PASF: PERSISTENCE-AWARE SPECTRAL FILTERING FOR TEST-TIME ADAPTATION IN SLEEP STAGING

Anonymous authors
Paper under double-blind review (Submission Number: 28109)

ABSTRACT

Physiological signal models generalize poorly across devices, sensor configurations, and recording protocols. Such acquisition shifts frequently manifest in the frequency domain, motivating spectral adaptation at test time. However, a recording's marginal spectrum reflects not only acquisition conditions but also the task-relevant physiological states it contains and their prevalence. Wholesale alignment can therefore suppress state-dependent physiology alongside domain nuisance, particularly under label shift. We study source-free test-time adaptation and identify **cross-state-persistent spectral shifts** as a conservative correction target. These persistent shifts remain consistent across physiological states, while state-specific disagreement serves as counterevidence against full correction. We instantiate this principle as **Persistence-Aware Spectral Filtering (PASF)**, which estimates probability-weighted state-conditioned spectral residuals relative to compact source templates. Their shared component shapes a recording-specific, phase-preserving filter, while the shared-energy fraction gates its strength. PASF processes each complete recording in two passes without target labels, raw source recordings, model updates, or gradients. Across four sleep-staging datasets comprising 740 recordings, PASF achieves four-domain mean balanced accuracies of 75.56% with U-Sleep and 74.30% with a CNN–Transformer. These results improve on source-only inference by 3.79 and 5.78 percentage points, respectively. Controlled comparisons show reduced composition-induced false correction and improved recovery under shared spectral distortion. Analyses of real recordings support cross-state persistence, while experiments beyond sleep staging identify the applicability and limits of the shared-shift principle.

1 INTRODUCTION

Physiological-signal models often generalize poorly under acquisition conditions that differ from those used for training. Changes in devices, sensor configurations, referencing conventions, recording protocols, and preprocessing pipelines can induce distribution shifts across physiological modalities, including electrocardiography, electroencephalography, and electromyography (Ballas & Diou, 2024; Oh & Bui, 2025). Because deployment domains may be unknown during training, adaptation must often operate on unlabeled physiological recordings at test time. We investigate this general problem through automated sleep staging, a state-structured classification task that assigns each 30-second epoch to Wake, N1, N2, N3, or REM. Models often transfer poorly across hardware, montages, protocols, and populations (Nasiri & Clifford, 2020; Guillot & Thorey, 2021; Dei Rossi et al., 2026).

The frequency domain is a natural space for test-time adaptation because acquisition changes often alter the spectral response of physiological signals. Input-level spectral adaptation methods therefore align target Fourier statistics or power spectra with source-domain references (Krishnamachari et al., 2023; Gnassounou et al., 2023). However, the marginal spectrum of a recording is determined not only by acquisition conditions but also by the task-relevant physiological states it contains and their prevalence. In sleep EEG, for example, both oscillatory peaks and aperiodic spectral structure vary systematically across stages (Brancaccio et al., 2020; Ameen et al., 2025). A marginal source–target discrepancy may therefore entangle acquisition-related variation with sleep-stage composition and meaningful state-dependent physiology, so correcting it wholesale may suppress predictive spectral

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.

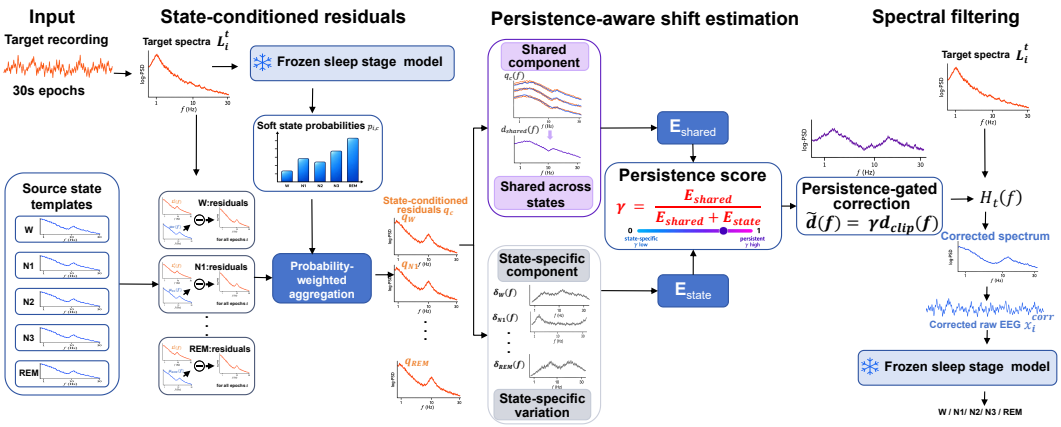

Figure 1: **PASF pipeline.** Given an unlabeled target recording, PASF computes epoch-wise log-power spectra and uses soft probabilities from a frozen source classifier to aggregate state-conditioned residuals relative to compact source templates. Their equal-state mean estimates a cross-state-persistent spectral shift and determines the correction shape, while state-specific deviations provide counterevidence against full correction. The relative energies of the shared and state-specific components define a persistence score γ that gates a recording-specific, phase-preserving spectral filter. The same frozen classifier then maps the corrected epochs to final sleep-stage predictions.

structure, particularly under label shift (Li et al., 2025). The central challenge is thus not to eliminate spectral discrepancy, but to identify a conservative correction target within it.

We therefore reformulate spectral adaptation from aligning marginal source–target spectra to identifying spectral shifts that persist across task-relevant physiological states. Recording-level factors may introduce spectral structure that recurs across states, whereas task-relevant physiology varies with state. We refer to this contrast as *cross-state persistence* and use it as a conservative correction cue: the component shared across state-conditioned source–target residuals is a candidate recording-level shift, while state-specific disagreement provides evidence against full correction. Importantly, cross-state persistence is not a causal identifier of acquisition effects: the shared component may also contain recording-specific physiology or other systematic mismatch. It instead narrows the correction target relative to the full marginal discrepancy.

Based on this principle, we propose **Persistence-Aware Spectral Filtering (PASF)**, an input-level framework for test-time spectral adaptation of state-structured physiological signals. For each target recording, PASF combines frozen source-model probabilities with compact source-derived state templates to estimate state-conditioned residuals and separate them into shared and state-specific components. The shared component shapes a phase-preserving spectral filter, while its fraction of residual energy defines a persistence score that controls correction strength and weakens correction when states disagree. PASF requires no target labels, raw source recordings, or model updates at test time.

Contributions. Our contributions are threefold:

- **A cross-state-persistence principle for conservative spectral adaptation.** We reformulate spectral test-time adaptation from marginal source–target alignment to estimating discrepancy components that persist across task-relevant physiological states. State-specific disagreement provides counterevidence that moderates correction strength, defining a conservative correction target for state-structured physiological signals.
- **Persistence-Aware Spectral Filtering (PASF).** We operationalize this principle through probability-weighted state-conditioned residuals, equal-state shared estimation, and a shared-energy gate. PASF produces a recording-specific, phase-preserving spectral filter in two passes, requiring no target labels, raw source recordings, model updates, or gradients.

- **Empirical gains, mechanistic evidence, and applicability boundaries.** Across four sleep-staging datasets comprising 740 recordings, leave-one-dataset-out evaluation shows the highest mean balanced accuracy among compared external baselines with both U-Sleep and a CNN–Transformer. Gains over source-only are 3.79 and 5.78 percentage points, respectively, with paired participant-cluster bootstrap intervals supporting aggregate improvements over CMMN and Tent. Controlled interventions and leave-one-state-out analyses support the persistence criterion and demonstrate reduced composition-induced false correction. Responsibility perturbations and studies of sedation, emotion recognition, and motor imagery characterize sensitivity and identify the applicability and limits of shared spectral correction.

2 Related Work

Cross-domain learning for physiological signals. Cross-domain physiological-signal learning spans representation transfer, domain-specific normalization, and statistics alignment (Kobler et al., 2022; Fan et al., 2022; Oh & Bui, 2025). Sleep-specific studies address montage-robust transfer, multi-level domain alignment, structured adversarial adaptation, and source-free personalization (Guillot & Thorey, 2021; Wang et al., 2024; Yoo et al., 2022; Zhou et al., 2025). Time-series benchmarks also expose sensitivity to evaluation and model selection (Ragab et al., 2023), while marginal statistical alignment can over-correct EEG under label shift (Li et al., 2025).

Test-time and source-free adaptation. Many test-time adaptation methods use unlabeled targets to update model parameters, normalization statistics, or prediction dynamics (Sun et al., 2020; Wang et al., 2021; 2022; Niu et al., 2022; 2023; Patel et al., 2025). Their effectiveness can depend on source-model quality, shift type, and unlabeled model selection (Zhao et al., 2023). Recent EEG-specific work such as BTTA-DG instead performs gradient-free online adaptation for cross-subject motor-imagery decoding using a Sinc-based backbone, Dirichlet feature projection, and GMM-driven Bayesian calibration (Luo et al., 2026). Like BTTA-DG, PASF avoids test-time gradient updates, but it targets recording-level sleep staging and uses state-conditioned input residuals to construct a phase-preserving correction for a frozen predictor.

Spectral normalization and frequency-domain adaptation. Frequency-domain adaptation either aligns temporal–spectral representations (Liu & Xue, 2021; He et al., 2023; Liu et al., 2024) or matches Fourier statistics at the input (Krishnamachari et al., 2023). The closest sleep-specific baseline, Convolutional Monge Mapping Normalization (CMMN), filters each target recording toward a source PSD barycenter without updating the predictor (Gnassounou et al., 2023). However, marginal spectra vary with physiological-state composition as well as acquisition conditions (Schneider et al., 2022; Li et al., 2025), so their discrepancy need not be wholly correctable. PASF shares CMMN's input-level, frozen-predictor deployment setting but replaces marginal PSD alignment with a state-conditioned estimate of the spectral component shared across physiological states as a conservative correction target.

3 Method

PASF performs recording-level test-time input adaptation with a frozen source classifier. Its objective is not to align the complete target spectrum, but to identify the component of the target–source spectral discrepancy that persists across physiological states. To operationalize this principle without target labels, PASF estimates state-conditioned residuals using frozen-model responsibilities, projects them into shared and state-dependent components, and uses the shared-energy fraction to control a phase-preserving correction. Figure 1 illustrates the procedure.

3.1 Problem Setup and Persistence Model

We consider recording-level adaptation in which a source-trained classifier and compact source statistics are available, while the target recording is unlabeled and the classifier remains fixed. Let

$$\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}, \qquad \mathcal{D}_t = \{x_i^t\}_{i=1}^{N_t}, \qquad f_{\theta_s} : \mathcal{X} \to \mathbb{R}^{|\mathcal{C}|}.$$

Here, $\mathcal{C} = \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ for sleep staging. PASF has transductive access to all N_t epochs in the current recording, but never accesses their labels or updates θ_s.

Each epoch is represented by its discretized natural-log power spectrum

$$L(x) = \ln(\max\{\mathrm{PSD}(x), \epsilon_{\mathrm{psd}}\}) \in \mathbb{R}^F,$$

with $\epsilon_{\mathrm{psd}} = 10^{-10}$, and we write $L_i^t = L(x_i^t)$. Before deployment, each leave-one-domain-out fold constructs source state templates $\mathcal{S} = \{\mu_c\}_{c\in\mathcal{C}}$ from its labeled source domains only. For each state c, let $\mathcal{A}_c$ contain the source log spectra after removing smooth recording-specific offsets; the template is their mean, $\mu_c = |\mathcal{A}_c|^{-1} \sum_{L\in\mathcal{A}_c} L$. Only $\mathcal{S}$, rather than the sampled source recordings, is retained for deployment; Appendix A.1 gives the construction details.

PASF models a correctable recording-level disturbance as an additive log-power component that recurs across physiological states, while task-relevant target–source mismatch may vary with state. Relative to the source template of its latent state, a target spectrum is therefore written as

$$L_i^t(f) = \mu_{c_i}(f) + d_t(f) + s_{t,c_i}(f) + \varepsilon_i(f), \tag{1}$$

where d_t is a state-invariant recording-level shift, s_{t,c_i} absorbs state-dependent mismatch, and ε_i is noise. Equation 1 motivates the estimator but is not an identifiable causal decomposition. PASF instead estimates the observable component shared across state-conditioned residuals and treats state-specific disagreement as evidence against full correction.

3.2 State-Conditioned Residual Estimation

Equation 1 motivates state-conditioned residuals, but the target states are unavailable at test time. Marginal aggregation would reintroduce physiological-state composition into the discrepancy, so PASF uses the frozen source classifier to obtain a soft responsibility for each uncorrected target epoch,

$$p_{i,c} = \mathrm{softmax}\big(f_{\theta_s}(x_i^t)\big)_c, \qquad \sum_{c\in\mathcal{C}} p_{i,c} = 1. \tag{2}$$

Using all states, PASF estimates a frequency-wise residual for state c with a weighted median,

$$q_c(f) = \mathrm{wmed}_{i=1}^{N_t}\left[L_i^t(f) - \mu_c(f);\, p_{i,c}\right], \qquad c \in \mathcal{C}, \tag{3}$$

Thus, $q_c(f)$ is the recording-level deviation supported by epochs associated with state c. Soft responsibilities avoid hard pseudo-labels, while the weighted median limits the influence of spectrally extreme epochs. Appendix A.8 compares mean-based alternatives under a matched protocol. Interpreting the estimated residuals as target-state-conditioned requires informative aggregate responsibilities from the frozen classifier.

3.3 Cross-State Persistence Decomposition

PASF operationalizes persistence as the component of the state-conditioned residuals that is constant across states. We define it by the least-squares projection

$$d_{\mathrm{shared}} = \arg\min_{d\in\mathbb{R}^F} \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \|q_c - d\|_2^2 = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} q_c, \qquad \delta_c = q_c - d_{\mathrm{shared}}. \tag{4}$$

Equal-state weighting encodes the principle that persistence requires support across states rather than support from the most prevalent state. It also gives $\sum_c \delta_c = 0$, so δ_c contains the variation not explained by the shared-across-state subspace.

The projection yields an exact energy decomposition over the F retained frequency bins:

$$\begin{aligned} E_{\mathrm{shared}} &= \frac{1}{F}\sum_f d_{\mathrm{shared}}(f)^2, \quad E_{\mathrm{state}} = \frac{1}{|\mathcal{C}|F}\sum_{c,f}\delta_c(f)^2, \\ \frac{1}{|\mathcal{C}|F}\sum_{c,f} q_c(f)^2 &= E_{\mathrm{shared}} + E_{\mathrm{state}}, \qquad \gamma = \frac{E_{\mathrm{shared}}}{E_{\mathrm{shared}} + E_{\mathrm{state}} + \epsilon_\gamma}. \end{aligned} \tag{5}$$

Here, ϵ_γ is machine epsilon. Up to this stabilizer, $\gamma \in [0, 1]$ is the fraction of residual energy explained by the shared-state subspace. Thus, $d_{\text{shared}}(f)$ determines the correction shape, while γ weakens its recording-level strength when state-dependent variation dominates. The score measures shared residual energy, not estimator confidence or the probability of downstream improvement. A high shared-energy fraction supports persistence but does not establish that the shared component is purely non-physiological.

3.4 Persistence-Gated Correction and Inference

PASF converts the shared projection into a recording-specific correction whose shape is given by $d_{\text{shared}}(f)$ and whose strength is controlled by γ. To bound extreme responses, we clip the natural-log-power curve to ± 20 dB, equivalently $d_{\max} = 20 \ln(10)/10$, and define

$$d_{\text{clip}}(f) = \text{clip}(d_{\text{shared}}(f), -d_{\max}, d_{\max}), \qquad \widetilde{d}(f) = \gamma d_{\text{clip}}(f), \qquad H_t(f) = \exp\Big[-\tfrac{1}{2}\widetilde{d}(f)\Big], \tag{6}$$

The 20-dB bound was selected on labeled source-validation data and fixed before target evaluation. Appendix A.8 reports a subsequent sensitivity analysis on the held-out target evaluation data; its results were not used to select or revise the bound. The shared curve is estimated directly on the frequency grid. The factor $1/2$ converts a log-power correction to a Fourier-amplitude multiplier. After interpolating $\widetilde{d}$ from the Welch grid to the real-FFT grid, PASF applies the same recording-specific response to every epoch:

$$x_i^{\text{corr}} = \text{Re}\big[\mathcal{F}^{-1}\big(H_t \odot \mathcal{F}(x_i^t)\big)\big]. \tag{7}$$

Because $H_t(f) > 0$ is real, this operation rescales Fourier magnitude without changing Fourier phase. A scalar shared-energy gate controls recording-level correction strength while $d_{\text{shared}}(f)$ specifies its frequency-dependent shape, avoiding a separate gate estimated from local agreement at each frequency bin.

For each recording, PASF first obtains $p_{i,c}$ from the uncorrected epochs, estimates one $\widetilde{d}$, filters all epochs, and then runs the same frozen classifier on the corrected recording:

$$\theta_t = \theta_s, \qquad \widehat{y}_i = \arg\max_{c \in \mathcal{C}} [f_{\theta_s}(x_i^{\text{corr}})]_c, \qquad i = 1, \ldots, N_t. \tag{8}$$

This is a one-shot, two-pass procedure over a complete recording, not online sample-by-sample adaptation. Deployment requires only the frozen classifier and compact templates: no gradients, optimizer states, target labels, or raw source recordings. Algorithm 1 in Appendix A.1 summarizes the procedure.

4 Experiments

4.1 Experimental Setup

Datasets and protocol. We evaluate HMC (Alvarez-Estevez & Rijsman, 2022), ISRUC-Sleep (Khalighi et al., 2016), 329 SHHS visit-1 recordings with an apnea–hypopnea index (AHI) below 5 events per hour (Quan et al., 1997), and Sleep-EDF Expanded (Kemp et al., 2000). SHHS therefore denotes a low-AHI subset, not the full cohort. In each leave-one-dataset-out fold, three datasets are labeled sources and the fourth is an unseen target. The benchmark contains 740 target recordings from approximately 664 unique participants; repeated nights are adapted independently. Target labels enter neither PASF's estimator nor the cap choice for their outer fold; they are used for evaluation and, as disclosed in Appendix A.8, a post hoc sensitivity analysis on the held-out target evaluation data that does not alter the frozen configuration. We train five source models with seeds $\{7, 17, 23, 42, 1337\}$.

Backbones. We evaluate a PyTorch implementation of fully convolutional U-Sleep (Perslev et al., 2021) and a single-EEG adaptation of the SleepBand student network (Lu et al., 2026). SleepBand is a domain-generalization framework; its student feeds learnable spectral and CNN epoch features to a Transformer sequence encoder. We disable its teacher and consistency objectives and train this SleepBand-based CNN–Transformer as a supervised backbone. This tests generalization beyond

U-Sleep's U-Net/BatchNorm design rather than reproducing the complete SleepBand system. Both backbones are trained from scratch with matched data, optimization, and seeds. Our code is available at https://anonymous.4open.science/r/PASF-Persistent-aware-spectral-filtering-0455.

Signal preprocessing and PASF frequency grid. We use C4–M1 for HMC, C4–A1/M1 for ISRUC, C3–A2 (EEG(sec)) for SHHS, and Fpz–Cz for Sleep-EDF. Signals are presented at 100 Hz in 30-s epochs and mapped to $\{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$. PASF computes log-power residuals over 0.3–40 Hz using Welch's method, estimates the shared correction on the resulting frequency grid, and interpolates it to the epoch Fourier grid for filtering. Appendices A.2 and A.1 give architecture details, exact spectral settings, and preprocessing provenance.

Table 1: Cross-domain balanced accuracy (BAcc, %; mean ± sample SD over five seeds). Avg. weights the four target domains equally; **bold** and underline denote the best and second-best result per backbone.

