TARGET-TO-SOURCE ALIGNMENT: A HIERARCHICAL GEOMETRIC FRAMEWORK WITH STRUCTURAL CONSTRAINTS FOR FEW-SHOT EEG ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot cross-subject EEG seizure prediction faces severe data heterogeneity and optimization instability. While prevalent "symmetric alignment" methods typically seek a compromised shared subspace, they often falter when domain discrepancies are vast, leading to mutual interference and negative transfer. To overcome this, we abandon the pursuit of a middle ground and advocate for a "target-to-source alignment" strategy that explicitly maps target distributions onto the rigid source manifold. To regulate the optimization dynamics of this asymmetric mapping, we present H-GSC (Hierarchical Geometric Framework with Structural Constraints). Specifically, H-GSC employs Feature-Space Guidance (FSG) as a geometric preconditioner to reorient adaptation gradients, and Hierarchical Gradient Alignment (H-GA) to filter destructive interference by strictly prioritizing source discriminability. To prevent manifold collapse, we preserve intrinsic manifold structure via a dual-constraint regularization scheme (synergizing Masked Structural Consistency with semantic anchors) and ensure robust early stopping with a Metric-Decoupled Validator. Theoretical analysis confirms that H-GSC aligns with generalization bound minimization. Extensive experiments on CHB-MIT demonstrate state-of-the-art performance (79.54% AUC). Crucially, in rigorous joint cross-dataset scenarios where prior methods suffer from negative transfer, H-GSC achieves a significant 9.65 pp AUC gain, validating that our rectified trajectory effectively bridges vast distributional shifts for scalable clinical deployment.

1 INTRODUCTION

Electroencephalography (EEG) serves as the cornerstone of non-invasive brain monitoring (Rivera et al., 2022), underpinning the widespread application of diverse Brain-Computer Interfaces (BCI), ranging from clinical tasks like seizure prediction (Koutsouvelis et al., 2024) and sleep staging to frontier fields such as affective computing (Shen et al., 2025). Despite its immense potential, the practical deployment of deep learning models faces severe challenges. Constrained by the inherent non-stationarity of neural signals and extreme inter-subject heterogeneity (e.g., variations in electrode impedance, brain geometry, and physiological states), significant distributional shifts occur, severely impeding the model's ability to generalize to unseen individuals (Wu et al., 2020; Saha & Baumert, 2020; Samek et al., 2012). While collecting and labeling massive datasets for every new user is theoretically feasible, the practical cost is prohibitively high. Consequently, recent research focus has shifted towards Domain Adaptation (DA) (Li et al., 2024a), Domain Generalization (DG) (Zhou et al., 2022), and emerging Foundation Models (Chen et al., 2025; Zhou et al., 2025), aiming to bridge this generalization gap and achieve robust cross-subject applications.

Despite the notable progress achieved by these approaches, they face significant challenges in balancing efficacy and efficiency. Traditional DA (Du et al., 2024) and DG (Yao et al., 2023) methods predominantly rely on the "common feature extraction" philosophy, assuming that a shared latent subspace invariant to subject-specific variations can be distilled from limited data. However, in the presence of severe EEG heterogeneity, the intersection of source and target manifolds is often negligible. Pursuing a common space in such cases may lead to feature collapse or negative transfer (Zhang et al., 2022), where discriminative source structures are compromised to accommodate the target. Foundation Models (Wang et al., 2024a) represent a substantial leap forward, conceptually

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.

amalgamating the DG and DA paradigms at scale: they leverage massive multi-task datasets to learn universal representations (DG) and subsequently adapt to downstream tasks via few-shot fine-tuning (DA). While powerful, FMs are predicated on the existence of universal commonalities across vast populations, and their high demand for computational resources poses considerable hurdles for deployment on portable clinical devices. This creates a pressing dilemma: existing methods often face a trade-off between maintaining robustness against severe domain shifts and reducing reliance on extensive data and computational resources.

To address this, we propose H-GSC, a Hierarchical Geometric framework with Structural Constraints. This framework departs from the traditional paradigm of seeking a compromised shared space, advocating instead for actively rectifying the optimization trajectory to explicitly align the target distribution onto the well-established source manifold, thereby successfully overcoming the limitations of recent EEG foundation models, which underperform in seizure prediction due to their inability to handle highly individualized heterogeneity, and achieving superior performance in this highly challenging cross-patient task (Guo et al., 2025). Specifically, we revisit Feature-Space Guidance (Zhang et al., 2026) as a geometric pre-conditioner that pulls target features towards the source convex hull, injecting unconstrained yet constructive geometric momentum to break the optimization stagnation typical of naive alignment. To resolve the conflicts induced by this unconstrained momentum and prevent geometric overshoot, we introduce Hierarchical Gradient Alignment, employing a non-symmetric operator to establish a Source-Protected Subspace that prevents catastrophic forgetting (Kirkpatrick et al., 2017). Furthermore, to preserve intrinsic manifold geometry, we devise a dual-constraint regularization scheme. This mechanism synergizes Masked Structural Consistency with sparse semantic anchors to preserve local neighborhood connectivity via Lipschitz continuity and ensure distinct class separation (Gouk et al., 2021). Finally, we design a Metric-Decoupled Unsupervised Validator that leverages label-free geometric proxies to guide robust early stopping.

Our main contributions are summarized as follows:

- **Geometric Optimization Paradigm:** We propose H-GSC, a framework that rectifies adaptation dynamics. By utilizing FSG as a geometric pre-conditioner to break optimization stagnation and H-GA to filter destructive interference within a Source-Protected Subspace, we rectify optimization trajectories to be conflict-mitigated, strictly prioritizing source discriminability.
- **Dual-Constraint Structural Regularization:** We introduce a compound mechanism that integrates self-supervised masked modeling with sparse target supervision. This approach promotes Lipschitz continuity constraints and semantic anchor alignment, effectively reducing the joint error term in the generalization bound to prevent manifold collapse.
- **Metric-Decoupled Unsupervised Validator:** We devise a Metric-Decoupled Unsupervised Validator that serves as an independent generalization proxy, thereby enabling robust early stopping when only limited target domain data is available.
- **Superior Performance with Theoretical Grounding:** We theoretically substantiate that H-GSC optimizes the core components of the generalization bound. H-GSC not only demonstrates leading performance in intra-dataset settings but also achieves significant positive transfer in joint cross-dataset scenarios where prior methods suffer from negative transfer, surpassing baselines by 9.65 pp in AUC.

2 Related Work

2.1 Domain Adaptation in EEG Analysis

Domain Adaptation strategies in EEG analysis are key for mitigating cross-subject and cross-device distribution shifts. TSMNet (Kobler et al., 2022) advanced this by introducing domain-specific momentum batch normalization on SPD covariance representations. Later work includes spectral normalization via optimal transport (Gnassounou et al., 2023), predictive shift adaptation for hybrid distribution shifts (Mellot et al., 2024), source-free adaptation addressing label shift (Li et al., 2024b), and explicit domain augmentation for temporal dynamics (Liu et al., 2025). Beyond these statistical alignment methods, Sera (Jia et al., 2025) introduces a separated coarse-to-fine framework that leverages variational source separation to facilitate robust distribution alignment in a shared latent space.

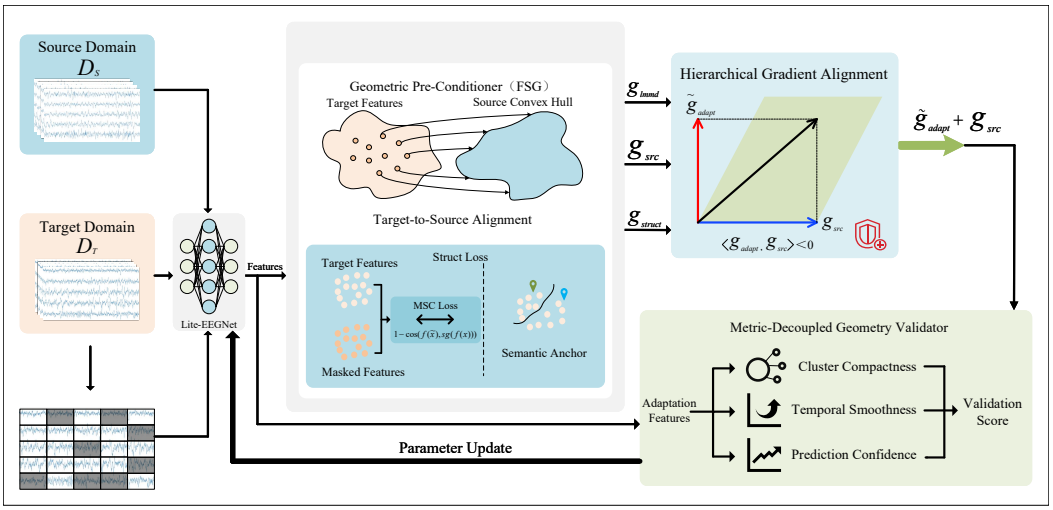


Figure 1: The overall framework of H-GSC. The architecture fundamentally reconfigures adaptation dynamics through three key components: (A) Geometric Pre-conditioner (FSG) reorients target features towards the source convex hull to break optimization stagnation. (B) H-GA filters destructive interference by projecting conflicting gradients onto a source-safe subspace. (C) Dual-Constraint Structural Regularization synergizes Masked Structural Consistency (MSC) with semantic anchors to prevent manifold collapse.

2.2 Domain Generalization for Subject-Independent BCI

In contrast to Domain Adaptation, DG achieves subject-independent classification on unseen individuals without accessing any target data. MAET (Jiang et al., 2023) advanced this by employing adversarial training to suppress subject discrepancy in multimodal emotion decoding. Later work includes denoising mixed mutual reconstruction (Wang et al., 2024d), multi-level domain alignment (Wang et al., 2024b), self-comparison learning (Tu et al., 2024), multi-scale representation learning (Jo et al., 2025), feature fusion with disentanglement (Zhao et al., 2026), and reprogramming large language models for generalizable prediction (Zhao et al., 2025).

2.3 EEG Foundation Models and Large-scale Pretraining

EEG foundation models for cross-subject analysis typically require large-scale pretraining on vast data to learn transferable representations, followed by linear probing or Parameter-Efficient Fine-Tuning for downstream classification and decoding tasks. BIOT (Yang et al., 2023) achieved robust generalization via cross-dataset biosignal Transformer pretraining; subsequent work includes Vector-Quantized EEG Tokenization in LaBraM (Jiang et al., 2024), universal objectives for reliable EEG representation in EEGPT (Wang et al., 2024a), criss-cross pretraining and scaling in CBraMod (Wang et al., 2024c), and Latent Unified Network Architecture addressing structural heterogeneity in LUNA (Döner et al., 2025).

3 Methodology

3.1 Preliminaries & Problem Formulation

We formulate Cross-Subject Seizure Prediction as a binary classification task distinguishing inter-ictal from pre-ictal states. Given an input space $\mathcal{X} \in \mathbb{R}^{C \times T}$ (C channels, T time points) and label space $\mathcal{Y} \in \{0, 1\}$, we utilize a multi-subject Source Domain $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and a heterogeneous Target Domain $\mathcal{D}_T$ where $P_S(X, Y) \neq P_T(X, Y)$. Adopting a strict chronological Few-Shot protocol, $\mathcal{D}_T$ is partitioned into an Adaptation Set $\mathcal{D}_T^{adapt}$ (first 20% for calibration) and a disjoint Test Set $\mathcal{D}_T^{test}$ (remaining 80%).

The goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ using $\mathcal{D}_S$ and $\mathcal{D}_T^{adapt}$ to minimize the expected risk on unseen future data:

$$\min_f \mathbb{E}_{(x,y)\sim\mathcal{D}_T^{test}}[\mathcal{L}_{src}(f(x), y)] \tag{1}$$

where $\mathcal{L}_{src}$ denotes the task-specific loss function.

3.2 Overview of H-GSC: A Geometric Perspective

Traditional DA strategies typically seek a shared latent subspace, aiming to locate the intersection $\mathcal{S} = \mathcal{M}_S \cap \mathcal{M}_T$. However, in cross-subject EEG analysis, severe physiological heterogeneity may result in mutually disjoint manifolds where $\mathcal{M}_S \cap \mathcal{M}_T \approx \emptyset$. Consequently, forcing alignment towards a compromised "middle ground" distorts source discriminative structures, leading to negative transfer.

To address this structural mismatch, H-GSC proposes a "Target-to-Source Alignment" paradigm, as illustrated in Figure 1. Establishing the source manifold $\mathcal{M}_S$ as a Rigid Anchor, we redefine FSG (Zhang et al., 2026) as a geometric pre-conditioner Φ to actively rectify the target manifold $\mathcal{M}_T$. This operator explicitly biases the target distribution P_T towards the convex hull of the source manifold (Conv($\mathcal{M}_S$)), promoting mapping within the discriminative validity region.

