coordinates have no direct physiological label. At inference, an edit such as $a \to a'$ recomputes the complete dictionary $Z_\phi(a', g, s)$: the age and age-interaction branches change, whereas the sex, stage, and sex–stage main inputs remain tied to their assigned values. The resulting operation is a structured model-level reassignment, not an estimate of a population-level causal effect.

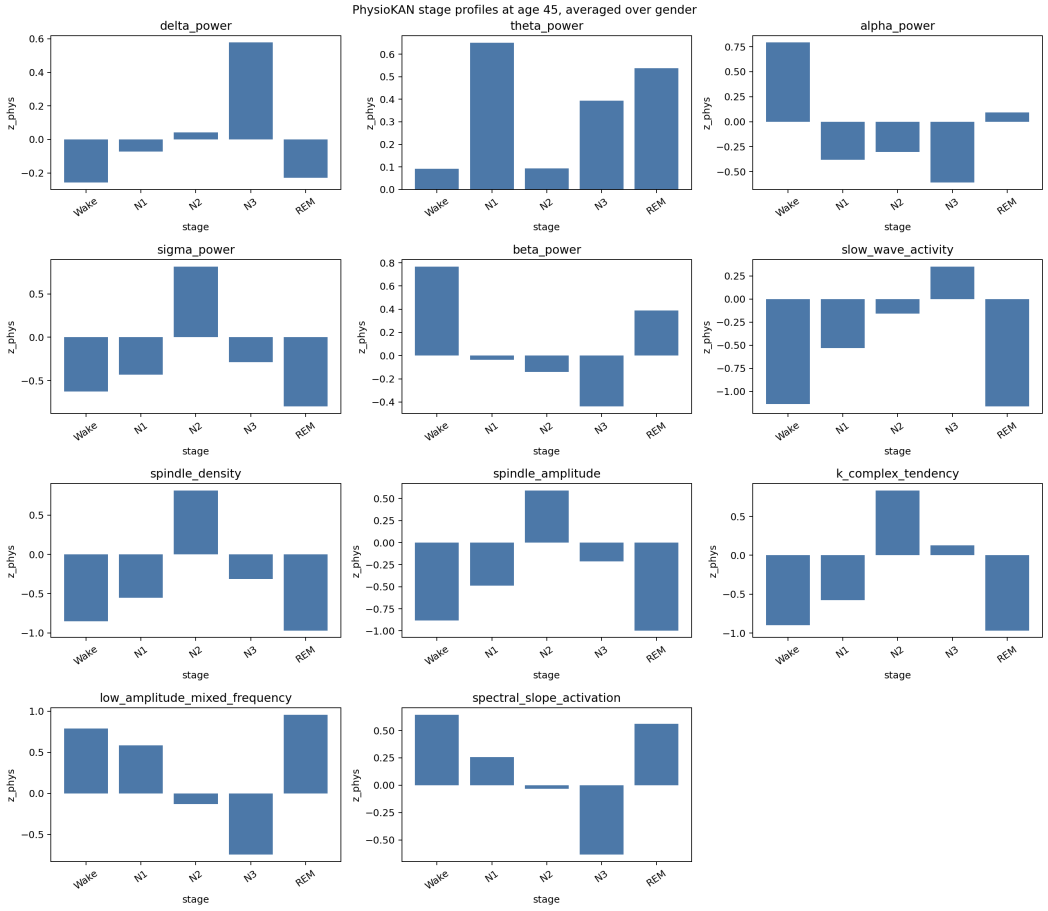

Figure 3: PhysioKAN stage profiles. N3 emphasizes delta and slow-wave activity; N2 emphasizes sigma, spindle, and K-complex coordinates; REM emphasizes low-amplitude mixed-frequency activity with weak spindle coordinates.

Table 7: Rule-distillation diagnostics. The single branch-level violation is an N2 age–stage sigma diagnostic; all combined-output behavior checks pass.

