# DLINK: DISTILLING LAYER-WISE AND DOMINANT KNOWLEDGE FROM EEG FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

EEG foundation models (EFMs) provide strong representations for diverse EEG decoding tasks through large-scale pretraining and downstream fine-tuning. Through empirical analysis, we observe that (i) while lightweight models remain competitive, they are highly task-dependent. Retaining high-capacity EFMs as inference backbones incurs substantial computational overhead, making knowledge distillation a natural route for model compression; and (ii) direct distillation from a fixed teacher representation underutilizes EFM knowledge, as task-discriminative information is distributed across intermediate layers rather than concentrated in the final layer. These observations motivate **DLink** (**D**istilling **L**ayer-w**i**se and Domi**n**ant **K**nowledge), a spectrally guided distillation framework with input-conditioned layer routing for transferring EFM knowledge into compact students. DLink uses a router to aggregate teacher layers for each input, and aligns magnitude and phase spectra to mitigate compression-induced spectral distortion in learned representations. The routed teacher knowledge is internalized by a project-then-compress student; the teacher and router are used only during training. Experiments across six EEG benchmarks show that DLink improves compact students while remaining lightweight at inference. Further analyses support learned layer aggregation and spectral distillation, establishing DLink as a practical framework for distilling layer-wise dominant knowledge into compact EEG models.

## 1 INTRODUCTION

Electroencephalogram (EEG) signals provide a non-invasive window into neural activity and have been used for brain–computer interfaces Mellot et al. (2024), affective computing, cognitive assessment, and neurological disorder analysis Pan et al. (2023); Jia et al. (2025); Ding & Guan (2023); Ding et al. (2025a). Recently, EEG foundation models (EFMs) have emerged as a promising paradigm for learning generalizable representations from large-scale heterogeneous EEG data Wu et al. (2025); Yang et al. (2026). By combining large-scale pretraining with downstream fine-tuning, EFMs provide strong representations for diverse EEG decoding tasks.

Through empirical analysis, we first observe that **(i) lightweight models remain competitive but are task-dependent, whereas retaining high-capacity EFMs as inference backbones incurs**

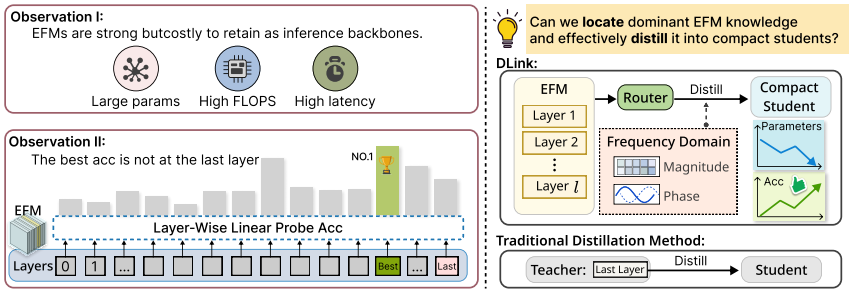


Figure 1: **Motivation and overview.** Compared with traditional final-layer distillation, DLink automatically aggregates layer-wise teacher knowledge and transfers it to a compact student.



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.

**substantial computational overhead**. Recent benchmarking further shows that increasing EEG-FM size does not necessarily improve generalization Liu et al. (2026). Directly using an adapted EFM keeps the entire massive backbone in the inference pipeline Lee et al. (2025). Moreover, EFM hidden states are typically organized as large latent representations, so attaching task-specific heads often requires additional pooling, projection, or flattening Yang et al. (2026), further increasing parameter and computational cost Döner et al. (2025). This cost motivates knowledge distillation as a natural route to transfer useful EFM knowledge into a compact student that is used alone after training.

Through layer-wise probing, we further observe that **(ii) task-discriminative EFM information is often more prominent in intermediate layers than in the final representation**. Although conventional distillation commonly uses the final logits or hidden representation as the teacher signal Kan et al. (2025), Fig. 1 shows that intermediate layers can provide stronger discriminative features for downstream EEG decoding. This suggests that effective EFM distillation should aggregate dominant intermediate-layer knowledge rather than rely on a fixed final-layer source.

