COLLABORATING SHARED AND ROUTED EXPERTS FOR CROSS-SUBJECT EEG GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject EEG generalization is challenging due to substantial heterogeneity across subjects. Existing methods typically learn either a shared subject-invariant model or multiple subject-specialized experts, but these two paradigms fail in complementary ways: the former may over-reduce subject-specific discriminative signals, while the latter may under-reduce transferable structure. We show that their suitability depends on the reducibility cost of branch-specific functions to branch-invariant ones, and we further provide a theory-to-method mapping that instantiates alignment principles in cross-subject EEG learning. Based on this insight, we propose Shared-Routed Expert Alignment (SREA), a collaborative framework that couples a shared expert for reducible invariant functions with routed experts for irreducible subject-specific functions. SREA trains the shared branch with joint embedding over augmented temporal neighbors, the routed branch with prototype-based sparse routing and expert specialization, and both branches with numerically stable mutual-guided reweighting based on cross-branch learnability gaps. Experiments on seven public EEG benchmarks across different tasks show that SREA consistently outperforms state-of-the-art methods and EEG foundation models.

1 INTRODUCTION

Electroencephalography (EEG) decoding plays an important role in brain-computer interfaces, computational neuroscience, and neurophysiological assessment (Schalk et al., 2004). Recent advances in deep learning have substantially improved EEG modeling across tasks such as auditory attention decoding, motor imagery classification, and sleep staging (Schirrmeister et al., 2017; Lawhern et al., 2018; Song et al., 2022; Fan et al., 2025). More recently, large-scale pretraining has led to EEG foundation models, such as LaBraM, EEGPT, CBraMod, and CSBrain, which learn transferable representations from multi-task and multi-dataset corpora (Jiang et al., 2024; Wang et al., 2024a; 2025; Zhou et al., 2025). These models have significantly strengthened generic EEG representation learning and improved transfer across tasks and datasets. However, despite these advances, cross-subject generalization remains a major bottleneck. Models trained on a set of source subjects often degrade markedly when transferred to unseen target subjects (Handiru & Prasad, 2016; Wang et al., 2024e).

Cross-subject generalization has long been a central challenge in EEG decoding, because neural signals vary substantially across individuals due to differences in brain anatomy, physiology, recording conditions, and behavioral strategies (Handiru & Prasad, 2016; Wang et al., 2024e; Bomatter & Gouk, 2025). Earlier efforts addressed this problem through handcrafted features, channel selection, and shallow transfer techniques (Handiru & Prasad, 2016). Existing approaches to cross-subject EEG learning largely follow two paradigms. The first learns a shared subject-invariant model, typically by encouraging representations to remove subject-specific variations so that a single classifier can transfer across subjects (Li et al., 2018; Ma et al., 2019; Shen et al., 2022; Wang et al., 2024c;e). This strategy is effective when the discriminative structure is largely shared, but it can become problematic when subject heterogeneity is strong: forcing all subjects into a single invariant representation may suppress subject-specific but task-relevant cues, leading to over-reduction. The second paradigm uses multiple specialized experts, such as ensembles or mixture-of-experts architectures, to preserve heterogeneous decision patterns across subjects or subject groups (Zhang et al., 2024a; Wang et al., 2024d; Zhao et al., 2024a; Li et al., 2023). While this strategy can better capture

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.

local specialization, it may rely too heavily on branch-specific structure, fail to extract transferable regularities, and introduce routing uncertainty on unseen subjects, resulting in under-reduction.

In this paper, we formalize cross-subject EEG generalization using algorithmic alignment under distribution shift (Xu et al., 2020). Our theoretical analysis shows that a shared expert is better aligned with branch-invariant functions, whereas routed experts are better aligned with branch-dependent functions. More importantly, their relative suitability depends on the reducibility cost of transforming branch-specific functions into branch-invariant ones. Based on this insight, we propose Shared-Routed Expert Alignment (SREA), a collaborative framework that couples a shared branch for reducible invariant structure with routed experts for irreducible subject-specific structure, and bridges them via mutual-guided reweighting driven by cross-branch learnability gaps. Unlike the first paradigm, SREA does not assume all useful variation should be absorbed into one common representation; unlike the second, it does not rely solely on independent expert aggregation. Instead, SREA lets the two branches explicitly supervise each other to dynamically generalize on different cases.

In summary, our contributions are as follows:

- We analyze cross-subject EEG generalization from an algorithmic alignment perspective and show that the suitability of shared and routed experts depends on the reducibility cost from branch-specific to branch-invariant functions.
- We propose SREA, a theory-driven framework that aligns a shared expert with reducible invariant structure and routed experts with irreducible subject-specific structure.
- We introduce a novel mutual-guided reweighting mechanism that transfers supervision across branches according to learnability gaps, thereby mitigating over-reduction and under-reduction.

2 Theoretical Analysis

In this section, we formalize cross-subject EEG generalization through algorithmic alignment (Xu et al., 2020), aiming to characterize how network architecture influences generalization under subject-induced distribution shift.

2.1 Architecture Alignment Impacts Cross-Subject Generalization

Definition 1. *Let $\mathcal{N}$ denote a neural network composed of n modules $\{\mathcal{N}_i\}_{i=1}^n$, and assume that a target function g can be decomposed into n sub-functions such that $g = f_n \circ \cdots \circ f_1$. We say that $\mathcal{N}$* aligns *with g if replacing each module $\mathcal{N}_i$ with f_i yields the same output as g. The alignment measure is defined as $\mathrm{Align}(\mathcal{N}, g, \epsilon, \delta) := n \cdot \max_{i\in[n]} \mathcal{M}(f_i, \mathcal{N}_i, \epsilon, \delta)$, where $\mathcal{M}(f_i, \mathcal{N}_i, \epsilon, \delta)$ denotes the sample complexity for $\mathcal{N}_i$ to learn f_i with precision ϵ and failure probability δ.*

Definition 1 shows that a well-aligned architecture decomposes the problem into simpler subproblems, reducing sample complexity. We now extend algorithmic alignment (Xu et al., 2020; Li et al., 2023) to domain generalization under distribution shift.

Assumption 1. *Let p_{tr} and p_{te} denote the training and test feature distributions at the output of $\mathcal{N}_1$. There exist constants $C \geq 1$ and $\eta \geq 0$ such that for any measurable set A, $\mathbb{P}_{D_{\mathrm{te}}}[\mathcal{N}_1(\boldsymbol{x}) \in A] \leq C \cdot \mathbb{P}_{D_{\mathrm{tr}}}[\mathcal{N}_1(\boldsymbol{x}) \in A] + \eta$.*

Remark. Subject-invariant physiological patterns ensure a bounded density ratio C, while training on diverse subjects yields a reasonable transfer error η (Berry et al., 2017; Bomatter & Gouk, 2025).

Assumption 2. *There exists a function g_c such that $g_c(\mathcal{N}_1(\boldsymbol{x})) = y$ for all $\boldsymbol{x} \in \mathcal{E}_{\mathrm{tr}}$, and $\mathbb{P}_{D_{\mathrm{te}}}[\|g_c(\mathcal{N}_1(\boldsymbol{x})) - y\| \leq \epsilon] \geq 1 - \delta$.*

Remark. The function g_c represents correlations that persist across both training and test distributions.

Assumption 3. *There exists a function g_s such that $g_s(\mathcal{N}_1(\boldsymbol{x})) = y$ for all $\boldsymbol{x} \in \mathcal{E}_{\mathrm{tr}}$, and $\mathbb{P}_{D_{\mathrm{te}}}[\|g_s(\mathcal{N}_1(\boldsymbol{x})) - y\| > \omega(\epsilon)] \geq 1 - \delta$.*

Remark. The function g_s represents correlations present in training but absent in test distributions.

Theorem 1. *Let $\mathcal{N}' = \{\mathcal{N}_2, \ldots, \mathcal{N}_n\}$. Suppose we train the neural network with ERM and Assumptions 1–3 hold. Then:*

(a) If $\text{Align}(\mathcal{N}', g_c, \epsilon, \delta) \leq |\mathcal{E}_{\text{tr}}|$*, then* $\mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - y\| \leq O(\epsilon)] \geq 1 - O(\delta) - \eta$.

(b) If $\text{Align}(\mathcal{N}', g_s, \epsilon, \delta) \leq |\mathcal{E}_{\text{tr}}|$*, then* $\mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - y\| > \omega(\epsilon)] \geq 1 - O(\delta) - \eta$.

Remark. Networks aligned with invariant correlations generalize robustly, while those aligned with spurious correlations fail on unseen domains.

2.2 Reducibility Cost Impacts the Choice of Shared and Routed Architecture

Definition 2. *Let g denote a target function with a branching structure $g(\boldsymbol{x}) = \sum_{j=1}^{M} \mathbf{1}_{I_j}(h_0(\boldsymbol{x})) \cdot h_j(\boldsymbol{x})$, where h_0 is a branching function that determines which branch to activate, $\{I_j\}_{j=1}^{M}$ are disjoint intervals partitioning the range of h_0, and $\{h_j\}_{j=1}^{M}$ are branch-dependent functions. A routed expert network $\mathcal{R}$ consists of M routed experts $\{R_j\}_{j=1}^{M}$ and a router R_0, where each routed expert R_j learns h_j and the router R_0 learns h_0. A shared expert network $\mathcal{S}$ consists of a single shared expert S_0 that learns the entire function g.*

Remark. The branching structure captures branch-dependent processing strategies, which commonly arise under distribution shift due to subject heterogeneity.