Diagnostic group	Passed/total	Pass rate
Stage ranking	37/37	100%
Age base branch	12/12	100%
Age–stage branch	12/13	92.3%
Sex base branch	9/9	100%
Sex–stage branch	7/7	100%
Age–sex branch	2/2	100%
Overall branch rules	79/80	98.75%
Final behavior	49/49	100%

A.6 Diffusion Training and Sampling

The full model and standard internal variants use seed 42 and 350 training epochs. The additional Random Age diagnostic uses its final epoch-500 checkpoint. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.02, and a cosine schedule with five warm-up epochs, base learning rate 5×10^{-4}, peak 10^{-3}, and minimum 10^{-5}. The per-device batch size is 128 with four-step gradient accumulation, giving an effective batch size of 512 on one process; gradients are clipped at 1.0. Classifier-free

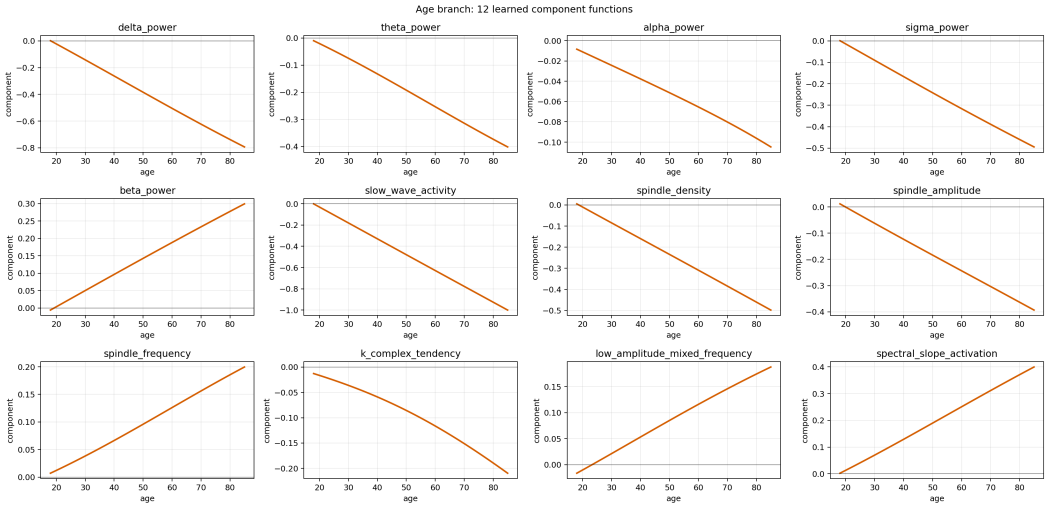

Figure 4: Age-branch component functions over the 12 physiological coordinates. Their directly viewable shapes make rule direction and saturation auditable.

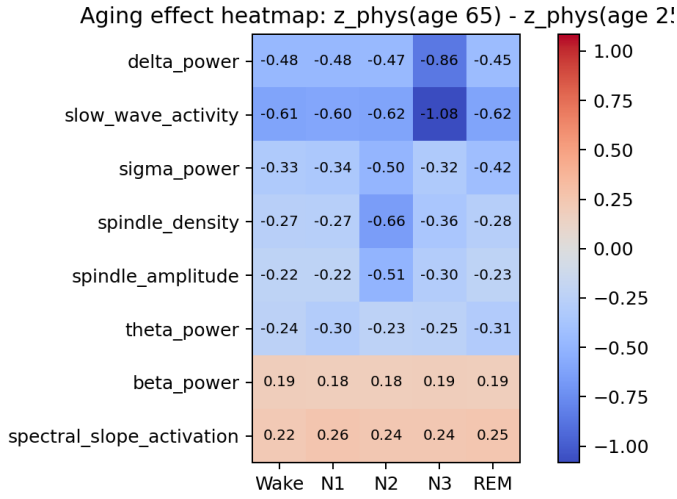

Figure 5: Stage-specific aging effects in the learned mechanism interface.

label dropout is 0.1. Mismatch and branch-drop weights increase linearly from zero at epoch 20 to 0.01 and 0.005 at epoch 80, with branch margin 0.02 and timestep exponent 1.0. EMA decays are 0.999 and 0.9996.

Reported samples use the final parameters tracked by the 0.9996 EMA, Heun sampling with 50 steps, CFG scale 1.0, and two-channel sequences of length 3000. Signal metrics use five sampling seeds and a 500-sample evaluation subset per seed; cell-level stress metrics use all available samples in each predefined condition cell. All retained external comparisons are conditional generators supplied with the same age, sex, and sleep-stage specification and evaluated at length 3000.

For a selected branch, mismatch shuffles its token across the batch and drop zeros the token. Their hinge penalties compare reconstruction difficulty with the matched condition. Drop encourages branch necessity; mismatch encourages the branch value to be condition-specific.

Internal variant definitions. Recon Only retains six-branch conditioning with reconstruction alone; w/o Drop Loss and w/o Mismatch Loss remove one branch-sensitive objective at a time.