Formally, we optimize model parameters θ by minimizing the following objective:

$$\mathcal{L}_{total} = \mathcal{L}_{src}(\mathcal{D}_S) + \lambda\mathcal{L}_{lmmd}(\Phi(\mathcal{D}_T), \mathcal{D}_S) + \beta\mathcal{L}_{struct}(\mathcal{D}_T) \tag{2}$$

where $\mathcal{L}_{src}$ maintains anchor discriminability; $\mathcal{L}_{lmmd}$, which quantifies the statistical discrepancy between domains, drives alignment via the geometric pre-conditioner Φ; and $\mathcal{L}_{struct}$ imposes structural constraints to prevent manifold collapse. The optimization trajectory is further rectified by H-GA to resolve gradient conflicts, while a Metric-Decoupled Validator guides robust early stopping. Furthermore, we theoretically substantiate that H-GSC optimizes the core components of the generalization bound (Ben-David et al., 2010), with detailed derivations provided in Appendix A.

3.3 Feature-Space Guidance as a Geometric Pre-conditioner

Although distribution alignment objectives (e.g., LMMD) aim to match domain statistics, their naive optimization often stagnates in cross-subject EEG analysis due to the quasi-disjoint nature of source and target manifolds, denoted as $\mathcal{M}_S = \{f_\theta(x^s)\}$ and $\mathcal{M}_T = \{f_\theta(x^t)\}$. Given the sparse target adaptation set, the gradient derived from the unmodified LMMD loss $\mathcal{L}_{lmmd}$, denoted as $\mathbf{g}_{raw} = \nabla_\theta\mathcal{L}_{lmmd}(\theta)$, typically exhibits gradient orthogonality or conflict with respect to the source discriminative gradient $\mathbf{g}_{src} = \nabla_\theta\mathcal{L}_{src}(\theta)$. This regime, characterized by $\langle\mathbf{g}_{raw}, \mathbf{g}_{src}\rangle \leq 0$, indicates that updates either drift along the null space of the source task or induce destructive interference, failing to reduce the generalization gap.

To break this impasse, we explicitly steer target features towards anchors lying within the source convex hull to inject geometric priors. For a target sample feature $\mathbf{z} = f_\theta(x^t)$, considering computational efficiency requirements, we adopt a Stochastic Pairwise Anchoring strategy, utilizing the mini-batch as an empirical proxy for the domain distribution instead of global retrieval. Specifically, we establish a randomized one-to-one correspondence within the batch: we directly assign the anchor $\mathbf{a}(\mathbf{z}_i)$ for the i-th target feature $\mathbf{z}_i$ as the source feature $\mathbf{s}_i$ at the same batch index. This stochastic approximation functions essentially as a Monte Carlo estimate (Robert & Casella, 1999); it avoids the overhead of maintaining global prototypes while effectively pulling target features towards the high-density regions of the source manifold $\mathcal{M}_S$ in expectation. Building upon this anchoring mechanism, we further define the affine contraction mapping $\Phi(\mathbf{z}) = (1-\alpha)\mathbf{z} + \alpha\mathbf{a}(\mathbf{z})$ (where $\alpha \in (0, 1)$), whose induced displacement vector $\boldsymbol{\delta} = \alpha(\mathbf{a}(\mathbf{z}) - \mathbf{z})$ points explicitly towards Conv($\mathcal{M}_S$) prior to loss evaluation, thereby completing the target-to-source trajectory rectification.

Crucially, this operator does not merely rescale a gradient but fundamentally alters the optimization landscape. Let $\ell(\mathbf{z})$ denote the batch-wise LMMD objective in the feature space. With FSG, we optimize the pre-conditioned objective $\ell(\Phi(\mathbf{z})) = \ell(\mathbf{z} + \boldsymbol{\delta})$. Assuming ℓ is twice differentiable and L-smooth, we apply the first-order Taylor expansion to the gradient function $\nabla\ell$ around $\mathbf{z}$:

$$\nabla\ell(\mathbf{z} + \boldsymbol{\delta}) = \nabla\ell(\mathbf{z}) + \mathbf{H}_\ell(\mathbf{z})\boldsymbol{\delta} + \mathbf{r} \tag{3}$$

where $\mathbf{H}_\ell(\mathbf{z}) = \nabla^2 \ell(\mathbf{z})$ is the Hessian tensor. Crucially, to ensure the geometric momentum acts as a fixed directional cue rather than introducing complex second-order dependency, we treat the displacement vector $\boldsymbol{\delta}$ as a constant during backpropagation. This is implemented via a stop-gradient operation: $\boldsymbol{\delta} \leftarrow \text{sg}(\boldsymbol{\delta})$. Notably, as rigorously derived in Appendix B, this detached affine rectification implicitly approximates a curvature-aware Hessian-vector product step, effectively injecting second-order geometric information without the prohibitive cost of explicit Hessian evaluation.

Lifting this to the parameter space via the Jacobian $\mathbf{J} = \frac{\partial \mathbf{z}}{\partial \theta}$, and noting that $\frac{\partial \text{sg}(\boldsymbol{\delta})}{\partial \theta} = \mathbf{0}$, the effective parameter gradient $\mathbf{g}_{fsg}$ decomposes into:

$$\begin{aligned} \mathbf{g}_{fsg} &= \nabla_\theta \ell(\mathbf{z} + \text{sg}(\boldsymbol{\delta})) = \mathbf{J}^\top \nabla \ell(\mathbf{z} + \boldsymbol{\delta}) \\ &= \underbrace{\mathbf{J}^\top \nabla \ell(\mathbf{z})}_{\mathbf{g}_{raw}} + \underbrace{\mathbf{J}^\top \mathbf{H}_\ell(\mathbf{z}) \boldsymbol{\delta}}_{\mathbf{m}_{geo}} + \mathbf{J}^\top \mathbf{r} \end{aligned} \tag{4}$$

This decomposition theoretically delineates the optimization dynamics. The first term, $\mathbf{g}_{raw}$, represents the Myopic Alignment Gradient derived from the original features, which typically suffers from orthogonality or conflict with the source task. The second term, $\mathbf{m}_{geo}$, signifies the Geometric Momentum Injection generated by the interaction between the alignment landscape's curvature (Hessian) and the externally imposed displacement $\boldsymbol{\delta}$. This term acts as a systematic escape force that leverages second-order information to redirect the trajectory and overcome optimization inertia. The final term, $\mathbf{J}^\top \mathbf{r}$, accounts for the Higher-Order Residual, capturing non-linear deviations bounded by the displacement magnitude.

However, the optimization dynamics driven by this injected momentum $\mathbf{m}_{geo}$ remain unconstrained. Its magnitude is bounded by $|\mathbf{m}_{geo}| \leq \alpha |\mathbf{J}||\mathbf{H}_\ell(\mathbf{z})||\mathbf{a}(\mathbf{z}) - \mathbf{z}|$. When the target feature is far from the source hull (large displacement) or the landscape is highly curved (large Hessian), FSG naturally produces large, fluctuating updates. Furthermore, the alignment stability is dominated by the inner product $\langle \mathbf{m}_{geo}, \mathbf{g}_{src} \rangle$. While often constructive ($\langle \mathbf{m}_{geo}, \mathbf{g}_{src} \rangle > 0$) due to the directional bias of $\boldsymbol{\delta}$, the momentum direction ultimately depends on local non-convex curvature and may occasionally conflict with the source task ($\langle \mathbf{m}_{geo}, \mathbf{g}_{src} \rangle < 0$), leading to "Geometric Overshoot." Consequently, FSG provides the necessary dynamics to break impasse but requires the subsequent H-GA to act as a safety valve, filtering destructive components to ensure consistent positive alignment.

3.4 Hierarchical Gradient Alignment via Nested Asymmetric Projection

The geometric momentum $\mathbf{m}_{geo}$ injected by FSG successfully overcomes the optimization inertia but remains unconstrained. Aggressive geometric pulling can lead to "Geometric Overshoot," where the update vector conflicts with the source discriminative boundaries. To tame this volatility, we propose H-GA, a nested optimization strategy that functions as a geometric safety valve. We impose a strict intrinsic hierarchy: the source discriminative structure serves as the invariant foundation, while adaptation objectives act as auxiliary regularizers that must operate within the Source-Protected Subspace (Farajtabar et al., 2020).

Before interacting with the primary source task, the diverse adaptation objectives must first achieve internal consistency. Let $\mathbf{g}_{lmmd}$ and $\mathbf{g}_{struct}$ denote the gradients derived from the pre-conditioned LMMD loss (incorporating $\mathbf{m}_{geo}$) and the target structural consistency loss, respectively. To prevent internal cancellation between these auxiliary tasks, we first consolidate them into a unified adaptation vector $\mathbf{g}_{adapt}$ using a symmetric conflict resolution strategy (PCGrad) (Yu et al., 2020): $\mathbf{g}_{adapt} = \text{PCGrad}(\mathbf{g}_{lmmd}, \mathbf{g}_{struct})$. This ensures that $\mathbf{g}_{adapt}$ represents a coherent adaptation force carrying the constructive geometric momentum.

The critical innovation lies in the interaction between this momentum-carrying vector $\mathbf{g}_{adapt}$ and the primary source gradient $\mathbf{g}_{src}$. Unlike symmetric methods that compromise both tasks, we implement a uni-directional rejection scheme to filter the destructive components of the unconstrained momentum. We decompose $\mathbf{g}_{adapt}$ into orthogonal components relative to the source tangent vector $\mathbf{g}_{src}$:

$$\begin{aligned} \mathbf{g}_{adapt} &= \mathbf{g}_{adapt}^{\parallel} + \mathbf{g}_{adapt}^{\perp} \\ &= \frac{\langle \mathbf{g}_{adapt}, \mathbf{g}_{src} \rangle}{\|\mathbf{g}_{src}\|^2} \mathbf{g}_{src} + \left(\mathbf{g}_{adapt} - \frac{\langle \mathbf{g}_{adapt}, \mathbf{g}_{src} \rangle}{\|\mathbf{g}_{src}\|^2} \mathbf{g}_{src} \right) \end{aligned} \tag{5}$$

A conflict arises specifically when $\langle \mathbf{g}_{adapt}, \mathbf{g}_{src} \rangle < 0$, indicating that the injected momentum is effectively pushing the optimization towards "overshoot" (eroding source discriminability). To

enforce the hierarchy, H-GA strictly projects such destructive updates onto the Null Space of the source task:

$$\tilde{\mathbf{g}}_{adapt} = \begin{cases} \mathbf{g}_{adapt}, & \text{if } \langle \mathbf{g}_{adapt}, \mathbf{g}_{src} \rangle \geq 0 \\ \mathbf{g}_{adapt}^{\perp}, & \text{if } \langle \mathbf{g}_{adapt}, \mathbf{g}_{src} \rangle < 0 \end{cases} \tag{6}$$

Geometrically, this operation retains the constructive component of the FSG-induced momentum while mathematically guaranteeing that the adaptation trajectory never traverses the ascent direction of the source loss.

This asymmetric construction ensures Pareto Stationarity with respect to the primary task. Consider the first-order Taylor expansion of the source loss $\mathcal{L}_{src}$ under the combined update direction $\mathbf{g}_{final} = \mathbf{g}_{src} + \tilde{\mathbf{g}}_{adapt}$:

$$\Delta\mathcal{L}_{src} \approx \langle \nabla\mathcal{L}_{src}, -\mathbf{g}_{final} \rangle = -\|\mathbf{g}_{src}\|^2 - \langle \mathbf{g}_{src}, \tilde{\mathbf{g}}_{adapt} \rangle \tag{7}$$

In the case of constructive interference ($\langle \mathbf{g}_{adapt}, \mathbf{g}_{src} \rangle \geq 0$), the second term is negative, accelerating convergence. Crucially, in the case of conflict (overshoot), the projection ensures $\langle \mathbf{g}_{src}, \tilde{\mathbf{g}}_{adapt} \rangle = 0$, reducing the update to $\Delta\mathcal{L}_{src} \approx -\|\mathbf{g}_{src}\|^2$. This proves that the source loss is monotonically non-increasing regardless of the volatility in the FSG momentum, effectively preventing negative transfer.

3.5 Dual-Constraint Structural Regularization

While FSG and H-GA rectify the optimization trajectory, aggressive distribution alignment risks Geometric Degeneracy, where distinct temporal patterns collapse into a low-dimensional subspace, minimizing statistical divergence at the expense of intrinsic discriminability. To preserve the structural integrity of the EEG manifold, we introduce a dual-constraint regularization scheme promoting both local perturbation invariance and global semantic anchoring.