Theorem 2. *Let $\mathcal{M}(R_j, h_j, \epsilon, \delta)$ and $\mathcal{M}(R_0, h_0, \epsilon, \delta)$ denote the sample complexity for routed expert R_j to learn h_j and for router R_0 to learn h_0, respectively, with precision ϵ and failure probability δ. Let $\mathcal{M}(S_0, g, \epsilon, \delta)$ denote the sample complexity for shared expert S_0 to learn g. Let α, β with $1 \leq \alpha \leq \beta$ satisfy $\alpha \cdot \max_j \mathcal{M}(R_j, h_j, \epsilon, \delta) \leq \mathcal{M}(S_0, g, \epsilon, \delta) \leq \beta \cdot \max_j \mathcal{M}(R_j, h_j, \epsilon, \delta)$. Suppose the branching function is no harder to learn than the branch-dependent functions, i.e., $\mathcal{M}(R_0, h_0, \epsilon, \delta) \leq \max_j \mathcal{M}(R_j, h_j, \epsilon, \delta)$. Then:*

(a) If $\alpha \geq M + 1$, then $\text{Align}(\mathcal{R}, g, \epsilon, \delta) \leq \text{Align}(\mathcal{S}, g, \epsilon, \delta)$.

(b) If $\beta \leq M + 1$, then $\text{Align}(\mathcal{R}, g, \epsilon, \delta) \geq \text{Align}(\mathcal{S}, g, \epsilon, \delta)$.

Remark. The ratio α (or β) reflects the reducibility cost of branch-specific functions to a branch-invariant one. When this cost is low (small β), the shared expert aligns better. When it is high (large α), routed experts achieve better alignment.

2.3 Collaboration Adapts to Different Cases

Assumption 4. *For any sample $\boldsymbol{x}$, at least one expert produces a prediction close to the target function, i.e.,* $\min\{\|S(\boldsymbol{x}) - g(\boldsymbol{x})\|, \|R(\boldsymbol{x}) - g(\boldsymbol{x})\|\} \leq \varepsilon$*, where ε is a small constant.*

Remark. This follows from Theorem 2: at least one expert aligns well for each sample.

Theorem 3. *Define the alignment errors of the shared expert and routed experts at epoch t as $\epsilon_S^{(t)} = \|S^{(t)}(\boldsymbol{x}) - g(\boldsymbol{x})\|^2$ and $\epsilon_R^{(t)} = \|R^{(t)}(\boldsymbol{x}) - g(\boldsymbol{x})\|^2$, respectively, where $g(\boldsymbol{x})$ denotes the target function. Under Assumption 4, the total alignment error satisfies $\mathbb{E}[\epsilon_S^{(t)} + \epsilon_R^{(t)}] \leq \mathbb{E}[\epsilon_S^{(0)} + \epsilon_R^{(0)}]$.*

Remark. The expert with higher loss learns from the better-performing one, avoiding over-reduction and under-reduction.

These results motivate a framework that aligns a shared branch with reducible invariant structure, a routed branch with irreducible subject-specific structure, and a collaborative mechanism that allows the two to correct each other.

3 Methodology

Fig. 1 illustrates the overall framework. Given an EEG segment $\boldsymbol{x}_t$ for subject U_i in trial j, SREA extracts representations through two separate feature extractors: $\mathcal{F}_1$ for the shared branch and $\mathcal{F}_2$ for the routed branch. The shared branch trains a shared expert E_0 to model stable cross-subject structure using empirical risk minimization together with a joint embedding objective over augmented temporal neighbors. The routed branch uses a router R with prototype-based sparse routing to dispatch samples to specialized routed experts $\{E_j\}_{j=1}^{M}$, while subject-level specialization and load balancing encourage consistent yet non-collapsed expert usage. Finally, the two branches are jointly optimized through mutual-guided reweighting, which emphasizes samples that are learned reliably by one branch but remain difficult for the other. During inference, the predictions of the shared and routed branches ($\hat{y}_S$ and $\hat{y}_R$) are fused to obtain the final output O.

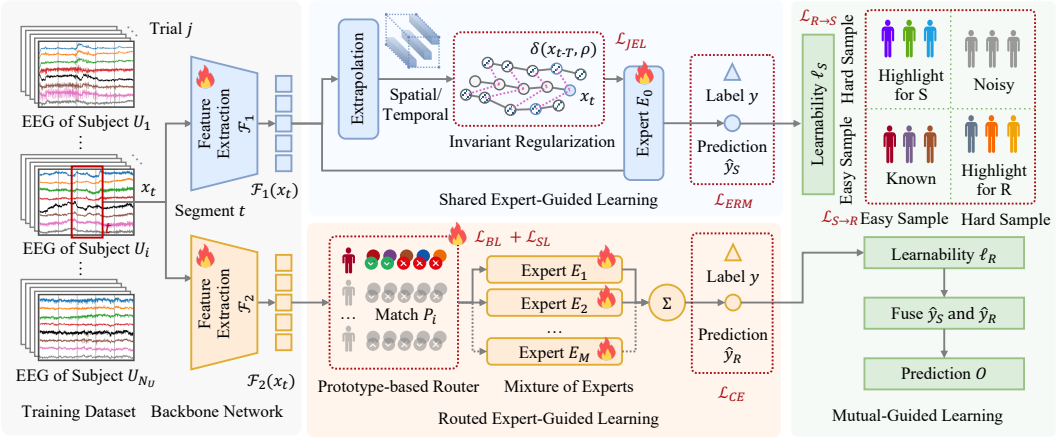


Figure 1: Overview of the proposed Shared-Routed Expert Alignment (SREA) framework.

3.1 Shared Branch for Reducible Invariant Structure

To satisfy the requirement of Theorem 2, we define a shared expert E_0 that models the reducible invariant structure across subjects. Given a batch of EEG signals, we train E_0 by minimizing the empirical risk across all subjects using cross-entropy loss:

$$\mathcal{L}_{\text{ERM}} = -\frac{1}{N_B}\sum_{i=1}^{N_B}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) \tag{1}$$

where N_B is the batch size, C is the number of classes, and $y_{ic} \in \{0, 1\}$ is the ground-truth indicator for sample i belonging to class c.

To bound the transfer error η under distribution shift, we construct augmented temporal neighbors as a proxy for unseen samples whose distributional distance from training data remains controlled. For each EEG sample $\boldsymbol{x}_t \in \mathbb{R}^{E\times L}$, where E is the number of electrodes and L is the segment length, we retrieve its preceding consecutive T-second segment $\boldsymbol{x}_{t-T}$ of the same class. We then apply a stochastic spatial masking operation $\delta(\cdot, \rho)$ that randomly masks electrodes with probability ρ, yielding $\tilde{\boldsymbol{x}}_{t-T} = \delta(\boldsymbol{x}_{t-T}, \rho)$.

Let $\mathcal{F}_1(\cdot)$ denote the feature extractor of the shared branch. We obtain the representations $\boldsymbol{z}_{t-T} = \mathcal{F}_1(\tilde{\boldsymbol{x}}_{t-T})$ and $\boldsymbol{z}_t = \mathcal{F}_1(\boldsymbol{x}_t)$. The joint embedding loss is defined as:

$$\mathcal{L}_{\text{JEL}} = \frac{1}{N_B}\sum_{i=1}^{N_B}\left(1 - \frac{\boldsymbol{z}_{i,t-T}\cdot\boldsymbol{z}_{i,t}}{\|\boldsymbol{z}_{i,t-T}\|\|\boldsymbol{z}_{i,t}\|}\right) \tag{2}$$

which minimizes the cosine distance between the representations of the current segment and its spatially-masked temporal neighbor.

The total loss for shared expert-guided learning is:

$$\mathcal{L}_S = \mathcal{L}_{\text{ERM}} + \mathcal{L}_{\text{JEL}} \tag{3}$$

3.2 Routed Branch for Irreducible Subject-Specific Structure

To satisfy the requirement of Theorem 2, we define a routed expert network that captures irreducible subject-specific structure. The routed expert network consists of M routed experts $\{E_j\}_{j=1}^{M}$ and a router R. Let $\boldsymbol{z} = \mathcal{F}_2(\boldsymbol{x}) \in \mathbb{R}^d$ denote the feature vector extracted from input $\boldsymbol{x}$ by the feature extractor $\mathcal{F}_2$. The output $\boldsymbol{o}$ is computed as a weighted combination of expert outputs:

$$\boldsymbol{o} = \sum_{j=1}^{M} r_j \cdot E_j(\boldsymbol{z}) \tag{4}$$

where M is the number of routed experts, $E_j(\cdot)$ denotes the j-th routed expert, and r_j represents the routing weight for expert j. The routing weights are sparse, with only the top-K experts receiving non-zero weights.

We adopt a prototype-based routing mechanism. Given feature $\boldsymbol{z} \in \mathbb{R}^d$, we first project it to a gate dimension space via $W\boldsymbol{z} \in \mathbb{R}^{d_r}$, then compute the routing weights using cosine similarity with learned prototype embeddings $\boldsymbol{P} = [\boldsymbol{P}_1, \ldots, \boldsymbol{P}_M] \in \mathbb{R}^{d_r \times M}$:

$$r_j = \begin{cases} \dfrac{\exp(\mathrm{sim}(\boldsymbol{P}_j, W\boldsymbol{z}))}{\sum_{l \in \mathcal{T}} \exp(\mathrm{sim}(\boldsymbol{P}_l, W\boldsymbol{z}))} & \text{if } j \in \mathcal{T} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathrm{sim}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^\top \boldsymbol{b} / (\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|)$ denotes the cosine similarity, and $\mathcal{T}$ denotes the set of top-K experts with the highest similarity scores. The prototype embedding matrix $\boldsymbol{P}$ can be interpreted as a codebook for neural activity patterns, and the cosine similarity acts as a matched filter for detecting these patterns. This design mitigates the issue where samples from similar subjects tend to cluster together, enabling each expert to specialize on its assigned cluster.

To optimize the selected experts' predictions toward the ground-truth labels, we employ the cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N_B} \sum_{i=1}^{N_B} \sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic}) \tag{6}$$

where N_B is the batch size, C is the number of classes, and $y_{ic} \in \{0, 1\}$ is the ground-truth indicator for sample i belonging to class c.