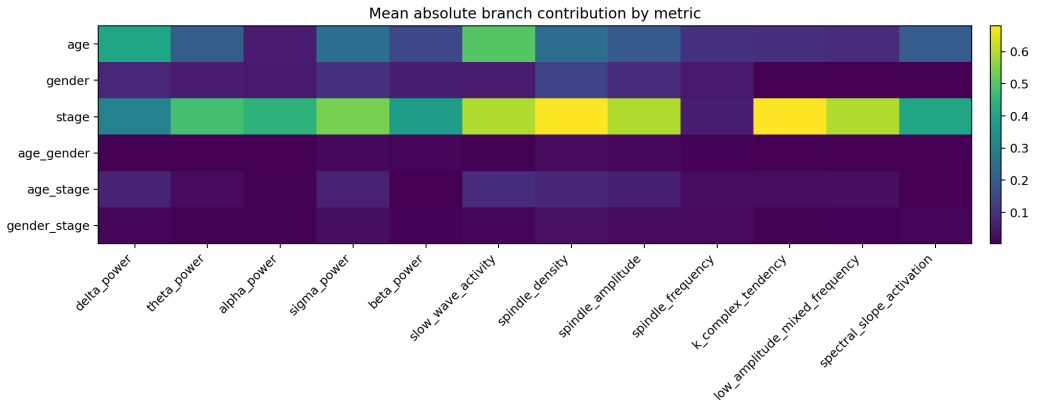

Figure 6: Coordinate-wise contributions of the six PhysioKAN branches.

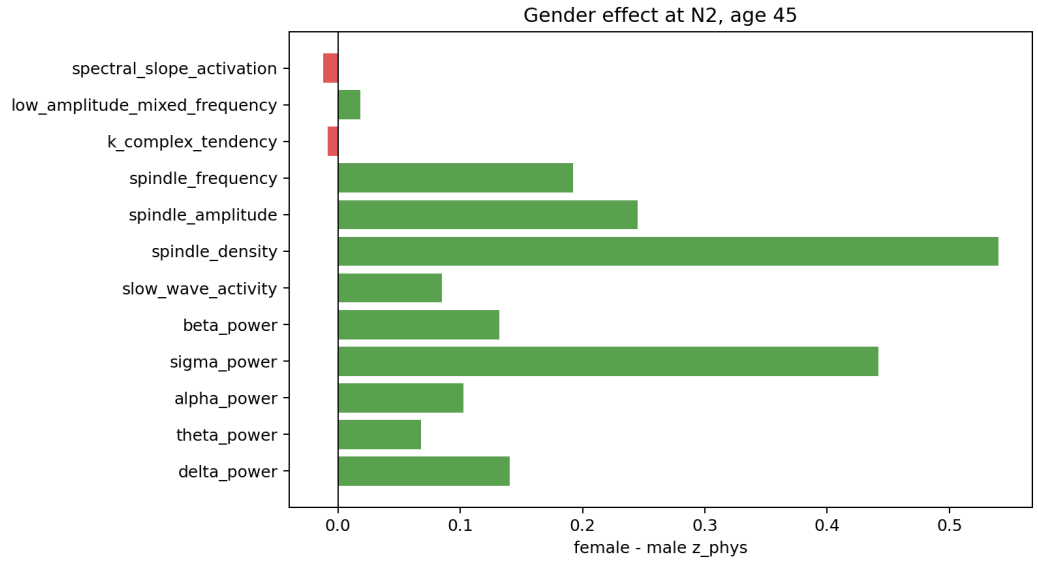

Figure 7: Example sex-related modulation for N2 at age 45. The legacy filename uses "gender," but the plotted variable is binary sex (0 male, 1 female).

Random KAN retains the KAN topology with random physiological mappings, Random $A \times S$ randomizes the age–stage branch, and Random Age permutes age-branch coordinates. MLP Cond. evaluates a shared MLP conditioning design, changing both factorization and encoder. Phy_MLP is the controlled encoder ablation: it retains the same six branches, physiological targets, branch objectives, diffusion architecture, and similar parameter count as Ours, replacing only KAN with MLP. Thus Phy_MLP isolates the KAN–MLP choice, while MLP Cond. measures an alternative conditioning design.

A.7 Metric Definitions

Signal evaluation reports classifier-feature FID, feature-manifold Precision, Recall and F1, LogPSD error, and band-level spectral errors. Let $\psi(x) \in \mathbb{R}^{256}$ denote the activation obtained after the frozen five-stage sleep classifier's convolutional stem, residual blocks, and global pooling, immediately before its final classification layer. The same frozen penultimate-layer feature extractor is applied to real and generated EEG.