Even when informative intermediate layers are identified, **their high-dimensional representations remain difficult to internalize compactly**. Unlike final logits, intermediate EFM features retain high-dimensional teacher-side representations. Directly flattening these features would require an oversized task-specific classifier, whereas naive dimensionality reduction may discard discriminative structure during transfer Dubois et al. (2020). Moreover, compact students often rely on aggressive projection or downsampling to reduce feature size, which can distort the spectrum of the ordered temporal-latent representation entering the compression block. This suggests that EFM distillation should not only identify which teacher layers to transfer, but also regularize how high-dimensional teacher knowledge is mapped into compact student representations.

These observations and transfer constraints raise a central question: ***Can we locate dominant EFM knowledge and effectively distill it into compact students?*** Addressing this question requires solving three technical challenges. **C1: Dominant knowledge localization and aggregation.** Because the useful teacher signal is not known a priori and may lie in intermediate layers, the distillation procedure should locate informative teacher-layer regions and aggregate them rather than rely on a fixed final representation. **C2: Compression-induced transfer loss.** The routed high-dimensional teacher representation must be transferred through projection and downsampling while preserving discriminative structure and limiting spectral distortion. **C3: Compact internalization.** The aggregated teacher knowledge must be absorbed into a lightweight inference path without a large classifier or teacher backbone.

To address these challenges, we propose **DLink** (**D**istilling **L**ayer-w**i**se and Domi**n**ant **K**nowledge), a spectrally guided distillation framework for transferring EFM knowledge into compact students. **For C1,** DLink introduces a lightweight layer router that learns soft weights over teacher layers from input-conditioned cues, allowing the student to receive supervision from dominant intermediate representations. **For C2,** DLink performs representation-level spectral distillation by aligning magnitude and phase spectra between the routed teacher representation and the student feature, which regularizes compact transfer and mitigates spectral distortion. **For C3,** DLink uses a compact project-then-compress student: it first maps the input into a feature space aligned with teacher features and then applies structured spatio-temporal compression before task prediction. Importantly, the teacher and router are discarded after training, leaving only the compact student for inference.

We evaluate DLink across six EEG benchmarks covering emotion recognition, mental-arithmetic decoding, depression diagnosis, motor-imagery decoding, and seizure prediction. DLink improves compact students while remaining lightweight at inference, with further analyses supporting the roles of learned layer aggregation and spectral distillation. The anonymized implementation is available at `https://anonymous.4open.science/r/DLink`.

## 2 Preliminaries and Problem Formulation

### 2.1 EEG Foundation Models

Let $X \in \mathcal{X}$ denote an EEG sample and $y \in \mathcal{Y}$ its corresponding label. An EFM teacher $\mathcal{T}$ represents $X$ as latent units and produces $L$ layer-wise hidden representations:

$$f_T^{(l)} = \mathcal{T}^{(l)}(X) \in \mathbb{R}^{N_l \times d_l}, \quad l = 1, \ldots, L, \tag{1}$$

where $N_l$ and $d_l$ denote the latent resolution and feature dimension. Representative EFMs, including LaBraM, CBraMod, and CSBrain, instantiate this hierarchy with different encoders Jiang et al. (2024); Wang et al. (2025); Zhou et al. (2025b). DLink uses hidden representations from the teacher and treats the final layer as only one candidate source within $\{f_T^{(l)}\}_{l=1}^{L}$.

In our implementation, DLink is instantiated separately for each teacher EFM. The router aggregates hidden states only within a single teacher, using the routed hidden-state list exposed at a common latent layout and hidden width after that teacher's tokenization. Accordingly, in the method section $L$ denotes this routed layer list, and we write $f_T^{(l)} \in \mathbb{R}^{N\times d}$ for all routed layers.