Given the strong temporal dependencies of EEG signals, the feature manifold must satisfy directional consistency. We promote this via MSC, formulated as an asymmetric self-distillation process. Utilizing the stochastic masking operator $\mathcal{M}$, we treat the original representation $f(x)$ as a fixed semantic anchor (teacher) to guide the perturbed view $f(\tilde{x})$ (student). The objective minimizes the angular distance in the embedding space:

$$\mathcal{L}_{msc} = \mathbb{E}_{x \sim \mathcal{D}_T} \left[1 - \cos(f(\tilde{x}), \text{sg}(f(x))) \right] \tag{8}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. Crucially, this asymmetric design prevents model collapse (Chen & He, 2021) by fixing the target distribution during the update. Theoretically, as derived in Appendix A, minimizing this objective strictly minimizes the upper bound of the prediction divergence, ensuring that the adaptation trajectory maintains local Lipschitz continuity relative to the unperturbed data manifold.

Complementing this local regularity, we address the "Boundary Diffusion" limitation of kernel-based alignment (e.g., LMMD), which aligns distributions but fails to enforce sharp decision boundaries. We leverage the sparse labeled subset $\mathcal{D}_T^{adapt}$ as semantic Anchors, formulated as a standard cross-entropy objective:

$$\mathcal{L}_{anchor} = -\mathbb{E}_{(x,y) \sim \mathcal{D}_T^{adapt}} \sum_{c=1}^{K} \mathbb{I}(y = c) \log p(y | f(x)) \tag{9}$$

Geometrically, while LMMD exerts a bulk force aligning probability masses, these anchors serve as explicit constraints, functioning conceptually as "discriminative pivots," to fix specific manifold coordinates to their precise semantic regions. This dual mechanism ensures that statistical alignment translates into robust, margin-preserving classification boundaries.

Integrating these constraints with the primary source task and LMMD alignment loss, we formulate the final objective function. Here, we formally define the composite structural loss term introduced in Eq. (2) as $\mathcal{L}_{struct} = \lambda_{msc}\mathcal{L}_{msc} + \lambda_{anchor}\mathcal{L}_{anchor}$. The total objective establishes a coherent optimization landscape:

$$\mathcal{L}_{total} = \mathcal{L}_{src}(\mathcal{D}_S) + \lambda_{lmmd}\mathcal{L}_{lmmd} + \lambda_{msc}\mathcal{L}_{msc} + \lambda_{anchor}\mathcal{L}_{anchor} \tag{10}$$

This formulation achieves a structure-preserving alignment, where MSC preserves local differential structure to prevent collapse, and semantic anchors maintain global semantic structure to prevent boundary diffusion.

3.6 METRIC-DECOUPLED GEOMETRY VALIDATOR

Standard validation on sparse anchors is circular and prone to overfitting. While accurate domain similarity estimation is pivotal for mitigating negative transfer (Zhang et al., 2022), the limited labels of the adaptation set are insufficient to characterize the holistic target distribution. Consequently, to prevent overfitting to this biased subset, we strategically abandon the use of adaptation labels during validation. Instead, we propose a Metric-Decoupled Geometry Validator that leverages intrinsic geometric properties as label-free proxies to robustly identify optimal states. We strictly adhere to an inductive setting: geometric quality is assessed utilizing only the data of the adaptation set ($\mathcal{D}_T^{adapt}$), while its labels and the test set ($\mathcal{D}_T^{test}$) remain strictly inaccessible.

To quantify spatial discriminability without supervision, we employ Soft Neighborhood Density (SND) (Saito et al., 2021). Given adaptation features F_T, we calculate the softmax-normalized affinity p_{ij} between samples using a temperature-scaled cosine kernel. Global clustering quality is measured by the mean entropy of these neighborhood distributions:

$$\mathcal{V}_{snd} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i} p_{ij}\log p_{ij} \tag{11}$$

Minimizing $\mathcal{V}_{snd}$ encourages the affinity distribution to become "peaked," serving as a proxy for compact, well-separated semantic clusters.

Complementing spatial structure, we verify temporal and probabilistic consistency. Temporal Consistency (TC) ensures trajectory smoothness by penalizing the divergence between consecutive predictions: $\mathcal{V}_{tc} = \mathbb{E}_t[\|p(y|x_{t+1}) - p(y|x_t)\|_2]$. Additionally, to penalize ambiguous decision boundaries, we incorporate Prediction Entropy ($\mathcal{V}_{ent}$), defined as $\mathcal{V}_{ent} = \mathbb{E}_x[-\sum p(y|x)\log p(y|x)]$, which encourages high-confidence predictions.

The final validation score synthesizes these geometric invariants:

$$\mathcal{V}_{total} = \mathcal{V}_{snd} + \lambda_{tc}\mathcal{V}_{tc} + \lambda_{ent}\mathcal{V}_{ent} \tag{12}$$

Minimizing this score selects a model that induces a spatially discriminative, temporally coherent, and confident manifold using only available adaptation data.

4 RESULTS

4.1 DATASETS AND EVALUATION PROTOCOLS

We evaluate H-GSC on two standard public benchmarks: CHB-MIT and Siena. Experiments are conducted under a strict Leave-One-Subject-Out (LOSO) cross-validation protocol to assess cross-subject generalization. To mimic realistic clinical calibration, we utilize only the chronologically first 20% of the target subject's data as the adaptation set ($\mathcal{D}_T^{adapt}$), reserving the remaining 80% for testing. Detailed protocols regarding data preprocessing, signal segmentation, and labeling definitions are provided in Appendix C.

4.2 EXPERIMENT SETTINGS AND IMPLEMENTATION

To rigorously evaluate cross-patient generalization, we employ a Leave-One-Subject-Out (LOSO) scheme across three progressive protocols: Intra-Dataset Adaptation on (1) CHB-MIT and (2) Siena; and (3) a Joint Cross-Dataset Adaptation scenario pooling all subjects to challenge the model with extreme heterogeneity in age, sampling rates, and electrode layouts. Consistent with the baselines (Zhang et al., 2026), we employ the lightweight LiteEEG-Net as the backbone feature extractor, adhering to an efficient end-to-end adaptation paradigm. Implementation details, including network architecture, hyperparameter configurations, and computational efficiency analysis, are provided in Appendix D. For performance assessment, we adopt standard metrics including Recall, Precision, Specificity, and AUC, whose formal definitions are detailed in Appendix E. Finally, a comprehensive hyperparameter sensitivity analysis is presented in Appendix F.

Table 1: Ablation Study on CHB-MIT. Method 1: Validator; Method 2: +Target CE; Method 3: +MSC; Method 4: H-GSC (Full).

Method	Pre(%)	Spe(%)	AUC
Baseline (FSG-Zero)	77.72	72.97	0.7207
Method 1	78.43	77.75	0.7463
Method 2	76.39	71.03	0.7653
Method 3	79.66	77.74	0.7888
Method 4 (H-GSC)	**81.00**	**79.16**	**0.7954**

4.3 Comparative Experiment Analysis

To comprehensively evaluate the effectiveness of H-GSC, we conducted extensive comparisons against state-of-the-art baselines, with detailed tabular results and in-depth analyses provided in Appendix G. In summary, H-GSC establishes new state-of-the-art benchmarks on both CHB-MIT and Siena datasets, consistently outperforming advanced methods like FSG-Zero and GSFDA in terms of AUC, Specificity, and Recall. Crucially, in the challenging Joint Cross-Dataset Adaptation scenario, where prior methods often suffer from negative transfer due to extreme physiological heterogeneity, H-GSC demonstrates robust positive transfer. For instance, our method achieves a 9.65% improvement in average AUC compared to FSG-Zero, validating that our hierarchical alignment strategy effectively bridges massive distributional shifts.

4.4 Ablation Study Analysis

As shown in Table 1, a systematic ablation study on the CHB-MIT dataset progressively reveals the independent contributions and synergistic mechanisms of the H-GSC components. First, deploying the Metric-Decoupled Geometry Validator (Method 1) establishes a robust foundation for adaptation, significantly lifting AUC (from 0.7207 to 0.7463) and specificity. Subsequently, the naive linear addition of Target CE (Method 2), while boosting AUC, precipitated a sharp decline in specificity. This phenomenon offers empirical evidence of the gradient conflicts induced by simple scalarization. Incorporating MSC (Method 3) as a geometric regularizer partially mitigated this instability and further enhanced AUC, yet the performance remained constrained by the bottlenecks inherent to linear optimization. Ultimately, the full H-GSC framework (Method 4) integrates the H-GA mechanism to enforce hierarchical gradient projection, which geometrically filters out conflicting components. Achieving a peak AUC of 0.7954, this result validates gradient calibration as the linchpin for resolving multi-objective interference and translating structural constraints into synergistic performance gains. Furthermore, the detailed results and analysis of the ablation study on the Siena dataset are provided in Appendix H.

5 Mechanistic Analysis and Geometric Verification

5.1 Verification of Gradient Rotation

To empirically verify the theoretical claims of Geometric Pre-conditioning (Section 3.3) and Hierarchical Gradient Alignment (Section 3.4), we tracked the cosine similarity between the source discriminative gradient $\mathbf{g}_{src}$ and adaptation gradients throughout training, comparing the Baseline, unconstrained FSG, and H-GSC. As shown in Appendix I (Figure 2), while naive alignment fluctuates around zero and unconstrained FSG suffers from "Geometric Overshoot" with severe negative conflicts, H-GSC maintains consistently positive cosine similarity ($\cos\theta > 0$). This confirms that H-GA acts as an effective safety valve by filtering destructive components within the Source-Protected Subspace.

5.2 Spectral Analysis of Manifold Structure

To investigate the role of MSC in preserving manifold geometry, we evaluated feature consistency under 30% random masking perturbation (He et al., 2022). As detailed in Appendix I (Figure 3), the variant without MSC displays a heavy-tailed similarity distribution dropping down to 0.850,

indicating localized overfitting. In contrast, H-GSC exhibits a concentrated density peak near 0.98. This confirms that MSC induces Lipschitz-like stability on the perturbation manifold, suppressing Jacobian energy to enforce global, redundant semantic representations.

5.3 Visualizing the Decision Boundary

To inspect whether semantic anchors resolve discriminative ambiguity, we analyze the learned feature manifold structures via t-SNE in Appendix I (Figure 4). As empirically visualized, relying solely on statistical alignment (w/o Anchor) precipitates severe boundary diffusion along classification margins. Conversely, H-GSC leverages sparse anchors as explicit discriminative pivots, untangling the manifold into compact, well-separated clusters with sharp margins and preventing boundary collapse.

6 Conclusion

In this paper, we propose H-GSC, a Hierarchical Geometric framework with Structural Constraints designed to rectify the severe negative transfer and optimization instability inherent in few-shot EEG adaptation. By synergizing structural constraints with Hierarchical Gradient Alignment to rectify conflicting gradient trajectories, H-GSC explicitly aligns target distributions onto the protected source manifold. H-GSC achieves state-of-the-art performance on the CHB-MIT benchmark (79.54% AUC). Notably, in rigorous joint cross-dataset scenarios, it delivers significant performance gains, successfully converting extreme data heterogeneity into positive transfer where recent studies have often struggled. Furthermore, H-GSC exhibits remarkable efficiency; by empowering a lightweight backbone ($\sim$35k parameters) with computationally efficient geometric constraints, it maintains low algorithmic complexity suitable for real-time inference. These advancements position H-GSC as a promising solution for implementing robust, patient-independent seizure prediction on resource-constrained wearable devices, paving the way for large-scale clinical deployment.

AI use statement

During the preparation of this manuscript, the authors utilized large language models exclusively for grammar verification, language polishing, and editorial formatting to improve readability. The authors independently conceived the research ideas, designed all geometric mechanisms, conducted all theoretical derivations, implemented the code, performed all empirical experiments, and reviewed the final manuscript, taking full scientific responsibility for the entire content.

Ethics statement

This study utilizes exclusively publicly available, anonymized benchmark scalp EEG datasets (CHB-MIT and Siena databases hosted on PhysioNet). No new human clinical experiments were conducted for this research, and all original patient identifiers were de-identified by the original data collectors. The algorithmic developments target patient-assistive seizure prediction for clinical warning systems, minimizing false alarms and fatal false-negative misdiagnoses. The research adheres strictly to the ICLR Code of Ethics.

Reproducibility statement

To ensure thorough reproducibility, all theoretical assumptions and rigorous step-by-step proofs of the generalization bounds and gradient flows are provided in Appendices A and B. Detailed dataset specifications, the non-overlapping chronological 20%/80% split protocol, and signal preprocessing pipelines are fully documented in Appendix C. The complete neural network architectural breakdown of LiteEEG-Net, hyperparameter settings, and inference profiling are reported in Appendix D. All metrics are formally defined in Appendix E.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Z. S. Chen, B. L. Lu, and W. Wu. Representation learning and foundation models for electroencephalography analyses: Current trends, fundamental insights, and future directions. *IEEE Signal Processing Magazine*, 42(5):36–57, 2025.

B. Döner, T. M. Ingolfsson, L. Benini, and Y. Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. *arXiv preprint arXiv:2510.22257*, 2025.

Q. Dong, Z. Wang, and M. Gao. Noise-aware epileptic seizure prediction network via self-attention feature alignment. *IEEE Journal of Biomedical and Health Informatics*, 2025. Early Access.