Classifier-feature FID. For real features $R = \{\psi(x_i)\}_{i=1}^{N_r}$ and generated features $G = \{\psi(\widetilde{x}_j)\}_{j=1}^{N_g}$, let (μ_r, Σ_r) and (μ_g, Σ_g) be their empirical means and covariances. We compute

$$\mathrm{FID}(R, G) = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big), \tag{26}$$

with a small diagonal covariance offset for numerical stability. Lower values indicate closer first- and second-order statistics in the classifier feature space.

Feature-manifold Precision, Recall, and F1. For each feature vector, its local radius is the Euclidean distance to its k-th nearest neighbor within the same set; we use $k = 3$. Let $\mathcal{M}_r$ and $\mathcal{M}_g$ be the unions of the resulting real and generated hyperspheres. Then

$$\begin{aligned} P &= \frac{1}{N_g}\sum_{j=1}^{N_g} \mathrm{I}[\psi(\widetilde{x}_j) \in \mathcal{M}_r], \\ R &= \frac{1}{N_r}\sum_{i=1}^{N_r} \mathrm{I}[\psi(x_i) \in \mathcal{M}_g]. \end{aligned} \tag{27}$$

Precision measures how much generated support lies inside the real manifold, whereas Recall measures how much real support is covered by the generated manifold. The reported feature-space F1 is their harmonic mean,

$$F1 = \begin{cases} \dfrac{2PR}{P+R}, & P + R > 0, \\ 0, & P + R = 0. \end{cases} \tag{28}$$

LogPSD and band errors. For domain $d \in \{r, g\}$, Welch's method with a Hann window, 256-sample segments, 128-sample overlap, and sampling rate 100 Hz gives $S_{n,c}^{(d)}(f)$ for sample n and channel c. The implementation first averages PSD in the linear domain across samples and channels and then converts it to decibels:

$$\begin{aligned} \overline{S}^{(d)}(f) &= \frac{1}{N_d C}\sum_{n=1}^{N_d}\sum_{c=1}^{C} S_{n,c}^{(d)}(f), \\ L_d(f) &= 10\log_{10}\Big(\overline{S}^{(d)}(f) + \varepsilon\Big). \end{aligned} \tag{29}$$

where $\varepsilon = 10^{-12}$. For the frequency-bin set $\mathcal{F}_B$ of band B,

$$E_{\mathrm{LogPSD}}(B) = \frac{1}{|\mathcal{F}_B|}\sum_{f\in\mathcal{F}_B} \left[L_r(f) - L_g(f)\right]^2. \tag{30}$$

The overall LogPSD error uses 0.5–45 Hz. Band-level errors use delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), sigma (12–16 Hz), beta (13–30 Hz), and gamma (30–45 Hz), with shared boundary bins included.

Classifier-based conditional metrics. For class q, target-stage recall is

$$\mathrm{Rec}_q = \frac{\mathrm{TP}_q}{\mathrm{TP}_q + \mathrm{FN}_q}, \qquad \mathrm{BalAcc} = \frac{1}{K}\sum_{q=1}^{K} \mathrm{Rec}_q\,. \tag{31}$$

Global five-stage balanced accuracy sets $K = 5$ and aggregates Wake, N1, N2, N3, and REM over every evaluated age and sex condition, covering the full stage set. Cell-level stress tests use the frozen five-stage classifier to report both target-stage recall and the destination of each error. The separate age-probe experiment uses a frozen binary classifier trained only on real young and old REM epochs with subject-disjoint training and validation splits; its balanced accuracy uses $K = 2$ and is not the global five-stage statistic.

Table 8: Classifier-independent direct waveform-marker audit, complementing the frozen-classifier evidence. Stage contrasts use matched age–sex conditions; correlations are condition-level Spearman ρ with real data.