Backbone	Method	HMC	ISRUC	SHHS	Sleep-EDF	Avg.
U-Sleep (U-Net)	Source-only	73.37±0.37	65.12±0.50	77.71±0.06	70.89±2.01	71.77±0.41
	PSDNorm‡	73.26±0.42	69.63±0.27	77.99±0.20	72.95±2.59	73.46±0.71
	CMMN	70.28±0.16	72.90±0.26	76.67±0.23	75.16±0.53	73.75±0.16
	Tent	73.19±0.26	74.18±0.34	76.87±0.76	70.62±0.61	73.72±0.31
	EATA	73.01±0.26	74.09±0.34	76.67±0.77	70.42±0.64	73.55±0.33
	SAR	73.10±0.25	74.16±0.33	76.80±0.76	70.50±0.62	73.64±0.32
	SF-UIDA§	**73.41±0.35**	64.96±0.47	77.78±0.07	71.06±1.97	71.80±0.41
	PASF	72.58±0.56	**74.45±0.34**	**78.37±0.11**	**76.84±0.21**	**75.56±0.18**
SleepBand-based CNN–Transformer	Source-only	72.33±0.27	63.46±0.58	77.50±0.26	60.77±4.16	68.51±1.12
	PSDNorm‡	70.04±0.12	67.35±0.51	75.32±0.25	69.26±0.73	70.49±0.09
	CMMN	68.61±0.47	72.43±0.24	76.03±0.22	71.05±0.47	72.03±0.14
	Tent	72.40±0.19	70.39±0.25	77.12±0.27	67.25±2.03	71.79±0.44
	EATA	72.33±0.19	70.39±0.27	77.04±0.25	66.94±2.04	71.67±0.45
	SAR	72.32±0.20	70.39±0.28	77.04±0.25	67.01±2.00	71.69±0.43
	SF-UIDA§	72.44±0.22	68.77±0.79	76.29±0.73	65.61±3.03	70.78±0.79
	PASF	**72.67±0.40**	**72.83±0.39**	**77.86±0.36**	**73.82±0.92**	**74.30±0.37**

‡ Five-seed models retrained with PSDNorm in place of BatchNorm.
§ Five-seed mechanism-consistent single-channel SF-UIDA ports (5 self-supervised epochs and 10 fine-tuning epochs), not the official multi-channel EEG/EOG implementation.

Baselines. We compare PASF with Source-only, the spectral methods PSDNorm and CMMN (Gnassounou et al., 2023), and the model-adaptation methods Tent (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), and a single-channel SF-UIDA-style port (Zhou et al., 2025). Tent, EATA, and SAR reset to the source checkpoint for every recording, process chronological non-overlapping target windows, and update only BatchNorm affine parameters. All rows for both backbones report five source-training seeds. PSDNorm and CMMN use method-specific retrained source models, while the SF-UIDA-style implementation is a mechanism-consistent single-channel port rather than a strict reproduction of the original multi-channel method. Appendix A.5 specifies target-batch access, the complete batch-count distribution, optimization settings, and deployment-access differences.

Metrics and statistical analysis. Balanced accuracy (BAcc) is primary; Macro-F1, accuracy, and Cohen's κ are secondary. For each seed, we first average recording-level metrics within each target domain and then average the four domains equally, without weighting by recordings or epochs. We report the mean and sample standard deviation of these domain-macro values across five seeds. The primary U-Sleep PASF–Source-only confidence interval is a two-sided Student's t-interval with four degrees of freedom computed from the five paired seed-level domain-macro differences; it reflects training-seed variation, not finite-participant sampling uncertainty.

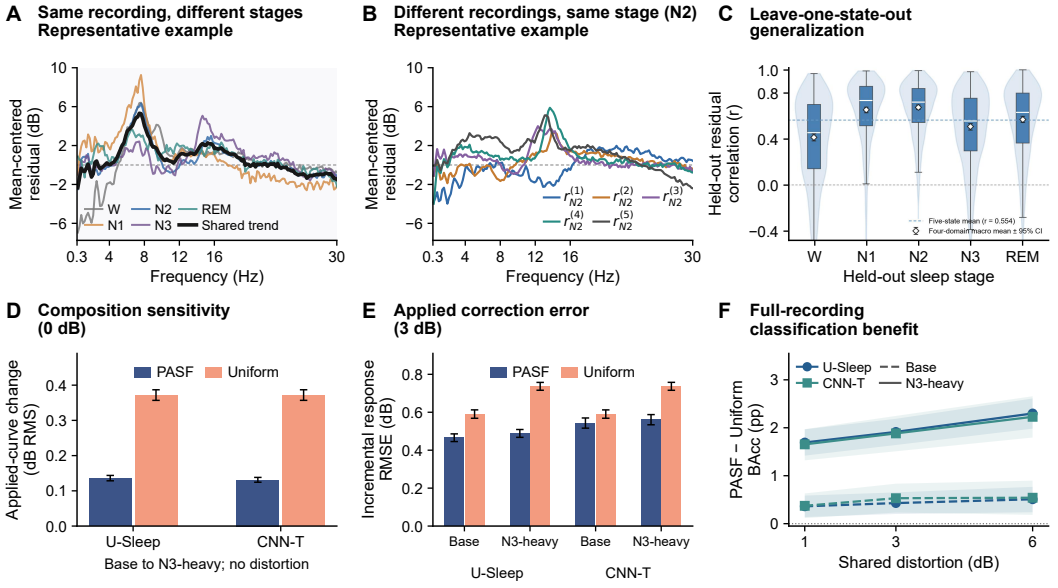

Figure 2: **Mechanistic evaluation of cross-state persistence.** **(A–B)** State residuals across stages within one recording and across recordings for N2. **(C)** Leave-one-state-out generalization. Panels A–C use true target labels for offline stage stratification. **(D)** Applied-curve change under N3 enrichment without added distortion. **(E)** Incremental applied-response RMSE at 3 dB, referenced to each method's clean Base curve. **(F)** Full-recording PASF-minus-Uniform BAcc. D–F use 656 recordings (616 participants), two seed-42 backbones, and three fixed contexts; intervals are pointwise 95% participant-cluster CIs. Appendix A.7 defines the metrics.

4.2 Cross-Domain Effectiveness

PASF achieves the highest four-domain mean BAcc among the external-method comparisons in Table 1, reaching 75.56% with U-Sleep and 74.30% with the CNN–Transformer. The paired gains over Source-only are 3.79 and 5.78 percentage points, respectively; for U-Sleep, the paired-seed 95% t-interval is 3.15–4.42 points. Improvements extend across seven of eight backbone–target combinations. Within this table, PASF ranks first on ISRUC, SHHS, and Sleep-EDF with U-Sleep and on all four targets with the CNN–Transformer. The only decrease is on HMC with U-Sleep, at 0.79 points. Section 4.4 separately reports the stronger U-Sleep Uniform control.

Among the external baselines, PSDNorm and CMMN both improve average performance, with CMMN ranking second in Table 1 for both backbones. Tent, EATA, and SAR remain competitive on ISRUC with U-Sleep, but their improvements over Source-only are less consistent elsewhere. The single-channel SF-UIDA-style port remains close to Source-only on average with U-Sleep and improves with the CNN–Transformer.

The U-Sleep decrease on HMC may partly reflect its clinically heterogeneous population, which was referred for polysomnography in the context of different sleep disorders (Alvarez-Estevez & Rijsman, 2022). Such disorders may induce stage-dependent spectral changes that are not fully addressed by PASF's shared-component correction.

4.3 Mechanistic Validation of the Persistence Principle

Existence and cross-state predictability. State-conditioned residuals exhibited a common spectral trend within a recording despite stage-dependent differences (Figure 2(A)). Across 574 eligible recordings, cross-state correlations were higher within recordings than between dataset- and state-matched recordings, with a paired contrast of +0.245 (95% CI, 0.230–0.260; Appendix A.6). This pattern held in every target dataset, supporting a recording-specific spectral component shared across sleep stages.

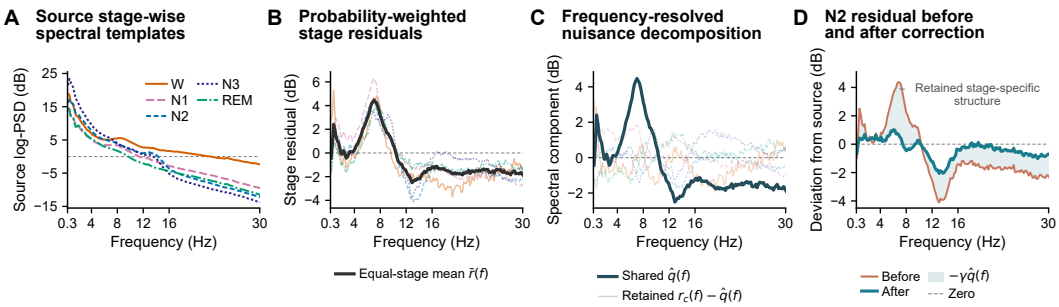

Figure 3: **Representative frequency-resolved PASF correction on SN094.** **(A)** Source stage templates; **(B)** probability-weighted residuals; **(C)** shared spectral component and retained stage-specific components; **(D)** N2 residual before and after correction.

Leave-one-state-out analysis tested whether the shared spectral structure extended to states excluded from estimation (Figure 2(C)). For the same 574 recordings, we removed dataset–stage population structure using leave-one-recording-out median centering. We then averaged residuals from the other four states and computed the Pearson correlation between this estimate and the held-out residual spectrum. The domain-macro correlations were positive for all five held-out states. These results support a recording-specific spectral component that can be estimated from other sleep states, providing an empirical basis for PASF's cross-state shared estimation.

Selective correction under composition change. We compare PASF and a uniform-responsibility control using identical estimation epochs and evaluate both filters on the same complete recording. Without added distortion, N3 enrichment changes PASF's applied curve 63–65% less than Uniform's across backbones (Figure 2(D)). Under Base composition and a 3-dB distortion, Uniform tracks the raw, unclipped response more accurately. With N3 enrichment, however, PASF yields lower incremental applied-response error (E) and improves BAcc over Uniform by 1.91 points with U-Sleep and 1.88 with CNN–Transformer (F). The advantage persists at 1 and 6 dB, supporting selective correction under composition–distortion confounding; domain effects remain heterogeneous (Appendix A.7).

Visual analysis of spectral correction. Figure 3 provides a frequency-resolved example. Distinct source templates serve as state-specific references, while the target residuals share a 6–8-Hz excess and a broad high-frequency deficit that PASF corrects. In N2, PASF reduces the shared offset while retaining a local 12–15-Hz discrepancy in the spindle-associated sigma band (Purcell et al., 2017), illustrating that it does not force each state to match its source template.

4.4 Component Ablations

Table 2 compares residual estimation, state aggregation, and correction strength. Marginal correction estimates a pseudo-prevalence-matched residual. Shared only uses state-conditioned residuals, equal-state aggregation, and $\gamma = 1$; prevalence-weighted uses adaptive γ and pseudo-prevalence aggregation. PASF combines equal-state aggregation with adaptive γ. Uniform removes target-state assignments while retaining source templates, equal-state aggregation, and the adaptive-strength rule; it differs from Marginal in both source reference and strength.

Shared only exceeds Marginal correction in all eight backbone–domain pairs ($+0.45/+0.52$ mean BAcc), testing target conditioning and source-reference weighting jointly. Adaptive γ adds $+0.65/+0.55$ BAcc points, while equal-state aggregation adds $+0.15/+0.20$ (from displayed means; Appendix A.8). Across five seeds, PASF-minus-Uniform BAcc is -0.28 points for U-Sleep and $+0.26$ for CNN–Transformer. Conditioning consistently helps on Sleep-EDF but reduces BAcc on ISRUC, indicating domain-dependent effects (Appendix A.12). Uniform is therefore a deliberate PASF variant and structural ablation: it preserves the state-balanced source anchor, shared projection, and adaptive correction while removing target-state assignments. Its strong performance shows

Table 2: Matched PASF component ablation under the final direct-residual protocol: four-domain BAcc (%, mean ± sample SD over five seeds).

Variant	Aggregation	γ	U-Sleep	CNN–Trans.
Marginal correction	Marginal	1	74.46 ± 0.17	73.23 ± 0.29
Shared only	Equal-state	1	74.91 ± 0.18	73.75 ± 0.30
Prevalence-weighted	Prevalence	Adaptive	75.41 ± 0.18	74.09 ± 0.35
Uniform responsibilities	Equal-state	Adaptive	$\mathbf{75.84 \pm 0.11}$	74.04 ± 0.40
PASF	Equal-state	Adaptive	75.56 ± 0.18	$\mathbf{74.30 \pm 0.37}$

that this simplified member of the PASF family is useful in its own right; the comparisons above isolate the additional, domain-dependent contribution of explicit target-state conditioning.

4.5 Generalization and Reliability

Cross-task generalization. Beyond sleep staging, Cambridge anesthesia/sedation analyses provide mechanistic evidence: the estimated persistent component generalizes to physiological states excluded from estimation, while correction improves state decodability and reduces subject predictability. On unseen multi-state DOSE-I sedation recordings, PASF improves recording-paired Macro-F1 by $+2.28$ points (95% CI, 0.65–3.90). We further transfer the principle to differential-entropy features for SEED emotion recognition: DE-PASF-Lite improves accuracy by $+3.56$ points with a DE–MLP, whereas its smaller gain with SOGNN (Li et al., 2021) is not statistically stable. Natural MI yields no consistent gains, consistent with the importance of electrode-specific spatial structure and cross-channel covariance outside PASF's shared multiplicative spectral model; approximately 88% recovery under matched synthetic frequency-response perturbations instead points to assumption mismatch. Together, these results support generalization when the representation and domain shift conform to the shared-shift assumption, with natural motor imagery marking an important boundary. Full protocols and results appear in Appendices A.14, A.15, and A.16.

Recording-level reliability. PASF improves domain-macro Macro-F1 by 4.16 and 5.81 points over Source-only with U-Sleep and the CNN–Transformer, respectively. The corresponding domain-macro negative-transfer rates are 39.72% and 40.10%; when all 740 recordings are pooled, the rates are 40.27% and 43.24%. Tent is less reliable on U-Sleep than on CNN–Transformer, consistent with reported sensitivity of model-updating TTA to source-model and normalization properties (Zhao et al., 2023; Niu et al., 2023). PASF's cross-backbone consistency does not preclude severe target-dependent losses (Appendix A.9).

5 Conclusion

PASF uses cross-state persistence to select conservative spectral corrections for state-structured physiological signals. Real-recording and leave-one-state-out analyses support a recording-specific shared component predictive across states. Matched interventions show less composition-induced false correction and better classification recovery under combined composition changes and shared spectral distortion. PASF improves performance across four sleep datasets and two backbones, while conditioning's incremental benefit varies by domain and backbone.

Limitations. Cross-state persistence does not uniquely identify acquisition nuisance; its state-conditioned interpretation requires informative aggregate responsibilities and sufficient state support. Sedation and feature-space emotion-recognition results suggest broader applicability, but inconsistent natural motor-imagery gains and negative transfer on HMC with U-Sleep limit reliability. Modalities beyond EEG remain untested. Future work should develop label-free eligibility checks and safeguards against violations of the correction assumptions.

AI use statement. Generative AI tools were used to refine the conceptual framework and hypotheses, provide feedback on the methodology and experimental design, assist with method implementation, translation of author-provided notes, and interpretation of results. They were not used to generate synthetic datasets, formulate mathematical claims, provide ingredients for proofs, write proofs, clean or reformat datasets, or conduct qualitative or thematic data analysis; proof-related and qualitative-analysis tasks were not applicable. They were also used for literature support, code editing, and manuscript organization and language editing. The authors reviewed and verified all AI-assisted work and take full responsibility for the final content.

REFERENCES

Diego Alvarez-Estevez and Roselyne M. Rijsman. Haaglanden medisch centrum sleep staging database. *PhysioNet*, 2022. doi: 10.13026/t79q-fr32. URL https://doi.org/10.13026/t79q-fr32. Version 1.1.

Mohamed S. Ameen, Joshua Jacobs, Manuel Schabus, Kerstin Hoedlmoser, and Thomas Donoghue. Temporally resolved analyses of aperiodic features track neural dynamics during sleep. *Communications Psychology*, 3(1):160, 2025. doi: 10.1038/s44271-025-00334-2. URL https://doi.org/10.1038/s44271-025-00334-2.

Aristotelis Ballas and Christos Diou. Towards domain generalization for ECG and EEG classification: Algorithms and benchmarks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):44–54, 2024. doi: 10.1109/TETCI.2023.3306253. URL https://doi.org/10.1109/TETCI.2023.3306253.

Arianna Brancaccio, Davide Tabarelli, Marco Bigica, and Daniel Baldauf. Cortical source localization of sleep-stage specific oscillatory activity. *Scientific Reports*, 10(1):6976, 2020. doi: 10.1038/s41598-020-63933-5. URL https://doi.org/10.1038/s41598-020-63933-5.

Alvise Dei Rossi, Matteo Metaldi, Michal Bechny, Irina Filchenko, Julia van der Meer, Markus H. Schmidt, Claudio L. A. Bassetti, Athina Tzovara, Francesca D. Faraci, and Luigi Fiorillo. SLEEPYLAND: Trust begins with fair evaluation of automatic sleep staging models. *npj Digital Medicine*, 9(1):55, 2026. doi: 10.1038/s41746-025-02237-2. URL https://doi.org/10.1038/s41746-025-02237-2.

Jiahao Fan, Hangyu Zhu, Xinyu Jiang, Long Meng, Chen Chen, Cong Fu, Huan Yu, Chenyun Dai, and Wei Chen. Unsupervised domain adaptation by statistics alignment for deep sleep staging networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:205–216, 2022. doi: 10.1109/TNSRE.2022.3144169. URL https://doi.org/10.1109/TNSRE.2022.3144169.

Théo Gnassounou, Rémi Flamary, and Alexandre Gramfort. Convolution monge mapping normalization for learning on sleep data. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/21718991f6acf19a42376b5c7a8668c5-Abstract-Conference.html.

Antoine Guillot and Valentin Thorey. RobustSleepNet: Transfer learning for automated sleep staging at scale. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1441–1451, 2021. doi: 10.1109/TNSRE.2021.3098968. URL https://doi.org/10.1109/TNSRE.2021.3098968.

Huan He, Owen Queen, Teddy Koker, Consuelo Cuevas, Theodoros Tsiligkaridis, and Marinka Zitnik. Domain adaptation for time series under feature and label shifts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12746–12774. PMLR, 2023. URL https://proceedings.mlr.press/v202/he23b.html.

Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A. C. Kamphuisen, and Josefien J. L. Oberyé. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928. URL https://doi.org/10.1109/10.867928.

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. doi: 10.1016/j.cmpb.2015.10.013. URL https://doi.org/10.1016/j.cmpb.2015.10.013.

Reinmar Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0450. URL https://papers.nips.cc/paper_files/paper/2022/hash/28ef7ee7cd3e03093acc39e1272411b7-Abstract-Conference.html.

Kiran Krishnamachari, See-Kiong Ng, and Chuan-Sheng Foo. Mitigating real-world distribution shifts in the fourier domain. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=lu4oAq55iK.

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

Jingcong Li, Shuqi Li, Jiahui Pan, and Fei Wang. Cross-subject EEG emotion recognition with self-organized graph neural network. *Frontiers in Neuroscience*, 15:611653, 2021. doi: 10.3389/fnins.2021.611653. URL https://doi.org/10.3389/fnins.2021.611653.

Shanglin Li, Motoaki Kawanabe, and Reinmar J. Kobler. SPDIM: Source-free unsupervised conditional and label shift adaptation in EEG. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fc6cd575b8340be6ee70beaed9ca4015-Abstract-Conference.html.

Mingyang Liu, Xinyang Chen, Yang Shu, Xiucheng Li, Weili Guan, and Liqiang Nie. Boosting transferability and discriminability for time series domain adaptation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100402–100427, 2024. doi: 10.52202/079017-3187. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b61da4f02b271cb7b5e3d538e2b78fb9-Abstract-Conference.html.

Qiao Liu and Hui Xue. Adversarial spectral kernel matching for unsupervised time series domain adaptation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2744–2750, 2021. doi: 10.24963/ijcai.2021/378. URL https://www.ijcai.org/proceedings/2021/378.

Zhi Lu, Yang Hu, and Yan Chen. SleepBand: Single-source domain generalization for sleep staging via physiologically structured spectral modeling. *arXiv preprint arXiv:2607.04851*, 2026. doi: 10.48550/arXiv.2607.04851. URL https://arxiv.org/abs/2607.04851.

Huan Luo, Na Lu, Xiaopeng Wang, and Xu Niu. Bayesian test-time adaptation via dirichlet feature projection and GMM-driven inference for motor imagery EEG decoding. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VDg6Pv4S3v.