### 2.2 Direct Teacher–Student Distillation

Given a compact student $\mathcal{S}$, teacher–student distillation aims to transfer knowledge from the high-capacity teacher $\mathcal{T}$ into $\mathcal{S}$ while preserving downstream prediction performance Park et al. (2021); Gou et al. (2021); Ko et al. (2024). Let $f_S = \mathcal{S}_{\text{feat}}(X)$ and $\hat{y}_S = h_S(f_S)$ denote the student feature and prediction. Direct distillation typically uses the teacher's final logits or final hidden representation as the supervision target:

$$\mathcal{L}_{\text{direct}} = \mathcal{L}_{\text{cls}}(\hat{y}_S, y) + \lambda \mathcal{L}_{\text{KD}}, \quad \mathcal{L}_{\text{KD}} = \begin{cases} \tau^2 D_{\text{KL}}(\sigma(z_T/\tau) \,\|\, \sigma(z_S/\tau)), & \text{logit KD}, \\ d\left(\tilde{f}_S, f_T^{(L)}\right), & \text{feature KD}, \end{cases} \tag{2}$$

where $z_T$ and $z_S$ denote teacher- and student-side logits when a logit-KD baseline is instantiated, $\tilde{f}_S$ is the dimension-matched student feature used by the feature-KD baseline, $\sigma(\cdot)$ is softmax, $\tau$ is the temperature, and $d(\cdot,\cdot)$ is a feature discrepancy measure. In DLink, the teacher is used as a frozen representation encoder and no teacher-side classifier is required. This fixed-target formulation is simple, but it can underuse EFMs whose task-relevant information is distributed across intermediate layers Kan et al. (2025); Liang et al. (2023); Zhou et al. (2025a). It also leaves open how a compact student should inherit high-dimensional teacher knowledge without losing discriminative structure during compression.

### 2.3 Problem Statement: Dominant Knowledge Distillation under Compression

We formulate EFM distillation as a problem of dominant knowledge distillation under compression. Given a teacher EFM $\mathcal{T}$ with layer-wise representations $\{f_T^{(l)}\}_{l=1}^{L}$, the goal is to train a compact student $\mathcal{S}$ that is used alone at inference time while retaining the task-relevant knowledge of the teacher. Instead of prescribing a fixed teacher layer, we introduce a dominant knowledge operator $\mathcal{G}$:

$$f_T^* = \mathcal{G}\left(\{f_T^{(l)}\}_{l=1}^{L}, X\right), \tag{3}$$

where $f_T^*$ denotes the aggregated teacher knowledge used to supervise the student. The student, however, cannot simply copy this representation: $f_T^*$ may be high-dimensional and teacher-specific, whereas the student must encode it into a much smaller carrier. We refer to the task-relevant structure lost in this high-to-low-dimensional transfer as *compression-induced transfer loss*. The general objective is therefore

$$\min_{\theta_S,\theta_G} \mathbb{E}_{(X,y)}\left[\mathcal{L}_{\text{cls}}(\hat{y}_S, y) + \lambda \mathcal{D}_{\text{comp}}\left(\psi_S(X), f_T^*\right)\right], \quad \text{s.t. } B(\mathcal{S}) \ll B(\mathcal{T}), \tag{4}$$

where $\psi_S(X)$ denotes the student representation used for distillation, $\mathcal{D}_{\text{comp}}(\cdot,\cdot)$ is the discrepancy used for compact transfer, and $B(\cdot)$ denotes an inference budget such as parameters or FLOPs.

This formulation gives three requirements: automatically aggregating dominant teacher layers (C1), preserving teacher-induced structure during compact transfer (C2), and keeping the inference path teacher-free and lightweight (C3).

## 3 Method: DLink

### 3.1 Framework Overview

Given a frozen teacher EFM $\mathcal{T}$ and a student $\mathcal{S}$, DLink distills layer-wise teacher representations into the student while discarding the teacher at inference. During DLink optimization, the teacher encoder