Direct marker summary	Real	Ours	DiffuSETS	ImagenTime	EEGDiffuser
Matched age–sex stage contrast					
N3–REM relative SWA	$+0.174$	$+0.229 \pm 0.011$	$+0.269 \pm 0.030$	$+0.156 \pm 0.013$	$+0.202 \pm 0.008$
N3–REM relative delta	$+0.058$	$+0.116 \pm 0.007$	$+0.083 \pm 0.018$	$+0.047 \pm 0.006$	$+0.103 \pm 0.011$
N2–REM relative sigma	$+0.007$	$+0.004 \pm 0.001$	-0.0002 ± 0.0001	-0.0008 ± 0.0005	-0.0023 ± 0.0015
Condition-level real–generated Spearman correlation					
Relative delta	–	0.664 ± 0.014	0.310 ± 0.087	0.438 ± 0.042	0.308 ± 0.052
Relative SWA	–	0.568 ± 0.026	0.327 ± 0.099	0.441 ± 0.045	0.334 ± 0.026
Relative sigma	–	0.553 ± 0.036	0.516 ± 0.055	0.358 ± 0.044	0.104 ± 0.052
RMS amplitude	–	0.770 ± 0.026	0.674 ± 0.039	0.729 ± 0.051	0.679 ± 0.031

A.8 Qualitative Signal Diagnostics

Figures 8 and 9 retain the diagnostics stored with each source folder. For DiffuSETS, ImagenTime, and EEGDiffuser. The left image in each panel is the marginal PSD and the right image is the t-SNE projection produced by the same Diff-KAN visualization routine.

The external generators likewise reproduce much of the aggregate spectral shape and occupy overlapping projected support. DiffuSETS shows the clearest low-frequency PSD offset, whereas ImagenTime and EEGDiffuser are closer in the marginal spectrum. These aggregate views are complemented by the direct waveform-marker evidence in Table 8 and the age–stage conflict analysis below, which holds N1 fixed and reports error destinations for both young and old requests.

A.9 Additional Age–Stage Conflict Tests

Together with the main REM/N2 analysis, this N1 age sweep evaluates a second requested stage across both age groups and tests the same consistency problem beyond REM.

External comparison: age changes the wrong-stage destination. With N1 fixed, ImagenTime moves in the wrong direction as age increases: N1 recall falls from 0.57 ± 0.10 to 0.52 ± 0.12, while REM leakage rises from 0.22 ± 0.06 to 0.25 ± 0.08. Its nearly zero Wake leakage therefore does not indicate good conditioning; the error is displaced toward REM. EEGDiffuser has the opposite age asymmetry: young N1 is split mainly toward Wake (0.29 ± 0.04) and REM (0.21 ± 0.09), but old N1 recall rises to 0.70 ± 0.09 and the remaining errors are led by N2. MLP conditioning also changes destination, from a nearly even young N2/REM split (0.23/0.22) to Wake (0.17) in the old group. These directional changes are hidden by aggregate PSD, t-SNE, FID, or global stage scores: the generator can match the marginal signal distribution while letting the age cue determine which neighboring stage wins.

Ablation: the proposed constraints reduce conditional shortcuts. Removing branch drop causes the strongest age dependence. At ages 25–39, requested N1 collapses to Wake in 0.59 ± 0.02 of samples and recall is only 0.25 ± 0.03; at ages 61–90, recall jumps to 0.75 ± 0.03. The 0.50 recall swing supports the role of branch drop: the denoiser must be unable to ignore the stage mechanism when demographic morphology supplies an easier shortcut. Removing mismatch loss produces the same failure at lower magnitude, with young Wake leakage of 0.39 ± 0.05, supporting the need to bind the specific token value to the waveform rather than merely make the branch nonzero. Random KAN also collapses young N1 toward Wake (0.52 ± 0.06), showing that architecture alone is insufficient without learned physiological coordinates.