Samaneh Nasiri and Gari D. Clifford. Attentive adversarial network for large-scale sleep staging. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pp. 457–478. PMLR, 2020. URL https://proceedings.mlr.press/v126/nasiri20a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

YongKyung Oh and Alex Bui. Multi-view contrastive learning for robust domain adaptation in medical time series analysis. In *Proceedings of the Sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 502–526. PMLR, 2025. URL https://proceedings.mlr.press/v287/oh25a.html.

Gaurav Patel, Christopher M. Sandino, Behrooz Mahasseni, Ellen Zippi, Erdrin Azemi, Ali Moin, and Juri Minxha. Efficient source-free time-series adaptation via parameter subspace disentanglement. In *International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e1fcd183ab33714a8464e4e9a20ac710-Abstract-Conference.html.

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-Sleep: Resilient high-frequency sleep staging. *npj Digital Medicine*, 4:72, 2021. doi: 10.1038/s41746-021-00440-5. URL https://doi.org/10.1038/s41746-021-00440-5.

S. M. Purcell, D. S. Manoach, C. Demanuele, B. E. Cade, S. Mariani, R. Cox, G. Panagiotaropoulou, R. Saxena, J. Q. Pan, J. W. Smoller, S. Redline, and R. Stickgold. Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource. *Nature Communications*, 8(1):15930, 2017. doi: 10.1038/ncomms15930. URL https://doi.org/10.1038/ncomms15930.

Stuart F. Quan, Barbara V. Howard, Conrad Iber, James P. Kiley, F. Javier Nieto, George T. O'Connor, David M. Rapoport, Susan Redline, John Robbins, Jonathan M. Samet, and Paul W. Wahl. The sleep heart health study: Design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997. doi: 10.1093/sleep/20.12.1077. URL https://doi.org/10.1093/sleep/20.12.1077.

Mohamed Ragab, Emadeldeen Eldele, Wee Ling Tan, Chuan-Sheng Foo, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. ADATIME: A benchmarking suite for domain adaptation on time series data. *ACM Transactions on Knowledge Discovery from Data*, 17(8):1–18, 2023. doi: 10.1145/3587937. URL https://doi.org/10.1145/3587937.

Bence Schneider, Orsolya Szalárdy, Péter P. Ujma, Péter Simor, Ferenc Gombos, Ilona Kovács, Martin Dresler, and Róbert Bódizs. Scale-free and oscillatory spectral measures of sleep stages in humans. *Frontiers in Neuroinformatics*, 16:989262, 2022. doi: 10.3389/fninf.2022.989262. URL https://doi.org/10.3389/fninf.2022.989262.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Michael Tangermann, Klaus-Robert M"uller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. M"uller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schl"ogl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055. URL https://doi.org/10.3389/fnins.2012.00055.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024. doi: 10.1609/aaai.v38i1.27779. URL https://doi.org/10.1609/aaai.v38i1.27779.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Chaehwa Yoo, Hyang Woon Lee, and Je-Won Kang. Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1273–1284, 2022. doi: 10.1109/JBHI.2021.3103614. URL https://doi.org/10.1109/JBHI.2021.3103614.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Shijian Li, Benyan Luo, Tao Li, and Gang Pan. Personalized sleep staging leveraging source-free unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14529–14537, 2025. doi: 10.1609/aaai.v39i13.33592. URL https://doi.org/10.1609/aaai.v39i13.33592.

A Appendix

A.1 Implementation and Source Artifacts

PSD and correction. PASF computes natural-log PSDs with `scipy.signal.welch`: a Hann window, 512-sample segments, 256-sample overlap, constant detrending, density scaling, and a numerical floor of 10^{-10}. At 100 Hz, the nominal 0.3–40 Hz range retains 203 bins from 0.390625 to 39.84375 Hz at 0.1953125 Hz resolution. State residuals remain on all 203 Welch bins when computing the shared component and γ. The shared curve is clipped to ± 20 dB before multiplication by γ; ϵ_γ is double-precision machine epsilon. At application time, linear interpolation uses the nearest boundary correction outside the retained Welch grid rather than forcing the response to zero.

Source artifacts. Each leave-one-domain-out fold constructs its artifact only from the three source datasets. We sample 32 source recordings per dataset and at most 24 epochs per state and recording, using artifact-sampling seed 42. Three alternating decomposition iterations estimate smooth recording nuisance curves, clipped to ± 8 dB, and update pooled epoch-level state templates. Thus, source datasets are balanced in sampled recording count, but templates are pooled epoch means rather than strict domain–state means. The four seed-42 artifacts are reused across the five source-model training seeds; reported across-seed variation therefore covers source-model training, not source-artifact sampling.

Source-template construction. Let $r = 1, \dots, R$ index the sampled source recordings, with log spectra $L_{r,i}$ and true source labels $y_{r,i}$. Initialize each $\mu_c^{(0)}$ by the pooled mean of sampled epochs labeled c. A state absent from the entire sampled source set falls back to the pooled mean of all sampled epochs. Let $B \in \mathbb{R}^{F \times 10}$ be an open-uniform cubic B-spline basis on the retained Welch grid, with six equally spaced internal knots and each endpoint repeated four times. Let D_2 denote the second-difference matrix on its ten coefficients. For iterations $k = 0, 1, 2$, estimate one source-recording offset by

$$\begin{aligned} a_r^{(k)}(f) &= \operatorname{med}_i \big[L_{r,i}(f) - \mu_{y_{r,i}}^{(k)}(f) \big], \\ \beta_r^{(k)} &= \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{10}} \|a_r^{(k)} - B\beta\|_2^2 + 10^{-3}\|\beta\|_2^2 + \|D_2\beta\|_2^2, \\ v_r^{(k)} &= \operatorname{clip}\Big(B\beta_r^{(k)}, -b_s, b_s \Big), \qquad b_s = 8\ln(10)/10. \end{aligned} \tag{9}$$

The median is frequency-wise over all sampled epochs in recording r, after subtracting each epoch's own labeled-state template. Let $N_c = \sum_r \#\{i : y_{r,i} = c\}$. Fix the additive ambiguity by centering the clipped offsets equally across source recordings, then update the templates using all corrected sampled epochs:

$$\begin{aligned} u_r^{(k)} &= v_r^{(k)} - \frac{1}{R} \sum_{r'=1}^{R} v_{r'}^{(k)}, \\ \mu_c^{(k+1)} &= \frac{1}{N_c} \sum_{r=1}^{R} \sum_{i : y_{r,i} = c} \big(L_{r,i} - u_r^{(k)} \big). \end{aligned} \tag{10}$$

An absent state retains its initialization. The centering enforces $\sum_r u_r^{(k)} = 0$; the 8-dB bound applies before centering and is not an additional bound on $u_r^{(k)}$. The deployed templates are $\mu_c =$

$\mu_c^{(3)}$. The artifact builder subsequently refits source-offset diagnostics against these final templates without updating them again; these diagnostics are not required by the deployed PASF estimator. Smoothing therefore belongs to source-template construction, whereas target residuals and their shared projection remain on the full Welch grid without a smoothing projection.

Weak or absent target states. At deployment, all five states enter the equal-state mean through frozen-model soft responsibilities. Equal weighting removes prevalence weighting across states, but does not guarantee adequate evidence for a rare or absent state: its weighted median can be determined by small probabilities assigned to epochs of other states. The final estimator does not apply a valid-state threshold or a label-free eligibility test. Its cross-state interpretation therefore assumes informative aggregate responsibilities and sufficient multi-state support; γ is a shared-energy fraction, not a certificate that every state is well represented.

Algorithm 1 PASF inference for one target recording.

Stored: frozen classifier f_{θ_s} and source templates $\mathcal{S} = \{\mu_c\}_{c\in\mathcal{C}}$.
Input: complete unlabeled recording $\mathcal{D}_t = \{x_i^t\}_{i=1}^{N_t}$.

1. Compute L_i^t and first-pass responsibilities $p_{i,c}$ with f_{θ_s}.
2. Estimate q_c, decompose them into d_{shared} and δ_c, and compute γ.
3. Form $\widetilde{d} = \gamma d_{\text{clip}}$ and apply the resulting response H_t to every epoch.
4. Run f_{θ_s} on the corrected epochs to obtain $\widehat{y}_i$.

Signal provenance. HMC, ISRUC, and SHHS were resampled to 100 Hz and converted with a zero-phase fourth-order 0.3–40 Hz Butterworth filter and a 50-Hz notch ($Q = 30$). The Sleep-EDF Fpz–Cz files were generated from the original PSG and hypnogram EDFs by resampling to 100 Hz, applying an MNE 0.3–40 Hz band-pass filter, merging stages N3 and N4, removing movement and unknown annotations, and retaining a 30-min Wake margin before the first and after the last non-Wake epoch. The original conversion exported EEG through `Raw.to_data_frame(scalings=100.0)`, which stores values as $100 \times \mathrm{V}$; the loader therefore applies a fixed factor of 10^4 to recover the equivalent μV scale. This deterministic conversion is used consistently whenever Sleep-EDF appears as a source or target domain. The four datasets retain their dataset-specific EEG derivations and acquisition histories rather than being treated as identically preprocessed recordings.

A.2 Backbone Architectures and Training Details

Common interface. Both backbones receive sequences of 35 consecutive 30-s single-channel EEG epochs sampled at 100 Hz, represented as $35 \times 1 \times 3{,}000$ samples, and output a 35×5 tensor of logits for Wake, N1, N2, N3, and REM. They nevertheless impose complementary inductive biases: U-Sleep performs dense temporal segmentation over the concatenated waveform through a fully convolutional encoder–decoder, whereas the SleepBand-based CNN–Transformer first learns an epoch-level spectral representation and then models dependencies among epochs by self-attention. Table 3 summarizes the implementations used in our experiments; neither uses pretrained weights.

Table 3: Backbone implementations used in the matched experiments. "SleepBand-based" denotes our supervised single-EEG port of the student architecture, not the original complete training system.

Property	U-Sleep backbone	SleepBand-based CNN–Transformer
Input	35×30-s single-EEG epochs	35×30-s single-EEG epochs
Sampling rate / samples per epoch	100 Hz / 3,000	100 Hz / 3,000
Output	35×5 logits	35×5 logits
Local representation	Raw-waveform 1D convolutions	Learnable constant-Q filters + 1D CNN
Sequence modeling	Fully convolutional; two kernel-5 output convolutions	One 8-head Transformer encoder layer
Main depth	12-level encoder + bottleneck + 12-level decoder	Four CNN blocks + one Transformer layer
Normalization	BatchNorm	BatchNorm in CNN; LayerNorm in Transformer
Activation	ELU; $\tanh$ in dense head	GELU
Explicit dropout	None	0.1
Trainable parameters	3,713,836	2,218,834
Pretrained weights	No	No
Mean Teacher	Not applicable	Disabled

U-Sleep backbone. Our primary backbone is a one-dimensional U-Net adapted from U-Sleep (Perslev et al., 2021), with a 12-level encoder, a bottleneck, a symmetric 12-level decoder, and an epoch-level classification head. We concatenate the epoch and time dimensions into a continuous single-channel waveform. Each encoder level applies a kernel-9 convolution, BatchNorm, ELU, and max pooling with kernel size and stride 2. The encoder widths are $(7, 9, 12, 16, 22, 31, 43, 60, 84, 118, 166, 234)$, followed by a 330-channel bottleneck. Each decoder level uses nearest-neighbor upsampling, a kernel-2 convolution, concatenation with the corresponding encoder skip representation, and a kernel-9 convolution with BatchNorm and ELU; its widths are $(234, 166, 118, 84, 60, 43, 31, 22, 16, 12, 9, 7)$. A 1×1 convolution and $\tanh$ produce a seven-channel dense representation, which is averaged over each 3,000-sample epoch. Two kernel-5 temporal convolutions then map the epoch sequence to five logits per epoch. This implementation has 3,713,836 trainable parameters and no explicit dropout. It was trained from scratch on the source datasets; we did not use the released pretrained U-Sleep model or its original multi-channel training data.

SleepBand-based CNN–Transformer backbone. To test architectural generalization beyond a U-Net-style network, we port the student architecture of SleepBand (Lu et al., 2026) to single-EEG input. For each epoch, a learnable spectral frontend applies 30 constant-Q, Morlet-like filters of length 101; their center frequencies are logarithmically initialized between 0.5 and 50 Hz, and both the center frequencies and shared quality factor are optimized. Four spectral-integration blocks with widths $(64, 128, 256, 512)$ then apply a depthwise temporal convolution, a 1×1 channel-mixing convolution, and squeeze-and-excitation-style calibration. The first block uses kernel size 49 and stride 6; the remaining blocks use kernel size 9 and stride 1. Each block is followed by BatchNorm, GELU, and max pooling with kernel size 9 and stride 2, with dropout 0.1 after the first block. Global average pooling yields one 512-dimensional embedding per epoch. After adding a learned positional embedding, one pre-normalized Transformer encoder layer with eight heads, model and feed-forward dimensions of 512, GELU, and dropout 0.1 processes the 35-epoch sequence. A final LayerNorm, a $512 \to 512$ projection, and a $512 \to 5$ classifier produce the logits. Padding masks exclude incomplete terminal positions from attention, loss, and evaluation. The model has 2,218,834 trainable parameters. We disable SleepBand's Mean Teacher and consistency-regularization options, so this experiment changes the feature and sequence architecture without introducing a semi-supervised objective.

Matched source-training protocol. For each target dataset, both backbones are trained from scratch on the other three complete source datasets; target samples and labels are excluded from source-model fitting and source-checkpoint selection. Both use 35-epoch sequences, a training stride of 10 epochs, and non-overlapping validation and test sequences with stride 35. No per-epoch or per-record z-score normalization is applied. Training examples are naturally sampled from the concatenated source datasets, while inverse-frequency class weights are computed over the pooled source labels. Both backbones use AdamW, an initial learning rate of 3×10^{-4}, weight decay 10^{-4}, a stepwise cosine schedule, and a batch size of 32 sequences. Early stopping is disabled. The fixed source-training budgets are 34, 24, 21, and 23 epochs when holding out HMC, ISRUC, SHHS, and Sleep-EDF, respectively; each fold therefore uses the same budget for the two backbones. We repeat every fold with seeds $\{7, 17, 23, 42, 1337\}$. At evaluation, invalid padded positions are discarded; accuracy, Macro-F1, BAcc, and Cohen's κ are computed per recording and then averaged within each target dataset. PASF uses one common set of inference hyperparameters across held-out domains and seeds; target labels are never used by PASF at inference.

A.3 Dataset Composition

A.4 Additional Main-Result Statistics

The five paired PASF–Source-only BAcc differences are 4.3220, 3.5266, 3.4678, 4.3503, and 3.2593 percentage points for seeds 7, 17, 23, 42, and 1337. Their mean and sample SD are 3.7852 ± 0.5128 points. The reported interval is

$$\bar{d} \pm t_{0.975,4} \frac{s_d}{\sqrt{5}} = [3.1485,\ 4.4219] \text{ points}, \qquad t_{0.975,4} = 2.776445.$$

Table 4: Target benchmark composition. Participant counts collapse repeated nights when the original identity can be recovered.

Dataset	Participants	Recordings	Notes
HMC	146	146	One recording per participant
ISRUC	111	112	Two subgroup-II recordings share one participant
SHHS	329	329	Visit-1, preselected AHI< 5 subset
Sleep-EDF	78	153	Sleep Cassette subset; mostly two nights
Total	664	740	Adaptation resets for every recording

Table 5: Additional four-domain macro results for the primary U-Sleep experiment (%, mean $\pm$ sample SD across five seeds).

Method	Accuracy	BAcc	Macro-F1	Cohen's κ
Source-only	70.99±0.72	71.77±0.41	65.10±0.55	61.22±0.65
Tent	72.41±0.36	73.72±0.31	67.11±0.34	63.14±0.43
EATA	72.22±0.37	73.55±0.33	66.94±0.35	62.90±0.44
SAR	72.31±0.37	73.64±0.32	67.03±0.34	63.02±0.44
PASF	**74.88±0.43**	**75.56±0.18**	**69.26±0.30**	**66.13±0.51**

The Source-only contrast above summarizes training-seed variation. We additionally test whether the final 20-dB PASF configuration improves over two stronger references under a participant-level estimand: CMMN, the closest input-level spectral baseline, and Tent, which has the highest four-domain mean BAcc and Macro-F1 among the evaluated model-level TTA methods for both backbones. For each method, we first average the five matched-seed metrics within each recording, average repeated nights within participant, average participants within target domain, and finally weight the four target domains equally. The paired intervals in Table 6 use 10,000 participant-cluster bootstrap replicates over all 740 recordings from 664 participants.

Table 6: Paired four-domain aggregate effects of PASF relative to strong baselines. Entries are PASF-minus-baseline changes in percentage points with pointwise 95% participant-cluster bootstrap intervals. Positive intervals excluding zero indicate a higher aggregate score for PASF.

Backbone	Paired comparison	ΔBAcc (pp), 95% CI	ΔMacro-F1 (pp), 95% CI
U-Sleep	PASF–CMMN	$+1.788$ [$+1.471$, $+2.113$]	$+2.302$ [$+1.950$, $+2.677$]
U-Sleep	PASF–Tent	$+1.824$ [$+1.462$, $+2.190$]	$+2.146$ [$+1.703$, $+2.606$]
CNN–Transformer	PASF–CMMN	$+2.243$ [$+1.926$, $+2.572$]	$+2.407$ [$+2.053$, $+2.772$]
CNN–Transformer	PASF–Tent	$+2.510$ [$+2.073$, $+2.969$]	$+2.536$ [$+1.981$, $+3.118$]

For all four comparisons, both the BAcc and Macro-F1 intervals exclude zero, supporting higher prespecified four-domain aggregate performance for PASF than for CMMN and Tent. This conclusion does not imply domain-wise dominance. For PASF–Tent Macro-F1, the U-Sleep intervals cross zero on HMC and ISRUC but are positive on SHHS and Sleep-EDF. With the CNN–Transformer, they are positive on HMC, ISRUC, and Sleep-EDF but negative on SHHS (-0.819 points; 95% CI, -1.109 to -0.528). The defensible claim is therefore that PASF is significantly better in the four-domain aggregate, not that it is consistently or significantly better on every target dataset.

These percentile intervals are pointwise and conditional on the existing checkpoints; they do not jointly resample training seeds and participants. We treat PASF–Tent as the principal strongest-TTA comparison and the CMMN contrasts as supportive rather than claiming family-wise confirmatory inference over every displayed baseline–metric contrast. If all such contrasts are instead designated as a confirmatory family, a multiplicity correction such as Holm's procedure is required.

A.5 Baseline Evaluation Protocols

Tent, EATA, and SAR operate on chronological target batches of 32 sequences. All 20 seed–target evaluations use a 35-epoch test window and a stride of 35 epochs, so adjacent sequences share no

valid epochs. A full batch therefore contains at most 1,120 unique 30-s epochs, corresponding to 9.33 h. Across the 740 target recordings, 654 (88.38%) require one batch, 82 (11.08%) require two, and four (0.54%) require three; the mean and median are 1.12 and 1 batch per recording, respectively. Thus, the parameter-adaptation baselines receive whole-recording unlabeled access for most recordings, close to the recording-level access used by PASF. Each recording starts from the source checkpoint; for the 86 longer recordings, adaptation proceeds chronologically, and predictions for the current batch are made before any later batch is accessed. All three methods use Adam, zero weight decay, target-batch normalization statistics, and BatchNorm affine updates only. Tent uses learning rate 10^{-3}. EATA uses learning rate 10^{-4}, Fisher coefficient 50, at most 2048 source examples for Fisher estimation, entropy margin $0.4 \ln 5$, redundancy margin 0.05, and probability momentum 0.9. SAR uses learning rate 10^{-3}, SAM radius 0.05, entropy margin $0.4 \ln 5$, and recovery threshold 0.2.

The negative-transfer comparison additionally includes BN Stats, which uses target-batch normalization statistics while keeping learned model parameters fixed. Like the gradient-based adaptation methods, it is reset to the corresponding source checkpoint at each recording boundary and is evaluated on the same chronological target inputs and source-model seeds.