Z. Du, X. Li, F. Li, K. Lu, L. Zhu, and J. Li. Domain-agnostic mutual prompting for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23375–23384, 2024.

M. Farajtabar, N. Azizan, A. Mott, and A. Li. Orthogonal gradient descent for continual learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3762–3773. PMLR, 2020.

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

T. Gnassounou, R. Flamary, and A. Gramfort. Convolution Monge mapping normalization for learning on sleep data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10457–10476, 2023.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xudong Guo, Lin Liu, Sha Lu, Jiuyong Li, Thuc Duy Le, and Jixue Liu. Can eeg foundation models help with epileptic seizure prediction? In *2025 IEEE International Conference on Big Data (BigData)*, pp. 1924–1933. IEEE, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Zhihao Jia, Meiyan Xu, Jingyuan Wang, Ziyu Jia, Yong Li, Xinliang Zhou, Chenyu Liu, Junfeng Yao, and Yi Ding. Sera: Separated coarse-to-fine representation alignment for cross-subject EEG-based emotion recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5509–5518, 2025.

Wei-Bang Jiang, Xing-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5975–5984, 2023.

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

S. Jo, J. S. Yoon, W. Jeong, K. Oh, and H.-I. Suk. MEASURE: Multi-scale minimal sufficient representation learning for domain generalization in sleep staging. *arXiv preprint arXiv:2510.12070*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

R. Kobler, J. I. Hirayama, Q. Zhao, and M. Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6219–6235, 2022.

P. Koutsouvelis, B. Chybowski, A. Gonzalez-Sulser, S. Abdullateef, and J. Escudero. Preictal period optimization for deep learning-based epileptic seizure prediction. *Journal of Neural Engineering*, 21(6):066040, 2024.

Jingjing Li, Zhengming Yu, Z. Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5743–5762, 2024a.

S. Li, M. Kawanabe, and R. J. Kobler. SPDIM: Source-free unsupervised conditional and label shift adaptation in EEG. *arXiv preprint arXiv:2411.07249*, 2024b.

J. Liao, Y. Chen, Y. He, K. Zhang, T. Ma, Y. Wang, and X. Shao. An EEG-based seizure prediction model encoding brain network temporal dynamics. *IEEE Journal of Biomedical and Health Informatics*, 2025. Early Access.

M. Liu, X. Chen, X. Li, W. Guan, and L. Nie. Breakthrough sensor-limited single view: Towards implicit temporal dynamics for time series domain adaptation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

T. Mao, C. Li, R. Song, G. Xu, and X. Chen. Online seizure prediction via fine-tuning and test-time adaptation. *IEEE Internet of Things Journal*, 11(11):20784–20796, 2024.

Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1): 153–179, 1997.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Alexandre Gramfort, and Denis A. Engemann. Geodesic optimization for predictive shift adaptation on EEG data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 32828–32855, 2024.

P. Peng, Y. Song, L. Yang, and H. Wei. Seizure prediction in EEG signals using STFT and domain adaptation. *Frontiers in Neuroscience*, 15:825434, 2022.

M. J. Rivera, A. Valera, A. Valera-Gonzalez, and H. Pomares. Diagnosis and prognosis of mental disorders by means of EEG and deep learning: a systematic mapping study. *Artificial Intelligence Review*, 55(2):1209–1251, 2022.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.

Sanwar Saha and Mathias Baumert. Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in Computational Neuroscience*, 13:87, 2020.

K. Saito, D. Kim, P. Teterwak, S. Sclaroff, T. Darrell, and K. Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9184–9193, 2021.

Wojciech Samek, Carmen Vidaurre, Klaus-Robert Müller, and Motoaki Kawanabe. Stationary common spatial patterns for brain–computer interfacing. *Journal of Neural Engineering*, 9(2): 026013, 2012.

X. Shen, R. Gan, K. Wang, S. Yang, Q. Zhang, Q. Liu, and S. Song. Dynamic-attention-based EEG state transition modeling for emotion recognition. *IEEE Transactions on Affective Computing*, 2025.

S. Tu, L. Cao, D. Zhang, J. Chen, L. Ma, Y. Zhang, and Y. Yang. DMNet: Self-comparison driven model for subject-independent seizure detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28254–28280, 2024.

G. Wang, W. Liu, Y. He, C. Xu, L. Ma, and H. Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024a.

J. Wang, S. Zhao, H. Jiang, S. Li, T. Li, and G. Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024b.

J. Wang, S. Zhao, Z. Luo, Y. Zhou, H. Jiang, S. Li, et al. CBraMod: A criss-cross brain foundation model for EEG decoding. *arXiv preprint arXiv:2412.07236*, 2024c.

Y. Wang, B. Zhang, and Y. Tang. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 628–636, 2024d.

Dongrui Wu, Yifei Xu, and Bao-Liang Lu. Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2020.

C. Yang, M. Westover, and J. Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023.

H. Yao, X. Yang, X. Pan, S. Liu, P. W. Koh, and C. Finn. Improving domain generalization with domain relations. *arXiv preprint arXiv:2302.02609*, 2023.

T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

H. Zhang, J. Zhang, W. Chen, and G. Du. FSG-zero: An efficient end-to-end fusion framework for robust continual adaptation in EEG seizure prediction. *Knowledge-Based Systems*, pp. 115203, 2026.

W. Zhang, L. Deng, L. Zhang, and D. Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.

S. Zhao, C. Zhang, J. Wang, H. Jiang, S. Li, T. Li, and G. Pan. SeiFuD: A novel deep learning framework leveraging domain generalization for cross-subject seizure prediction. *Neurocomputing*, pp. 132227, 2026.

Y. Zhao, S. Feng, C. Li, R. Song, D. Liang, and X. Chen. Source-free domain adaptation for privacy-preserving seizure prediction. *IEEE Transactions on Industrial Informatics*, 20(2):2787–2798, 2023.

Y. Zhao, A. Liu, C. Li, L. Wang, R. Qian, and X. Chen. Generalizable seizure prediction with LLMs: Converting EEG to textual representations. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

X. Zhou et al. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *IEEE Signal Processing Magazine*, 42(5):22–35, 2025.

APPENDIX ORGANIZATION

The Appendix is organized as follows:

- The theoretical derivation and analysis of the generalization bound are provided in Appendix A.
- Theoretical clarifications regarding the gradient flow of Feature-Space Guidance are discussed in Appendix B.
- Dataset specifications, preprocessing pipelines, and the few-shot chronological adaptation protocol are outlined in Appendix C.
- Implementation details, including the LiteEEG-Net architecture, masking operator, and computational complexity analysis, are presented in Appendix D.
- Definitions of the evaluation metrics used in this study are listed in Appendix E.
- A comprehensive hyperparameter sensitivity analysis is conducted in Appendix F.
- Detailed comparative experimental results and in-depth analyses for intra-dataset and joint cross-dataset scenarios are documented in Appendix G.
- A supplementary ablation study demonstrating the generalizability of our framework on the independent Siena dataset is provided in Appendix H.
- Empirical mechanistic analyses and geometric verification are detailed in Appendix I.

A GENERALIZATION BOUND ANALYSIS

A.1 NOTATION, PROTOCOL, AND OBJECTIVE

Source domain: Joint distribution $P_S(X, Y)$.
Target domain: $P_T(X, Y)$. We follow a strict chronological few-shot protocol: the target stream is split into an adaptation set $\mathcal{D}_T^{adapt}$ (first 20%) and a disjoint test set $\mathcal{D}_T^{test}$ (remaining 80%).
Hypothesis class: $h = g \circ f$, where f is the feature extractor and g is the classifier head.

The loss $\ell(h(x), y)$ is bounded in $[0, 1]$. We define expected risks as:

$$R_S(h) = \mathbb{E}_{(x,y)\sim P_S}[\ell(h(x), y)] \tag{13}$$

$$R_T(h) = \mathbb{E}_{(x,y)\sim P_T}[\ell(h(x), y)] \tag{14}$$

H-GSC optimizes the following composite objective:

$$\mathcal{L}_{total} = \mathcal{L}_{src}(\mathcal{D}_S) + \lambda\mathcal{L}_{lmmd}(\Phi(\mathcal{D}_T), \mathcal{D}_S) + \beta\mathcal{L}_{struct}(\mathcal{D}_T) \tag{15}$$

where $\mathcal{L}_{src}$ preserves source discriminability, $\mathcal{L}_{lmmd}$ drives alignment via Φ (FSG), and $\mathcal{L}_{struct}$ enforces structural constraints (comprising Masked Structural Consistency $\mathcal{L}_{msc}$ and Semantic Anchors $\mathcal{L}_{anchor}$).

Implementation correspondence:

- **FSG rectification** is exactly an affine contraction in feature space: $f_t^{guided} = f_t + w(f_s - f_t) = (1 - w)f_t + wf_s$. This matches $\Phi(z) = (1 - \alpha)z + \alpha a(z)$ with $\alpha = w$ and anchor $a(z) = f_s$ (or a class-conditional prototype).
- **H-GA** is implemented as a nested projection: an internal PCGrad between $(\mathcal{L}_{lmmd}, \mathcal{L}_{struct})$, followed by an asymmetric projection that removes components conflicting with the source CE gradient.

These components are used below to connect each module to a term in Ben-David's bound.

A.2 THE ORIGINAL BEN-DAVID DOMAIN ADAPTATION BOUND

Define the $H\Delta H$-divergence:

$$d_{H\Delta H}(P_S, P_T) = 2 \sup_{h,h'\in\mathcal{H}} |\Pr_{x\sim P_S}[h(x) \neq h'(x)] - \Pr_{x\sim P_T}[h(x) \neq h'(x)]| \tag{16}$$

Theorem A.1 (Ben-David et al., 2010). For any $h \in \mathcal{H}$:

$$R_T(h) \leq R_S(h) + \frac{1}{2}d_{H\Delta H}(P_S, P_T) + \lambda^* \tag{17}$$

where $\lambda^* = \min_{h\in\mathcal{H}}(R_S(h) + R_T(h))$ is the ideal joint risk.

Our goal is to show that the optimization procedure minimizes:

- $R_S(h)$
- $d_{H\Delta H}(P_S, P_T)$
- λ^* (or a provable upper bound of λ^* that is monotone in our losses)

The rest of Appendix A provides explicit "bridge assumptions" so that each of these three quantities becomes provably controlled by our implemented losses.

A.3 Bridging $d_{H\Delta H}$ to Wasserstein/IPM on the learned representation

Let $z = f(x)$ be the representation and define induced feature marginals: $\mu_S = f_{\#}P_S^X, \mu_T = f_{\#}P_T^X$. We make the following bridge assumption:

Assumption A.1 (Lipschitz-controlled discriminator family). There exists a function class $\mathcal{F}_L$ consisting of L-Lipschitz functions on feature space $\mathcal{Z}$ such that for all $h, h' \in \mathcal{H}$, the symmetric-difference indicator $\mathbf{1}[h(\cdot) \neq h'(\cdot)]$ can be upper bounded by a scaled element of $\text{conv}(\mathcal{F}_L)$, i.e.,

$$\mathbf{1}[h(z) \neq h'(z)] \leq \kappa \cdot \phi(z) \tag{18}$$

for some $\phi \in \text{conv}(\mathcal{F}_L)$, with a constant $\kappa \geq 1$. This assumption aligns with the Lipschitz continuity constraints established in Wasserstein GANs (Arjovsky et al., 2017).

Lemma A.2 ($H\Delta H$ is an IPM and is controlled by Wasserstein). Under Assumption A.1,

$$d_{H\Delta H}(P_S, P_T) = d_{H\Delta H}(\mu_S, \mu_T) \leq 2\kappa \cdot \text{IPM}_{\mathcal{F}_L}(\mu_S, \mu_T) \leq 2\kappa L \cdot W_1(\mu_S, \mu_T) \tag{19}$$

Proof sketch. $H\Delta H$ is an integral probability metric (IPM) over the class $\{\mathbf{1}[h \neq h']\}$. The first inequality is Assumption A.1; the second inequality uses Kantorovich–Rubinstein duality (Gulrajani et al., 2017): IPM over L-Lipschitz functions is upper bounded by $L \cdot W_1$. □

A.4 Target-to-Source Alignment as a contraction that strictly shrinks the divergence term

Our paper explicitly defines the affine contraction mapping (FSG):

$$\Phi(z) = (1-\alpha)z + \alpha a(z), \quad \text{with } a(z) \in \text{Conv}(\mathcal{M}_S) \tag{20}$$

Let $\mu_T^\alpha = (\Phi)_{\#}\mu_T$ be the rectified target feature distribution.