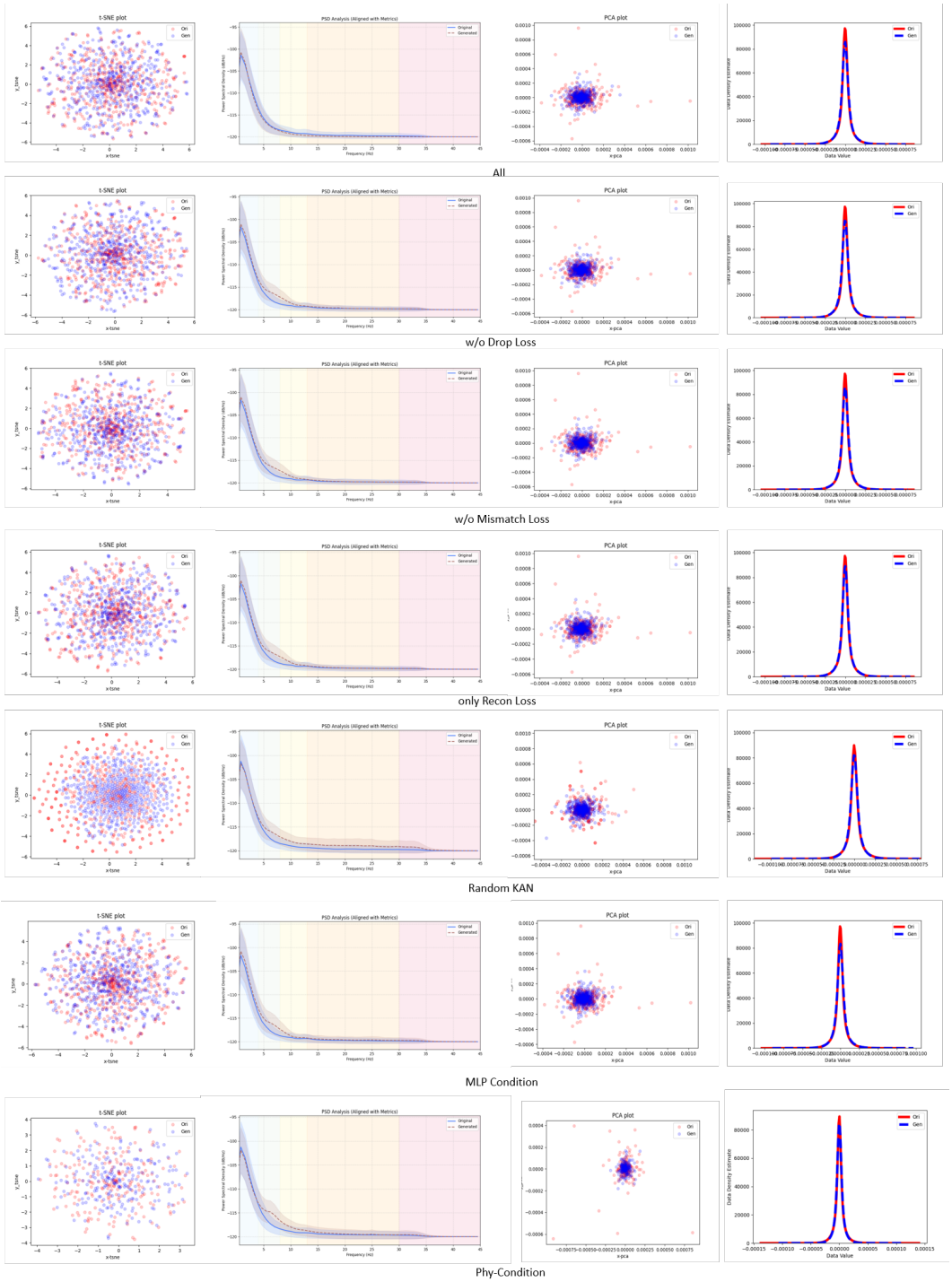

Figure 8: Ablation visualization results.

Recon Only illustrates why an isolated strong cell is not decisive. It obtains the second-highest old-N1 recall (0.77 ± 0.04) yet reaches only 0.41 ± 0.05 for young N1, with 0.33 ± 0.03 Wake leakage. With the updated results, the full model gives the best recall in both groups: 0.69 ± 0.08 for young N1 and 0.88 ± 0.03 for old N1. Its cross-age recall change is 0.19, compared with 0.50 for w/o Drop Loss, 0.36 for Recon Only, 0.33 for MLP conditioning, 0.30 for EEGDiffuser, and 0.29 for Random KAN. The full model also has low visible Wake leakage (0.06/0.03) and N2/REM leakage