The current EATA evaluation reloads raw source examples to estimate Fisher information at evaluation startup. PASF instead loads a compact source spectral artifact, so the two runs do not have identical deployment access despite sharing the target protocol and source checkpoints. For both backbones, PSDNorm replaces BatchNorm with a PSD-normalization module and retrains each fold for all five seeds. CMMN constructs an equal-recording Bures–Wasserstein source PSD barycenter, normalizes both source and target recordings with a 128-sample filter, and retrains the source model for every fold and seed. Its deployment requires the saved barycenter and corresponding source model, not raw source epochs.

The SF-UIDA-style rows for both backbones are mechanism-consistent single-channel ports evaluated over five source-model seeds with recording reset, five self-supervised epochs, and ten fine-tuning epochs. They are not strict reproductions of the official implementation, which uses multi-channel EEG/EOG.

A.6 Cross-State Persistence Protocol

Offline residuals and inclusion. The cross-state analyses use source templates from the seed-42 U-Sleep folds and target labels only for offline stage stratification. A recording is eligible if every true stage contains at least 20 epochs, retaining 574 of 740 recordings: 132 HMC, 104 ISRUC, 234 SHHS, and 104 Sleep-EDF. For each recording i, stage c, and draw $b \in \{1, \dots, 50\}$, we sample 20 epochs uniformly without replacement within that stage, using experiment seed 42. The diagnostic residual $q_{i,c}^{(b)}(f)$ is their frequency-wise median natural-log PSD minus $\mu_c(f)$. These label-stratified residuals are distinct from the probability-weighted deployment estimator in Equation 3. All correlations use the 203 retained Welch bins without smoothing; the 50 draws are repeated measurements of one recording.

Matched cross-state contrast. For each of the ten unordered stage pairs $p = (c, c')$ and each draw, the within-recording statistic is $w_{i,p}^{(b)} = \mathrm{corr}_f(q_{i,c}^{(b)}, q_{i,c'}^{(b)})$, where corr_f denotes Pearson correlation across frequency bins, with each curve centered by its frequency mean. The matched between-recording statistic $v_{i,p}^{(b)}$ averages the correlations of $q_{i,c}^{(b)}$ with $q_{j,c'}^{(b)}$, and of $q_{i,c'}^{(b)}$ with $q_{k,c}^{(b)}$. Partners $j \neq i$ and $k \neq i$ are sampled independently and uniformly from eligible recordings in the same domain for each pair and draw; partner reuse is allowed. We summarize within, between, and paired differences separately:

$$W_i = \mathrm{med}_p \, \mathrm{med}_b \, w_{i,p}^{(b)}, \qquad V_i = \mathrm{med}_p \, \mathrm{med}_b \, v_{i,p}^{(b)}, \qquad D_i = \mathrm{med}_p \, \mathrm{med}_b \big(w_{i,p}^{(b)} - v_{i,p}^{(b)} \big).$$

Each reported summary averages its recording values within domains and then weights the four domains equally. Median aggregation is nonlinear, so the reported mean of D_i, 0.245, need not equal $0.663 - 0.400$. Its 95% interval uses 10,000 domain-stratified recording bootstrap replicates. The one-sided paired test independently flips the sign of each recording's D_i, retains equal domain weighting, and uses 20,000 randomizations with the plus-one correction for the Monte Carlo p-value.

The four-domain macro summaries of W_i, V_i, and D_i were 0.663, 0.400, and $+0.245$, respectively; the paired contrast had a 95% CI of 0.230–0.260 and a one-sided $p < 5 \times 10^{-5}$.

Leave-one-state-out prediction. Figure 2(C) removes dataset–stage population structure before evaluating cross-state prediction. Within each draw, we subtract the frequency-wise median residual of all other eligible recordings in the same domain and stage:

$$z_{i,c}^{(b)}(f) = q_{i,c}^{(b)}(f) - \mathrm{med}_{j \neq i,\, j \in \mathcal{I}_d}\, q_{j,c}^{(b)}(f), \qquad \widehat{z}_{i,c}^{(b)}(f) = \frac{1}{4} \sum_{k \neq c} z_{i,k}^{(b)}(f).$$

Here $\mathcal{I}_d$ is the eligible set in the domain of recording i. We compute $\mathrm{corr}_f(\widehat{z}_{i,c}^{(b)}, z_{i,c}^{(b)})$, take its median across the 50 draws for each recording and held-out stage, and then average recordings within domains and domains equally. The estimate uses neither the held-out stage of recording i nor a smoothing projection. Stage-specific 95% intervals use 5,000 domain-stratified recording bootstrap replicates. These are offline diagnostics, with intervals conditional on fixed source artifacts and the sampling seed; repeated nights are not clustered by participant, and training seeds are not resampled.

Four-domain macro mean Pearson correlations (95% CIs) were 0.406 (0.376–0.437) for Wake, 0.642 (0.618–0.666) for N1, and 0.664 (0.644–0.684) for N2. The corresponding values were 0.498 (0.470–0.526) for N3 and 0.560 (0.533–0.587) for REM, giving a five-state mean of $\bar{r} = 0.554$.

Representative frequency-resolved example. For N2 in recording SN094, source-referenced residual RMS fell from 2.137 to 0.743 dB (65.2%), while local structure remained (Figure 3(D)). This illustrates the intended selectivity: reducing the common spectral discrepancy without requiring complete alignment of each state to its source template. The example supports this correction mechanism, not a population-level accuracy gain or a guarantee of physiological-feature preservation.

A.7 Controlled-Intervention Protocol

Matched methods, cohort, and contexts. Figure 2(D–F) compares final-protocol PASF with Uniform responsibilities, $p_{i,c} = 1/5$, using both seed-42 backbones, the same source templates, a 20-dB cap, and no target-side curve smoothing. Target labels determine eligibility, context composition, and evaluation only; PASF estimates residuals from frozen-model probabilities. In Wake/N1/N2/N3/REM order, Base proportions are $(0.10, 0.05, 0.50, 0.20, 0.15)$ and N3-heavy proportions are $(0.10, 0.05, 0.30, 0.40, 0.15)$. We choose the largest $K \in \{100, 120, 140, 160, 180, 200\}$ allowing both compositions to be sampled without replacement. This requires at least $K(0.10, 0.05, 0.50, 0.40, 0.15)$ epochs by stage and retains 656 recordings from 616 participants: 134 HMC, 106 ISRUC, 325 SHHS, and 91 Sleep-EDF. No record is excluded by method performance.

Three deterministic coupled draws, indexed 0/1/2, are fixed from the experiment seed, target identifier, and recording identifier. Wake, N1, and REM selections are identical within each paired Base/N3-heavy draw; the smaller N2 and N3 selections are nested in the larger selections. Both methods use these same epochs at every amplitude, without choosing a representative draw based on estimated curves or classification performance. Changing context composition leaves each selected waveform unchanged. First-pass probabilities are computed on the complete chronological recording and then restricted to the context. Each estimated filter is applied to that same complete recording for final frozen-classifier inference; the classification evaluation set is never resampled to N3-heavy.

Shared distortion. For each recording, Gaussian coefficients on the stored ten-function spline basis define a smooth profile $u_i(f)$, centered over frequency and scaled to $\max_f |u_i(f)| = 1$. A deterministic seed fixes the profile across methods, contexts, compositions, and amplitudes. We inject $a_{i,\alpha}(f) = \alpha u_i(f)$ dB for $\alpha \in \{0, 1, 3, 6\}$; the amplitude denotes peak absolute response, not RMS. After interpolation to the FFT grid, the same multiplier $10^{a_{i,\alpha}(f)/20}$ is applied to every epoch, preserving Fourier phase. Frequencies outside the retained band inherit the nearest boundary

response. Welch spectra are recomputed from these waveforms on all 203 retained bins (0.390625–39.84375 Hz). The spline basis defines the injected distortion only; it does not smooth the target estimator. Zero amplitude leaves the waveform unchanged.

Composition sensitivity and incremental response error. Let $b^{\mathrm{raw}}_{m,i,k,c,\alpha}(f)$ be the unbounded shared estimate in dB, for method m, recording i, context draw k, composition c, and amplitude α. The applied curve is $b^{\mathrm{app}} = \gamma\,\mathrm{clip}(b^{\mathrm{raw}}, -20, 20)$, with each estimate's own γ. With $\mathrm{RMS}_f(v) = [F^{-1}\sum_f v(f)^2]^{1/2}$, Figure 2(D) reports

$$S_{m,i,k} = \mathrm{RMS}_f\left(b^{\mathrm{app}}_{m,i,k,\mathrm{N3},0} - b^{\mathrm{app}}_{m,i,k,\mathrm{Base},0}\right).$$

Smaller S means less composition-induced change, not a smaller absolute correction or proof of better distortion recovery. Figure 2(E) separately reports

$$E_{m,i,k,c,\alpha} = \mathrm{RMS}_f\left(b^{\mathrm{app}}_{m,i,k,c,\alpha} - b^{\mathrm{app}}_{m,i,k,\mathrm{Base},0} - a_{i,\alpha}\right).$$

Thus Base measures distortion tracking; N3-heavy additionally includes composition confounding. These errors subtract each method's own clean Base estimate and do not identify the absolute natural recording shift. Raw-curve errors replace b^{app} by b^{raw}. A same-composition diagnostic instead subtracts $b_{m,i,k,c,0}$, isolating distortion tracking at fixed composition; it must not be substituted for the mixed total error. Supplementary machine-readable summaries also compare incremental estimates with the realized context-median Welch change to quantify finite-window effects.

Table 7: Composition-only sensitivity (dB RMS; lower is better). Values are domain-macro means with pointwise 95% participant-cluster CIs; the contrast is paired.

Backbone	Curve	PASF	Uniform	PASF−Uniform
U-Sleep	Raw	0.176 [0.169, 0.184]	0.591 [0.574, 0.608]	−0.415 [−0.433, −0.397]
U-Sleep	Applied	0.136 [0.128, 0.144]	0.372 [0.357, 0.387]	−0.236 [−0.251, −0.221]
CNN–Transformer	Raw	0.175 [0.168, 0.183]	0.591 [0.574, 0.609]	−0.416 [−0.434, −0.398]
CNN–Transformer	Applied	0.131 [0.124, 0.138]	0.372 [0.357, 0.387]	−0.241 [−0.256, −0.226]

Table 8: Incremental response error at 3 dB (RMSE in dB). Each method is referenced to its own clean Base curve; lower is better. The final column gives paired effects and pointwise 95% participant-cluster CIs.

Backbone	Composition	Curve	PASF	Uniform	Difference [95% CI]
U-Sleep	Base	Raw	0.252	0.015	+0.237 [+0.224, +0.251]
U-Sleep	Base	Applied	0.465	0.590	−0.125 [−0.143, −0.108]
U-Sleep	N3-heavy	Raw	0.306	0.591	−0.285 [−0.308, −0.263]
U-Sleep	N3-heavy	Applied	0.488	0.737	−0.248 [−0.268, −0.228]
CNN–Transformer	Base	Raw	0.377	0.015	+0.362 [+0.335, +0.391]
CNN–Transformer	Base	Applied	0.542	0.590	−0.049 [−0.074, −0.022]
CNN–Transformer	N3-heavy	Raw	0.416	0.591	−0.175 [−0.209, −0.141]
CNN–Transformer	N3-heavy	Applied	0.560	0.737	−0.177 [−0.203, −0.150]

Classification and uncertainty. Figure 2(F) reports paired PASF-minus-Uniform BAcc on the complete recording at the indicated distortion amplitude. Gains over Source-only use the same perturbed recording, not each method's clean-to-corrupted loss; they are not recovery fractions. Metrics are computed for each context first, then averaged over the three contexts within recording, recordings within domain, and the four domains equally. Pointwise 95% percentile intervals use 10,000 domain-stratified participant-cluster bootstrap replicates; repeated recordings from one participant remain together. They quantify participant sampling conditional on the seed-42 checkpoints, fixed contexts, and one distortion profile per recording, without resampling training seeds, contexts, profiles, or target domains.

Table 9: Classification gains over Source-only on the same 3-dB perturbed recordings (pp). These are gains over corrupted Source-only, not fractions of clean-to-corrupted loss recovered.

Backbone	Context	Metric	PASF [95% CI]	Uniform [95% CI]
U-Sleep	Base	BAcc	$+5.84$ $[+4.88, +6.84]$	$+5.41$ $[+4.47, +6.41]$
U-Sleep	Base	Macro-F1	$+6.58$ $[+5.56, +7.65]$	$+6.06$ $[+5.06, +7.10]$
U-Sleep	N3-heavy	BAcc	$+5.75$ $[+4.84, +6.74]$	$+3.84$ $[+2.91, +4.82]$
U-Sleep	N3-heavy	Macro-F1	$+6.56$ $[+5.58, +7.62]$	$+5.06$ $[+4.11, +6.09]$
CNN–Transformer	Base	BAcc	$+7.73$ $[+6.64, +8.88]$	$+7.20$ $[+6.13, +8.32]$
CNN–Transformer	Base	Macro-F1	$+8.20$ $[+7.01, +9.47]$	$+7.77$ $[+6.61, +9.01]$
CNN–Transformer	N3-heavy	BAcc	$+7.74$ $[+6.66, +8.88]$	$+5.86$ $[+4.77, +7.00]$
CNN–Transformer	N3-heavy	Macro-F1	$+8.25$ $[+7.07, +9.51]$	$+6.91$ $[+5.73, +8.14]$

Findings and scope. PASF reduces no-distortion applied-curve sensitivity by 63.4% with U-Sleep and 64.8% with CNN–Transformer. Uniform's raw-curve tracking at Base/3 dB is nearly exact (0.015 dB RMSE), whereas PASF's corresponding errors are 0.252/0.377 dB. Under N3-heavy/3 dB, PASF's incremental applied-response errors are 0.488/0.560 dB versus Uniform's 0.737 dB, and its paired BAcc gains are 1.914/1.883 pp. Positive domain-macro classification contrasts persist at 1 and 6 dB, but ISRUC does not share this advantage. The complete classification tables below include zero distortion, every tested amplitude, and each target domain. The intervention tests a specified shared frequency-response family and composition change, not arbitrary acquisition shifts or causal removal of nonphysiological activity.

Table 10: U-Sleep: full-recording PASF-minus-Uniform classification effects (pp) at every amplitude and composition. Intervals are pointwise 95% participant-cluster intervals, conditional on seed 42.

Target	Composition	dB	BAcc [95% CI]	Macro-F1 [95% CI]
HMC	Base	0	-0.81 $[-1.29, -0.36]$	-0.63 $[-1.15, -0.14]$
HMC	Base	1	-0.85 $[-1.32, -0.40]$	-0.65 $[-1.18, -0.16]$
HMC	Base	3	-0.57 $[-1.04, -0.12]$	-0.30 $[-0.83, +0.19]$
HMC	Base	6	-0.10 $[-0.60, +0.37]$	$+0.33$ $[-0.21, +0.85]$
HMC	N3-heavy	0	$+0.68$ $[+0.17, +1.17]$	$+0.83$ $[+0.25, +1.35]$
HMC	N3-heavy	1	$+0.74$ $[+0.26, +1.21]$	$+0.88$ $[+0.33, +1.41]$
HMC	N3-heavy	3	$+1.13$ $[+0.65, +1.61]$	$+1.39$ $[+0.85, +1.93]$
HMC	N3-heavy	6	$+1.85$ $[+1.31, +2.39]$	$+2.28$ $[+1.67, +2.88]$
ISRUC	Base	0	-0.60 $[-0.94, -0.27]$	-0.80 $[-1.19, -0.42]$
ISRUC	Base	1	-0.54 $[-0.87, -0.21]$	-0.74 $[-1.13, -0.35]$
ISRUC	Base	3	-0.56 $[-0.95, -0.17]$	-0.72 $[-1.19, -0.25]$
ISRUC	Base	6	-0.58 $[-1.08, -0.10]$	-0.64 $[-1.21, -0.08]$
ISRUC	N3-heavy	0	-0.19 $[-0.67, +0.28]$	-0.57 $[-1.16, +0.01]$
ISRUC	N3-heavy	1	-0.11 $[-0.60, +0.38]$	-0.51 $[-1.11, +0.09]$
ISRUC	N3-heavy	3	-0.24 $[-0.75, +0.30]$	-0.61 $[-1.25, +0.05]$
ISRUC	N3-heavy	6	-0.26 $[-0.89, +0.34]$	-0.55 $[-1.28, +0.18]$
SHHS	Base	0	-0.03 $[-0.23, +0.18]$	-0.34 $[-0.56, -0.13]$
SHHS	Base	1	-0.01 $[-0.22, +0.21]$	-0.33 $[-0.55, -0.13]$
SHHS	Base	3	$+0.03$ $[-0.22, +0.29]$	-0.27 $[-0.51, -0.05]$
SHHS	Base	6	$+0.13$ $[-0.21, +0.48]$	-0.10 $[-0.43, +0.20]$
SHHS	N3-heavy	0	$+1.41$ $[+1.09, +1.74]$	$+0.63$ $[+0.32, +0.93]$
SHHS	N3-heavy	1	$+1.45$ $[+1.12, +1.78]$	$+0.66$ $[+0.33, +0.97]$
SHHS	N3-heavy	3	$+1.65$ $[+1.28, +2.02]$	$+0.83$ $[+0.48, +1.17]$
SHHS	N3-heavy	6	$+1.91$ $[+1.46, +2.35]$	$+1.23$ $[+0.79, +1.64]$
Sleep-EDF	Base	0	$+2.79$ $[+2.13, +3.46]$	$+3.46$ $[+2.76, +4.17]$
Sleep-EDF	Base	1	$+2.84$ $[+2.18, +3.52]$	$+3.46$ $[+2.78, +4.19]$
Sleep-EDF	Base	3	$+2.83$ $[+2.20, +3.46]$	$+3.41$ $[+2.70, +4.14]$
Sleep-EDF	Base	6	$+2.57$ $[+1.80, +3.32]$	$+3.08$ $[+2.23, +3.94]$
Sleep-EDF	N3-heavy	0	$+4.61$ $[+3.79, +5.44]$	$+4.06$ $[+3.27, +4.85]$
Sleep-EDF	N3-heavy	1	$+4.68$ $[+3.88, +5.47]$	$+4.10$ $[+3.33, +4.87]$
Sleep-EDF	N3-heavy	3	$+5.11$ $[+4.39, +5.85]$	$+4.40$ $[+3.60, +5.19]$
Sleep-EDF	N3-heavy	6	$+5.69$ $[+4.82, +6.52]$	$+4.81$ $[+3.93, +5.70]$
Average	Base	0	$+0.34$ $[+0.12, +0.57]$	$+0.42$ $[+0.18, +0.67]$
Average	Base	1	$+0.36$ $[+0.14, +0.59]$	$+0.44$ $[+0.20, +0.68]$
Average	Base	3	$+0.43$ $[+0.20, +0.65]$	$+0.53$ $[+0.27, +0.78]$
Average	Base	6	$+0.51$ $[+0.24, +0.77]$	$+0.67$ $[+0.37, +0.96]$
Average	N3-heavy	0	$+1.63$ $[+1.35, +1.91]$	$+1.24$ $[+0.94, +1.53]$
Average	N3-heavy	1	$+1.69$ $[+1.42, +1.97]$	$+1.28$ $[+0.99, +1.57]$
Average	N3-heavy	3	$+1.91$ $[+1.64, +2.19]$	$+1.50$ $[+1.20, +1.80]$
Average	N3-heavy	6	$+2.30$ $[+1.98, +2.61]$	$+1.94$ $[+1.60, +2.28]$

Table 11: CNN–Transformer: full-recording PASF-minus-Uniform classification effects (pp) at every amplitude and composition. Intervals are pointwise 95% participant-cluster intervals, conditional on seed 42.