A.4.1 The key bridge assumption: anchors approximate the OT map

Assumption A.2 (Stochastic OT Proxy via Measure Concentration). In the paradigm of mini-batch training via Stochastic Gradient Descent (SGD), mandating a precise pointwise correspondence between every randomly sampled anchor $a(z, y)$ and the global optimal transport map T^* is theoretically incompatible with the inherent stochasticity of the sampling process. Consequently, rather than imposing such an intractable constraint, we shift our theoretical grounding from deterministic alignment to statistical consistency. We leverage the concentration of measure phenomenon in deep feature spaces, assuming that for a class y, the source feature distribution $\mu_S(\cdot|y)$ is unimodal and compactly concentrated (e.g., exhibits sub-Gaussian concentration properties).

Under this probabilistic framework, the proposed Stochastic Pairwise Anchoring strategy is reformulated as a Monte Carlo estimator for the optimal transport map, consistent with theoretical findings on mini-batch optimal transport (Fatras et al., 2021). For any given target sample z, the randomly selected source anchor $a(z)$ satisfies the condition that its expectation asymptotically aligns with the probability mass center (barycenter) of the source domain:

$$\mathbb{E}[a(z)] \approx \mathbb{E}_{s\sim\mu_S}[s] \approx T^*_{barycenter}(z) \tag{21}$$

It follows that the anchor approximation error ε_{anchor} is strictly bounded by the inherent variance of the source distribution (i.e., sampling noise) rather than by systemic geometric misalignment:

$$\mathbb{E}_z\|a(z) - T_y^*(z)\|^2 \leq \text{Var}(\mu_S) + \text{Bias}(T^*) := \varepsilon_{anchor} \tag{22}$$

Justification: This statistically grounded assumption provides a rigorous theoretical basis for the proposed Stochastic Pairwise Anchoring strategy. While individual anchors may introduce variance in a single iteration, the expected gradient trajectory, accumulated over the optimization process, effectively simulates the Wasserstein gradient flow converging towards the high-density support of the source manifold. The error term $\alpha \cdot \varepsilon_{anchor}$ in Theorem A.5 is thus explicitly interpreted as the variance cost incurred by employing this computationally efficient stochastic estimator, rather than an intrinsic model bias.

A.4.2 Strict shrinkage in Wasserstein

Lemma A.3 (Displacement interpolation shrinkage (McCann, 1997)). Assume the conditional OT maps T_y^* exist and have finite second moments. Define the displacement interpolation $\Phi^{*(z,y)} = (1-\alpha)z + \alpha T_y^*(z)$. Let $\mu_T^{\alpha,*} = (\Phi^*)_{\#}\mu_T$. Then for each class y,

$$W_2(\mu_S(\cdot|y), \mu_T^{\alpha,*}(\cdot|y)) = (1-\alpha) \cdot W_2(\mu_S(\cdot|y), \mu_T(\cdot|y)) \tag{23}$$

Consequently, for the mixture, $W_2(\mu_S, \mu_T^{\alpha,*}) \leq (1-\alpha) \cdot W_2(\mu_S, \mu_T)$.

Lemma A.4 (Approximate anchors preserve shrinkage up to ε_{anchor}). Under Assumption A.2,

$$W_2(\mu_S, \mu_T^{\alpha}) \leq (1-\alpha) \cdot W_2(\mu_S, \mu_T) + \alpha \cdot \varepsilon_{anchor} \tag{24}$$

Proof sketch. Couple μ_T^{α} with $\mu_T^{\alpha,*}$ through the same z and use the triangle inequality: $W_2(\mu_S, \mu_T^{\alpha}) \leq W_2(\mu_S, \mu_T^{\alpha,*}) + W_2(\mu_T^{\alpha,*}, \mu_T^{\alpha})$. The first term is Lemma A.3; the second term is bounded by $\alpha \cdot \mathbb{E}\|a(z,y) - T_y^*(z)\| \leq \alpha\varepsilon_{anchor}$. □

A.4.3 From Wasserstein shrinkage to a strict reduction of Ben-David's divergence term

Theorem A.5 (Target-to-Source Alignment reduces the divergence control term). Although the generalization bound is naturally controlled by the W_1 distance (via Kantorovich-Rubinstein duality), our geometric rectification acts on the Euclidean transport cost (corresponding to the W_2 distance). Utilizing the fundamental inequality $W_1(\mu, \nu) \leq W_2(\mu, \nu)$, which holds for any distributions with finite second moments, we can bound the relevant divergence term using the geometric shrinkage properties of the W_2 space.

Under Assumptions A.1–A.2, we derive:

$$\begin{aligned} d_{H\Delta H}(P_S, P_T^{\alpha}) &\leq 2\kappa L \cdot W_1(\mu_S, \mu_T^{\alpha}) \\ &\leq 2\kappa L \cdot W_2(\mu_S, \mu_T^{\alpha}) \qquad (25) \\ &\leq 2\kappa L \left[(1-\alpha)W_2(\mu_S, \mu_T) + \alpha\varepsilon_{anchor}\right] \qquad (26) \end{aligned}$$

This implies a trade-off: increasing α (where $0 < \alpha < 1$) linearly shrinks the domain discrepancy $W_2(\mu_S, \mu_T)$, but introduces an anchor approximation error scaled by α. Provided that the anchors are reasonably accurate (i.e., $\varepsilon_{anchor} < W_2(\mu_S, \mu_T)$), a non-zero α strictly tightens the bound compared to the unrectified baseline ($\alpha = 0$). In practice, α typically takes an intermediate value (e.g., $w = 0.5$ in implementation) to balance strict geometric contraction against the accumulation of anchor noise.

A.5 LMMD minimizes the post-rectification discrepancy that upper bounds $d_{H\Delta H}$

Assumption A.3 (Characteristic RKHS kernel for LMMD). The Gaussian kernel family used in LMMD is characteristic on $\mathcal{Z}$, and the LMMD estimator is consistent for the conditional MMD.

Lemma A.6 (MMD as an IPM (Gretton et al., 2012)). For the unit ball of an RKHS $\mathcal{H}_k$, $\text{MMD}_k(\mu_S, \mu_T^{\alpha}) = \text{IPM}_{\mathcal{F}_k}(\mu_S, \mu_T^{\alpha})$. Thus minimizing LMMD reduces $\text{IPM}_{\mathcal{F}_k}(\mu_S, \mu_T^{\alpha})$.

Assumption A.4 (Function-class domination). There exists c_k such that $\text{IPM}_{\mathcal{F}_L}(\cdot,\cdot) \leq c_k \cdot \text{IPM}_{\mathcal{F}_k}(\cdot,\cdot)$. Then:

$$d_{H\Delta H}(P_S, P_T^{\alpha}) \leq 2\kappa \cdot \text{IPM}_{\mathcal{F}_L}(\mu_S, \mu_T^{\alpha}) \leq 2\kappa c_k \cdot \text{MMD}_k(\mu_S, \mu_T^{\alpha}) \tag{27}$$

So LMMD directly minimizes an explicit upper bound of the original $d_{H\Delta H}$ term on the rectified target distribution.

A.6 H-GA (HIERARCHICAL PROJECTION) MINIMIZES THE SOURCE-RISK TERM WITHOUT SACRIFICING ADAPTATION

Assumption A.5 (Small-step smoothness). $\mathcal{L}_{src}(\theta)$ is β-smooth, and learning rate η is small enough that first-order descent guarantees hold up to $O(\eta^2)$.

Lemma A.7 (Source-risk monotonicity under H-GA). Let $g_{src} = \nabla_\theta \mathcal{L}_{src}$ and g_{adapt} be the composed adaptation gradient. Let $\hat{g}_{adapt}$ be the outer projected adaptation gradient used by H-GA such that $\langle g_{src}, \hat{g}_{adapt} \rangle \geq 0$ always. Then the update $\theta^+ = \theta - \eta(g_{src} + \hat{g}_{adapt})$ satisfies:

$$\mathcal{L}_{src}(\theta^+) \leq \mathcal{L}_{src}(\theta) - \eta \|g_{src}\|^2 + O(\eta^2) \tag{28}$$

H-GA minimizes the R_S term via its empirical surrogate while allowing adaptation in the orthogonal subspace.

A.7 MSC AS A CONSISTENCY UPPER BOUND UNDER STOP-GRADIENT

While MSC formally resembles positive pair alignment in contrastive learning, structurally, it is implemented as a self-distillation process with an asymmetric gradient flow. Methodologically, the MSC loss employs a stop-gradient (sg) operator on the unmasked view, treating it as a fixed semantic target ("teacher") to guide the masked view ("student").

Formally, let f_θ be the current network. The MSC loss is defined as:

$$\mathcal{L}_{msc} = \mathbb{E}_{x \sim \mathcal{D}_T} \left[1 - \cos(f_\theta(\tilde{x}), sg(f_\theta(x))) \right] \tag{29}$$

where $sg(\cdot)$ implies that the term produces no gradients during backpropagation.

To bridge this loss to the generalization error, we introduce an assumption regarding the local consistency of the prediction function on the manifold.

Assumption A.6 (Directional Consistency on Projected Space). Assume that the divergence in the predictions of the hypothesis $h = g \circ f$ is primarily dominated by deviations in feature direction. For a fixed anchor $z = f_\theta(x)$ and its perturbation $\tilde{z} = f_\theta(\tilde{x})$, there exists a constant C such that the prediction divergence is bounded by the angular distance:

$$|\ell(h(\tilde{x}), y) - \ell(h(x)_{fixed}, y)| \leq C \cdot \sqrt{1 - \cos(\tilde{z}, z)} \tag{30}$$

where $h(x)_{fixed}$ denotes the fixed prediction based on the current parameter snapshot.

Lemma A.9 (Consistency Bound). Under the stop-gradient operator, minimizing $\mathcal{L}_{msc}$ minimizes the upper bound of the expected prediction divergence between the perturbed view and the anchor view established by the current parameters:

$$\begin{aligned} \mathbb{E}_x[|\ell(h(\tilde{x})) - \ell(h(x)_{fixed})|] &\leq C \cdot \mathbb{E}_x \left[\sqrt{1 - \cos(f_\theta(\tilde{x}), sg(f_\theta(x)))} \right] \\ &\leq C \cdot \sqrt{\mathcal{L}_{msc}} \end{aligned} \tag{31}$$

(via Jensen's inequality).

Significance: By enforcing this bound, MSC ensures that the adaptation trajectory maintains local semantic consistency relative to the unperturbed data manifold. The inclusion of the sg operator prevents model collapse (Chen & He, 2021), where both $f(x)$ and $f(\tilde{x})$ degenerate to trivial solutions, and compels the network to explicitly leverage global context to reconstruct semantic representations consistent with the original view, despite structural missingness induced by masking.

A.8 SEMANTIC ANCHORS MINIMIZE AN EXPLICIT UPPER BOUND ON THE IDEAL JOINT RISK λ^*

Let $n_a = |\mathcal{D}_T^{adapt}|$.

Lemma A.10 (λ^* is upper bounded by source risk + target anchor risk). For any hypothesis h,

$$\lambda^* = \min_{h'} (R_S(h') + R_T(h')) \leq R_S(h) + R_T(h) \tag{32}$$

Assumption A.7 (Uniform convergence on the anchor set). With probability $\geq 1-\delta$, for all $h \in \mathcal{H}$,

$$R_T(h) \leq \hat{R}_{T^{adapt}}(h) + \text{Gen}_T(\mathcal{H}, n_a, \delta) \tag{33}$$

where Gen_T is a complexity term. Thus:

$$\lambda^* \leq R_S(h) + \hat{R}_{T^{adapt}}(h) + \text{Gen}_T(\mathcal{H}, n_a, \delta) \tag{34}$$

Minimizing target CE on $\mathcal{D}_T^{adapt}$ minimizes the bound on λ^*, while MSC reduces Gen_T by improving stability.

A.9 Final theorem: H-GSC tightens the original Ben-David bound term-by-term

Theorem A.11 (H-GSC minimizes each component of Ben-David's bound). Under Assumptions A.1–A.7, with probability $\geq 1-\delta$, for the hypothesis h learned by H-GSC,

$$R_T(h) \leq \underbrace{\hat{R}_S(h) + \text{Gen}_S(\mathcal{H}, n_s, \delta)}_{(1)} + \underbrace{\frac{1}{2} d_{H\Delta H}(P_S, P_T)}_{(2)} + \underbrace{\lambda^*}_{(3)} \tag{35}$$

Each term admits an explicit monotone upper bound minimized by our objective:

1. **Source Risk:** $\mathcal{L}_{src}$ + H-GA $\Rightarrow$ minimizes $R_S(h)$ (Lemma A.7).
2. **Divergence Term:**

   $$d_{H\Delta H}(P_S, P_T) \leq C_{div} \cdot [(1-\alpha) W_2(\mu_S, \mu_T) + \alpha \varepsilon_{anchor}] \tag{36}$$

   and $d_{H\Delta H} \leq C_{mmd} \cdot \text{LMMD}(\mu_S, \mu_T^\alpha)$. Choosing an appropriate $\alpha \in (0,1)$ (implemented as FSG strength $w = 0.5$) strictly tightens the divergence control term by contracting the feature geometry, effectively reducing the upper bound relative to standard alignment ($\alpha = 0$).
3. **Joint Risk:**

   $$\lambda^* \leq R_S(h) + \hat{R}_{T^{adapt}}(h) + \text{Gen}_T(\mathcal{H}, n_a, \delta) \tag{37}$$

   Minimizing target anchor CE reduces $\hat{R}_{T^{adapt}}$, and MSC reduces Gen_T (Lemma A.9).