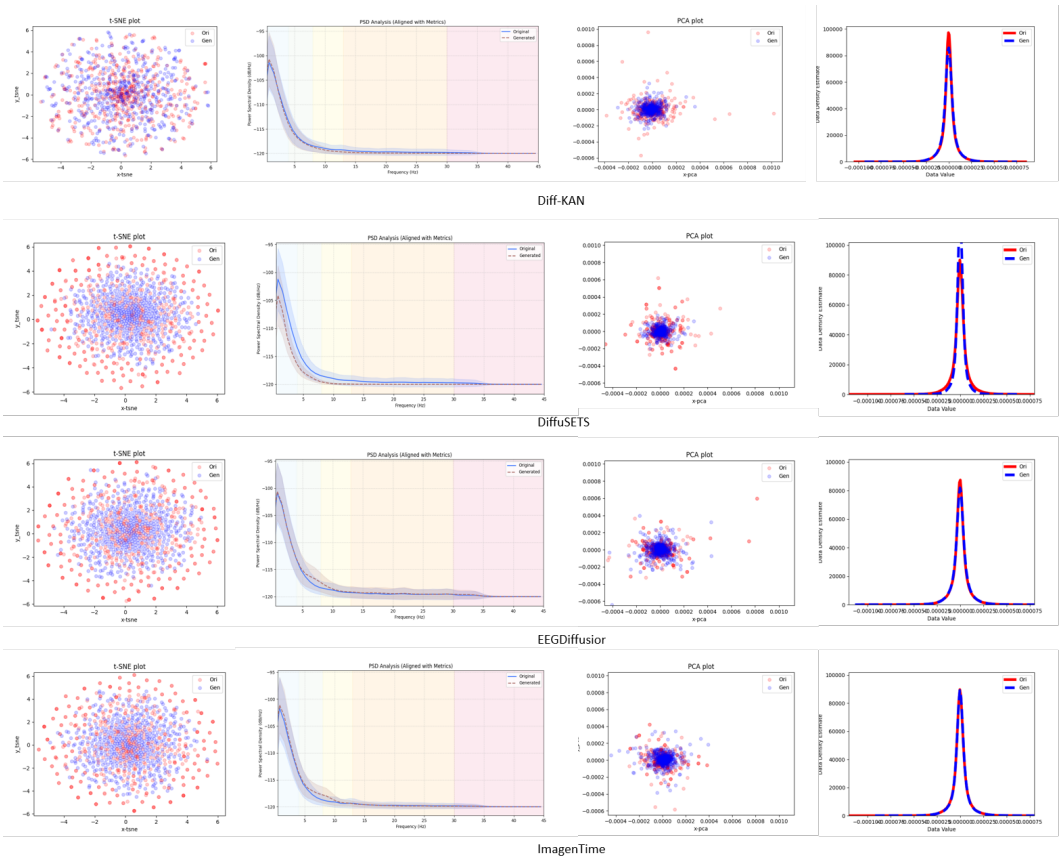

Figure 9: Comparison visualization results.

(0.02/0.05 in the young group and 0.05/0.01 in the old group); the omitted young N3 residual is 0.18. Thus, the updated table supports stronger target-stage recognizability across both age groups, while also showing that the remaining young-group errors are concentrated in the omitted N3 category rather than Wake. This pattern is consistent with encoding age and stage as explicit physiological coordinates and enforcing both branch necessity and value specificity.

Random Age keeps the pretrained physiological encoder and full denoiser but permutes age-branch coordinates during training. Its competitive N1 recalls do not prove correct age use because the requested age–coordinate correspondence has been broken. It is therefore evaluated by the positive-control age probe below rather than by another marginal signal table.

Positive control: age remains decodable within REM. Strong stage preservation can be obtained trivially by ignoring age. Random Age illustrates this failure: despite competitive N1 recall in Table 9, its age-probe balanced accuracy is only 0.52 ± 0.04, near chance. The frozen probe therefore supplies an essential positive requirement. The full model reaches old/young recall of 0.69/0.74 and balanced accuracy of 0.72, close to the real-REM reference of 0.71/0.74 and 0.73. External generators reach only 0.46–0.49 balanced accuracy, while MLP conditioning falls to 0.36. The conjunction of the signal diagnostics, cell-level table, and age probe tests two requirements: the requested sleep stage must remain recognizable, and age information must remain present rather than being discarded to obtain that stage fidelity.

A.10 Dataset Split Details

The study uses a condition-balanced subset of Sleep-EDF Expanded Sleep Cassette with Fpz-Cz and Pz-Oz EEG. The subject-disjoint training split contains 21 subjects, 42 nights, and 35,503 epochs: seven subjects in each age band, with 10 female and 11 male subjects. The held-out test split con-

Table 9: N1 age-sweep classification destinations. Entries are mean±SD over five seeds. Arrow columns give the fraction classified as the indicated wrong stage. Omitted N3 leakage is the residual to one (up to rounding) and reaches 0.18 in the updated young-Ours cell. Bold and underline mark the best and second-best unrounded means; exact ties share rank.