Target	Composition	dB	BAcc [95% CI]	Macro-F1 [95% CI]
HMC	Base	0	-0.18 $[-0.63, +0.26]$	-0.10 $[-0.58, +0.38]$
HMC	Base	1	-0.18 $[-0.64, +0.27]$	-0.11 $[-0.60, +0.38]$
HMC	Base	3	$+0.09$ $[-0.42, +0.62]$	$+0.23$ $[-0.32, +0.78]$
HMC	Base	6	$+0.21$ $[-0.43, +0.81]$	$+0.43$ $[-0.30, +1.10]$
HMC	N3-heavy	0	$+1.39$ $[+0.84, +1.93]$	$+1.43$ $[+0.85, +2.02]$
HMC	N3-heavy	1	$+1.40$ $[+0.82, +1.97]$	$+1.44$ $[+0.81, +2.05]$
HMC	N3-heavy	3	$+1.62$ $[+1.03, +2.21]$	$+1.76$ $[+1.10, +2.41]$
HMC	N3-heavy	6	$+2.11$ $[+1.35, +2.85]$	$+2.40$ $[+1.51, +3.26]$
ISRUC	Base	0	-1.20 $[-1.64, -0.77]$	-1.67 $[-2.16, -1.20]$
ISRUC	Base	1	-1.09 $[-1.51, -0.69]$	-1.60 $[-2.08, -1.15]$
ISRUC	Base	3	-0.79 $[-1.29, -0.30]$	-1.28 $[-1.86, -0.72]$
ISRUC	Base	6	-0.59 $[-1.21, +0.01]$	-1.02 $[-1.75, -0.35]$
ISRUC	N3-heavy	0	-1.07 $[-1.64, -0.51]$	-1.78 $[-2.43, -1.16]$
ISRUC	N3-heavy	1	-0.93 $[-1.49, -0.39]$	-1.69 $[-2.34, -1.07]$
ISRUC	N3-heavy	3	-0.58 $[-1.20, +0.05]$	-1.38 $[-2.11, -0.65]$
ISRUC	N3-heavy	6	-0.33 $[-1.08, +0.42]$	-1.08 $[-1.99, -0.18]$
SHHS	Base	0	$+0.51$ $[+0.27, +0.75]$	$+0.15$ $[-0.05, +0.35]$
SHHS	Base	1	$+0.37$ $[+0.10, +0.61]$	$+0.08$ $[-0.15, +0.29]$
SHHS	Base	3	$+0.33$ $[-0.08, +0.68]$	$+0.11$ $[-0.24, +0.39]$
SHHS	Base	6	$+0.40$ $[-0.06, +0.83]$	$+0.28$ $[-0.12, +0.65]$
SHHS	N3-heavy	0	$+2.11$ $[+1.74, +2.48]$	$+1.15$ $[+0.84, +1.45]$
SHHS	N3-heavy	1	$+2.04$ $[+1.66, +2.40]$	$+1.10$ $[+0.77, +1.41]$
SHHS	N3-heavy	3	$+2.16$ $[+1.68, +2.63]$	$+1.25$ $[+0.83, +1.63]$
SHHS	N3-heavy	6	$+2.56$ $[+1.98, +3.11]$	$+1.77$ $[+1.24, +2.26]$
Sleep-EDF	Base	0	$+2.38$ $[+1.66, +3.13]$	$+2.63$ $[+1.74, +3.54]$
Sleep-EDF	Base	1	$+2.39$ $[+1.60, +3.20]$	$+2.69$ $[+1.74, +3.62]$
Sleep-EDF	Base	3	$+2.49$ $[+1.63, +3.38]$	$+2.69$ $[+1.70, +3.70]$
Sleep-EDF	Base	6	$+2.14$ $[+1.11, +3.14]$	$+2.24$ $[+1.03, +3.38]$
Sleep-EDF	N3-heavy	0	$+4.11$ $[+3.16, +5.03]$	$+3.64$ $[+2.52, +4.72]$
Sleep-EDF	N3-heavy	1	$+4.11$ $[+3.14, +5.07]$	$+3.64$ $[+2.51, +4.78]$
Sleep-EDF	N3-heavy	3	$+4.33$ $[+3.25, +5.41]$	$+3.76$ $[+2.49, +5.01]$
Sleep-EDF	N3-heavy	6	$+4.57$ $[+3.31, +5.78]$	$+3.73$ $[+2.26, +5.14]$
Average	Base	0	$+0.38$ $[+0.13, +0.63]$	$+0.25$ $[-0.03, +0.54]$
Average	Base	1	$+0.37$ $[+0.12, +0.63]$	$+0.26$ $[-0.03, +0.55]$
Average	Base	3	$+0.53$ $[+0.23, +0.83]$	$+0.44$ $[+0.10, +0.77]$
Average	Base	6	$+0.54$ $[+0.18, +0.90]$	$+0.48$ $[+0.07, +0.87]$
Average	N3-heavy	0	$+1.63$ $[+1.31, +1.95]$	$+1.11$ $[+0.75, +1.46]$
Average	N3-heavy	1	$+1.65$ $[+1.32, +1.97]$	$+1.12$ $[+0.74, +1.48]$
Average	N3-heavy	3	$+1.88$ $[+1.51, +2.25]$	$+1.35$ $[+0.93, +1.76]$
Average	N3-heavy	6	$+2.23$ $[+1.80, +2.66]$	$+1.70$ $[+1.21, +2.20]$

A.8 Design Choices and Sensitivity

Table 12: Target-side residual-processing comparison under a matched 20-dB protocol (domain-macro BAcc in %, mean $\pm$ sample SD over five source-training seeds).

Target residual processing	U-Sleep	CNN–Transformer
Direct residuals	$\mathbf{75.56 \pm 0.18}$	$\mathbf{74.30 \pm 0.37}$
B-spline projection	75.23 ± 0.27	74.11 ± 0.39

Direct residual processing improves mean BAcc over B-spline projection by 0.33 percentage points with U-Sleep and 0.19 points with the CNN–Transformer. The descriptive paired 95% t-intervals are $[0.15,\ 0.50]$ and $[0.01,\ 0.37]$ points, respectively; these intervals summarize variation across the five matched training seeds.

We next compare the weighted median with a probability-weighted mean and a 10% weighted trimmed mean, which removes 10% of responsibility mass from each frequency-wise tail using fractional boundary weights. The fully matched study comprises two backbones, five source-training seeds, and four target domains (40 backbone–seed–domain jobs). Within each job, the checkpoint, source template, target recording, direct residual processing without B-spline projection, 20-dB cap, and frozen inference are fixed; only the frequency-wise target-epoch aggregator changes. Each aggregator recomputes both the state residuals and the resulting adaptive γ, and target labels are used only for final evaluation.

Table 13: Residual-aggregation ablation under the matched final PASF protocol. The four-domain BAcc is the mean ± sample SD over five matched source-training seeds; domain columns report alternative-minus-median changes in percentage points. Checkpoints, source templates, target recordings, the adaptive-strength rule, the 20-dB cap, and frozen inference are fixed; only the frequency-wise target-epoch aggregation changes.

Backbone	Aggregator	Mean BAcc (%)	ΔHMC	ΔISRUC	ΔSHHS	ΔSleep-EDF
U-Sleep	Weighted median	75.56 ± 0.18	–	–	–	–
	Weighted mean	75.22 ± 0.18	-1.16	-0.13	-0.14	+0.08
	10% trimmed mean	75.39 ± 0.18	-0.57	-0.06	-0.05	+0.03
CNN–Trans.	Weighted median	74.30 ± 0.37	–	–	–	–
	Weighted mean	74.05 ± 0.38	-0.98	-0.05	-0.12	+0.15
	10% trimmed mean	74.17 ± 0.37	-0.49	-0.04	-0.07	+0.12

Table 13 shows that the weighted median gives the highest four-domain mean BAcc for both backbones. At the seed-level estimand, replacing it with the weighted mean changes BAcc by -0.34 points for U-Sleep and -0.25 points for the CNN–Transformer; the trimmed mean narrows these differences to -0.16 and -0.12 points.

For the participant-level estimand, metrics are first averaged over the five seeds within each recording, with multiple nights from one participant kept in the same cluster. In 10,000 participant-cluster bootstrap replicates, weighted mean minus median is -0.33 points for U-Sleep (95% CI, -0.40 to -0.26) and -0.24 points for the CNN–Transformer (-0.31 to -0.17); trimmed mean minus median is -0.16 points for U-Sleep (-0.20 to -0.11) and -0.12 points for the CNN–Transformer (-0.16 to -0.07). The small differences from the seed-level contrasts reflect the different estimands rather than inconsistent results.

The domain-level changes expose the robustness trade-off hidden by the aggregate. The weighted mean improves Sleep-EDF slightly, by $+0.08$ and $+0.15$ BAcc points, but decreases HMC by -1.16 and -0.98 points and SHHS by -0.14 and -0.12 points for U-Sleep and the CNN–Transformer, respectively. The trimmed mean shows the same but attenuated pattern. Thus, the experiment does not establish pointwise dominance in every domain; it supports the median as the more conservative cross-domain aggregator because its protection against extreme epochs avoids substantially larger losses in HMC at the cost of small gains in Sleep-EDF.

The scalar shared-energy gate summarizes the fraction of residual energy shared across states at the recording level; $d_{\mathrm{shared}}(f)$ determines the frequency-dependent correction shape. It is not a calibrated confidence or reliability estimate. Relevant additional sensitivity axes for the final estimator are the maximum correction magnitude, source recordings sampled per domain, and fixed versus adaptive γ. The final method has no cosine temperature or valid-state threshold.

Fixed versus adaptive correction strength. We hold each recording's shared correction curve fixed under the final direct-residual, 20-dB protocol and compare fixed strengths $s \in \{0.25, 0.5, 0.75, 1\}$ with recording-adaptive γ. The source checkpoints, source templates, target recordings, and frozen inference are matched across conditions. Metrics follow the main-result aggregation: recording means within each target domain, followed by an equal four-domain mean for each seed. Table 14 reports mean and sample standard deviation across the five source-training seeds.

Adaptive γ achieves the highest four-domain mean for both metrics and backbones. Relative to the best tested fixed mean in each metric, it improves BAcc by 0.48 and 0.55 percentage points and Macro-F1 by 0.56 and 0.48 points for U-Sleep and the CNN–Transformer, respectively. Fixed-strength preferences depend on the metric: U-Sleep BAcc peaks at $s = 0.75$, whereas Macro-F1 peaks at $s = 1$ for both backbones; CNN–Transformer BAcc is tied at the displayed precision between $s = 0.75$ and $s = 1$. These results support a recording-adaptive refinement beyond the tested fixed shrinkage levels, rather than a benefit attributable only to avoiding full correction. The reported standard deviations describe training-seed variation; they are not paired confidence intervals for these differences.

Table 14: Fixed versus adaptive correction strength under the final 20-dB protocol. BAcc and Macro-F1 are percentages (four-domain mean ± sample SD across five seeds). Bold marks the highest mean in each column.

	U-Sleep		CNN–Transformer	
Correction strength	BAcc (%)	Macro-F1 (%)	BAcc (%)	Macro-F1 (%)
Fixed $s = 0.25$	73.49 ± 0.08	66.84 ± 0.25	70.69 ± 0.87	64.91 ± 0.93
Fixed $s = 0.5$	74.64 ± 0.18	68.01 ± 0.30	72.62 ± 0.63	66.74 ± 0.66
Fixed $s = 0.75$	75.08 ± 0.20	68.60 ± 0.30	73.75 ± 0.37	68.03 ± 0.43
Fixed $s = 1$	74.91 ± 0.18	68.70 ± 0.30	73.75 ± 0.30	68.17 ± 0.38
Adaptive γ	$\mathbf{75.56 \pm 0.18}$	$\mathbf{69.26 \pm 0.30}$	$\mathbf{74.30 \pm 0.37}$	$\mathbf{68.65 \pm 0.45}$

Cap-selection separation and post hoc sensitivity. The 20-dB cap was selected using labeled source-domain validation data and fixed before evaluation on the held-out target domain. Within each outer leave-one-dataset-out fold, the held-out target dataset was excluded from this selection. The subsequent sensitivity analysis uses the four held-out target domains as evaluation data: their labels quantify performance at alternative cap values, but those results were not used to select, revise, or transfer a cap choice between folds. Thus, source-validation-based cap selection and held-out target-domain sensitivity evaluation are distinct stages; target-domain sensitivity results are not used for hyperparameter selection.

After freezing the reported configuration, we varied only the symmetric cap while holding fixed the classifier checkpoints, source templates, label-free target estimator, and adaptive strength γ. The sweep evaluated 12, 15, 18, 19, 20, 21, 22, 24, and 30 dB and an unbounded condition, using source-training seed 42 for both backbones across all four target domains. Multiple nights from one participant were averaged before analysis, the domains were weighted equally, and pointwise 95% intervals used 10,000 participant-cluster bootstrap resamples. Target labels are used here only to measure how the already-frozen method would have behaved under counterfactual cap values. These labeled comparisons therefore provide neither a selection criterion nor independent target-domain validation, and none was used to choose, update, or justify the cap in the primary experiments.

Table 15: Post hoc sensitivity of participant-mean Macro-F1 to the symmetric correction-amplitude cap. Results use seed 42; domains are weighted equally after participant-level aggregation. Brackets give pointwise 95% participant-cluster bootstrap intervals. Bold marks the setting selected on source-domain validation data and frozen before target evaluation; it does not denote a target-selected optimum.

Cap (dB)	U-Sleep	CNN–Transformer	Two-backbone mean
12	0.6877 [0.6790, 0.6961]	0.6779 [0.6687, 0.6868]	0.6828 [0.6741, 0.6913]
15	0.6907 [0.6822, 0.6989]	0.6829 [0.6741, 0.6914]	0.6868 [0.6783, 0.6949]
18	0.6943 [0.6860, 0.7024]	0.6867 [0.6779, 0.6951]	0.6905 [0.6821, 0.6986]
19	0.6956 [0.6871, 0.7036]	0.6873 [0.6785, 0.6957]	0.6914 [0.6831, 0.6996]
20	**0.6963 [0.6878, 0.7045]**	**0.6876 [0.6788, 0.6960]**	**0.6920 [0.6836, 0.7002]**
21	0.6964 [0.6879, 0.7046]	0.6874 [0.6786, 0.6958]	0.6919 [0.6834, 0.7001]
22	0.6962 [0.6877, 0.7044]	0.6868 [0.6780, 0.6951]	0.6915 [0.6830, 0.6997]
24	0.6949 [0.6864, 0.7032]	0.6853 [0.6765, 0.6938]	0.6901 [0.6817, 0.6983]
30	0.6938 [0.6853, 0.7021]	0.6846 [0.6758, 0.6931]	0.6892 [0.6807, 0.6975]
Unbounded	0.6939 [0.6853, 0.7022]	0.6847 [0.6758, 0.6932]	0.6893 [0.6808, 0.6976]

As a descriptive result, the two-backbone mean increased from 0.6828 at 12 dB to 0.6920 at the frozen 20-dB setting, a paired difference of 0.916 percentage points (95% CI, 0.581–1.273). The difference between 21 and 20 dB was −0.005 points (−0.041 to 0.033), whereas 20 dB exceeded 18 dB by 0.147 points (0.058–0.248); performance then decreased modestly to 0.6901 at 24 dB and 0.6893 without clipping. Thus, the post hoc curve shows a broad 19–22-dB plateau around the pre-existing choice, rather than evidence that target evaluation sharply identifies 20 dB as an optimum.

The 20-versus-12-dB effect was heterogeneous across domains: averaged over the two backbones, Macro-F1 changed by −0.079 points in HMC (95% CI, −0.157 to −0.015), +3.632 in ISRUC (2.275–5.061), −0.001 in SHHS (−0.055 to 0.064), and +0.112 in Sleep-EDF (0.031–0.199). At

20 dB, approximately 11.4% of ISRUC frequency points remained clipped, compared with approximately 0–0.03% in the other domains, indicating that the wider cap mainly relaxed an active constraint in ISRUC. Participant-level NTR was 39.89% at 12 dB and 40.23% at 20 dB; the corresponding rates of losses exceeding one Macro-F1 point were 25.24% and 25.75%. Thus, the higher aggregate Macro-F1 reflects a change in effect magnitude and domain distribution, not a general reduction in harm or evidence that 20 dB is universally optimal.

Multiseed characterization at the adopted cap. With the same source checkpoints, source artifacts, target recordings, and five training seeds, increasing the direct-residual cap from 12 to the frozen 20-dB setting raises four-domain mean BAcc from 74.70% to 75.56% with U-Sleep and from 73.22% to 74.30% with the CNN–Transformer. Stable NTR remains nearly unchanged (39.79% to 39.72% and 40.17% to 40.10%, respectively). Like the seed-42 sweep, this target-label-informed comparison is descriptive: it characterizes the source-validation-selected cap on the evaluation domains but neither selects the cap nor supplies independent target-domain validation.

A.9 Negative-Transfer Protocol and Additional Results

All negative-transfer analyses use ordinary LODO predictions from the corresponding U-Sleep or CNN–Transformer backbone and the same target inputs as the main experiment, without artificial interventions or corruption. For method m, target recording i, and the common seed set $\mathcal{R}$, the five-seed paired change is

$$\overline{\Delta\mathrm{F1}}_i^{(m)} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}}\left(\mathrm{F1}_{i,r}^{(m)} - \mathrm{F1}_{i,r}^{(\mathrm{src})}\right), \qquad |\mathcal{R}| = 5.$$

A recording undergoes negative transfer when $\overline{\Delta\mathrm{F1}}_i^{(m)} < 0$, and severe negative transfer when the loss exceeds one Macro-F1 point, i.e., $\overline{\Delta\mathrm{F1}}_i^{(m)} < -1$. The former is less stringent than requiring a decrease for all five seeds. Let $\mathcal{I}_d$ denote the recording indices in target domain d, with $n_d = |\mathcal{I}_d|$ and $N = \sum_{d=1}^{4} n_d = 740$. We compute

$$\mathrm{NTR}_{\mathrm{macro}}^{(m)} = \frac{1}{4}\sum_{d=1}^{4}\frac{1}{n_d}\sum_{i\in\mathcal{I}_d}\mathbf{1}\left[\overline{\Delta\mathrm{F1}}_i^{(m)} < 0\right], \qquad \mathrm{NTR}_{\mathrm{pool}}^{(m)} = \frac{1}{N}\sum_{d=1}^{4}\sum_{i\in\mathcal{I}_d}\mathbf{1}\left[\overline{\Delta\mathrm{F1}}_i^{(m)} < 0\right].$$

The reported mean Δ Macro-F1 likewise averages recording-level paired changes within each domain and then weights the four domains equally. Source-only is identically zero under this paired definition and is excluded from method rankings. The 740 point-estimate units are recordings rather than 740 independent participants: Sleep-EDF contributes 153 nights from 78 participants, and ISRUC contains one repeated participant (Table 4). For PASF's stable NTR intervals, participants are the bootstrap clusters, so all recordings from the same participant are resampled together.

Table 16: Complete recording-level negative-transfer comparison for both backbones. Domain-macro quantities weight the four target domains equally; pooled NTR merges all 740 recordings.

	U-Sleep				CNN–Transformer			
Method	Macro NTR (%) ↓	Pooled NTR (%) ↓	Neg. rec.	Mean Δ F1 (pp) ↑	Macro NTR (%) ↓	Pooled NTR (%) ↓	Neg. rec.	Mean Δ F1 (pp) ↑
PASF	**39.72**	**40.27**	**298/740**	**+4.16**	40.10	43.24	320/740	**+5.81**
PSDNorm	47.05	48.51	359/740	+1.42	47.78	47.16	349/740	+2.92
CMMN	52.88	57.84	428/740	+1.86	53.08	58.92	436/740	+3.40
Tent	53.02	56.62	419/740	+2.01	**37.02**	**36.89**	**273/740**	+3.27
SAR	53.11	56.62	419/740	+1.93	37.28	37.97	281/740	+3.12
BN Stats	53.79	57.16	423/740	+1.83	37.61	38.24	283/740	+3.08
EATA	53.79	57.16	423/740	+1.84	37.54	38.11	282/740	+3.10

Table 17: Target-domain PASF negative transfer. Any negative transfer denotes a negative five-seed mean paired Macro-F1 change; severe negative transfer denotes a loss exceeding one Macro-F1 point. Four-domain macro values weight the target domains equally.

Backbone	Target domain	Recordings	Mean ΔF1 (pp)	Any negative transfer (%)	Severe negative transfer (%)
U-Sleep	HMC	146	−0.63	60.96	44.52
	ISRUC	112	+9.12	50.00	35.71
	SHHS	329	+0.30	45.29	22.19
	Sleep-EDF	153	+7.87	2.61	1.31
	Four-domain macro	740	+4.16	**39.72**	**25.93**
CNN–Transformer	HMC	146	+0.52	46.58	32.19
	ISRUC	112	+9.72	55.36	41.96
	SHHS	329	−0.07	57.14	28.27
	Sleep-EDF	153	+13.09	1.31	0.65
	Four-domain macro	740	+5.81	**40.10**	**25.77**

Table 18: Seedwise variability of recording-level negative transfer. Each entry is the mean ± sample SD across five source-model seeds after equal weighting of the four target domains.