B Theoretical Clarifications on Feature-Space Guidance

This section provides a rigorous mathematical justification for the implementation of FSG, bridging the gap between the second-order dynamics discussed in Section 3.3 and the computationally efficient affine rectification used in practice.

B.1 The Gradient Flow of Affine Rectification

In Section 3.3, we characterized the FSG mechanism as optimizing a pre-conditioned objective $\ell(\Psi_\alpha(\mathbf{z})) = \ell(\mathbf{z} + \boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a displacement vector pointing towards the source convex hull. A critical implementation detail is treating $\boldsymbol{\delta}$ as a constant vector during backpropagation (implemented via `.detach()` in PyTorch). Here, we analyze the gradient flow to demonstrate that this operation serves as a first-order proxy for curvature-aware optimization.

Let $\mathbf{z} = f_\theta(x)$ be the feature representation and $\boldsymbol{\delta} = \alpha(\mathbf{a} - \mathbf{z})$ be the guidance vector. The implemented gradient update, denoted as $\mathbf{g}_{impl}$, is computed by backpropagating the loss evaluated at the rectified point $\tilde{\mathbf{z}} = \mathbf{z} + \boldsymbol{\delta}$:

$$\mathbf{g}_{impl} = \nabla_\theta \ell(\tilde{\mathbf{z}}) = \left(\frac{\partial \tilde{\mathbf{z}}}{\partial \theta}\right)^\top \nabla_{\tilde{\mathbf{z}}} \ell(\tilde{\mathbf{z}}) \tag{38}$$

Crucially, because $\boldsymbol{\delta}$ is detached from the computation graph, $\frac{\partial \boldsymbol{\delta}}{\partial \theta} = \mathbf{0}$. Thus, the Jacobian simplifies to the standard feature Jacobian: $\frac{\partial \tilde{\mathbf{z}}}{\partial \theta} = \frac{\partial (\mathbf{z} + \boldsymbol{\delta})}{\partial \theta} = \mathbf{J}$.

Now, we perform a first-order Taylor expansion of the gradient term $\nabla_{\tilde{\mathbf{z}}} \ell(\tilde{\mathbf{z}})$ around the original feature $\mathbf{z}$:

$$\nabla \ell(\mathbf{z} + \boldsymbol{\delta}) = \nabla \ell(\mathbf{z}) + \nabla^2 \ell(\mathbf{z}) \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^2) \tag{39}$$

where $\nabla^2 \ell(\mathbf{z})$ corresponds to the Hessian matrix $\mathbf{H}_\ell(\mathbf{z})$ discussed in the main text.

Substituting Eq. (39) back into Eq. (38), we obtain the effective parameter update dynamics:

$$\mathbf{g}_{impl} \approx \mathbf{J}^\top \left(\nabla \ell(\mathbf{z}) + \mathbf{H}_\ell(\mathbf{z})\boldsymbol{\delta}\right) = \underbrace{\mathbf{J}^\top \nabla \ell(\mathbf{z})}_{\mathbf{g}_{raw}} + \underbrace{\mathbf{J}^\top \mathbf{H}_\ell(\mathbf{z})\boldsymbol{\delta}}_{\mathbf{m}_{geo}} \tag{40}$$

Theoretical Implication: This derivation confirms that the simple affine mixup implemented in our code explicitly recovers the Geometric Momentum ($\mathbf{m}_{geo}$) term defined in Eq. (4) of the main paper. By evaluating the gradient at the shifted location $\mathbf{z} + \boldsymbol{\delta}$, the optimization trajectory *implicitly* incorporates local curvature information (the Hessian-vector product) to rectify the update direction. This validates that our lightweight implementation achieves the theoretical benefits of second-order guidance without incurring the prohibitive $O(N^2)$ computational cost of explicit Hessian evaluation.

C Datasets and Preprocessing Details

In this section, we provide a comprehensive description of the datasets, the signal preprocessing pipeline, the label definition standards, and the chronological data partitioning protocol used in our experiments. These details are critical for ensuring the reproducibility of our results.

C.1 Dataset Specifications

We evaluate our framework on two widely recognized public benchmarks for scalp EEG seizure prediction namely the CHB-MIT Scalp EEG Database covering pediatric subjects and the Siena Scalp EEG Database covering adult subjects. The distinct characteristics of these two datasets ranging from patient age and sampling rates to electrode montages provide a rigorous testbed for assessing cross-subject generalization.

C.1.1 CHB-MIT Dataset

Source: This authoritative dataset was collected at the Children's Hospital Boston in collaboration with the Massachusetts Institute of Technology and is hosted on the reputable PhysioNet repository.

Subjects: Contains long-term EEG recordings from 24 cases (originally described as 23 cases from 22 pediatric subjects, with Chb24 later added as an independent supplementary case), including both males and females. All subjects were monitored for up to several days following the withdrawal of anti-seizure medication.

Signal Acquisition: Recorded using the International 10-20 electrode system with a sampling rate of 256 Hz. The signals utilize a **bipolar montage** to enhance spatial localization.

Subject Exclusion: Following standard protocols to ensure data quality and balance:

- **Excluded:** Chb12 and Chb24 were excluded due to severe class imbalance between pre-ictal and inter-ictal data, whereas Chb14 was excluded due to insufficient data quality for the proposed method.
- **Final Cohort:** 21 cases were used for evaluation.

Statistics: The complete dataset includes over 960 hours of continuous recording containing 198 seizure events. Detailed demographic and seizure information for all cases is provided in Table 2. After applying the exclusion criteria, the final cohort contains 134 seizure events.

C.1.2 Siena Dataset

Source: This dataset was acquired at the Unit of Neurology and Neurophysiology of the University of Siena in Italy and is publicly accessible via the PhysioNet platform.

Subjects: Consists of EEG recordings from 14 adult patients aged between 20 and 71.

Signal Acquisition: Recorded using the International 10-20 system with a sampling rate of 512 Hz. Unlike CHB-MIT, this dataset uses a monopolar montage.

Subject Exclusion:

Table 2: Summary of the CHB-MIT Dataset

Patient	Gender	Age	Seizure count
Chb01	Female	11	7
Chb02	Male	11	3
Chb03	Female	14	7
Chb04	Male	22	4
Chb05	Female	7	5
Chb06	Female	1.5	10
Chb07	Female	14.5	3
Chb08	Male	3.5	5
Chb09	Female	10	4
Chb10	Male	3	7
Chb11	Female	12	3
Chb12	Female	2	40
Chb13	Female	3	12
Chb14	Female	9	8
Chb15	Male	16	20
Chb16	Female	7	10
Chb17	Female	12	3
Chb18	Female	18	6
Chb19	Female	19	3
Chb20	Female	6	8
Chb21	Female	13	4
Chb22	Female	9	3
Chb23	Female	6	7
Chb24	-	-	16
Total	**-**	**-**	**198**

- **Excluded:** PN00 was excluded because recordings consisted of discontinuous or short-duration files which prevented the extraction of valid inter-ictal segments defined by our gap criteria.
- **Final Cohort:** 13 subjects were used for evaluation.

Statistics: Contains over 128 hours of recording with 47 seizure events. Detailed demographic and seizure information is provided in Table 3.

Table 3: Summary of the Siena Dataset

Patient	Gender	Age	Seizure count
PN00	Male	55	5
PN01	Male	46	2
PN03	Male	54	2
PN05	Female	51	3
PN06	Male	36	5
PN07	Female	20	1
PN09	Female	27	3
PN10	Male	25	10
PN11	Female	58	1
PN12	Male	71	4
PN13	Female	34	3
PN14	Male	49	4
PN16	Female	41	2
PN17	Male	42	2

C.2 Signal Preprocessing Pipeline

To mitigate artifacts and standardize the input for the neural network, we applied a uniform preprocessing pipeline to all raw EEG signals before segmentation.

1. **High-Pass Filtering:** A high-pass filter with a cut-off frequency of 1 Hz was applied to eliminate low-frequency drift and DC offsets known as Direct Current artifacts.
2. **Power Line Interference Removal (Notch Filtering):** Distinct band-stop filters were employed based on the geographical origin of the data to remove power line noise and its harmonics.
   - **For CHB-MIT (USA):** Noise was removed within the 57 to 63 Hz and 117 to 123 Hz bands which targets 60 Hz and its 120 Hz harmonic.
   - **For Siena (Europe):** Noise was removed within the 47 to 53 Hz and 97 to 103 Hz bands which targets 50 Hz and its 100 Hz harmonic.
3. **Windowing (Segmentation):** The continuous, filtered EEG signals were segmented into non-overlapping 30-second windows. Each 30-second segment serves as a single input sample $x \in \mathbb{R}^{C \times T}$ for the model.

C.3 Label Definition and Sample Selection

We formulate seizure prediction as a binary classification task: distinguishing between **Pre-ictal** (pre-seizure) and **Inter-ictal** (normal) states. We adhere to the standard SOP/SPH protocol:

- **Seizure Prediction Horizon (SPH):** Set to 5 minutes. This is the "blind interval" immediately preceding the seizure onset where prediction is considered imminent but no warning is triggered to avoid anxiety.
- **Seizure Occurrence Period (SOP):** Set to 30 minutes.
- **Pre-ictal Definition (Positive Class):** Data segments falling within the window from 35 minutes to 5 minutes prior to the seizure onset (T_{onset}), i.e., $[T_{onset} - 35\text{min}, T_{onset} - 5\text{min}]$.
- **Inter-ictal Definition (Negative Class):** To prevent label ambiguity and data contamination, inter-ictal segments are extracted from periods strictly separated from any seizure activity (at least 4 hours away on CHB-MIT, and at least 1.5 hours away on Siena due to shorter total recording times).

C.4 Few-Shot Chronological Adaptation Protocol

To simulate a realistic clinical deployment scenario (Online/Continual Adaptation), we strictly prohibit random shuffling of the target subject's data.

- **Dataset Partitioning:** For the target subject in the Leave-One-Subject-Out (LOSO) scheme:
  - **Adaptation Set ($\mathcal{D}_T^{adapt}$):** Comprises strictly the chronologically first 20% of the available pre-ictal and inter-ictal samples. This simulates the initial calibration phase for a new patient.
  - **Test Set ($\mathcal{D}_T^{test}$):** Comprises the remaining chronologically later 80% of the data. This set is fully unseen during the adaptation phase.
- **Zero Overlap:** There is no temporal overlap between the adaptation and test sets. While this chronological split might partition a single seizure event's sequence (early segments for adaptation, later for testing), it rigorously tests the model's ability to leverage immediate historical context to predict future states.

D Implementation Details and Computational Complexity

To ensure reproducibility and validate the suitability of H-GSC for resource-constrained clinical deployment (e.g., wearable EEG monitoring devices), we provide detailed specifications of the network architecture, computational efficiency analysis, and hyperparameter configurations.

D.1 Network Architecture: LiteEEG-Net

The backbone network, LiteEEG-Net, is a highly optimized 1D Convolutional Neural Network designed to extract temporal-spectral features from multi-channel EEG signals. The detailed architecture is presented in Table 4. The model takes raw EEG signals as input ($C = 22, T = 7680$) and maps them to a compact embedding space ($D = 40$). We explicitly incorporate Batch Normalization (BN) and Dropout layers to enhance generalization capabilities on limited subject-specific data.

Table 4: Detailed architecture of LiteEEG-Net. The input tensor shape is $(B, 22, 7680)$. The parameter counts include weights and biases for convolutional, linear, and normalization layers. The total parameter count is strictly verified via the `thop` library.