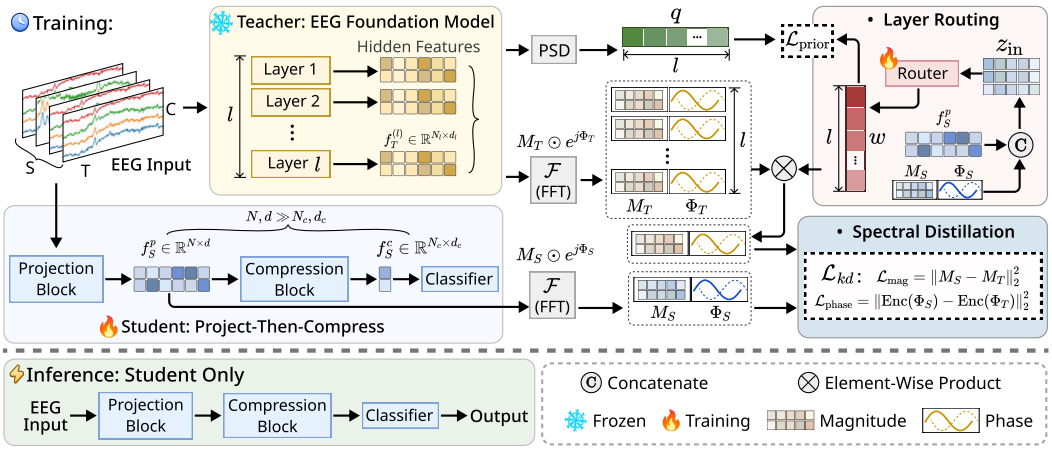


Figure 2: **Overview of the DLink pipeline.** DLink transfers knowledge from a frozen EEG foundation-model teacher to a compact student. During training, an input-conditioned layer router softly aggregates dominant teacher-layer representations, and spectral distillation aligns the resulting routed teacher representation with the student feature in learned representation spectra. After distillation, both the teacher and router are discarded, leaving only the compact student for inference.

is fixed and no teacher-side classifier is used. As shown in Fig. 2, DLink combines a layer router, representation-level spectral distillation, and a compact project-then-compress student. The student exposes two internal representations:

$$f_S^p = P_{\theta_S}(X) \in \mathbb{R}^{N\times d}, \qquad f_S^c = C_{\theta_S}(f_S^p) \in \mathbb{R}^{N_c\times d_c}, \tag{5}$$

where $P_{\theta_S}(\cdot)$ is the projection block for teacher–student alignment, $C_{\theta_S}(\cdot)$ is the compression block, and $N_c d_c \ll Nd$. The target shape $(N, d)$ is teacher-specific and is matched to the common shape of the routed hidden layers for the current teacher. The representation $f_S^p$ is used for distillation, while $f_S^c$ is used for compact prediction.

### 3.2 Layer-wise Knowledge Aggregation

DLink avoids assuming that the final teacher layer is always the most informative supervision source. Instead, it learns a lightweight layer router that assigns soft importance weights to teacher layers from student-side cues:

$$z_{\text{in}} = \left[\text{Pool}(f_S^p) \parallel \text{Pool}\left(|\mathcal{F}_N(f_S^p)|\right)\right], \tag{6}$$

where $\mathcal{F}_N(\cdot)$ denotes a one-dimensional Fourier transform along the ordered temporal-latent axis $N$, $|\mathcal{F}_N(\cdot)|$ denotes the corresponding representation spectrum, $\text{Pool}(\cdot)$ denotes lightweight pooling over the resulting feature map, and $\parallel$ denotes concatenation. The router maps $z_{\text{in}}$ to layer-wise logits:

$$a = R_{\theta_R}(z_{\text{in}}) \in \mathbb{R}^L, \tag{7}$$

which are normalized by a temperature-scaled softmax:

$$w_l = \frac{\exp(a_l/\tau)}{\sum_{k=1}^{L}\exp(a_k/\tau)}, \qquad l = 1, \ldots, L. \tag{8}$$

The resulting vector $w = [w_1, \ldots, w_L]$ defines an input-conditioned distribution over the routed teacher layers. The routed teacher representation is obtained by weighted aggregation:

$$f_T^* = \sum_{l=1}^{L} w_l f_T^{(l)}. \tag{9}$$

To stabilize routing, we further introduce a representation-level spectral-concentration prior over teacher layers. For each input and teacher representation $f_T^{(l)}$, we compute aggregated spectral energy over representation-frequency bins induced by the temporal-latent axis:

$$e_l(\omega) = \left\|\mathcal{F}_N(f_T^{(l)})(\omega)\right\|_2^2, \qquad p_l(\omega) = \frac{e_l(\omega)}{\sum_{\omega'\in\Omega} e_l(\omega') + \epsilon}, \tag{10}$$