To encourage each expert to specialize on specific subjects, we compute the average routing probability distribution across all samples from each subject, then minimize the entropy of this distribution:

$$\mathcal{L}_{\text{SL}} = \frac{1}{N_U} \sum_{k=1}^{N_U} H_k, \quad H_k = -\sum_{j=1}^{M} \bar{r}_{k,j} \log \bar{r}_{k,j} \tag{7}$$

where N_U is the number of unique subjects and $\bar{r}_{k,j} = \frac{1}{|\mathcal{D}_k|} \sum_{\boldsymbol{x} \in \mathcal{D}_k} r_j(\boldsymbol{x})$ is the average routing weight for expert j across all samples from subject k, with $\mathcal{D}_k$ denoting the set of samples belonging to subject k. This loss encourages the router to consistently route samples from the same subject to the same expert.

To prevent routing collapse where the model consistently selects only a few experts while leaving others undertrained, we employ an auxiliary balance loss:

$$\mathcal{L}_{\text{BL}} = M \sum_{j=1}^{M} f_j \cdot \bar{r}_j \tag{8}$$

where $f_j = \frac{1}{N_B} \sum_{i=1}^{N_B} \mathbb{I}[r_j(\boldsymbol{x}_i) > 0]$ denotes the fraction of samples routed to expert j and $\bar{r}_j = \frac{1}{N_B} \sum_{i=1}^{N_B} r_j(\boldsymbol{x}_i)$ represents the average routing weight of expert j over the batch. This loss penalizes configurations where a small subset of experts receives disproportionately high routing weights and selection frequencies.

The total loss for routed expert-guided learning is:

$$\mathcal{L}_R = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SL}} + \mathcal{L}_{\text{BL}} \tag{9}$$

3.3 Mutual-Guided Reweighting

For a given sample $\boldsymbol{x}^{(i)}$ with ground-truth label $y^{(i)}$, let $\ell_S^{(i)} = -\log P_S(y^{(i)}|\boldsymbol{x}^{(i)})$ and $\ell_R^{(i)} = -\log P_R(y^{(i)}|\boldsymbol{x}^{(i)})$ denote the cross-entropy losses from the shared expert E_0 and the routed experts $\{E_j\}_{j=1}^M$, respectively. Following Theorem 3, the mutual-guided reweighting mechanism couples the two branches by transferring supervision according to cross-branch learnability gaps. Specifically, we derive the reweighting formulas in Eq. equation 10 and Eq. equation 11 to ensure that the

expert with higher loss learns from the better-performing one, thereby reducing the total alignment error.

When training the routed experts guided by the shared expert, we compute the weighted cross-entropy loss:

$$\mathcal{L}_{R\leftarrow S} = \frac{1}{N_B}\sum_{i=1}^{N_B}(1+\exp(\ell_R^{(i)} - \ell_S^{(i)}))\cdot \ell_R^{(i)} \tag{10}$$

Samples that the shared expert finds easier (lower $\ell_S^{(i)}$) but the routed experts find challenging (higher $\ell_R^{(i)}$) receive higher weights, constraining the routed experts from under-reduction with guidance from the shared expert.

Similarly, when training the shared expert guided by the routed experts:

$$\mathcal{L}_{S\leftarrow R} = \frac{1}{N_B}\sum_{i=1}^{N_B}(1+\exp(\ell_S^{(i)} - \ell_R^{(i)}))\cdot \ell_S^{(i)} \tag{11}$$

which assigns higher weights to samples that the routed experts handle well, constraining the shared expert from over-reduction.

The total loss for mutual-guided learning combines both directions:

$$\mathcal{L}_M = \mathcal{L}_{R\leftarrow S} + \mathcal{L}_{S\leftarrow R} \tag{12}$$

3.4 Inference and Prediction Fusion

During inference, the shared branch and the routed branch each produce a prediction for the input EEG sample. Let $\hat{y}_S$ and $\hat{y}_R$ denote the predicted class probabilities from the shared expert E_0 and the routed experts $\{E_j\}_{j=1}^{M}$, respectively. The final fused output is $O = \frac{1}{2}(\hat{y}_S + \hat{y}_R)$.

4 Experiments

4.1 Experimental Settings

We evaluate the proposed SREA on seven publicly available EEG datasets covering three representative tasks, including motor imagery (MI), sleep stage detection (SSD), and auditory attention decoding (AAD). Specifically, for MI, we use BCI IV-2a (Tangermann et al., 2012) and BCI IV-2b (Leeb et al., 2007). For SSD, we use Sleep-EDFx (Goldberger et al., 2000) and ISRUC-3 (Khalighi et al., 2016). For AAD, we use the KUL dataset (Das et al., 2016), the DTU dataset (Fuglsang et al., 2017), and the AVED dataset (Fan et al., 2024).

We follow the standard experimental protocols used in the corresponding prior studies. For BCI IV-2a and BCI IV-2b, leave-one-subject-out cross-validation is employed (Zhao et al., 2024b). For Sleep-EDFx, subject-wise 10-fold cross-validation is used, while leave-one-subject-out cross-validation is used on ISRUC-3 (Wang et al., 2024b). For the KUL, DTU, and AVED datasets, leave-one-subject-out cross-validation is adopted (Fan et al., 2025). In leave-one-subject-out evaluation, one subject is held out for testing and the remaining subjects are used for training. In subject-wise k-fold cross-validation, one fold of subjects is used for testing and the remaining folds are used for training.

We train both the shared expert and the routed experts using the Adam optimizer. The batch size is set to 256, and the maximum number of training epochs is 100. Early stopping with a patience of 10 epochs is adopted to alleviate overfitting. The learning rate is set to 10^{-4}, and the weight decay is set to 5×10^{-4}. Since different tasks involve distinct EEG topologies and temporal resolutions, we adopt task-specific backbones, i.e., DeepConvNet (Schirrmeister et al., 2017) for MI, SleepWaveNet (Wang et al., 2024b) for SSD, and ListenNet (Fan et al., 2025) for AAD. Unless otherwise specified, all loss coefficients are set to 1. The number of routed experts is set to $M = 5$, and the top $K = 3$ experts are activated for each sample. The masking ratio is set to $\rho = 10\%$. For datasets with fewer than 10 electrodes, temporal masking is used instead of electrode masking.

Table 1: Comparison under subject-independent conditions. Accuracy (%) is reported. * indicates significance (paired t-test, $p < 0.05$).

(a) Motor Imagery		
Model	IV-2a	IV-2b
ShallowConvNet	56.8*	74.3*
DeepConvNet	60.2*	75.2*
EEGNet	56.9*	75.1*
Conformer	53.4*	73.5*
CTNet	58.6*	76.3*
Proposed	**61.3**	**77.5**

(b) Sleep Stage Detection		
Model	EDFx	ISRUC
DeepSleepNet	81.8*	76.5*
Utime	80.6*	73.5*
TinySleepNet	82.7*	75.8*
SalientSleepNet	82.9*	76.9*
SleepWaveNet	83.2*	79.2*
Proposed	**84.1**	**79.8**

(c) Auditory Attention Decoding			
Model	KUL	DTU	AVED
SSF-CNN	59.3*	52.3*	51.7*
MBSSFCC	62.7*	52.5*	52.2*
DBPNet	61.1*	55.5*	52.1*
DARNet	69.9*	55.6*	51.3*
ListenNet	78.1*	56.8*	52.8*
Proposed	**81.7**	**58.3**	**53.9**

4.2 Comparison with Task-Specific Methods

We compare with several representative task-specific models on the motor imagery task using the BCI IV-2a and BCI IV-2b datasets. The compared methods include ShallowConvNet (Schirrmeister et al., 2017), DeepConvNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), Conformer (Song et al., 2022), and CTNet (Zhao et al., 2024b). The results in Table 1 show that the proposed method achieves the highest accuracy, obtaining 61.3% on BCI IV-2a and 77.5% on BCI IV-2b. Notably, the gain is achieved under a highly challenging setting with only nine subjects available for training and testing, indicating that the proposed framework can effectively alleviate cross-subject discrepancy even when the number of training subjects is limited.

We then evaluate the proposed method on sleep stage detection using the Sleep-EDFx and ISRUC-3 datasets. We compare with strong baselines, including DeepSleepNet (Supratak et al., 2017), Utime (Perslev et al., 2019), TinySleepNet (Supratak & Guo, 2020), SalientSleepNet (Jia et al., 2021), and SleepWaveNet (Wang et al., 2024b). As shown in Table 1, our method obtains the best performance on both datasets, achieving 84.1% on Sleep-EDFx and 79.8% on ISRUC-3. Although the improvements are relatively smaller than those on MI and AAD, the proposed framework still provides consistent gains. A possible reason is that SSD usually contains stronger subject-shared physiological patterns, such as sleep spindles and K-complexes, making the cross-subject gap less severe than in other EEG tasks.

Next, we evaluate auditory attention decoding. The compared methods include SSF-CNN (Cai et al., 2021), MBSSFCC (Jiang et al., 2022), DBPNet (Ni et al., 2024), DARNet (Yan et al., 2024), and ListenNet (Fan et al., 2025). As reported in Table 1, the proposed method achieves the best performance on all three datasets, reaching 81.7% on KUL, 58.3% on DTU, and 53.9% on AVED. These results verify the advantage of the proposed framework for cross-subject auditory attention decoding, even under short 1-second decision windows.

4.3 Comparison with EEG Foundation Methods

We further evaluate the proposed framework in a multi-task multi-dataset setting under subject-independent protocols. Following prior work (Zhou et al., 2025), we report balanced accuracy and compare with recent EEG foundation models, including FFCL, ST-Trans, BIOT, LaBraM, and CBraMod. We use CBraMod as the backbone for our method. As shown in Table 2, foundation models already provide strong cross-dataset generalization, while the proposed framework further improves the best-performing baseline across most datasets. This result suggests that SREA is not limited to lightweight task-specific architectures, but can also serve as a general collaboration paradigm that complements powerful pre-trained EEG representations.