Age 25–39, N1 requested				
Method	**N1 ↑**	**→W ↓**	**→N2 ↓**	**→REM ↓**
Internal variants				
Random KAN	0.34±0.07	0.52±0.06	<u>0.03±0.04</u>	<u>0.11±0.03</u>
Random Age	0.56±0.03	0.19±0.05	0.13±0.04	0.12±0.06
MLP Cond.	0.39±0.03	0.12±0.06	0.23±0.04	0.22±0.02
w/o Drop Loss	0.25±0.03	0.59±0.02	0.03±0.02	0.13±0.02
w/o Mismatch Loss	0.44±0.08	0.39±0.05	0.04±0.03	0.13±0.04
Recon Only	0.41±0.05	0.33±0.03	0.08±0.02	0.18±0.05
Ours	**0.69±0.08**	**0.06±0.02**	**0.02±0.01**	**0.05±0.02**
External conditional generators				
ImagenTime	<u>0.57±0.10</u>	**0.01±0.01**	0.18±0.03	0.22±0.06
EEGDiffuser	0.40±0.07	0.29±0.04	0.11±0.07	0.21±0.09
DiffuSETS	0.48±0.04	0.11±0.06	0.08±0.02	0.17±0.06

Age 61–90, N1 requested				
Method	**N1 ↑**	**→W ↓**	**→N2 ↓**	**→REM ↓**
Internal variants				
Random KAN	0.63±0.07	0.30±0.04	**0.03±0.02**	0.04±0.03
Random Age	0.68±0.05	0.13±0.04	0.12±0.04	0.06±0.03
MLP Cond.	0.72±0.02	0.17±0.05	<u>0.06±0.04</u>	0.04±0.01
w/o Drop Loss	<u>0.75±0.03</u>	0.13±0.07	0.09±0.03	0.02±0.02
w/o Mismatch Loss	0.71±0.04	0.14±0.05	0.11±0.02	0.04±0.01
Recon Only	0.77±0.04	0.14±0.04	<u>0.06±0.03</u>	<u>0.03±0.03</u>
Ours	**0.88±0.03**	**0.03±0.02**	0.05±0.02	**0.01±0.01**
External conditional generators				
ImagenTime	0.52±0.12	0.04±0.01	0.19±0.05	0.25±0.08
EEGDiffuser	0.70±0.09	0.09±0.02	0.11±0.07	0.08±0.04
DiffuSETS	0.61±0.12	0.03±0.03	0.17±0.04	0.13±0.06

tains nine subjects, 18 nights, and 6,805 epochs: three subjects per band, with four female and five male subjects. Mean selected-epoch contributions are 1,690.6 per training subject and 756.1 per test subject. The generator uses a fixed training schedule without a model-selection validation split. Validation is used only for the binary age probe, whose validation subjects are drawn from the real training partition; the probe is frozen before evaluation on real test and generated data.

For the 25–39, 40–60, and 61–90 bands, the female/male subject counts are 3/4, 4/3, and 3/4 in training and 1/2, 1/2, and 2/1 in testing. Every subject in each demographic cell contributes at least one selected epoch to each of Wake, N1, N2, N3, and REM, so the stage-coverage count equals that cell's subject count. All selected subjects contribute two nights; these counts refer to this subset, not the full corpus.

Table 10: Age decodability under a fixed REM request. Bold and underline mark the best and second-best generated-model results; held-out real EEG is an unranked reference.

Method	**Old-age Recall ↑**	**Young-age Recall ↑**	**Age-probe Bal. Acc. ↑**
Held-out real REM			
Real reference	0.71 ± 0.01	0.74 ± 0.00	0.73 ± 0.02
External conditional generators			
ImagenTime	0.39 ± 0.02	0.53 ± 0.01	0.46 ± 0.00
EEGDiffuser	0.42 ± 0.00	0.55 ± 0.02	0.49 ± 0.01
DiffuSETS	0.41 ± 0.01	0.54 ± 0.02	0.47 ± 0.02
Internal variants			
MLP Cond.	0.24 ± 0.02	0.48 ± 0.01	0.36 ± 0.00
Random Age	0.53 ± 0.03	0.51 ± 0.07	0.52 ± 0.04
Random $A \times S$	0.62 ± 0.01	0.62 ± 0.00	0.62 ± 0.02
w/o Drop Loss	0.60 ± 0.00	0.70 ± 0.02	0.65 ± 0.01
Random KAN	0.60 ± 0.02	0.66 ± 0.01	0.63 ± 0.00
Recon Only	0.59 ± 0.01	0.67 ± 0.02	0.63 ± 0.01
w/o Mismatch Loss	$\underline{0.66 \pm 0.02}$	$\underline{0.71 \pm 0.00}$	$\underline{0.69 \pm 0.01}$
Ours	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.74 \pm 0.02}$	$\mathbf{0.72 \pm 0.00}$

Table 11: Split-level cohort and recording summary.

Split	Subjects	Nights	Epochs	Epochs/subject
Train	21	42	35,503	1,690.6
Test	9	18	6,805	756.1