Stage	Layer type	Kernel / stride / pad	Input shape	Output shape	Params	Components
Input	-	-	$(22, 7680)$	-	-	-
Block 1	Conv1d	15 / 2 / 7	$(22, 7680)$	$(32, 3840)$	10,656	Conv+BN
	MaxPool	4 / 4 / 0	$(32, 3840)$	$(32, 960)$	0	-
	Dropout	$p = 0.2$	-	-	0	ELU activation
Block 2	Conv1d	7 / 2 / 3	$(32, 960)$	$(64, 480)$	14,464	Conv+BN
	MaxPool	2 / 2 / 0	$(64, 480)$	$(64, 240)$	0	-
	Dropout	$p = 0.2$	-	-	0	ELU activation
Block 3	Conv1d	3 / 1 / 1	$(64, 240)$	$(40, 240)$	7,800	Conv+BN
	AdaptiveAvgPool	Output=48	$(40, 240)$	$(40, 48)$	0	-
	BatchNorm	-	-	-	(Included)	-
Classifier	Temporal mean	-	$(40, 48)$	(40)	0	-
	LayerNorm	-	(40)	(40)	80	-
	Linear (proj)	-	(40)	(40)	1,640	GELU, dropout(0.3)
	Linear (pred)	-	(40)	(2)	82	-
Total	-	-	-	-	**34,786**	-

D.2 Implementation Details of the Masking Operator $\mathcal{M}$

To ensure reproducibility, we specify the exact configuration of the masking operator $\mathcal{M}$. Our implementation adopts a Time-Patch Zero-filling Strategy. Given the input EEG signal with a temporal dimension of $T = 7680$, we set the patch length to $L = 40$, thereby discretizing the signal into a sequence of $N = T/L = 192$ non-overlapping patches. During training, we generate a binary Bernoulli mask $m \in \{0, 1\}^N$ with a probability of $r = 0.3$ and strictly zero-fill all time points within the selected patches. This block-wise, rather than point-wise, masking strategy effectively disrupts local short-range correlations, compelling the model to leverage the remaining 70% of context for global semantic reconstruction, thereby significantly reinforcing the local Lipschitz continuity of the feature representation.

D.3 Computational Efficiency Analysis

A critical requirement for real-world seizure prediction is the ability to perform real-time inference on edge hardware without reliance on cloud computing. We benchmarked our model using both a high-end GPU (NVIDIA RTX 4090 D) and a Laptop CPU in Single-Thread Mode to simulate resource-constrained portable environments.

As reported in Table 5, LiteEEG-Net demonstrates exceptional efficiency. With a model size of only **0.13 MB** and **0.0999 GFLOPs**, it is well within the operating limits of low-power embedded processors (e.g., ARM Cortex-M or mobile chips).

Table 5: Computational efficiency benchmarks. Inference time is averaged over 100 runs with a batch size of 1. "Laptop CPU (single thread)" restricts execution to a single core to mimic the sequential processing capability of portable/wearable microcontrollers.

Metric	Value	Unit	Implication
Total parameters	**34,786**	#	Fits in L1/L2 cache
FLOPs	**0.0999**	GFLOPs	Ultra-low battery consumption
Model storage	**0.13**	MB	Negligible memory footprint
Latency (RTX 4090 D)	0.49	ms	High-throughput cloud scenario
Latency (Laptop CPU single-thread)	**3.93**	ms	Real-time wearable scenario

Deployment Feasibility: In a clinical setting, EEG signals are typically analyzed in 30-second windows. Our model requires only **3.93 ms** to process one window on a single CPU core. This

represents a CPU utilization rate of approximately $\mathbf{0.013}\%$ ($3.93\text{ms}/30,000\text{ms}$), leaving ample computational resources for signal acquisition, Bluetooth transmission, and system background tasks. This confirms that H-GSC is ready for immediate deployment on portable, battery-powered BCI devices.

D.4 Hyperparameter Configuration

We maintain a fixed set of hyperparameters across all subjects (CHB-MIT and Siena) and all experimental protocols (Intra- and Cross-Dataset) to ensure fair evaluation and reproducibility. The detailed configuration used in our experiments is listed in Table 6.

Table 6: Hyperparameter configuration for H-GSC.

Category	Hyperparameter	Value
Optimization	Optimizer	AdamW
	Learning rate	1×10^{-4}
	Weight decay	5×10^{-4}
	Batch size	32
Adaptation	Adaptation start epoch	40
	Total epochs	100
	Guidance scale (w)	0.5
	Loss weight λ_{lmmd}	0.6
	Loss weight λ_{msc} (MSC)	0.5
	Loss weight λ_{anchor} (target CE)	1.0
Structural	Mask ratio	0.3
	Validator temperature (τ)	0.05
	Dropout (feature / head)	0.2 / 0.3

E Evaluation Metrics

All performance evaluations in this study are conducted within a segment-based framework. We employ four standard metrics derived from the confusion matrix components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

- **Recall:** Also known as Sensitivity, this metric measures the ability of the model to correctly identify true preictal samples: Recall $= \frac{TP}{TP+FN}$.
- **Precision:** Measures the reliability of alarms generated: Precision $= \frac{TP}{TP+FP}$.
- **Specificity:** Measures the ability to correctly identify true interictal samples: Specificity $= \frac{TN}{TN+FP}$.
- **AUC:** The Area Under the Receiver Operating Characteristic Curve provides threshold-independent assessment.

F Hyperparameter Sensitivity Analysis

To verify the robustness of H-GSC against hyperparameter variations and to empirically validate the theoretical claims regarding "Geometric Overshoot" and structural constraints, we conducted a comprehensive sensitivity analysis on the CHB-MIT dataset using the LOSO protocol. We specifically examined five critical hyperparameters: the Guidance Scale (α), the Mask Ratio (r), the Semantic Anchor Weight (λ_{anchor}), the Batch Size, and the Anchor Ratio.

F.1 Impact of Guidance Scale (α)

The guidance scale α (implemented as `guidance_scale`) controls the intensity of the affine contraction mapping in the FSG module. As presented in Table 7, the model exhibits a clear trade-off

curve. At low values ($\alpha = 0.1$), the geometric momentum is insufficient to fully rectify the trajectory, yielding an AUC of 0.7887. The performance remains robust within the range $\alpha \in [0.5, 0.7]$, achieving its empirical peak AUC of 0.8016 at $\alpha = 0.7$. However, we deliberately select $\alpha = 0.5$ as the default configuration across all experiments. This is because aggressively tuning the parameter to the absolute maximum on CHB-MIT ($\alpha = 0.7$) risks overfitting to this specific data distribution, potentially compromising the model's stability and generalization in more severely heterogeneous scenarios, such as the independent Siena dataset and the joint cross-dataset evaluations. Furthermore, when α is aggressively increased to 0.9, although Recall peaks at 82.35%, Specificity notably declines to 75.10%. This decline confirms our theoretical analysis regarding Geometric Overshoot: excessive guidance forces target features too aggressively across decision boundaries, increasing false alarms despite higher sensitivity.

Table 7: Sensitivity Analysis of Guidance Scale (α). The optimal range is observed between 0.5 and 0.7, validating that moderate geometric rectification is superior to both weak and excessive guidance.

Scale (α)	Rec	Pre	Spe	AUC
0.1	0.8001	0.7903	0.7774	0.7887
0.3	0.7641	0.7991	0.7841	0.7741
0.5	0.7991	0.8100	0.7912	0.7954
0.7	0.8110	0.8128	0.7922	0.8016
0.9	0.8235	0.7902	0.7510	0.7872

F.2 Impact of Mask Ratio (r)

The mask ratio r determines the difficulty of the pretext task in MSC. Table 8 demonstrates that a ratio of $r = 0.3$ yields the best trade-off (AUC 0.7954), effectively promoting the learning of robust global semantics. A low ratio ($r = 0.1$) provides insufficient regularization (AUC 0.7843). Conversely, aggressive masking ($r \geq 0.7$) degrades overall AUC (≈ 0.78). While Recall remains high in these settings ($> 82\%$), Specificity drops significantly, indicating that destroying critical temporal dependencies biases the model towards the positive class at the expense of discriminative precision.

Table 8: Sensitivity Analysis of Mask Ratio (r). Performance peaks at $r = 0.3$, indicating that the network benefits most from moderate structural perturbations.

Ratio (r)	Rec	Pre	Spe	AUC
0.1	0.8090	0.7882	0.7597	0.7843
0.3	0.7991	0.8100	0.7912	0.7954
0.5	0.8055	0.8028	0.7803	0.7929
0.7	0.8218	0.7783	0.7384	0.7801
0.9	0.8233	0.7745	0.7362	0.7797

F.3 Impact of Target Anchor Weight (λ_{anchor})

The parameter λ_{anchor} (implemented as `lambda_target_ce`) balances global statistical alignment (LMMD) with instance-level semantic supervision. As detailed in Table 9, removing anchors ($\lambda_{anchor} = 0.0$) results in high Recall (82.55%) but poor Specificity (74.84%), validating the necessity of semantic anchors for maintaining boundary sharpness. The optimal performance is observed at $\lambda_{anchor} = 1.0$, which achieves the best synergy. However, increasing the weight beyond 1.5 leads to a gradual performance decline; notably, Recall drops sharply to 76.99% at $\lambda = 2.0$. This indicates that excessive weight on the few-shot adaptation set leads to overfitting, hindering the model's ability to generalize to unseen seizure events.

Table 9: Sensitivity Analysis of Target Anchor Weight (λ_{anchor}). A balanced weight of 1.0 achieves the best synergy between alignment and discrimination.

λ_{anchor}	Rec	Pre	Spe	AUC
0.0	0.8255	0.7819	0.7484	0.7870
0.5	0.8205	0.7827	0.7457	0.7831
1.0	0.7991	0.8100	0.7912	0.7954
1.5	0.8086	0.7823	0.7556	0.7821
2.0	0.7699	0.8093	0.7903	0.7801

F.4 Optimization Sensitivity to Batch Size

Since LMMD and FSG rely on batch-level statistics, we evaluated optimization stability across different batch sizes. As shown in Table 10, H-GSC is highly robust, performing consistently from a batch size of 8 to 64, with the optimal AUC observed at the default size of 32. This indicates that randomized mapping effectively approximates domain distributions even with smaller mini-batches.

Table 10: Sensitivity Analysis of Batch Size. H-GSC maintains robust optimization stability across varying batch sizes, peaking at 32.

Batch Size	Rec	Pre	Spe	AUC
8	0.7915	0.8064	0.7790	0.7853
16	0.8347	0.7906	0.7496	0.7922
32	0.7991	0.8100	0.7916	0.7954
64	0.7918	0.7994	0.7895	0.7906

F.5 Performance Scaling with Semantic Anchors

H-GSC scales robustly under sparse supervision. As demonstrated in Table 11, performance degrades gracefully as the anchor ratio decreases. Remarkably, with only 5% calibration data, the model maintains a functional AUC of 0.7330, validating the efficacy of our geometric and structural modules. We selected 20% as the default adaptation split to ensure high clinical reliability across heterogeneous populations.

Table 11: Performance Scaling with Anchor Ratio. The model degrades gracefully under sparse supervision, remaining functional even with only 5% calibration data.

Anchor Ratio	Rec	Pre	Spe	AUC
5%	0.7547	0.7565	0.7113	0.7330
10%	0.8092	0.7644	0.7230	0.7661
15%	0.7785	0.7971	0.7758	0.7772
20%	0.7991	0.8100	0.7916	0.7954

G Detailed Comparative Experiment Analysis

In this section, we provide the full numerical results and a detailed analysis of H-GSC compared to baseline methods, utilizing the metrics defined in Appendix E.

G.1 Intra-Dataset Performance

As presented in Table 12, H-GSC establishes new state-of-the-art benchmarks across both CHB-MIT and Siena datasets. This superiority is demonstrated on the most comprehensive patient cohorts

available, including 21 subjects for CHB-MIT and 13 for Siena. Unlike methods such as GSFDA and Meta+Adv VAE which validate on smaller subsets to reduce variability, H-GSC effectively overcomes subject-specific variations through hierarchical alignment despite the increased challenge posed by the severe physiological heterogeneity of these larger populations.

This advantage is most evident in the substantial improvement in Recall, which serves as the premier safety metric in seizure prediction. On the CHB-MIT dataset, H-GSC achieves a Recall of 79.91%, significantly outperforming the 71.16% of FSG-Zero. On the Siena dataset, we observe a distinct performance trade-off where H-GSC outperforms the baseline FSG-Zero in both AUC by 1.63 percentage points and Recall by 14.24 percentage points, while exhibiting lower Precision and Specificity. This phenomenon represents a fundamental shift in the decision paradigm from passive conservation to active sensitivity. The baseline FSG-Zero exhibits high Specificity of 80.81% but a dangerously low Recall of 53.08%, suggesting it adopts a conservative strategy under high heterogeneity that tends to reject ambiguous samples to minimize errors at the cost of missing nearly half of the seizure events. In contrast, H-GSC employs explicit Feature-Space Guidance to implement an aggressive geometric rectification that actively pulls target features towards the source seizure manifold. This strategy effectively activates the model response to potential seizure patterns and results in a massive boost in Recall. Although this aggressive alignment inevitably pulls some hard negative samples near the boundary into the positive class and consequently lowers Precision and Specificity, the overall improvement in AUC confirms that the global capacity of the model to separate classes has actually increased. Given that a false negative carries life-threatening risks in clinical warning systems whereas a false positive is generally tolerable, the operating point selected by H-GSC is clinically more viable than the conservative baseline.

Table 12: Detailed comparison of different methods on CHB-MIT and Siena datasets. H-GSC improves AUC and Recall, with a trade-off in Precision and Specificity on Siena.