4.4 Case Studies

To further understand when the proposed framework is effective, we conduct case studies on the KUL dataset. We follow the same experimental settings as in Sec. 4.2 and report the subject-wise test accuracy of the full model, together with two ablated variants: using only the shared expert (w/o R) and using only the routed experts (w/o S). The results are shown in Fig. 2.

Table 2: Comparison with EEG foundation models on multi-task multi-dataset scenarios under subject-independent conditions. Balanced accuracy (%) is reported. * indicates significance (paired t-test, $p < 0.05$).

(a) Motor Imagery		
Model	IV-2a	IV-2b
FFCL	44.7*	74.8*
ST-Trans	45.8*	75.0*
BIOT	47.5*	76.0*
LaBraM	48.7*	76.4*
CbraMod	51.4*	76.9*
Proposed	**56.2**	**77.6**

(b) Sleep Stage Detection		
Model	EDFx	ISRUC
FFCL	77.3*	72.8*
ST-Trans	77.9*	73.8*
BIOT	80.8*	75.3*
LaBraM	81.7*	76.3*
CbraMod	82.1*	78.7*
Proposed	**83.2**	**79.6**

(c) Auditory Attention Decoding			
Model	KUL	DTU	AVED
FFCL	73.4*	52.0*	51.4*
ST-Trans	73.6*	52.4*	51.5*
BIOT	74.8*	54.6*	52.5*
LaBraM	75.5*	55.3*	52.9*
CbraMod	76.5*	55.6*	53.8*
Proposed	**79.8**	**57.0**	**53.9**

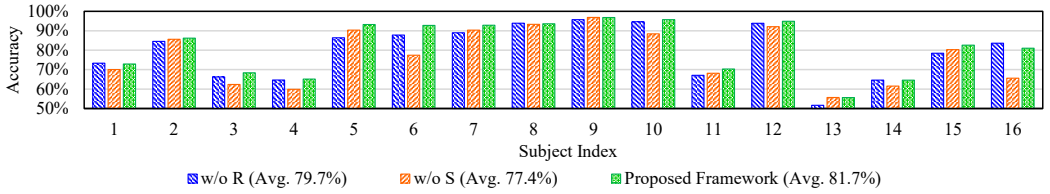


Figure 2: Case studies of the proposed framework on the KUL dataset. The x-axis represents the subject index, and the y-axis represents the accuracy.

The proposed framework surpasses both ablated variants on 10 of 16 subjects and achieves the highest mean accuracy (81.7%, versus 79.7% for w/o R and 77.4% for w/o S). This pattern is consistent with the two branches capturing complementary information. The shared expert focuses on subject-invariant patterns, while the routed experts preserve subject-specific discriminative cues. Their collaboration improves average performance across subjects.

4.5 Ablation Study

To validate the contribution of each component, we conduct systematic ablation experiments on the KUL and DTU datasets. We consider two levels of ablation: (1) branch-level variants that remove entire architectural components, and (2) component-level variants that disable individual loss terms within the full model. The results are reported in Table 3.

Several observations can be made from Table 3. First, removing either branch reduces average accuracy on both datasets. On KUL, removing the shared branch causes the larger drop (81.7% to 77.4%, versus 79.7% without the routed branch), indicating a larger contribution from the shared branch in this setting. On DTU, the two drops are similar, with a slightly larger drop when removing the shared branch. Second, the w/o Mutual variant shows that independent branch training followed by prediction fusion underperforms the full model, supporting the value of mutual-guided reweighting during training. Third, disabling either the joint embedding loss or the expert routing losses also lowers accuracy, consistent with their intended roles in the model.

4.6 Empirical Verification of Theoretical Assumptions

To validate that the assumptions underlying Theorems 1–3 hold on real EEG data, we conduct quantitative verification experiments on all three benchmark datasets. For each dataset, we train a single-fold subject-independent model and compute three metrics tied to the theoretical assumptions.

As shown in Table 4, the raw-feature MMD values are small on all three datasets, confirming that the cross-subject distribution shift required by Assumption 1 is bounded. CKA between shared and routed branch representations on held-out test subjects is consistently high, supporting the existence of invariant correlations posited by Assumption 2. The generalization gap of the routed branch

Table 3: Ablation study on the KUL and DTU datasets. Average accuracy (%) is reported under leave-one-subject-out cross-validation. *w/o Routed*: shared branch only. *w/o Shared*: routed branch only. *w/o Mutual*: both branches trained independently and fused at inference. *w/o JEL*: no joint embedding loss. *w/o SL+BL*: no expert routing losses.

(a) Branch-level			(b) Component-level		
Variant	KUL	DTU	Variant	KUL	DTU
Proposed	**81.7**	**58.3**	Proposed	**81.7**	**58.3**
w/o Routed	79.7	57.2	w/o JEL	81.1	57.9
w/o Shared	77.4	57.0	w/o SL+BL	81.0	57.7
w/o Mutual	80.8	57.8			

Table 4: Empirical verification of theoretical assumptions on real EEG datasets. MMD is computed with multi-scale RBF kernels between training and test features. CKA measures the representational similarity between the shared and routed branches. The generalization gap is defined as the difference between training accuracy and test accuracy.

(a) Assumption 1: Bounded distribution shift				(b) Assumptions 2&3: Correlations			
Metric	KUL	IV-2a	ISRUC	Metric	KUL	IV-2a	ISRUC
Raw MMD ($\times 10^{-3}$)	1.47	1.41	1.13	CKA (test set)	0.737	0.623	0.856
Shared MMD ($\times 10^{-3}$)	21.5	3.59	6.54	Gap (shared branch)	0.220	0.170	0.021
Routed MMD ($\times 10^{-3}$)	12.7	4.88	3.90	Gap (routed branch)	0.231	0.188	0.035

exceeds that of the shared branch on all datasets, indicating that routed experts capture more subject-specific correlations that do not transfer, consistent with Assumption 3.

4.7 Limitation

The proposed framework has been evaluated on classification tasks. Extending it to regression-based EEG decoding tasks, such as vigilance estimation (Wu et al., 2020) and speech envelope reconstruction (Défossez et al., 2023), remains an open direction for future work.

5 Conclusion

In this paper, we show that shared and routed experts are suited to different components of the target function, with their relative advantage determined by the reducibility cost from branch-specific to branch-invariant structure. Based on this insight, we proposed SREA. Experiments under subject-independent settings show that SREA consistently outperforms strong task-specific baselines, standard ensemble methods, and recent EEG foundation models. We further conduct case studies, ablations, complexity analysis, and validation on assumptions, all of which support the proposed design and its underlying alignment principle.

Reproducibility Statement

Section 2 presents the theoretical analysis, Section 3 describes SREA, and Section 4 specifies the seven public EEG benchmarks, subject-independent evaluation protocol, baselines, and metrics. The appendix provides dataset and preprocessing details, implementation settings, ensemble comparisons, synthetic verification, qualitative and complexity analyses, and proofs. The companion project code contains the model implementation and experiment scripts.

Ethics Statement

This study uses publicly available EEG datasets. It did not involve new data collection or participant recruitment. Dataset sources and preprocessing are documented in the appendix.

AI Use Statement

Generative AI assisted with polishing and clarifying prose. The authors reviewed the AI-assisted work and take responsibility for the final content.

References

Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. Aasm scoring manual updates for 2017, 2017.

Philipp Bomatter and Henry Gouk. Is limited participant diversity impeding EEG-based machine learning? *NeurIPS*, 2025.

Siqi Cai, Pengcheng Sun, Tanja Schultz, and Haizhou Li. Low-latency auditory spatial attention detection based on spectro-spatial features from eeg. In *EMBC*, pp. 5812–5815, 2021.

Neetha Das, Wouter Biesmans, Alexander Bertrand, and Tom Francart. The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *Journal of Neural Engineering*, 13(5):056014, 2016.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023.

Cunhang Fan, Jingjing Zhang, Hongyu Zhang, Wang Xiang, Jianhua Tao, Xinhui Li, Jiangyan Yi, Dianbo Sui, and Zhao Lv. MSFNet: Multi-scale fusion network for brain-controlled speaker extraction. In *MM*, pp. 1652–1661, 2024.

Cunhang Fan, Xiaoke Yang, Hongyu Zhang, Ying Chen, Lu Li, Jian Zhou, and Zhao Lv. ListenNet: A lightweight spatio-temporal enhancement nested network for auditory attention detection. *IJCAI*, 2025.

Søren Asp Fuglsang, Torsten Dau, and Jens Hjortkjær. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage*, 156:435–444, 2017.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Vikram Shenoy Handiru and Vinod A Prasad. Optimized bi-objective EEG channel selection and cross-subject generalization with brain–computer interfaces. *Transactions on Human-Machine Systems*, 46(6):777–786, 2016.

Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. SalientSleepNet: Multimodal salient wave detection network for sleep staging. *IJCAI*, 2021.

Weibang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *ICLR*, 2024.

Yifan Jiang, Ning Chen, and Jing Jin. Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of EEG. *Journal of Neural Engineering*, 19(5):056035, 2022.

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

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Robert Leeb, Felix Lee, Claudia Keinrath, Reinhold Scherer, Horst Bischof, and Gert Pfurtscheller. Brain–computer communication: motivation, aim, and impact of exploring a virtual apartment. *Transactions on Neural Systems and Rehabilitation Engineering*, 15(4):473–482, 2007.

Bo Li, Yifei Shen, Jingkang Yang, Yezhen Wang, Jiawei Ren, Tong Che, Jun Zhang, and Ziwei Liu. Sparse mixture-of-experts are domain generalizable learners. In *ICLR*, 2023.