	U-Sleep		CNN–Transformer	
Method	NTR (%) ↓	Δ F1 (pp) ↑	NTR (%) ↓	Δ F1 (pp) ↑
PASF	**39.76 ± 3.58**	**+4.16 ± 0.60**	39.57 ± 2.73	**+5.81 ± 0.73**
PSDNorm	46.33 ± 6.45	+1.42 ± 1.06	46.69 ± 2.10	+2.92 ± 0.95
CMMN	52.70 ± 2.06	+1.86 ± 0.63	51.70 ± 1.38	+3.40 ± 1.18
Tent	51.95 ± 5.86	+2.01 ± 0.78	**37.97 ± 3.42**	+3.27 ± 0.61
SAR	52.75 ± 5.90	+1.93 ± 0.79	39.48 ± 3.31	+3.12 ± 0.64
BN Stats	53.31 ± 5.85	+1.83 ± 0.80	39.57 ± 3.25	+3.08 ± 0.62
EATA	53.22 ± 5.83	+1.84 ± 0.80	39.32 ± 3.17	+3.10 ± 0.62

Average benefit and backbone dependence. Table 16 reveals pronounced backbone dependence among model-updating TTA methods. The domain-macro NTRs of Tent, EATA, and SAR increase from approximately 37% with the CNN–Transformer to 53% with U-Sleep. Tent shows the strongest reversal, achieving the lowest NTR with the former backbone but harming more than half of the recordings with the latter. This sensitivity is consistent with the dependence of model-updating TTA on source-model and normalization properties (Zhao et al., 2023; Niu et al., 2023). PASF is more consistent across backbones, maintaining an NTR of approximately 40% while delivering the largest mean Macro-F1 gains ($+4.16$ and $+5.81$ points). Its NTR decreases to approximately one quarter under the stricter one-point loss criterion (Table 17). These results support a more backbone-consistent reliability profile for PASF, although Tent remains preferable by NTR on the CNN–Transformer alone.

Target-dependent failure modes. PASF's aggregate reliability varies substantially across target domains (Table 17). On Sleep-EDF, its NTR falls to 2.61% and 1.31%, accompanied by large mean gains for both backbones. In contrast, U-Sleep on HMC and the CNN–Transformer on SHHS show negative mean changes with NTRs above 57%. ISRUC presents a different failure pattern: large average gains coexist with high NTR, indicating that the benefit is unevenly distributed across recordings. Thus, PASF is more consistent across the tested backbones than the model-updating baselines, but it is not uniformly reliable across targets or individual recordings. These failures motivate a target- or recording-aware abstention criterion rather than unconditional correction.

Variation across training seeds. The seedwise estimates preserve the aggregate conclusions: PASF has the largest mean gain for both backbones and the lowest mean NTR for U-Sleep, whereas Tent has the lowest mean NTR for the CNN–Transformer. Variation in PASF's gain is modest relative to its mean (4.16 ± 0.60 and 5.81 ± 0.73 points), while several U-Sleep baselines show wider NTR variation across seeds. Seedwise NTR and stable-negative NTR answer different questions—the former averages five seed-specific binary outcomes, while the latter thresholds each recording only after averaging its five paired changes—so their values need not coincide exactly.

A.10 Spectral-Interpretation Protocol

Figure 3 follows the deployed direct-residual correction pipeline for a representative recording. It shows the source state templates, the corresponding state-conditioned target residuals, their shared and retained components, and the residual before and after applying PASF. The same frequency range is used throughout the panels so that the applied correction can be compared directly with the structure selected by the estimator. This end-to-end visualization complements the aggregate evidence for cross-state persistence and controlled selectivity in Section 4.3; it is not used to assign a unique physiological or hardware cause to the estimated shift.

A.11 Oracle-State Diagnostic

Protocol. We replace the frozen model's soft state responsibilities with one-hot target labels and re-estimate the final direct-residual PASF correction under the same 20-dB cap, without B-spline projection or smoothing. The classifier remains frozen. The paired analysis uses seed 42, both backbones, four equally weighted target domains, and the 716 recordings (653 participants) containing all five sleep stages; uncertainty is estimated by 10,000 paired participant-cluster bootstrap replicates.

Finding. Oracle states improve participant-mean Macro-F1 over predicted states by 0.71 points for U-Sleep (95% CI, 0.55–0.92) and 0.63 points for the CNN–Transformer (0.46–0.82). A matched control that replaces the residual curve while retaining predicted-state γ preserves nearly all of this gain, whereas replacing γ alone has an interval spanning zero. The same conclusion holds when every stage is required to have at least ten epochs and when results are inspected by target domain. Thus, the remaining state-estimation headroom lies primarily in the frequency-resolved residual curve rather than the scalar correction strength.

Scope. Clipping remains active under the final configuration and is concentrated in ISRUC. Because the diagnostic uses target labels and a single training seed, it is neither deployable PASF nor a guaranteed performance upper bound; the bootstrap quantifies participant-sampling uncertainty conditional on the seed-42 checkpoint, not training-seed uncertainty.

A.12 Sensitivity to Responsibility Corruption and Removal of Target-State Conditioning

Five-seed matched Uniform control. We extend the seed-42 Uniform control to seeds $\{7, 17, 23, 42, 1337\}$, using the same source checkpoints as Source-only and PASF without retraining. For each backbone and seed, all three methods cover the same 740 recordings from 664 participants, yielding 22,200 paired recording–method results. Uniform sets all target responsibilities to $1/5$, while retaining source-state templates, equal-state projection, the 20-dB cap, and the adaptive-strength rule. Target curves are not smoothed. This isolates removal of target-state assignments within the retained correction construction; it is distinct from the pseudo-prevalence-referenced Marginal baseline.

BAcc first averages recordings within each target; Macro-F1 first averages repeated nights within participant and participants within target, matching the original diagnostic below. Both metrics then weight the four targets equally. Table 19 reports absolute BAcc means and sample SDs across the five seed-level aggregates. Table 20 reports paired PASF-minus-Uniform seed means and SDs; its pointwise 95% percentile intervals use 10,000 participant-cluster bootstrap replicates after averaging each recording's paired effects over seeds. BAcc retains recording weights within resampled clusters, whereas Macro-F1 weights participants equally. These intervals condition on the five checkpoints and do not jointly resample training seeds and participants.

Table 19: Matched natural-data BAcc (%, mean ± sample SD across five seeds). Recordings are averaged within target and targets equally.

Backbone	Method	HMC	ISRUC	SHHS	Sleep-EDF	Avg.
U-Sleep	Source-only	73.37 ± 0.37	65.12 ± 0.50	77.71 ± 0.06	70.89 ± 2.01	71.77 ± 0.41
U-Sleep	PASF	72.58 ± 0.56	74.45 ± 0.34	78.37 ± 0.11	76.84 ± 0.21	75.56 ± 0.18
U-Sleep	Uniform	73.51 ± 0.46	75.19 ± 0.24	78.36 ± 0.12	76.29 ± 0.36	75.84 ± 0.11
CNN–Transformer	Source-only	72.33 ± 0.27	63.46 ± 0.58	77.50 ± 0.26	60.77 ± 4.16	68.51 ± 1.12
CNN–Transformer	PASF	72.67 ± 0.40	72.83 ± 0.39	77.86 ± 0.36	73.82 ± 0.92	74.30 ± 0.37
CNN–Transformer	Uniform	72.98 ± 0.32	73.53 ± 0.29	77.87 ± 0.31	71.76 ± 1.38	74.04 ± 0.40

Table 20: Five-seed PASF-minus-Uniform effects (pp). BAcc uses recording means; Macro-F1 uses participant means. Both weight domains equally. SD describes paired seed effects; 95% CIs cluster participants after averaging seeds and are conditional on the checkpoints.

Backbone	Metric	Target	Mean	Seed SD	95% CI
U-Sleep	BAcc	HMC	-0.936	0.104	$[-1.349, -0.556]$
U-Sleep	BAcc	ISRUC	-0.743	0.139	$[-1.073, -0.421]$
U-Sleep	BAcc	SHHS	$+0.013$	0.118	$[-0.150, +0.176]$
U-Sleep	BAcc	Sleep-EDF	$+0.550$	0.297	$[+0.218, +0.873]$
U-Sleep	BAcc	Average	-0.279	0.094	$[-0.436, -0.124]$
U-Sleep	Macro-F1	HMC	-0.369	0.138	$[-0.869, +0.127]$
U-Sleep	Macro-F1	ISRUC	-0.802	0.127	$[-1.260, -0.337]$
U-Sleep	Macro-F1	SHHS	-0.130	0.167	$[-0.312, +0.032]$
U-Sleep	Macro-F1	Sleep-EDF	$+1.544$	0.310	$[+1.215, +1.875]$
U-Sleep	Macro-F1	Average	$+0.061$	0.072	$[-0.129, +0.255]$
CNN–Transformer	BAcc	HMC	-0.316	0.217	$[-0.694, +0.043]$
CNN–Transformer	BAcc	ISRUC	-0.703	0.173	$[-1.114, -0.301]$
CNN–Transformer	BAcc	SHHS	-0.009	0.135	$[-0.171, +0.152]$
CNN–Transformer	BAcc	Sleep-EDF	$+2.060$	0.682	$[+1.607, +2.523]$
CNN–Transformer	BAcc	Average	$+0.258$	0.094	$[+0.075, +0.441]$
CNN–Transformer	Macro-F1	HMC	$+0.162$	0.221	$[-0.254, +0.587]$
CNN–Transformer	Macro-F1	ISRUC	-0.970	0.209	$[-1.589, -0.355]$
CNN–Transformer	Macro-F1	SHHS	-0.074	0.155	$[-0.251, +0.090]$
CNN–Transformer	Macro-F1	Sleep-EDF	$+2.277$	0.531	$[+1.775, +2.798]$
CNN–Transformer	Macro-F1	Average	$+0.349$	0.086	$[+0.120, +0.581]$

Table 21: Individual-seed PASF-minus-Uniform effects (pp), using the metric-specific averaging rules of Table 20.

Backbone	Metric	Seed	HMC	ISRUC	SHHS	Sleep-EDF	Avg.
U-Sleep	BAcc	7	-0.804	-0.821	$+0.164$	$+0.641$	-0.205
U-Sleep	BAcc	17	-0.974	-0.950	$+0.001$	$+0.368$	-0.389
U-Sleep	BAcc	23	-0.985	-0.675	$+0.004$	$+0.496$	-0.290
U-Sleep	BAcc	42	-0.855	-0.647	-0.162	$+1.009$	-0.164
U-Sleep	BAcc	1337	-1.061	-0.624	$+0.059$	$+0.237$	-0.347
U-Sleep	F1	7	-0.140	-0.979	$+0.072$	$+1.548$	$+0.125$
U-Sleep	F1	17	-0.402	-0.891	-0.079	$+1.410$	$+0.009$
U-Sleep	F1	23	-0.395	-0.744	-0.129	$+1.425$	$+0.039$
U-Sleep	F1	42	-0.393	-0.695	-0.391	$+2.069$	$+0.147$
U-Sleep	F1	1337	-0.516	-0.699	-0.123	$+1.266$	-0.018
CNN–Transformer	BAcc	7	-0.136	-0.439	$+0.067$	$+0.991$	$+0.121$
CNN–Transformer	BAcc	17	-0.625	-0.894	-0.010	$+2.797$	$+0.317$
CNN–Transformer	BAcc	23	-0.373	-0.698	-0.162	$+2.332$	$+0.275$
CNN–Transformer	BAcc	42	-0.363	-0.816	$+0.173$	$+1.867$	$+0.215$
CNN–Transformer	BAcc	1337	-0.081	-0.668	-0.111	$+2.313$	$+0.363$
CNN–Transformer	F1	7	$+0.340$	-0.710	$+0.124$	$+1.354$	$+0.277$
CNN–Transformer	F1	17	-0.129	-1.266	-0.105	$+2.626$	$+0.282$
CNN–Transformer	F1	23	$+0.223$	-0.943	-0.264	$+2.544$	$+0.390$
CNN–Transformer	F1	42	-0.004	-1.061	$+0.036$	$+2.297$	$+0.317$
CNN–Transformer	F1	1337	$+0.379$	-0.868	-0.159	$+2.563$	$+0.479$

The four-domain PASF-minus-Uniform BAcc effects are -0.279 pp with U-Sleep and $+0.258$ pp with CNN–Transformer; Macro-F1 effects are $+0.061$ and $+0.349$ pp. Sleep-EDF benefits in both metrics for all ten backbone–seed combinations, whereas ISRUC effects are negative throughout. SHHS effects are close to zero and HMC depends on backbone and metric. Thus, target-state conditioning provides domain-dependent refinement rather than explaining most of the average gain over Source-only. Section 4.3 independently tests its role under controlled composition changes. The

temperature and responsibility-redistribution experiments below remain seed-42 diagnostics; their original values and uncertainty must not be interpreted as five-seed estimates.

Correction objective and state estimation. PASF uses cross-state persistence as a cue for selecting a candidate spectral correction; state probabilities supply the target-state evidence used to estimate it. **Uniform responsibilities remove target-state conditioning while retaining PASF's shared-correction construction and state-balanced source reference.** Under this control, PASF reduces to a state-balanced recording-level shared-shift estimator, retaining source-state templates and adaptive correction rather than becoming the ordinary Marginal baseline. Complementary evidence comes from the matched Marginal-versus-Shared-only ablation: introducing state-conditioned, equal-state shared estimation improves four-domain BAcc in every matched seed for both backbones (Table 2). The analyses below distinguish sensitivity to probability calibration, removal of target-state evidence, and the domain-dependent refinement provided by explicit state conditioning.

Motivation and protocol. The oracle-state diagnostic above tests the favorable direction in which state information improves. To test the complementary concern that PASF could fail when its frozen classifier supplies poorly calibrated or incorrect responsibilities, we perturb only the $p_{i,c}$ supplied to the final no-B-spline PASF estimator. Target signals, source state templates, residual aggregation, the 20-dB correction cap, the adaptive-strength rule, and the frozen evaluation classifier remain fixed; the model is not retrained. We evaluate seed 42 with both backbones on all 740 recordings in each backbone, comprising 664 participant clusters across the four targets. Point estimates first average repeated recordings within participant and target, then average participants within each target, and finally weight the four targets equally. All intervals use 10,000 paired participant-cluster bootstrap replicates and are conditional on the seed-42 checkpoints.

Controlled perturbations. We distinguish perturbations of responsibility calibration from disruptions of epoch-wise state identity. Temperature scaling changes confidence and entropy without changing the predicted state,

$$\widetilde{p}_{i,c}(T) = \frac{p_{i,c}^{1/T}}{\sum_{k\in\mathcal{C}} p_{i,k}^{1/T}}, \qquad T \in \{0.5, 2, 4\}. \tag{11}$$

To corrupt state identity, we symmetrically redistribute each state's mass to the other four states,

$$\widetilde{\mathbf{p}}_i(\rho) = (1-\rho)\mathbf{p}_i + \rho\mathbf{p}_i M, \qquad \rho \in \{0.25, 0.5, 0.75, 1\}, \tag{12}$$

where M has zero diagonal and off-diagonal entries $1/(|\mathcal{C}|-1)$. At $\rho = 1$, the original maximum-probability state becomes the minimum-probability state for every epoch. The implementation checks confirm zero argmax changes for every temperature condition, 100% changes for $\rho = 1$, and normalized entropy one for the uniform control (numerical error below 3×10^{-16}).

Uniform responsibilities as a structural ablation. Uniform responsibilities, $\widetilde{p}_{i,c} = 1/|\mathcal{C}|$, remove all epoch-wise target-state assignments while preserving the source-state templates, equal-state projection, and adaptive-strength rule. This is a structural ablation of the target estimator. Let the resulting target-recording marginal median be

$$m_t(f) = \mathrm{wmed}_{i=1}^{N_t}\left[L_i^t(f);\, \frac{1}{|\mathcal{C}|}\right].$$

Because every state uses the same target-epoch weights and the weighted median is translation equivariant, Equation 3 reduces to

$$\begin{aligned} q_c(f) &= m_t(f) - \mu_c(f), \\ d_{\mathrm{shared}}(f) &= m_t(f) - \overline{\mu}(f), \quad \overline{\mu}(f) = \frac{1}{|\mathcal{C}|}\sum_{k\in\mathcal{C}} \mu_k(f), \\ \delta_c(f) &= \overline{\mu}(f) - \mu_c(f). \end{aligned}$$

Thus, the only target-dependent term is the recording-level marginal spectrum m_t, referenced to the equal-state source anchor $\overline{\mu}$. The control retains a shared correction informed by source-state

structure, but its state-varying residuals δ_c are determined entirely by the source templates. Its adaptive-strength rule therefore no longer measures observed target-side cross-state disagreement. Strong performance under this control can demonstrate robustness to removal of target-state assignments, but does not independently establish cross-state persistence in the target recording or identify a nonphysiological cause for the correction.

Table 22: Responsibility perturbation and target-state removal. Macro-F1 is the participant mean followed by an equal four-domain mean (%). Parentheses report paired changes from the original PASF in percentage points with 95% participant-cluster bootstrap intervals. $\overline{H}$ is normalized responsibility entropy. Uniform responsibilities remove target-state assignments while retaining source-state templates, the equal-state shared projection, and adaptive correction; they are a structural control, not another calibration setting.

	U-Sleep			CNN–Transformer		
Condition	$\overline{H}$	Macro-F1	Δ vs. PASF	$\overline{H}$	Macro-F1	Δ vs. PASF
Source-only	–	64.92	-4.72 $[-5.72, -3.76]$	–	62.85	-5.91 $[-7.07, -4.81]$
PASF	0.285	69.63	reference	0.153	68.76	reference
$T = 0.5$	0.137	69.57	-0.07 $[-0.13, -0.01]$	0.075	68.67	-0.10 $[-0.18, -0.02]$
$T = 2$	0.544	69.70	$+0.07$ $[-0.02, +0.17]$	0.319	68.80	$+0.04$ $[-0.01, +0.09]$
$T = 4$	0.807	69.66	$+0.03$ $[-0.14, +0.21]$	0.617	68.88	$+0.12$ $[+0.01, +0.23]$
$\rho = 0.25$	0.686	69.76	$+0.13$ $[-0.02, +0.28]$	0.631	68.87	$+0.11$ $[-0.02, +0.25]$
$\rho = 0.5$	0.902	69.60	-0.04 $[-0.25, +0.17]$	0.883	68.67	-0.10 $[-0.30, +0.11]$
$\rho = 0.75$	0.997	69.49	-0.15 $[-0.39, +0.10]$	0.996	68.47	-0.29 $[-0.53, -0.05]$
$\rho = 1$	0.921	69.54	-0.09 $[-0.33, +0.15]$	0.895	68.43	-0.33 $[-0.57, -0.09]$
Uniform responsibilities (no target-state information)	1.000	69.49	-0.15 $[-0.40, +0.10]$	1.000	68.44	-0.32 $[-0.56, -0.08]$

Calibration sensitivity and retained average benefit. Temperature scaling has a modest effect: across $T \in \{0.5, 2, 4\}$, the absolute change from PASF is at most 0.12 Macro-F1 points, indicating limited sensitivity to the tested confidence changes. The separate uniform control retains most of the average benefit while removing target-state assignments: Macro-F1 changes from 69.63% to 69.49% for U-Sleep and from 68.76% to 68.44% for the CNN–Transformer, remaining 4.57 and 5.60 points above Source-only. This result concerns the retained source-state-informed correction design, not generic marginal normalization. Explicit target-state conditioning nevertheless matters on Sleep-EDF: removing it reduces Macro-F1 by 2.07 and 2.30 points, and complete responsibility inversion at $\rho = 1$ causes larger reductions of 2.57 and 3.26 points (Table 23). Balanced accuracy also cautions against universal benefit from target-state conditioning: Source-only, PASF, $\rho = 1$, and uniform obtain $71.48/75.83/76.05/75.99$ for U-Sleep and $68.62/74.39/74.09/74.17$ for the CNN–Transformer.