Method	**Dataset**	**No. of patients**	**Rec(%)**	**Pre(%)**	**Spe(%)**	**AUC**
Domain-Adapt STFT (Peng et al., 2022)	CHB-MIT	16	76.00	-	-	0.7500
GSFDA (Zhao et al., 2023)	CHB-MIT	15	70.50	-	61.90	0.6830
FT3A (Mao et al., 2024)	CHB-MIT	20	54.40	-	66.00	0.6510
NSFA-Net (Dong et al., 2025)	CHB-MIT	19	-	-	-	0.6460
FSG-Zero (Zhang et al., 2026)	CHB-MIT	21	71.16	77.72	72.97	0.7207
H-GSC (Ours)	**CHB-MIT**	**21**	**79.91**	**81.00**	**79.16**	**0.7954**
Meta+Adv VAE+MMD (Liao et al., 2025)	Siena	6	-	-	76.30	-
FSG-Zero (Zhang et al., 2026)	Siena	13	53.08	74.07	80.81	0.6694
H-GSC (Ours)	**Siena**	**13**	**67.32**	65.96	69.82	**0.6857**

G.2 Joint Cross-Dataset Adaptation Performance

Furthermore, the Joint Cross-Dataset Adaptation results presented in Table 13 serve as a rigorous stress test that highlights the robustness of our framework under extreme distributional shifts involving variations in patient age demographics, sampling rates, and electrode montages. While the state-of-the-art FSG-Zero succumbs to severe negative transfer due to its inability to reconcile these disparities, H-GSC demonstrates robust positive transfer.

The quantitative results reveal comprehensive improvements across all key metrics where gains in Recall and AUC are particularly pronounced. As evidenced by the average performance metrics, H-GSC boosts Recall by over 14 percentage points as it surges from 60.96% to 75.17% and elevates the average AUC from 0.6701 to 0.7666. Notably, unlike the single-dataset scenario, the average Specificity also improves from 75.10% to 78.14% in this cross-dataset setting. This confirms that by strictly enforcing a rectified target-to-source trajectory, our method successfully bridges massive distribution gaps and effectively translates the increased scale of diverse data into significant generalization gains rather than interference.

Table 13: Joint Cross-Dataset Adaptation Results. H-GSC significantly outperforms FSG-Zero on average, turning negative transfer into positive gains.

Dataset	Method	Rec(%)	Pre(%)	Spe(%)	AUC
CHB-MIT	FSG-Zero	67.59	59.21	74.18	0.7089
	Ours	**82.74**	**67.05**	**80.76**	**0.8175**
Siena	FSG-Zero	50.26	70.81	76.02	0.6314
	Ours	**62.95**	**72.00**	73.91	**0.6843**
Average	FSG-Zero	60.96	65.01	75.10	0.6701
	Ours	**75.17**	**68.94**	**78.14**	**0.7666**

H Ablation Study on Siena Dataset

To further demonstrate the generalizability of our framework, we conducted a comprehensive ablation study on the independent Siena dataset, as detailed in Table 14. While the baseline FSG-Zero suffers from a severe conservative bias yielding an unacceptable recall of 53.08%, the full model ultimately improves Recall and AUC, although the performance of intermediate variants is not monotonic. The complete integration of our proposed validator and structural constraint components effectively corrects this decision boundary imbalance. Ultimately, the full H-GSC framework improves the recall to 67.32% and achieves an AUC of 0.6857, successfully preventing majority-class collapse. These consistent dataset-agnostic improvements mirror our CHB-MIT findings, proving that the model does not overfit to a single dataset but fundamentally enhances cross-subject generalization under severe heterogeneity.

Table 14: Ablation Study on the Siena dataset. Method 1: Validator; Method 2: +Target CE; Method 3: +MSC; Method 4: H-GSC (Full).

Method	Spe(%)	Pre(%)	Rec(%)	AUC
Baseline (FSG-Zero)	80.81	74.07	53.08	0.6694
Method 1	75.60	65.78	51.55	0.6358
Method 2	75.70	68.34	60.24	0.6797
Method 3	71.85	65.89	60.87	0.6637
Method 4 (H-GSC)	**69.82**	**65.96**	**67.32**	**0.6857**

I Mechanistic Analysis and Geometric Verification

While the quantitative results demonstrate the superiority of H-GSC, this section aims to empirically verify the theoretical claims established in Section 3. We analyze the underlying mechanisms to confirm that the performance gains stem fundamentally from rectified gradient geometry and preserved manifold structure.

I.1 Verification of Gradient Rotation

To empirically verify the theoretical claims of Geometric Pre-conditioning (Section 3.3) and Hierarchical Gradient Alignment (Section 3.4), we tracked the cosine similarity between the source discriminative gradient $\mathbf{g}_{src}$ and adaptation gradients throughout the training process, visualizing three distinct trajectories: the Baseline (Orange) comparing $\mathbf{g}_{src}$ vs. $\mathbf{g}_{raw}$; the FSG-only Ablation (Gray) comparing $\mathbf{g}_{src}$ vs. $\mathbf{g}_{fsg}$; and the Proposed H-GSC (Blue) comparing $\mathbf{g}_{src}$ vs. the final projected $\mathbf{g}_{adapt}$ (Figure 2).

Observations reveal that while the Baseline fluctuates significantly around zero with frequent negative spikes, thereby empirically confirming the gradient orthogonality or conflict between naive distribution alignment and source discriminability, the FSG-only variant critically exhibits increased volatility and occasionally deeper negative conflicts. This counter-intuitive phenomenon validates the nuance of our geometric perspective: while FSG successfully acts as a Geometric Pre-conditioner by

actively rotating the trajectory to overcome optimization inertia, unconstrained geometric pulling on non-convex manifolds leads to "Geometric Overshoot," where aggressive guidance forces target features across decision boundaries into incorrect semantic regions.

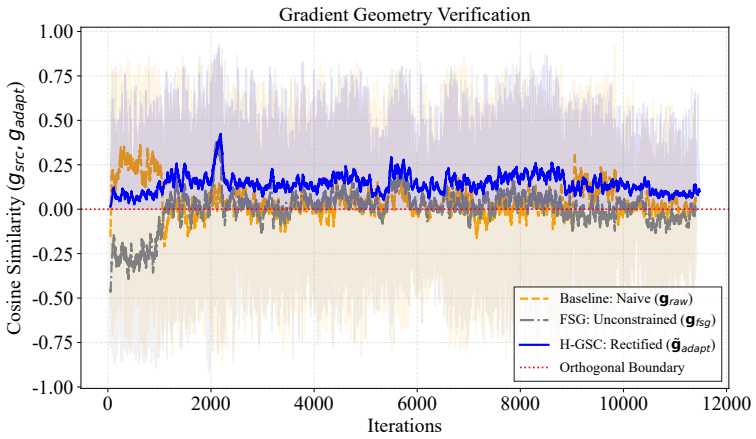


Figure 2: Gradient geometry verification throughout the adaptation process. Compared to the baseline ($\mathbf{g}_{raw}$) which exhibits gradient orthogonality or conflict and the unconstrained FSG ($\mathbf{g}_{fsg}$) which suffers from geometric overshoot, H-GSC maintains a consistently positive cosine similarity with the source discriminative gradients.

In stark contrast, the Full H-GSC framework maintains a consistently positive alignment ($\cos\theta > 0$), proving that the H-GA mechanism functions as an effective safety valve. By projecting the high-momentum updates generated by FSG onto the Source-Protected Subspace, H-GSC filters out the destructive components observed in the Gray line while retaining constructive guidance, thus confirming that FSG provides the necessary optimization dynamics while H-GA ensures structural safety, validating the indispensability of the hierarchical design.

I.2 Spectral Analysis of Manifold Structure

To investigate the role of MSC in preserving manifold geometry, we compared H-GSC against its ablation variant (w/o MSC). The Masked Consistency Analysis (Figure 3) reveals the critical impact of MSC on local stability. This histogram estimates the distribution of cosine similarities between original test features and their counterparts under 30% random masking perturbation (He et al., 2022), where the vertical axis represents the probability density of samples.

As shown, the model without MSC (Blue) displays a heavy-tailed distribution with similarities dropping as low as approximately 0.850, suggesting a high sensitivity to local evidence and potential overfitting to specific patches. In stark contrast, the distribution of H-GSC (Red) is concentrated near 0.98, effectively eliminating the long tail. This empirical evidence confirms that MSC induces a Lipschitz-like local stability on the perturbation manifold. By suppressing the Jacobian energy of the representation mapping with respect to structured missingness, MSC compels the model to learn global, redundant semantic anchors, thereby achieving superior Geometric Rigidity and robustness against input corruption.

I.3 Visualizing the Decision Boundary

To empirically verify the necessity of semantic anchors in resolving Discriminative Ambiguity, we visualize the learned feature manifolds using t-SNE. Specifically, we compare H-GSC against its ablation variant (w/o Anchor). As shown in Figure 4, the manifold of w/o Anchor (Figure 4a) exhibits severe Boundary Diffusion: despite statistical alignment, the inter-ictal (Blue) and pre-ictal (Red) classes remain intertwined at the margins. This confirms that global statistical matching lacks the instance-level force to establish strict margins, resulting in decision boundaries traversing high-density regions.

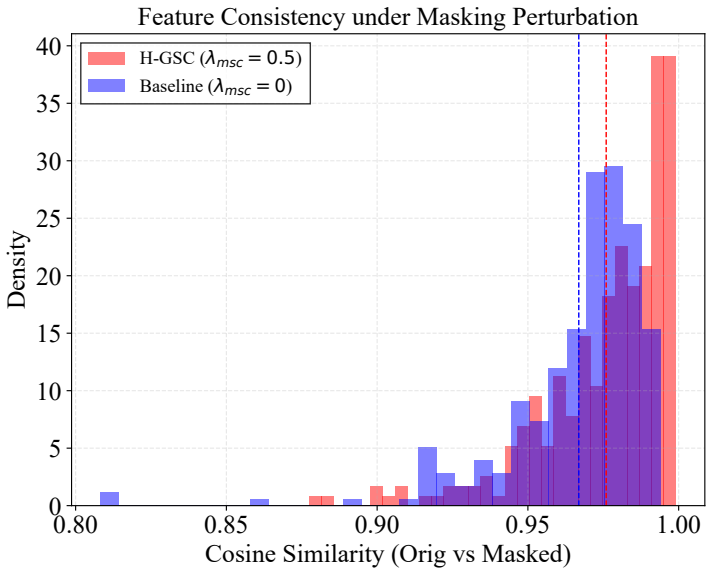
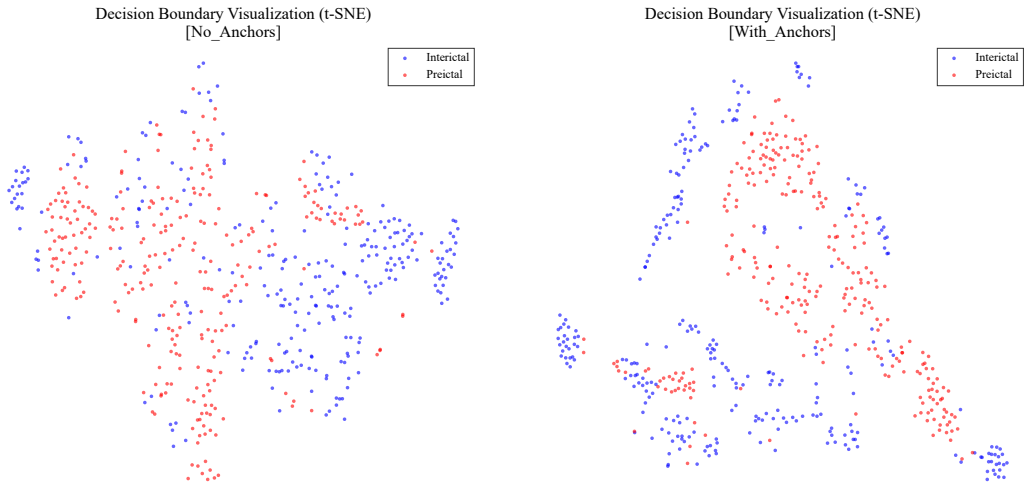


Figure 3: Feature consistency under 30% masking perturbation. H-GSC exhibits superior geometric rigidity with a concentrated distribution near 0.98, eliminating the heavy-tailed instability found in the baseline without MSC.

In contrast, the full H-GSC (Figure 4b) demonstrates well-separated clusters. The introduction of sparse semantic anchors, functioning conceptually as "discriminative pivots," effectively locks samples to their semantic regions, untangling the manifold into clear, large-margin clusters. These results validate that explicit discriminative supervision is a critical constraint for ensuring structure-preserving alignment and preventing the semantic ambiguity common in pure distribution adaptation.



(a) Without Anchors

(b) With Anchors (H-GSC)

Figure 4: t-SNE visualization of learned feature manifolds. (a) Boundary diffusion and class intertwining are prevalent in the variant without semantic anchors. (b) In the full H-GSC framework, sparse semantic anchors function as discriminative pivots to ensure sharp margins and structurally distinct clusters.