He Li, Yi-Ming Jin, Wei-Long Zheng, and Bao-Liang Lu. Cross-subject emotion recognition using deep adaptation networks. In *International Conference on Neural Information Processing*, pp. 403–413, 2018.

Bo-Qun Ma, He Li, Wei-Long Zheng, and Bao-Liang Lu. Reducing the subject variability of EEG signals with adversarial domain generalization. In *ICONIP*, pp. 30–42, 2019.

Qinke Ni, Hongyu Zhang, Cunhang Fan, Shengbing Pei, Chang Zhou, and Zhao Lv. DBPNet: Dual-branch parallel network with temporal-frequency fusion for auditory attention detection. In *IJCAI*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.

Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *NeurIPS*, 32, 2019.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *Transactions on Affective Computing*, 14(3):2496–2511, 2022.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Akara Supratak and Yike Guo. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *EMBC*, pp. 641–644, 2020.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. *NeurIPS*, 37:39249–39280, 2024a.

Jing Wang, Xuehui Wang, Xiaojun Ning, Youfang Lin, Huy Phan, and Ziyu Jia. Subject-adaptation salient wave detection network for multimodal sleep stage classification. *Journal of Biomedical and Health Informatics*, 2024b.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *AAAI*, volume 38, pp. 265–273, 2024c.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *ICLR*, 2025.

Xiaotian Wang, Min Dang, Kunkuo Yang, Xinyu Cui, Doudou Zhang, and Chao Chen. The ensemble multi-scale convolution neural network for visual target detection EEG-based brain-computer interfaces. *Biomedical Signal Processing and Control*, 96:106583, 2024d.

Yiming Wang, Bin Zhang, and Yujiao Tang. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *AAAI*, volume 38, pp. 628–636, 2024e.

Wei Wu, Wei Sun, QM Jonathan Wu, Yimin Yang, Hui Zhang, Wei-Long Zheng, and Bao-Liang Lu. Multimodal vigilance estimation using deep learning. *Transactions on Cybernetics*, 52(5): 3097–3110, 2020.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? In *ICLR*, 2020.

Sheng Yan, Cunhang Fan, Hongyu Zhang, Xiaoke Yang, Jianhua Tao, and Zhao Lv. DARNet: Dual attention refinement network with spatiotemporal construction for auditory attention detection. *NeurIPS*, 37:31688–31707, 2024.

Hanzhong Zhang, Tienyu Zuo, Zhiyang Chen, Xin Wang, and Poly ZH Sun. Evolutionary ensemble learning for eeg-based cross-subject emotion recognition. *Journal of Biomedical and Health Informatics*, 28(7):3872–3881, 2024a.

Zhi Zhang, Sheng-hua Zhong, and Yan Liu. Torcheegemo: A deep learning toolbox towards eeg-based emotion recognition. *Expert Systems with Applications*, 249:123550, 2024b.

Peng Zhao, Ruicong Wang, Zijie Lin, Zexu Pan, Haizhou Li, and Xueyi Zhang. Ensemble deep learning models for EEG-based auditory attention decoding. In *ISCSLP*, pp. 339–343, 2024a.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. CTNet: a convolutional transformer network for EEG-based motor imagery classification. *Scientific reports*, 14(1):20237, 2024b.

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

A Appendix

A.1 Benchmark Datasets

In this section, we detail the benchmark datasets and preprocessing used in the main text. We evaluate the proposed framework on seven publicly available datasets spanning three EEG analysis tasks. The statistics of all datasets are summarized in Table 5.

For preprocessing, we apply a bandpass filter with cutoff frequencies of 0.1–50 Hz to preserve task-relevant neural activity, followed by a 50 Hz notch filter to suppress power-line interference. For BCI IV-2a and BCI IV-2b, the signals are segmented into 4-second windows without overlap. For Sleep-EDFx and ISRUC-3, we use 30-second windows without overlap. For the KUL, DTU, and AVED datasets, EEG signals are segmented into 1-second decision windows with 50% overlap.

Table 5: Statistics of seven publicly available datasets. We report the task type, total number of subjects, total number of categories, and total number of electrodes. AAD denotes auditory attention decoding, MI denotes motor imagery, and SSD denotes sleep stage detection.

Dataset	Task	Subjects	Categories	Electrodes
BCI IV-2a (2012)	MI	9	4	22
BCI IV-2b (2007)	MI	9	2	3
Sleep-EDFx (2000)	SSD	79	5	2
ISRUC-3 (2016)	SSD	9	5	6
KUL (2016)	AAD	16	2	64
DTU (2017)	AAD	18	2	64
AVED (2024)	AAD	20	2	32

For motor imagery classification, we use the BCI IV-2a dataset (Tangermann et al., 2012) and the BCI IV-2b dataset (Leeb et al., 2007).

The BCI IV-2a dataset contains EEG recordings from 9 subjects (A01–A09). EEG signals were recorded using 22 Ag/AgCl electrodes at a sampling rate of 250 Hz. Each subject participated in two sessions on separate days, with the first session allocated for training and the second for testing. Each session consisted of 288 trials, with 72 trials per class. We used the temporal segment from 2 to 6 seconds in our experiments. Labels indicate one of four motor imagery tasks, including left hand, right hand, both feet, and tongue.

The BCI IV-2b dataset contains EEG recordings from 9 subjects (B01–B09). EEG signals were recorded using 3 bipolar channels (C3, Cz, and C4) at a sampling rate of 250 Hz. Each subject participated in five sessions, where the first three sessions were used for calibration and the remaining two sessions were used for testing. There are approximately 400 trials in the training set and 320 trials in the test set. We used the temporal segment from 3 to 7 seconds in our experiments. Labels indicate one of two motor imagery tasks, including left hand and right hand.

For sleep stage classification, we use the Sleep-EDFx dataset (Goldberger et al., 2000) and the ISRUC-3 dataset (Khalighi et al., 2016).

The Sleep-EDFx dataset contains whole-night polysomnographic recordings from 78 healthy individuals aged 25 to 101 years. Each recording includes EEG, EOG, and chin EMG signals. Each participant has two day-night recordings, with the exception of three individuals who have only one recording due to equipment issues. The dataset comprises 195,479 sleep epochs in total. Labels indicate one of five sleep stages annotated in 30-second epochs, including Wake, N1, N2, N3, and REM.

The ISRUC-3 dataset contains polysomnographic recordings from 10 healthy subjects. Each recording includes EEG, EOG, and EMG signals, segmented into 30-second epochs. We use data from 30 minutes before and after the in-bed period. Following the latest AASM sleep standard, we merge the N3 and N4 stages into a single N3 stage. Labels indicate one of five sleep stages, including Wake, N1, N2, N3, and REM.

For auditory attention decoding, we use the KUL dataset (Das et al., 2016), the DTU dataset (Fuglsang et al., 2017), and the AVED dataset (Fan et al., 2024).

The KUL dataset contains EEG recordings from 16 normal-hearing subjects. EEG signals were recorded using a 64-channel BioSemi ActiveTwo system. Each subject was instructed to attend to one of two competing speech streams, with speakers positioned at 90° to the left or right. Each subject completed 8 trials, each lasting 6 minutes. Labels indicate the attended direction.

The DTU dataset contains EEG recordings from 18 normal-hearing subjects. EEG signals were recorded using a 64-channel BioSemi system. Each subject performed a target speaker tracking task in an environment with reverberation and dynamic background noise, attending to one of two competing speakers positioned at 60° relative to the subject. Each subject completed 60 trials, each lasting approximately 50 seconds. Labels indicate the attended speaker.

The AVED dataset contains EEG recordings from 20 normal-hearing subjects. EEG signals were recorded using a 32-channel system. Subjects were evenly divided into two experimental conditions: audio-only and audio-visual, with 10 subjects in each condition. Each subject was instructed to attend to one of two competing speech streams, with speakers positioned at 90° to the left or right. In the audio-visual condition, subjects also watched a video of the narrator they were instructed to focus on. Each subject completed 16 trials, each lasting 152 seconds. Labels indicate the attended direction.

A.2 Implementation Details

We use the Adam optimizer for both the shared expert and the routed experts. The batch size N_B is set to 256, and the maximum number of training epochs is set to 100. Early stopping with a patience of 10 epochs is employed to prevent overfitting. The learning rate is set to 0.0001 for the shared expert and the routed experts. Weight decay is set to 0.0005. Since task-specific models cannot accommodate different tasks with varying topologies and decision windows, we use ListenNet (Fan et al., 2025) for AAD, DeepConvNet (Schirrmeister et al., 2017) for MI, and SleepWaveNet (Wang et al., 2024b) for SSD. We adopt these network architectures but do not use the training methods from these papers. Instead, we use the proposed SREA for training.

We conduct experiments with foundation brain models, including LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), and EEGPT (Wang et al., 2024a). To adapt these foundation brain models for AAD, MI, and SSD, we prepend two layers before the foundation brain models: a 1D convolutional layer that maps the electrode channels from the dataset to the number of channels used during pretraining, and a resampling layer that resamples the EEG signals to match the patch size used in pretraining. We replace the classification head with the proposed shared expert and routed experts.

The proposed loss function is insensitive to the weighting coefficients, so all coefficients are set to 1. We use $M = 5$ experts and activate the top $K = 3$ experts for each sample. We set $\rho = 10\%$. When the number of electrodes is fewer than 10, we randomly mask temporal segments within the electrodes. We conduct experiments on 2 Tesla A100 GPUs.

The source code is implemented based on PyTorch (Paszke et al., 2019) and TorchEEG (Zhang et al., 2024b). We provide the code in the supplementary materials for reviewers to reproduce all experimental results. We will release the code on GitHub upon paper acceptance for follow-up studies.