Matched component evidence. The five-seed component control in Table 2 supports the incremental value of state-conditioned, equal-state shared estimation. At the same fixed correction strength, replacing the pseudo-prevalence-matched Marginal estimator with Shared only improves four-domain BAcc by 0.45 points for U-Sleep and 0.52 points for the CNN–Transformer. This contrast is positive for every matched seed, ranging from $+0.38$ to $+0.53$ points for U-Sleep and from $+0.43$ to $+0.57$ points for the CNN–Transformer. It evaluates target conditioning and source-reference weighting jointly, rather than isolating either factor. Final PASF exceeds Marginal correction by 1.10 and 1.07 BAcc points, respectively. Uniform-responsibility PASF is a different control: it retains the equal-state source anchor and adaptive strength, whereas Marginal uses a pseudo-prevalence-matched source reference and fixed strength. Its strong performance therefore preserves evidence for the retained state-informed correction design without establishing equivalence to ordinary marginal normalization.

Domain-dependent protection. Table 23 shows no consistent degradation on HMC, ISRUC, or SHHS, whereas both removal and inversion of target-state information substantially reduce performance on Sleep-EDF. Target-state responsibilities therefore provide measurable additional protection in this domain. Nevertheless, even the fully redistributed variants remain 7.23 and 9.28 Macro-F1 points above their corresponding Source-only models on Sleep-EDF. The perturbations are also not numerically inert: relative to PASF, the estimated correction curve changes by approximately 1.24 dB RMS under $\rho = 1$ and 1.30–1.33 dB under uniform responsibilities. Together,

Table 23: Domain heterogeneity under the two strongest responsibility-information disruptions. Entries are paired Macro-F1 changes from the original PASF in percentage points with 95% participant-cluster bootstrap intervals.

Backbone	Target	$\rho = 1$ vs. PASF	Uniform vs. PASF
U-Sleep	HMC	$+0.77$ $[+0.28, +1.32]$	$+0.39$ $[-0.14, +0.95]$
	ISRUC	$+0.95$ $[+0.48, +1.45]$	$+0.69$ $[+0.14, +1.25]$
	SHHS	$+0.48$ $[+0.32, +0.66]$	$+0.39$ $[+0.20, +0.60]$
	Sleep-EDF	$\mathbf{-2.57}$ $[\mathbf{-3.20}, \mathbf{-1.95}]$	$\mathbf{-2.07}$ $[\mathbf{-2.65}, \mathbf{-1.50}]$
CNN–Transformer	HMC	$+0.42$ $[+0.01, +0.84]$	$+0.00$ $[-0.45, +0.46]$
	ISRUC	$+1.38$ $[+0.95, +1.83]$	$+1.06$ $[+0.53, +1.59]$
	SHHS	$+0.14$ $[-0.03, +0.32]$	-0.04 $[-0.21, +0.15]$
	Sleep-EDF	$\mathbf{-3.26}$ $[\mathbf{-3.94}, \mathbf{-2.60}]$	$\mathbf{-2.30}$ $[\mathbf{-2.92}, \mathbf{-1.72}]$

these results show that a shared correction retaining source-state structure can remain useful under substantial responsibility changes, while explicit target-state evidence refines that correction in a domain-dependent manner.

Scope. These controls assess the state estimator used to implement PASF's correction objective. Temperature perturbations support limited sensitivity to the tested calibration changes; uniform responsibilities show that the retained source-state-informed correction can preserve much of the average benefit without target-state assignments. Neither result establishes that state information is dispensable or that the uniform control directly measures target cross-state persistence. The real-recording and controlled-intervention analyses in Section 4.3 provide separate evidence for the persistence-based correction cue. Here, the matched component gains and Sleep-EDF losses establish the value and domain dependence of explicit state-conditioned estimation. Because the responsibility diagnostic uses predefined perturbations and one training seed, its intervals quantify participant sampling rather than training-seed uncertainty, and its conclusions do not extend to arbitrary responsibility errors or guarantee safe correction.

A.13 Downstream Outcomes Stratified by Shared-Shift Energy Fraction

Interpretation of γ. The coefficient γ in Equation 5 is the estimated fraction of residual-shift energy in the state-shared component targeted for correction. Under the final direct-residual protocol, the shared curve is the equal-state mean of the five residual curves, and the state-specific energy measures deviations from that mean. This energy fraction is not defined as estimator confidence, prediction reliability, or the probability of downstream improvement. Apart from the numerical stabilizer, multiplying all residual curves by a common factor leaves γ unchanged while changing their absolute magnitude. Its definition therefore implies neither monotonically increasing F1 gains nor monotonically decreasing negative transfer. Stratification assesses downstream outcomes across estimated shift compositions, rather than calibrating γ as a reliability score.

Recording-level aggregation and bins. For each backbone, we use all 740 recordings and seeds $\{7, 17, 23, 42, 1337\}$. Within each recording, we first average γ and the paired Macro-F1 difference between PASF and the same-seed Source-only model across seeds. Each recording then contributes once to one of the intervals $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, or $[0.8, 1]$. Internal boundaries belong to the right bin; the last bin includes 1. Macro-F1 follows the evaluation implementation: an unweighted class mean over the union of true and predicted classes, with zero division set to zero. For these seed-averaged observations, NTR is the fraction with $\overline{\Delta F1} < -10^{-12}$; zero changes are not counted as negative transfer. Bin means pool recordings and are not domain-macro means. Their 95% intervals use 10,000 participant-cluster bootstrap replicates stratified by target, retaining each bin's observed domain composition and fixed seed averages. Empty bins are missing, not zero; the intervals condition on the five completed training seeds.

Pooled association and domain adjustment. Larger γ accompanies larger average gains in the pooled data (Table 24), but the association depends on the target domain. We compute participant-level Spearman correlations on the 740 unbinned, seed-averaged recording observations, rather than on the five bin means. For domain-adjusted partial Spearman correlations, we remove domain fixed effects from the global ranks of both variables and correlate the residual ranks. Two-sided p-values

Table 24: Downstream outcomes by five-seed mean γ for the final no-smoothing, 20-dB PASF configuration. Each backbone contributes 740 recordings. F1 changes are paired against Source-only; NTR thresholds the seed-mean change per recording. Intervals are participant-cluster 95% bootstrap intervals.

γ bin	n	Mean ΔF1 (pp)	95% CI (pp)	NTR (%)	95% CI (%)
U-Sleep					
$[0, 0.2)$	8	$+0.131$	$[-0.040, +0.328]$	25.0	$[0.0, 60.0]$
$[0.2, 0.4)$	40	$+0.264$	$[-0.101, +0.638]$	27.5	$[14.0, 42.1]$
$[0.4, 0.6)$	101	$+0.145$	$[-0.464, +0.777]$	50.5	$[40.9, 60.0]$
$[0.6, 0.8)$	256	$+0.971$	$[+0.478, +1.452]$	42.6	$[36.7, 48.6]$
$[0.8, 1]$	335	$+4.517$	$[+3.649, +5.415]$	37.6	$[32.9, 42.2]$
CNN–Transformer					
$[0, 0.2)$	9	$+0.031$	$[-0.328, +0.496]$	66.7	$[33.3, 100.0]$
$[0.2, 0.4)$	45	$+0.004$	$[-0.358, +0.441]$	66.7	$[52.3, 80.8]$
$[0.4, 0.6)$	122	$+0.189$	$[-0.339, +0.751]$	51.6	$[43.0, 60.2]$
$[0.6, 0.8)$	256	$+1.708$	$[+1.035, +2.397]$	48.0	$[42.1, 54.1]$
$[0.8, 1]$	308	$+6.906$	$[+5.830, +8.015]$	32.1	$[27.3, 37.0]$

use the prespecified participant-level correlation procedure; Holm correction covers all 24 tests: two backbones, six scopes (pooled, domain-adjusted, and four individual domains), and two outcomes (F1 change and the recording-level negative-transfer indicator).

Table 25: Pooled and domain-adjusted participant-level associations with γ, computed on 740 recordings per backbone. I_{neg} is the binary negative-transfer indicator, not a bin-level NTR. Holm-adjusted p-values account for all 24 correlation tests, including within-domain tests.

Backbone	Scope	Outcome	ρ	Two-sided p	Holm p
U-Sleep	Pooled	ΔF1	$+0.161$	3.20×10^{-5}	5.45×10^{-4}
	Pooled	I_{neg}	-0.029	0.457	1.000
	Domain-adjusted	ΔF1	$+0.062$	0.137	0.959
	Domain-adjusted	I_{neg}	$+0.044$	0.276	1.000
CNN–Transformer	Pooled	ΔF1	$+0.230$	2.08×10^{-9}	4.79×10^{-8}
	Pooled	I_{neg}	-0.184	1.85×10^{-6}	3.52×10^{-5}
	Domain-adjusted	ΔF1	$+0.133$	0.00145	0.0217
	Domain-adjusted	I_{neg}	-0.107	0.00825	0.0907

Table 26: Within-domain participant-level Spearman $\rho(\gamma, \overline{\Delta\text{F1}})$. Signs differ across domains for both backbones. Each recording contributes its five-seed mean values.

Backbone	HMC	ISRUC	SHHS	Sleep-EDF
U-Sleep	-0.252	$+0.486$	-0.125	$+0.495$
CNN–Transformer	-0.101	$+0.474$	-0.041	$+0.347$

Evidence and limitations. The pooled F1 correlations are positive for both backbones ($\rho = 0.161$ and 0.230), and the domain-adjusted associations are smaller for U-Sleep (0.062, Holm $p = 0.959$) but remain positive and statistically detectable for the CNN–Transformer (0.133, Holm $p = 0.0217$; Table 25). The pooled negative-transfer association is not significant for U-Sleep, whereas it is negative for the CNN–Transformer after multiplicity correction; the domain-adjusted negative-transfer association does not survive Holm correction for either backbone. Within-domain F1 correlations are positive in ISRUC and Sleep-EDF and negative in HMC and SHHS for both backbones (Table 26). Even in the highest-γ bin, 37.6% of U-Sleep recordings and 32.1% of CNN–Transformer recordings undergo negative transfer. U-Sleep NTR is not monotonic across bins, and the lowest γ bins contain only 8–9 recordings per backbone. These findings support a positive pooled association between shared-shift energy fraction and average gain, but they do not establish γ as a

domain-general safety or benefit score. Benefits can also depend on absolute shift energy, spectral location, estimation error, and classifier sensitivity. Testing the energy-fraction mechanism itself would require controlled variation of shared and state-specific energy, separately from variation in total energy; such an experiment is not part of this analysis.

A.14 Generalization to Anesthesia and Sedation EEG

Motivation and applicability. We assess whether PASF's cross-state-persistence principle extends beyond sleep staging using DOSE-I sedation decoding and Cambridge anesthesia/sedation EEG analyses. PASF relies on observing state-varying physiology alongside a state-persistent recording component: multiple physiological states within the same target recording provide the contrasts used to estimate the shared residual. We therefore evaluate recordings containing multiple states with sufficient clean EEG support, consistent with the applicability condition of the cross-state estimator. The experiments assess frozen-source decoding, the role of state and context coverage, and generalization across physiological states and datasets.

DOSE-I protocol and estimands. The strict three-state primary cohort contains 12 recording-groups with Awake, Light, and Moderate/Deep sedation, each supported by at least three clean 10-s epochs. A broader cohort contains 52 recording-groups with at least two coarse sedation states under the same minimum per-state support requirement. Source-only and PASF use the same source checkpoint, and the backbone and classifier remain frozen during PASF inference; target ground-truth labels are not used to estimate or apply the correction. We report classification scores separately from the mean recording-level paired effect,

$$\overline{\Delta \mathrm{F1}} = \frac{1}{R} \sum_{r=1}^{R} \left(\mathrm{F1}_r^{\mathrm{PASF}} - \mathrm{F1}_r^{\mathrm{Source}} \right),$$

where R is the number of recording-groups. Confidence intervals for paired effects use a recording-group cluster bootstrap, rather than treating epochs as independent observations. The paired effect is a different estimand from the difference between the reported aggregate Macro-F1 scores and should not be reconstructed by subtracting those scores. Label-free correction also does not imply label-free verification of cohort eligibility.

Table 27: Frozen-source sedation decoding on unseen DOSE-I recording-groups. Classification scores are on a 0–1 scale; higher is better. The final column reports the mean recording-paired Macro-F1 change and its recording-cluster bootstrap 95% CI in percentage points (pp), separately from aggregate scores.

Cohort	Method	Macro-F1	Bal. acc.	Accuracy	Kappa	Paired ΔF1 [95% CI]
Three states ($R = 12$)	Source-only	0.525	0.556	0.574	0.363	—
	PASF	0.566	0.595	0.609	0.407	$+2.47$ $[-0.46, +5.56]$
At least two ($R = 52$)	Source-only	0.573	0.585	0.705	0.437	—
	PASF	**0.589**	**0.601**	**0.726**	**0.470**	$+2.28$ $[+0.65, +3.90]$

Effectiveness on unseen multi-state recordings. PASF improves the point estimates of all four classification metrics in both DOSE-I cohorts (Table 27). The strict primary cohort yields a positive mean paired Macro-F1 effect of 2.47 pp, but its interval crosses zero; with only 12 recording-groups, this is suggestive evidence. The broader multi-state cohort provides more precise evidence: PASF improves recording-paired Macro-F1 by 2.28 pp with a 95% CI of $[0.65, 3.90]$ pp. These results support frozen-source sedation decoding in unseen DOSE-I recordings with adequate multi-state coverage, rather than unconditional effectiveness on arbitrary target recordings.

Dependence on state coverage and EEG context. Controlled ablations on the strict three-state eligible cohort support the role of the target evidence available for persistence estimation (Table 28). Restricting estimation to one state produces a negative paired effect, whereas two and three states yield successively larger positive effects. Increasing available target context likewise moves the effect from negative at 30 s to approximately zero at 60 s and positive with longer context. The three-state/full-context ablation reports Macro-F1 0.560, separately from the primary evaluation's 0.566;

these summaries are not interchangeable. The observed trends are consistent with state diversity and sufficient clean EEG improving residual estimation, but do not establish a universal minimum duration. In particular, the eligibility threshold of 30 s per state is a minimum support rule; it should not be equated with the 30-s target-context ablation or interpreted as an optimal duration.

Table 28: State-coverage and target-context ablations on the strict three-state DOSE-I cohort. State coverage denotes the number of states available for persistence estimation. Paired changes are reported against the corresponding Source-only evaluation; uncertainty intervals are not reported for these ablations.

Ablation	Available target evidence	Macro-F1	Paired ΔF1 (pp)
State coverage	One state	0.517	-1.8
	Two states	0.549	$+1.3$
	Three states	0.560	$+2.5$
Target context	30 s	0.502	-3.3
	60 s	0.536	0.0
	120 s	0.538	$+0.3$
	Full available context	0.560	$+2.5$

Cambridge evidence for cross-state persistence. Cambridge analyses provide complementary mechanistic evidence across Baseline, Mild, Moderate, and Recovery. Within-subject cross-state residual similarity exceeds between-subject similarity: cosine similarity is 0.675 versus -0.029, with an estimated difference of 0.705 (95% CI, $[0.552, 0.859]$); Pearson similarity is 0.566 versus -0.036, with a difference of 0.602. Estimating the persistent component from three states also predicts the residual of a fourth state excluded from estimation. In Baseline, Mild, Moderate, and Recovery order, PASF's held-out residual prediction errors are $(0.368, 0.335, 0.565, 0.367)$, below both zero-component errors $(0.702, 0.763, 0.771, 0.661)$ and single-state errors $(0.466, 0.454, 0.613, 0.462)$ in every state. This supports generalization of the estimated component across physiological states within the tested protocol.

State information and subject information after correction. The Cambridge probe analyses further support selective removal of subject-associated structure. State Macro-F1 increases from approximately 0.39 to 0.48, with an estimated change of 0.0876 (95% CI, $[0.0568, 0.1192]$), while subject-ID recall decreases from approximately 0.82 to 0.35, with a change of -0.4710 (95% CI, $[-0.5255, -0.4148]$). Together with held-out-state prediction, these results are consistent with reducing subject-specific spectral structure while retaining useful sedation-state information. They do not identify the persistent component as purely nonphysiological nuisance, nor establish that all state information is preserved.

Cross-dataset generalization on multi-state recordings. Cambridge-to-DOSE-I transfer changes population, hardware, montage, recording environment, and the label definition from propofol conditions to MOAA/S responsiveness. We report a retrospective secondary analysis of 30 target recording-groups containing both Awake and Sedated states. Ground-truth labels establish subgroup membership after evaluation but do not enter the PASF correction. Within this multi-state subgroup, median aggregation increases aggregate Macro-F1 from 0.468 to 0.494 and yields a mean recording-paired improvement of 5.05 pp (95% CI, $[0.12, 10.26]$). Mean aggregation yields a consistent improvement of 4.86 pp (95% CI, $[1.10, 8.65]$). These results provide additional evidence of cross-dataset generalization within the evaluated multi-state recordings.

Generalization beyond sleep staging. Together, the sedation experiments support PASF's effectiveness beyond sleep staging when target recordings contain multiple physiological states with sufficient clean EEG support. Cambridge analyses show that the estimated persistent component generalizes to held-out states and that correction reduces subject predictability while improving state decoding. DOSE-I experiments establish improved classification with a frozen source model on unseen multi-state recordings, complemented by positive effects in the retrospective cross-dataset subgroup. This evidence connects the cross-state-persistence principle to useful adaptation in a second EEG task.

A.15 DE-PASF-LITE: PASF-INSPIRED FEATURE-SPACE ADAPTATION FOR EEG EMOTION RECOGNITION

Purpose and scope. We test whether PASF's nuisance-centric adaptation principle generalizes to a different EEG task, representation, and domain boundary. The SEED experiment replaces sleep-stage waveforms with movie-trial differential-entropy (DE) features, predicts three emotion classes, and holds out subjects rather than datasets. We call the resulting task-specific simplification *DE-PASF-Lite*. It is exactly the previously implemented `mean_alignment` branch, not an additional training run or a new generic alignment algorithm. DE-PASF-Lite retains four PASF principles: a fixed source-training reference, estimation from a complete unlabeled target adaptation unit, partial rather than mandatory correction, and a frozen target-time classifier. Unlike the waveform operator in Equation 7, however, it performs neither Fourier-domain inverse filtering nor frequency-response estimation. The experiment therefore tests transfer of the underlying adaptation principle, not algorithmic identity with waveform PASF.

Table 29: Conceptual correspondence between waveform PASF and its task-aligned DE-PASF-Lite instantiation. The estimators and operators differ; the shared object is the source-reference–persistent-shift–partial-correction design principle.

Design element	Sleep-staging waveform PASF	SEED DE-PASF-Lite
Source reference	State-balanced log-power anchor	Subject–session-balanced DE anchor
Target adaptation unit	Complete target recording	One target-subject session
Persistent residual	Across-epoch log-PSD component	Across-trial channel–band DE shift
Correction operator	Partial smooth inverse response	Partial additive DE translation
Strength control	Source-validation-selected cap; recording-adaptive γ	Source-validation-selected γ, including zero
Target predictor	Frozen sleep classifier	Frozen DE–MLP or SOGNN
Target-label boundary	Labels used only for evaluation	Labels loaded only after predictions are saved

Working model and the role of class balance. For subject u, session s, and movie trial j, let $Z_{u,s,j} \in \mathbb{R}^{62\times 5}$ denote the channel–band DE map. After the unlabeled normalization defined below, DE-PASF-Lite uses the working decomposition

$$\widetilde{Z}_{u,s,j} = M_{\mathrm{src}} + b_{u,s} + e_{y_{u,s,j}} + \epsilon_{u,s,j}, \tag{13}$$

where $b_{u,s}$ is a subject–session shared shift, $e_{y_{u,s,j}}$ is the emotion-related trial response, and $\epsilon_{u,s,j}$ collects trial-specific variation and estimation noise. Every SEED session contains five negative, five neutral, and five positive trials. Hence its class contribution averages to the same balanced marginal, $15^{-1}\sum_j e_{y_{u,s,j}} = 3^{-1}\sum_{k=1}^{3} e_k$, which is also represented in a source anchor built from complete sessions. Under this balanced design, the target session's mean residual is therefore a plausible estimator of $b_{u,s}$. This interpretation is design-dependent: class imbalance, missing trials, or natural continuous emotion streams can mix genuine prevalence changes into the session mean.