A.3 Comparison with Ensemble Methods

To further demonstrate that the gains of the proposed framework arise from collaborative shared-routed alignment rather than simple model aggregation, we compare with two standard ensemble strategies: Bagging and Boosting. Both ensemble baselines are built on the same task-specific backbone used by the proposed method (ListenNet for AAD, DeepConvNet for MI, and SleepWaveNet for SSD). Table 6 reports the results on all seven datasets.

As shown in Table 6, although Bagging and Boosting provide modest improvements over the corresponding single-backbone baselines, the proposed framework consistently outperforms both ensemble strategies across all seven datasets. For example, on KUL, the proposed method achieves

Table 6: Comparison with ensemble methods under subject-independent conditions. Accuracy (%) is reported. * indicates significance (paired t-test, $p < 0.05$).

(a) Auditory Attention Decoding			
Model	KUL	DTU	AVED
ListenNet	78.1*	56.8*	52.8*
Bagging (ListenNet)	79.1*	57.2*	53.2*
Boosting (ListenNet)	79.6*	57.5*	53.1*
Proposed	**81.7**	**58.3**	**53.9**

(b) Motor Imagery		
Model	IV-2a	IV-2b
DeepConvNet	60.2*	75.2*
Bagging (DeepConvNet)	60.6*	76.6*
Boosting (DeepConvNet)	60.5*	76.8*
Proposed	**61.3**	**77.5**

(c) Sleep Stage Detection		
Model	EDFx	ISRUC
SleepWaveNet	83.2*	79.2*
Bagging (SleepWaveNet)	83.3*	79.4*
Boosting (SleepWaveNet)	83.5*	79.4*
Proposed	**84.1**	**79.8**

81.7%, compared to 79.6% for Boosting and 79.1% for Bagging. On BCI IV-2b, the proposed method reaches 77.5%, exceeding both Bagging (76.6%) and Boosting (76.8%). These results confirm that the gains of SREA stem from the principled collaboration between shared and routed experts via mutual-guided reweighting, rather than from simple model aggregation. Standard ensemble methods aggregate independently trained models without distinguishing reducible invariant structure from irreducible subject-specific structure, limiting their ability to address the complementary failure modes identified in our theoretical analysis.

A.4 Synthetic Verification

To empirically validate the theoretical analysis in Sec. 2, we construct a synthetic domain generalization benchmark with hierarchical domain structures. Specifically, we organize domains into groups to simulate subject groups with similar neural patterns. Each group center is first sampled from a Gaussian distribution, and each domain center is then obtained by perturbing its corresponding group center. Samples are finally drawn from Gaussian distributions centered at the domain-specific means.

Based on this construction, we define a domain-independent function shared across all domains and a domain-dependent function generated by hierarchical perturbations of group-level parameters. The target function of each domain is formulated as a weighted combination of these two components, where the weighting coefficient controls the proportion of reducible components.

We use five source domains, each containing 500 samples with an input dimension of 128, and conduct leave-one-domain-out cross-validation. By varying the proportion of reducible components from 0 to 1, we evaluate the performance of the full model and its ablated variants. The results are shown in Fig. 3.

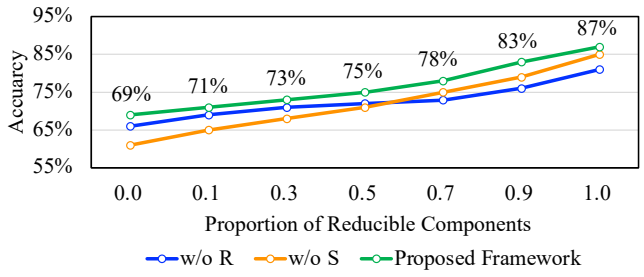


Figure 3: Experiments on synthetic verification under varying proportions of reducible components. We report the target domain test accuracy (%) with different ablation settings.

Several observations can be made. When the proportion of reducible components is low, the routed experts perform better than the shared expert, since domain-specific processing dominates. As the reducible proportion increases, the advantage of routed experts gradually diminishes, while the shared expert becomes increasingly effective. Importantly, the proposed full framework maintains the best performance across different settings, indicating that it can adaptively benefit from both paradigms. This result strongly supports our theoretical claim that the effectiveness of shared and

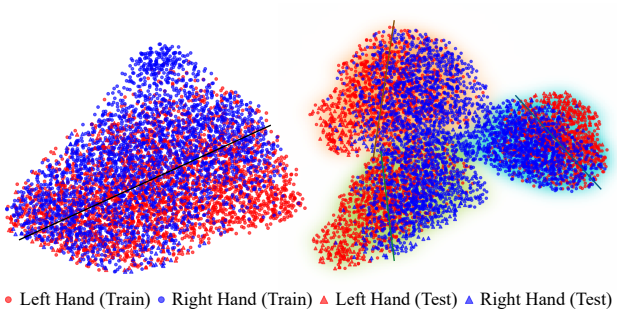


Figure 4: Qualitative analysis on the BCI-IV-2b dataset. The left subfigure visualizes the embeddings from the shared expert model using UMAP for the 1st fold of the training and test sets, and the right subfigure shows the embeddings from the routed expert model using UMAP for the 1st fold of the training and test sets.

routed experts depends on the reducibility cost of domain-specific functions to domain-invariant ones.

A.5 Qualitative Analysis

We further perform qualitative analysis to visualize the representations learned by the shared expert and routed experts. Specifically, we train the model on the BCI IV-2b dataset and use UMAP to project the learned embeddings of both training and test samples into a two-dimensional space. The visualization results are presented in Fig. 4.

As shown in the left panel of Fig. 4, the embeddings produced by the shared expert exhibit a relatively unified decision boundary, indicating that the shared branch captures common task-relevant structures across subjects. In contrast, the right panel shows that the routed experts form multiple specialized decision regions. Moreover, the routed assignments reveal that different samples are adaptively handled by different experts, suggesting that the routed branch indeed models heterogeneous subject-specific patterns. These observations are consistent with our theoretical analysis: the shared expert is suitable for reducible and invariant components, whereas the routed experts are advantageous for irreducible and branch-dependent components.

A.6 Complexity Analysis

To demonstrate the practical efficiency of the proposed framework, we analyze the parameter overhead and per-batch training time introduced by each component. Fig. 5 visualizes both metrics on three representative datasets using the corresponding task-specific backbones.

As shown in Fig. 5, the parameter overhead of the routed MoE head scales with the backbone's feature dimension, ranging from 3.9K (+0.3%) for SleepWaveNet to 686.6K for DeepConvNet, while the mutual-guided mechanism introduces zero additional parameters. In terms of training time, each component adds moderate per-batch cost, and the total per-batch time ranges from 180 ms to 349 ms across different backbones.

A.7 Proof for Theorems

In this section, we provide complete proofs for the theorems stated in the main text. We first state the algorithmic alignment theorem in the i.i.d. setting (Xu et al., 2020).

Theorem 4. *Fix $\epsilon > 0$ and $\delta \in (0, 1)$. Let g be a target function and $\mathcal{N}$ be a neural network with n modules. Suppose $\{\boldsymbol{x}_i\}_{i=1}^{M}$ are i.i.d. samples drawn from a distribution D, and let $y_i = g(\boldsymbol{x}_i)$. Assume the following conditions hold:*

(a) Algorithmic Stability. Let A be a learning algorithm for each module $\mathcal{N}_i$. Suppose $f = A(\{(\boldsymbol{x}_i, y_i)\})$ and $\hat{f} = A(\{(\hat{\boldsymbol{x}}_i, y_i)\})$. Then for any $\boldsymbol{x}$, $\|f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\| \leq L_0 \cdot \max_i \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|$.

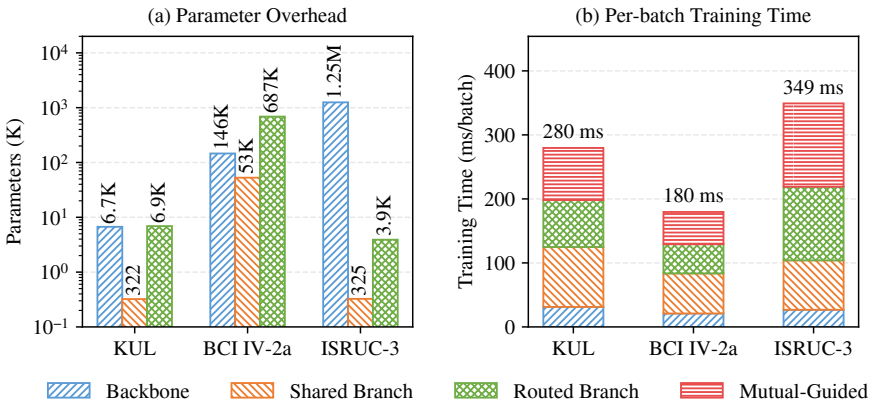


Figure 5: Complexity analysis. (a) Parameter overhead per component (log scale). The mutual-guided reweighting introduces zero extra parameters. (b) Per-batch training time breakdown (batch size 256, NVIDIA A100).

(b) Sequential Training. The modules $\mathcal{N}_1, \ldots, \mathcal{N}_n$ are trained sequentially. Module $\mathcal{N}_1$ is trained on samples $\{(\boldsymbol{x}_i^{(1)}, f_1(\boldsymbol{x}_i^{(1)}))\}_{i=1}^M$, where $\boldsymbol{x}_i^{(1)}$ are obtained from the training dataset. For $j > 1$, the inputs $\hat{\boldsymbol{x}}_i^{(j)}$ to module $\mathcal{N}_j$ are the outputs from the previous modules, while the labels are generated by the ground-truth functions $f_1, \ldots, f_{j-1}$ applied to $\boldsymbol{x}_i^{(1)}$.

(c) Lipschitz Continuity. Each learned function $\hat{f}_j$ satisfies $\|\hat{f}_j(\boldsymbol{x}) - \hat{f}_j(\hat{\boldsymbol{x}})\| \le L_1 \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|$ for some constant $L_1 > 0$.

Under conditions (a)–(c), if $\mathrm{Align}(\mathcal{N}, g, \epsilon, \delta) \le M$, then there exists a learning algorithm A such that $\mathbb{P}_{\boldsymbol{x}\sim D}[\|\mathcal{N}(\boldsymbol{x}) - g(\boldsymbol{x})\| \le O(\epsilon)] \ge 1 - O(\delta)$, where $\mathcal{N}$ denotes the network produced by A on the training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^M$.

Next, we state a lemma that transfers probability bounds from the training distribution to the test distribution, given Assumption 1.

Lemma 1. *Under Assumption 1, for any function f and measurable set A, if $\mathbb{P}_{D_{\mathrm{tr}}}[f(\mathcal{N}_1(\boldsymbol{x})) \in A] \le \delta$, then $\mathbb{P}_{D_{\mathrm{te}}}[f(\mathcal{N}_1(\boldsymbol{x})) \in A] \le C\delta + \eta$.*

Proof. Define the preimage $B = \{\boldsymbol{s} : f(\boldsymbol{s}) \in A\}$. Then

$$\mathbb{P}_{D_{\mathrm{te}}}[f(\mathcal{N}_1(\boldsymbol{x})) \in A] = \mathbb{P}_{D_{\mathrm{te}}}[\mathcal{N}_1(\boldsymbol{x}) \in B]. \tag{13}$$

Applying Assumption 1 with the measurable set B, we obtain

$$\begin{aligned} &\mathbb{P}_{D_{\mathrm{te}}}[\mathcal{N}_1(\boldsymbol{x}) \in B] \\ &\le C \cdot \mathbb{P}_{D_{\mathrm{tr}}}[\mathcal{N}_1(\boldsymbol{x}) \in B] + \eta \\ &= C \cdot \mathbb{P}_{D_{\mathrm{tr}}}[f(\mathcal{N}_1(\boldsymbol{x})) \in A] + \eta \\ &\le C\delta + \eta. \end{aligned} \tag{14}$$

□

We now prove Theorem 1 based on Theorem 4. We address the distribution shift in $\boldsymbol{x}$ with Lemma 1 and analyze the distribution shift in y with Assumptions 1, 2, and 3.

Proof. **Condition 1:** From Theorem 4, since $\mathrm{Align}(\mathcal{N}', g_c, \epsilon, \delta) \le |\mathcal{E}_{\mathrm{tr}}|$, we have

$$\mathbb{P}_{D_{\mathrm{tr}}}[\|\mathcal{N}(\boldsymbol{x}) - g_c(\mathcal{N}_1(\boldsymbol{x}))\| > O(\epsilon)] \le O(\delta). \tag{15}$$

By Lemma 1, this implies

$$\begin{aligned} &\mathbb{P}_{D_{\mathrm{te}}}[\|\mathcal{N}(\boldsymbol{x}) - g_c(\mathcal{N}_1(\boldsymbol{x}))\| > O(\epsilon)] \\ &\le C \cdot O(\delta) + \eta \\ &= O(\delta) + \eta. \end{aligned} \tag{16}$$

Equivalently,

$$\mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - g_c(\mathcal{N}_1(\boldsymbol{x}))\| \le O(\epsilon)] \ge 1 - O(\delta) - \eta. \tag{17}$$

Combining with Assumption 2, which states $\mathbb{P}_{D_{\text{te}}}[\|g_c(\mathcal{N}_1(\boldsymbol{x})) - y\| \le \epsilon] \ge 1 - \delta$, we obtain

$$\begin{aligned}
&\mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - y\| \le O(\epsilon)] \\
&\ge \mathbb{P}_{D_{\text{te}}}[\|g_c(\mathcal{N}_1(\boldsymbol{x})) - y\| \le \epsilon] \\
&\quad \cdot \mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - g_c(\mathcal{N}_1(\boldsymbol{x}))\| \le O(\epsilon)] \\
&\ge (1-\delta)(1 - O(\delta) - \eta) \\
&= 1 - O(\delta) - \eta,
\end{aligned} \tag{18}$$

where the first inequality follows from the triangle inequality $O(\epsilon) + O(\epsilon) = O(\epsilon)$.

Condition 2: Similarly, from Theorem 4, we have

$$\mathbb{P}_{D_{\text{tr}}}[\|\mathcal{N}(\boldsymbol{x}) - g_s(\mathcal{N}_1(\boldsymbol{x}))\| > O(\epsilon)] \le O(\delta). \tag{19}$$

By Lemma 1,

$$\mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - g_s(\mathcal{N}_1(\boldsymbol{x}))\| \le O(\epsilon)] \ge 1 - O(\delta) - \eta. \tag{20}$$

Combining with Assumption 3, which states $\mathbb{P}_{D_{\text{te}}}[\|g_s(\mathcal{N}_1(\boldsymbol{x})) - y\| > \omega(\epsilon)] \ge 1 - \delta$, we obtain

$$\begin{aligned}
&\mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - y\| > \omega(\epsilon)] \\
&\ge \mathbb{P}_{D_{\text{te}}}[\|g_s(\mathcal{N}_1(\boldsymbol{x})) - y\| > \omega(\epsilon)] \\
&\quad \cdot \mathbb{P}_{D_{\text{te}}}[\|\mathcal{N}(\boldsymbol{x}) - g_s(\mathcal{N}_1(\boldsymbol{x}))\| \le O(\epsilon)] \\
&\ge (1-\delta)(1 - O(\delta) - \eta) \\
&= 1 - O(\delta) - \eta,
\end{aligned} \tag{21}$$

where the first inequality follows from the reverse triangle inequality $\omega(\epsilon) - O(\epsilon) = \omega(\epsilon)$. □

We now prove Theorem 2, which compares the alignment of shared expert networks and routed expert networks.

Proof. We first compute the alignment for the routed expert network $\mathcal{R}$. The network $\mathcal{R}$ consists of M routed experts $\{R_j\}_{j=1}^{M}$ and a router R_0, yielding $M+1$ modules in total. Each routed expert R_j learns the branch-dependent function h_j, and the router R_0 learns the branching function h_0. By Definition 1, the alignment of $\mathcal{R}$ is given by

$$\begin{aligned}
&\text{Align}(\mathcal{R}, g, \epsilon, \delta) \\
&= (M+1) \cdot \max\Big\{ \max_{j \in [M]} \mathcal{M}(R_j, h_j, \epsilon, \delta), \\
&\qquad\qquad \mathcal{M}(R_0, h_0, \epsilon, \delta) \Big\}.
\end{aligned} \tag{22}$$

Define $\mathcal{C}^* := \max_{j \in [M]} \mathcal{M}(R_j, h_j, \epsilon, \delta)$. By the assumption that the branching function is no harder to learn than the branch-dependent functions, we have $\mathcal{M}(R_0, h_0, \epsilon, \delta) \le \mathcal{C}^*$. Therefore,

$$\begin{aligned}
&\max\Big\{ \max_{j \in [M]} \mathcal{M}(R_j, h_j, \epsilon, \delta), \\
&\qquad \mathcal{M}(R_0, h_0, \epsilon, \delta) \Big\} = \mathcal{C}^*,
\end{aligned} \tag{23}$$

and hence

$$\text{Align}(\mathcal{R}, g, \epsilon, \delta) = (M+1) \cdot \mathcal{C}^*. \tag{24}$$

For the shared expert network $\mathcal{S}$, there is a single module S_0 that learns the entire function g. By Definition 1, the alignment of $\mathcal{S}$ is

$$\begin{aligned}
&\text{Align}(\mathcal{S}, g, \epsilon, \delta) \\
&= 1 \cdot \mathcal{M}(S_0, g, \epsilon, \delta) = \mathcal{M}(S_0, g, \epsilon, \delta).
\end{aligned} \tag{25}$$

Condition 1: Suppose $\alpha \geq M+1$. By the lower bound condition in the theorem statement,

$$\begin{aligned}\mathrm{Align}(\mathcal{S}, g, \epsilon, \delta) &= \mathcal{M}(S_0, g, \epsilon, \delta) \\ &\geq \alpha \cdot \mathcal{C}^* \\ &\geq (M+1) \cdot \mathcal{C}^* \\ &= \mathrm{Align}(\mathcal{R}, g, \epsilon, \delta). \end{aligned} \tag{26}$$

Therefore, $\mathrm{Align}(\mathcal{R}, g, \epsilon, \delta) \leq \mathrm{Align}(\mathcal{S}, g, \epsilon, \delta)$.

Condition 2: Suppose $\beta \leq M+1$. By the upper bound condition in the theorem statement,

$$\begin{aligned}\mathrm{Align}(\mathcal{S}, g, \epsilon, \delta) &= \mathcal{M}(S_0, g, \epsilon, \delta) \\ &\leq \beta \cdot \mathcal{C}^* \\ &\leq (M+1) \cdot \mathcal{C}^* \\ &= \mathrm{Align}(\mathcal{R}, g, \epsilon, \delta). \end{aligned} \tag{27}$$

Therefore, $\mathrm{Align}(\mathcal{R}, g, \epsilon, \delta) \geq \mathrm{Align}(\mathcal{S}, g, \epsilon, \delta)$. □

Next, we state a lemma that characterizes the effect of the collaborative objective.