DE construction and unlabeled normalization. The audited cache contains 15 subjects, three sessions per subject, 15 trials per session, and 675 trials in total. A trial is represented as $X_{u,s,j} \in \mathbb{R}^{62\times 5\times T_j}$, padded to 265 time steps for batching; only its T_j valid one-second windows enter the trial mean $Z_{u,s,j} = T_j^{-1}\sum_{t=1}^{T_j} X_{u,s,j,:,:,t}$. DE uses 200-Hz signals, non-overlapping one-second periodic-Hann windows, a 512-point FFT, and the five bands 1–3, 4–7, 8–13, 14–30, and 31–50 Hz, with $\mathrm{DE}_b = \frac{1}{2}\ln(2\pi e P_b)$ for band-mean power P_b. These locally reconstructed features do not use moving-average or linear-dynamical-system smoothing; accordingly, the SOGNN result is not presented as a strict reproduction of the published feature pipeline.

Every subject is standardized independently without labels. For DE–MLP, each channel–band coordinate is standardized across that subject's 45 trial means, $\widetilde{Z}_{u,s,j,d} = (Z_{u,s,j,d} - \mu_{u,d})/\max(\sigma_{u,d}, 10^{-6})$. SOGNN analogously standardizes each channel–band–time coordinate across the subject's 45 padded sequences. All anchors, shifts, and classifications below are defined in this standardized space. Target-wide statistics are therefore accessed, but target emotion labels are not; "Source-only" denotes a frozen source classifier with unlabeled subject-wise target normalization.

Source anchor and target-session correction. For source-training subjects i, the fixed reference gives equal weight to subjects, sessions, and trials,

$$M_{\mathrm{src}} = \frac{1}{N_{\mathrm{src}}}\sum_i \frac{1}{S_i}\sum_s \frac{1}{N_{i,s}}\sum_j \widetilde{Z}_{i,s,j}. \tag{14}$$

For an unlabeled target subject t, the session-specific shift and corrected trial are

$$D_{t,s} = \frac{1}{N_{t,s}} \sum_{j} \big(\widetilde{Z}_{t,s,j} - M_{\mathrm{src}}\big), \qquad \widetilde{Z}'_{t,s,j} = \widetilde{Z}_{t,s,j} - \gamma D_{t,s}. \tag{15}$$

The same vector is applied to all trials in a session, so pairwise trial differences are preserved exactly: $\widetilde{Z}'_{t,s,i} - \widetilde{Z}'_{t,s,j} = \widetilde{Z}_{t,s,i} - \widetilde{Z}_{t,s,j}$. If fewer than three trials are available, the implementation sets $D_{t,s} = 0$; no SEED session triggers this fallback. For SOGNN, $D_{t,s}$ is estimated from valid-time trial means and broadcast over the full standardized sequence, $\widetilde{X}'_{t,s,j,c,b,\tau} = \widetilde{X}_{t,s,j,c,b,\tau} - \gamma D_{t,s,c,b}$. Neither estimator uses target labels, pseudo-labels, target probabilities, or confidence weights.

Protocol and backbones. Each leave-one-subject-out fold assigns 12 subjects to source training, two whole subjects to source validation, and one whole subject to target testing; all sessions from a subject remain in one role. The DE–MLP applies a $310 \to 64 \to 3$ network with ReLU and dropout 0.1, trained with Adam at 10^{-3}, weight decay 10^{-4}, batch size 128, and validation patience 25 for at most 200 epochs. SOGNN retains each $62 \times 5 \times 265$ sequence and uses the convolutional, self-organized graph, graph-convolution, and classification architecture of Li et al. (2021); the adapted public network has 386,787 parameters, top-$k = 10$, and dropout 0.1, and is trained with Adam at 10^{-5}, weight decay 10^{-4}, batch size 16, and patience 10 for at most 80 epochs. Both use one prescribed deterministic fold-level seed schedule, $42 +$ target-subject ID, and select checkpoints only on source validation.

Source-only model selection and conservative correction. The correction strength is selected before target inference using only the two source-validation subjects: candidates are ranked by validation Macro-F1, then accuracy, then the smaller γ. DE–MLP averages validation metrics across the two held-out validation subjects and searches $\{0, 0.1, 0.25, 0.5, 0.75, 1\}$; the selected counts across 15 folds are 1, 0, 5, 2, 4, and 3. The archived SOGNN implementation instead computes Macro-F1 on the 90 pooled validation trials and searches $\{0, 0.25, 0.5, 0.75, 1\}$, yielding counts 9, 1, 2, 1, and 2. Equal trial counts make pooled and subject-averaged accuracy equivalent here, but not Macro-F1; this implementation asymmetry limits a strict cross-backbone comparison. DE-PASF-Lite nevertheless actively corrects 14 of 15 DE–MLP folds and abstains with $\gamma = 0$ in 9 of 15 SOGNN folds. The no-op candidate is prespecified and limits negative transfer without consulting target inputs or labels during hyperparameter selection.

Target inference, access boundary, and complexity. For each target subject, the procedure first computes subject-wise statistics from all 45 unlabeled trials and saves Source-only probabilities. It then estimates one shift from the 15 unlabeled trials in each session, applies the source-selected γ, saves corrected probabilities from the same frozen classifier, and only then loads labels for evaluation. Model parameters, optimizer state, and normalization-layer state are never updated. The method is therefore an offline transductive input/feature correction, not an online or target-distribution-free Source-only procedure. For a session with N trials, C channels, and B bands, estimating and storing the shift costs $O(NCB)$ time and $O(CB)$ memory; broadcasting it over an SOGNN sequence of length T costs $O(NCBT)$. Target inference requires no backpropagation, covariance inversion, singular-value decomposition, iterative optimization, or pseudo-label update.

Table 30: SEED leave-one-subject-out results under one prescribed fold-level seed schedule rooted at 42. Accuracy is mean $\pm$ sample SD across 15 target subjects; Macro-F1 is averaged across subjects. ΔAcc is relative to the matched Source-only backbone, and the last column counts subjects whose trial accuracy improves, ties, or decreases. Full RP-PASF-T is retained only as an extended ablation. Bold marks the best mean within each backbone.

Backbone	Method	Accuracy (%)	Macro-F1	ΔAcc (pp)	Subjects $\uparrow$ / $=$ / $\downarrow$
DE–MLP	Source-only	74.37 ± 9.68	0.7417	—	—
	DE-PASF-Lite	$\mathbf{77.93 \pm 9.29}$	**0.7765**	$+3.56$	9/3/3
	Full RP-PASF-T	77.19 ± 9.52	0.7697	$+2.81$	8/6/1
SOGNN	Source-only	84.15 ± 6.82	0.8406	—	—
	DE-PASF-Lite	$\mathbf{84.74 \pm 7.36}$	**0.8470**	$+0.59$	4/9/2
	Full RP-PASF-T	$\mathbf{84.74 \pm 7.22}$	0.8467	$+0.59$	5/9/1

Cross-task and cross-backbone results. Table 30 shows a clear DE–MLP improvement: DE-PASF-Lite increases accuracy by 3.56 pp, with a 10,000-replicate paired subject-bootstrap 95% interval of $[+1.04, +6.07]$ pp, and increases Macro-F1 by 3.48 pp, with interval $[+0.94, +6.06]$ pp. Nine subjects improve, three tie, and three decrease. The stronger SOGNN starts 9.78 pp higher and changes by only $+0.59$ pp in accuracy (95% interval $[-0.59, +1.78]$ pp) and $+0.64$ pp in Macro-F1 (95% interval $[-0.58, +1.86]$ pp); both intervals cross

zero, and nine folds select no correction. These intervals describe paired target-subject variation conditional on the trained checkpoints; overlapping source sets make the 15 outer folds dependent, and the bootstrap does not replace multi-seed replication. Thus the DE–MLP result supplies positive evidence that the feature-space correction transfers to emotion recognition, whereas the present SOGNN result establishes plug-in compatibility but not a statistically stable gain on a stronger backbone.

Why the lite variant is the appropriate main comparison. The full RP-PASF-T extension adds 27 left–right relation constraints, a source-derived discriminant subspace, coordinate-wise sign-persistence weighting, and reprojection. These components do not improve the average result: DE-PASF-Lite exceeds full RP-PASF-T by 0.74 pp in accuracy and 0.68 pp in Macro-F1 with DE–MLP, while SOGNN accuracy is identical and DE-PASF-Lite is 0.0003 higher in Macro-F1. Within the evaluated protocol, the gain is therefore attributable to the conservative shared session-shift correction rather than to relation protection or persistence weighting. This finding motivates treating the full construction as a negative complexity ablation, rather than presenting it as the primary emotion-recognition method.

Table 31: Representation-interface control on the frozen SOGNN pipeline under the same prescribed fold-level seed schedule. Direct waveform-space ports are not plug-compatible with the pretrained DE pipeline.

Input/correction interface	Accuracy (%)
Source-only DE	84.15
Waveform PASF, fixed original normalization	50.22
Waveform PASF, refitted normalization	66.37
DE-PASF-Lite	**84.74**

Representation control, verification, and boundary. A direct waveform-space PASF port performs poorly when connected to the frozen SOGNN DE pipeline (Table 31); refitting normalization recovers part, but not all, of the loss. This negative control indicates that transfer requires matching the correction to the downstream representation and normalization interface, rather than applying an unchanged waveform operator mechanically. The archived audit covers all 15 target subjects and independently reconstructs 5,400 trial predictions; the maximum probability discrepancy is 2.98×10^{-8}, and checks of source standardization, reference construction, hyperparameter grids, label isolation, and frozen-classifier inference pass. Nevertheless, the results use only one prescribed fold-level seed schedule rooted at 42, SOGNN uses a different validation aggregation rule, its preprocessing reconstructs unsmoothed DE sequences rather than exactly reproducing the original publication pipeline, and target-wide normalization makes the experiment offline transductive. Moreover, DE-PASF-Lite was specialized after observing an interface mismatch on SEED, so this is exploratory cross-task evidence rather than a fully external confirmatory validation. Accordingly, this appendix supports a bounded claim: PASF's unlabeled shared-shift correction principle can be re-expressed in DE space and yields a reliable gain for the evaluated DE–MLP setting, but stronger cross-backbone, cross-seed, and cross-dataset efficacy requires further replication.

A.16 Generalization Boundary on Motor-Imagery EEG

Purpose and setup. We evaluate whether PASF extends to motor-imagery (MI) brain–computer interfaces using BCI Competition IV datasets 2a and 2b (Tangermann et al., 2012). Dataset 2a contains nine subjects, 22 EEG channels, and four classes (left hand, right hand, both feet, and tongue); dataset 2b contains nine subjects, three EEG channels, and two classes (left and right hand). The primary protocol is leave-one-subject-out evaluation in which the target subject is excluded from classifier training and model selection. A standard EEGNet-8,2 backbone (Lawhern et al., 2018) obtains subject-macro accuracies of $58.83\% \pm 15.65\%$ on 2a and $76.91\% \pm 8.57\%$ on 2b. An alternative UnifiedEEGNet implementation obtains $60.73\% \pm 11.82\%$ and $75.04\% \pm 7.76\%$, respectively. The large 2a dispersion reflects substantial subject heterogeneity even before adaptation.

Natural cross-subject shifts. With frozen EEGNet-8,2, a single shared frequency-correction curve does not improve natural LOSO performance (Table 32). Accuracy decreases by 0.81 pp on 2a and 1.25 pp on 2b, while Macro-F1 decreases by 1.54 and 1.33 pp. On 2a, five of nine subjects improve, but subject A08 gains 8.68 pp while A09 loses 15.63 pp; on 2b, only two subjects improve and seven decline. Subject-bootstrap 95% intervals for the accuracy changes are $[-5.19, 2.76]$ pp and $[-2.52, -0.12]$ pp. Estimating a separate correction curve per channel reduces the mean losses

Table 32: Supplementary PASF results for motor-imagery EEG. Accuracies weight the nine target subjects equally; parentheses give the change from the matched Source-only model. For synthetic shifts, Source-only denotes evaluation after corruption, and the clean accuracies before corruption are 58.83% on 2a and 76.91% on 2b. "Re-z" applies trial- and channel-wise standardization after correction.

Evaluation	Dataset / direction	PASF configuration	Source-only (%)	Adapted (%; Δ pp)
Natural LOSO	BCIC-IV-2a	Shared curve	58.83	58.02 (-0.81)
	BCIC-IV-2a	Channel-wise	58.83	58.80 (-0.04)
	BCIC-IV-2b	Shared curve	76.91	75.66 (-1.25)
	BCIC-IV-2b	Channel-wise	76.91	75.84 (-1.07)
Spatially protected LOSO	BCIC-IV-2a	Shared PASF	60.725	60.108 (-0.617)
	BCIC-IV-2a	Full SP-PASF	60.725	60.745 ($+0.019$)
	BCIC-IV-2a	Inner-selected SP-PASF	60.725	60.648 (-0.077)
Cross-dataset	2a→2b	Shared curve	59.33	58.75 (-0.58)
	2b→2a	Shared curve	61.38	62.38 ($+1.00$)
	2a→2b	Channel-wise, $\alpha = 0.25$	59.33	59.53 ($+0.20$)
	2b→2a	Channel-wise, $\alpha = 0.25$	61.38	61.81 ($+0.42$)
Synthetic 4-dB RMS shift	BCIC-IV-2a	PASF	32.06	55.59 ($+23.53$)
	BCIC-IV-2a	PASF + re-z	32.06	57.94 ($+25.88$)
	BCIC-IV-2b	PASF	63.40	75.37 ($+11.96$)
	BCIC-IV-2b	PASF + re-z	63.40	76.70 ($+13.30$)

to 0.04 and 1.07 pp, but does not make the effects reliably positive: the corresponding intervals are $[-6.66, 5.94]$ and $[-2.32, 0.15]$ pp. Moreover, on 2a, a 15.97-pp improvement for A05 coexists with a 20.66-pp loss for A09. Additional channel freedom can therefore reduce the average bias while increasing subject-specific risk.

Spatial protection and abstention. On 2a with a frozen UnifiedEEGNet, Source-only accuracy is 60.725%, shared PASF obtains 60.108%, and Full SP-PASF, which protects selected spatially discriminative structure, obtains 60.745%. The Full SP-PASF change is only $+0.019$ pp with a subject-bootstrap interval of $[-0.772, 0.887]$ pp; its Macro-F1 is 59.066%, slightly below Source-only at 59.080%, and four of nine subjects decline. A source-inner pseudo-target selector then chooses among 72 configurations and a no-adaptation option without target labels. It selects no adaptation in six of nine outer folds and yields 60.648%, or -0.077 pp relative to Source-only. Although this selector does not improve the mean, it reduces the number of harmed subjects to two and bounds the worst loss at 0.347 pp, illustrating the benefit and cost of conservative abstention.

Natural cross-dataset shifts. We next reduce both datasets to C3/Cz/C4 and left- versus right-hand imagery, and transfer in both directions. Shared PASF decreases accuracy by 0.58 pp for 2a→2b but increases it by 1.00 pp for 2b→2a, where seven of nine target subjects improve; both uncertainty intervals include zero. Source-validation selects a weak channel-wise correction, $\alpha = 0.25$, in both directions. This produces small same-direction changes of $+0.20$ and $+0.42$ pp, with 95% intervals of $[-0.51, 0.88]$ and $[-0.69, 1.43]$ pp. The cross-dataset experiment therefore gives a weak positive signal for shrunk channel-wise correction, not evidence of reliable improvement.

Recovery under matched synthetic frequency-response shifts. The same estimator succeeds when the perturbation matches its assumed mechanism. We independently apply a smooth 4-dB RMS frequency response to each channel. Accuracy falls from 58.83% to 32.06% on 2a and from 76.91% to 63.40% on 2b. PASF recovers accuracy to 55.59% and 75.37%, restoring 87.90% and 88.58% of the aggregate loss; trial- and channel-wise re-standardization raises the recovered accuracies to 57.94% and 76.70%, or 96.66% and 98.44% loss recovery. PASF improves over the corrupted Source-only model for all nine target subjects in both datasets. Thus, the negative natural-shift results do not indicate an inability to estimate and invert smooth frequency responses; they indicate a mismatch between that mechanism and important components of natural MI shift.

A.16.1 Why Natural Motor-Imagery Shifts Violate the PASF Assumptions

Spatial reference and covariance structure. Natural MI shifts are not limited to smooth multiplicative frequency responses. Even after restricting both datasets to C3/Cz/C4, 2a uses monopolar recordings whereas 2b uses local bipolar derivations; amplitude-only per-channel correction cannot reconstruct changes in spatial mixing or cross-channel covariance. Representation probes support the importance of this structure. Preserving channel identity in band-power features improves accuracy over channel-averaged power by 6.02 pp on 2a and 13.67 pp on 2b, while randomly permuting electrode identity within each trial decreases accuracy by 6.91 and 13.57 pp. On 2a, band-specific log-covariance features add another 3.22 pp over channel-wise band power. These findings explain why reducing marginal spectral discrepancy can simultaneously damage MI-discriminative spatial–spectral relations.

Target correction exceeds source-supported nuisance directions. Across the LOSO folds, approximately three source nuisance principal components explain 90% of source subject/session variation, but only 64.5% of the target persistent-residual energy lies in that source nuisance subspace. The remaining 35.5% lacks corresponding source-domain evidence and may include target-specific neurophysiology, reference mismatch, or responsibility error. Moreover, 95.7% of source nuisance energy is channel shared and only 4.3% is channel differential. Fully releasing the channel dimension therefore gives the adapter substantially more freedom than the source data can support, consistent with the high variance of channel-wise correction.

Protected templates do not match the frozen network's decision directions. The SVD subspace constructed from source class-mean spectra is nearly orthogonal to the UnifiedEEGNet spatial–frequency gradient-sensitivity subspace: the mean principal angle is 79.65°, with mean minimum and maximum angles of 74.57° and 83.57°. Correspondingly, adding SVD protection to the Mask branch decreases accuracy from 59.510% to 58.507%. By contrast, an ERD/ERS discriminant map has Spearman correlation 0.510 with the model-gradient map, and their top-10% channel–frequency units overlap by 53.6%. Mathematical preservation of a class-template subspace therefore does not guarantee preservation of the nonlinear classifier's actual decision features; physiology- and model-sensitivity-aware protection is better aligned with the observed discriminant structure.

Persistence is not an adaptation-safety score. Mean session-level γ is 0.9734 on 2a and 0.9876 on 2b, and the channel-wise version remains approximately 0.968, so the original gate applies nearly full correction to most natural MI sessions. SP-PASF residuals are reproducible (split-half correlation 0.976; mean persistence $\rho = 0.952$), yet the exploratory Spearman association between ρ and accuracy gain is -0.469. Correction RMS, maximum magnitude, and the C3/C4 curve difference also correlate negatively with gain (Pearson r approximately -0.36, -0.46, and -0.51). These correlations involve only nine subjects and are descriptive, but consistently show that a stable or cross-class-consistent residual need not be safe to remove.

Responsibility error is not the primary bottleneck. Replacing frozen-model soft responsibilities with true target classes raises channel-wise PASF accuracy only from 58.80% to 58.93% on 2a and from 75.84% to 76.48% on 2b. A09 remains at 49.31%, far below its 69.97% Source-only accuracy. On source validation, the soft-assignment residual RMS is 0.1430 dB versus 0.1339 dB with true labels, a ratio of 1.067. Responsibility error modestly inflates the residual but does not explain the major subject-level failures; improving pseudo-label quality alone is unlikely to resolve them.

Supported conclusion and deployment boundary. PASF reliably reverses controlled, class-independent smooth frequency-response shifts in MI EEG, but it does not provide stable average gains under the evaluated natural cross-subject or cross-dataset shifts. The discrepancy is consistent with natural MI variation containing reference, spatial-covariance, subject-physiology, and protocol components outside the multiplicative spectral-shift model. The results motivate restricting correction to source-observed nuisance directions, shrinking channel-differential components, protecting regions supported jointly by physiology and frozen-model sensitivity, bounding correction magnitude to a source-calibrated range, and adding an explicit no-adaptation gate. Until these changes are validated across training seeds and external MI datasets, Source-only should remain the default for natural MI deployment.