Lemma 2. *Let the collaborative objective functions for the shared expert and routed experts be $\mathcal{L}_S(\boldsymbol{x}) = \exp(\ell_S(\boldsymbol{x}) - \ell_R(\boldsymbol{x})) \cdot \ell_S(\boldsymbol{x})$ and $\mathcal{L}_R(\boldsymbol{x}) = \exp(\ell_R(\boldsymbol{x}) - \ell_S(\boldsymbol{x})) \cdot \ell_R(\boldsymbol{x})$, where ℓ_S and ℓ_R denote the cross-entropy losses with respect to ground-truth labels. Then the expected losses over the dataset can be rewritten as weighted knowledge distillation: $\mathcal{E}_S = \sum_{\boldsymbol{x}\in\mathcal{D}} a_S(\boldsymbol{x}) \cdot \mathrm{KL}(p^R(\boldsymbol{x}) \| q^S(\boldsymbol{x})) + C_S$ and $\mathcal{E}_R = \sum_{\boldsymbol{x}\in\mathcal{D}} a_R(\boldsymbol{x}) \cdot \mathrm{KL}(q^S(\boldsymbol{x}) \| p^R(\boldsymbol{x})) + C_R$, where $a_S(\boldsymbol{x}) = \exp(\ell_S(\boldsymbol{x}) - \ell_R(\boldsymbol{x}))$, $a_R(\boldsymbol{x}) = \exp(\ell_R(\boldsymbol{x}) - \ell_S(\boldsymbol{x}))$, and C_S, C_R are constants independent of the optimization variables.*

Proof. We derive the result for the shared expert. The derivation for the routed experts is symmetric. The expected collaborative objective over the dataset is $\mathcal{E}_S = \sum_{\boldsymbol{x}\in\mathcal{D}} \exp(\ell_S(\boldsymbol{x}) - \ell_R(\boldsymbol{x})) \cdot \ell_S(\boldsymbol{x})$. Let $i^*(\boldsymbol{x})$ denote the ground-truth label index for sample $\boldsymbol{x}$, let $q^S(\boldsymbol{x})$ denote the predicted probability distribution of the shared expert, and let $p^R(\boldsymbol{x})$ denote the predicted probability distribution of the routed experts. The cross-entropy losses can be expressed as $\ell_S(\boldsymbol{x}) = -\log q^S_{i^*}(\boldsymbol{x})$ and $\ell_R(\boldsymbol{x}) = -\log p^R_{i^*}(\boldsymbol{x})$.

Substituting the cross-entropy losses into the collaborative objective, we obtain

$$\begin{aligned}\mathcal{E}_S &= \sum_{\boldsymbol{x}\in\mathcal{D}} \exp(-\log q^S_{i^*}(\boldsymbol{x}) + \log p^R_{i^*}(\boldsymbol{x})) \\ &\qquad \cdot (-\log q^S_{i^*}(\boldsymbol{x})) \\ &= \sum_{\boldsymbol{x}\in\mathcal{D}} \frac{p^R_{i^*}(\boldsymbol{x})}{q^S_{i^*}(\boldsymbol{x})} \cdot (-\log q^S_{i^*}(\boldsymbol{x})). \end{aligned} \tag{28}$$

The KL divergence is defined as $\mathrm{KL}(p^R \| q^S) = \sum_{i=1}^{C} p^R_i \log(p^R_i / q^S_i) = -H(p^R) + H(p^R, q^S)$, where $H(p^R) = -\sum_i p^R_i \log p^R_i$ is the entropy of p^R and $H(p^R, q^S) = -\sum_i p^R_i \log q^S_i$ is the cross-entropy between p^R and q^S. The entropy term $H(p^R)$ is independent of the shared expert parameters. Under the hard-label approximation, the predicted probability of the routed experts concentrates near the ground-truth label i^*, i.e., $p^R_{i^*} \approx 1$. In this regime, the cross-entropy is dominated by $H(p^R, q^S) \approx -p^R_{i^*} \log q^S_{i^*}$.

Defining $a_S(\boldsymbol{x}) = \exp(\ell_S(\boldsymbol{x}) - \ell_R(\boldsymbol{x})) = p^R_{i^*}(\boldsymbol{x}) / q^S_{i^*}(\boldsymbol{x})$, the expected loss can be rewritten as

$$\begin{aligned}\mathcal{E}_S &= \sum_{\boldsymbol{x}\in\mathcal{D}} a_S(\boldsymbol{x}) \cdot (-\log q^S_{i^*}(\boldsymbol{x})) \\ &= \sum_{\boldsymbol{x}\in\mathcal{D}} a_S(\boldsymbol{x}) \cdot \mathrm{KL}(p^R(\boldsymbol{x}) \| q^S(\boldsymbol{x})) + C_S, \end{aligned} \tag{29}$$

where $C_S = \sum_{\boldsymbol{x}} a_S(\boldsymbol{x}) \cdot H(p^R(\boldsymbol{x}))$ is independent of q^S. The derivation for the routed experts is symmetric. □

We now prove Theorem 3, which establishes the effect of mutual-guided collaboration between the shared expert and routed experts.

Proof. By Lemma 2, the collaborative objective is equivalent to a weighted knowledge distillation loss. Taking the gradient with respect to the shared expert output $S(\boldsymbol{x})$ in function space, we have $\nabla_S \mathrm{KL}(p^R \| q^S) \propto -(R(\boldsymbol{x}) - S(\boldsymbol{x}))$, which gives $\nabla_S \mathcal{E}_S \propto -a_S(\boldsymbol{x}) \cdot (R(\boldsymbol{x}) - S(\boldsymbol{x}))$. The gradient descent updates are

$$\begin{aligned} S^{(t+1)}(\boldsymbol{x}) &= S^{(t)}(\boldsymbol{x}) \\ &\quad + \eta \cdot a_S(\boldsymbol{x}) \cdot (R^{(t)}(\boldsymbol{x}) - S^{(t)}(\boldsymbol{x})), \end{aligned} \tag{30}$$

$$\begin{aligned} R^{(t+1)}(\boldsymbol{x}) &= R^{(t)}(\boldsymbol{x}) \\ &\quad + \eta \cdot a_R(\boldsymbol{x}) \cdot (S^{(t)}(\boldsymbol{x}) - R^{(t)}(\boldsymbol{x})), \end{aligned} \tag{31}$$

where $\eta > 0$ is the learning rate.

Define $\delta_S^{(t)}(\boldsymbol{x}) = S^{(t)}(\boldsymbol{x}) - g(\boldsymbol{x})$ and $\delta_R^{(t)}(\boldsymbol{x}) = R^{(t)}(\boldsymbol{x}) - g(\boldsymbol{x})$. Subtracting $g(\boldsymbol{x})$ from the update rule for the shared expert yields

$$\begin{aligned} \delta_S^{(t+1)}(\boldsymbol{x}) &= \delta_S^{(t)}(\boldsymbol{x}) + \eta \cdot a_S(\boldsymbol{x}) \cdot (R^{(t)}(\boldsymbol{x}) - S^{(t)}(\boldsymbol{x})) \\ &= \delta_S^{(t)}(\boldsymbol{x}) + \eta \cdot a_S(\boldsymbol{x}) \cdot (\delta_R^{(t)}(\boldsymbol{x}) - \delta_S^{(t)}(\boldsymbol{x})) \\ &= (1 - \eta \cdot a_S(\boldsymbol{x})) \cdot \delta_S^{(t)}(\boldsymbol{x}) \\ &\quad + \eta \cdot a_S(\boldsymbol{x}) \cdot \delta_R^{(t)}(\boldsymbol{x}). \end{aligned} \tag{32}$$

Similarly, $\delta_R^{(t+1)}(\boldsymbol{x}) = (1 - \eta \cdot a_R(\boldsymbol{x})) \cdot \delta_R^{(t)}(\boldsymbol{x}) + \eta \cdot a_R(\boldsymbol{x}) \cdot \delta_S^{(t)}(\boldsymbol{x})$.

Taking the squared norm and applying the triangle inequality, for η sufficiently small such that $\eta \cdot a_S(\boldsymbol{x}) < 1$ and $\eta \cdot a_R(\boldsymbol{x}) < 1$, we have

$$\begin{aligned} \|\delta_S^{(t+1)}(\boldsymbol{x})\| &\le (1 - \eta \cdot a_S(\boldsymbol{x}))\|\delta_S^{(t)}(\boldsymbol{x})\| \\ &\quad + \eta \cdot a_S(\boldsymbol{x})\|\delta_R^{(t)}(\boldsymbol{x})\|, \end{aligned} \tag{33}$$

$$\begin{aligned} \|\delta_R^{(t+1)}(\boldsymbol{x})\| &\le (1 - \eta \cdot a_R(\boldsymbol{x}))\|\delta_R^{(t)}(\boldsymbol{x})\| \\ &\quad + \eta \cdot a_R(\boldsymbol{x})\|\delta_S^{(t)}(\boldsymbol{x})\|. \end{aligned} \tag{34}$$

Consider the case $\ell_S(\boldsymbol{x}) > \ell_R(\boldsymbol{x})$. Then $a_S(\boldsymbol{x}) > 1$ and $a_R(\boldsymbol{x}) < 1$. By Assumption 4, $\|\delta_R^{(t)}(\boldsymbol{x})\| \le \varepsilon$. Substituting this bound, we obtain $\|\delta_S^{(t+1)}(\boldsymbol{x})\| \le (1 - \eta \cdot a_S(\boldsymbol{x}))\|\delta_S^{(t)}(\boldsymbol{x})\| + \eta \cdot a_S(\boldsymbol{x}) \cdot \varepsilon$. Since $a_S(\boldsymbol{x}) > a_R(\boldsymbol{x})$, the contraction rate for the shared expert exceeds the expansion rate for the routed experts. The case $\ell_R(\boldsymbol{x}) > \ell_S(\boldsymbol{x})$ is symmetric.

Define $\epsilon_S^{(t)}(\boldsymbol{x}) = \|\delta_S^{(t)}(\boldsymbol{x})\|^2$ and $\epsilon_R^{(t)}(\boldsymbol{x}) = \|\delta_R^{(t)}(\boldsymbol{x})\|^2$. Adding the squared-norm recursions and using the convexity of the squared norm, for each sample $\boldsymbol{x}$, the total error $\epsilon_S^{(t)}(\boldsymbol{x}) + \epsilon_R^{(t)}(\boldsymbol{x})$ is non-increasing in t. Taking the expectation over the dataset completes the proof. □