where $\Omega$ indexes representation-frequency components along the temporal-latent axis. This prior is computed from learned teacher representations and should not be interpreted as raw-EEG spectral supervision. We reduce this distribution to a scalar layer score by its spectral concentration:

$$s_l = \sum_{\omega \in \Omega} p_l(\omega)^2. \tag{11}$$

For each input, the layer scores are normalized into a soft guidance distribution with guidance temperature $\tau_q$:

$$q_l = \frac{\exp(s_l/\tau_q)}{\sum_{k=1}^{L} \exp(s_k/\tau_q)}. \tag{12}$$

The router is regularized by

$$\mathcal{L}_{\text{prior}} = D_{\text{KL}}(w \,\|\, q). \tag{13}$$

### 3.3 Representation-Spectrum Consistency for Compact Transfer

DLink transfers the routed teacher representation to the student through representation-spectrum consistency. For the teacher–student feature pairs used by DLink, we write $f \in \mathbb{R}^{N \times d}$ as a notational view: $N$ denotes the ordered temporal-latent axis used for spectral alignment, while $d$ denotes the remaining feature indices. We apply a one-dimensional Fourier transform $\mathcal{F}_N$ along the $N$ axis, independently over the remaining feature indices. Thus, the spectral bins describe coarse-to-fine variation across adjacent temporal latent units in the learned representation; they are not raw EEG frequencies in Hz. This axis is meaningful for compact transfer because the compression block reduces temporal latent resolution through projection and spatio-temporal downsampling. We therefore apply spectral distillation before compression as a feature-level regularizer.

A direct feature-distillation baseline would align the same routed teacher representation by pointwise matching, which constrains Euclidean discrepancy but does not explicitly regulate the temporal-latent spectrum entering the compression block. DLink instead constrains the magnitude and phase spectra of $f_S^p$ to follow those of $f_T^*$.

Given the student projection feature $f_S^p$ and the routed teacher representation $f_T^*$, we compute their spectral representations:

$$\mathcal{F}_N(f_S^p) = M_S \odot e^{j\Phi_S}, \qquad \mathcal{F}_N(f_T^*) = M_T \odot e^{j\Phi_T}, \tag{14}$$

where $M_S, M_T$ are magnitude spectra and $\Phi_S, \Phi_T$ are phase spectra. The magnitude term aligns representation-level spectral magnitudes:

$$\mathcal{L}_{\text{mag}} = \|M_S - M_T\|_F^2 . \tag{15}$$

Directly matching raw phase values is unstable due to the $2\pi$ periodicity. Following a continuous phase encoding, we represent phase by sine and cosine components:

$$\text{Enc}(\Phi) = [\cos\Phi, \sin\Phi]. \tag{16}$$

The phase consistency term is then defined as

$$\mathcal{L}_{\text{phase}} = \|\text{Enc}(\Phi_S) - \text{Enc}(\Phi_T)\|_F^2 . \tag{17}$$

The spectral distillation objective is

$$\mathcal{L}_{\text{kd}} = \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{phase}}. \tag{18}$$

### 3.4 Compact Project-then-Compress Student

The student is designed as a compact carrier for internalizing routed teacher knowledge. It follows a project-then-compress principle. The CNN–Transformer projection block first produces $f_S^p$ for spectral distillation against $f_T^*$. A CNN-based compression block then applies spatio-temporal downsampling to obtain the compact feature $f_S^c$, which is fed into a compact MLP head for prediction, i.e., $\hat{y}_S = h_{\theta_S}(f_S^c)$. This design separates knowledge alignment from compact prediction. The CNN–Transformer projection block provides sufficient capacity to absorb teacher-side supervision, while CNN downsampling prevents the MLP classifier from directly operating on oversized teacher-like features. Detailed layer configurations and student size variants are provided in Appendix B.1.

### 3.5 Training Objective and Inference Path

The final training objective combines task supervision, spectral distillation, and router guidance:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}}(\hat{y}_S, y) + \lambda \mathcal{L}_{\text{kd}} + \mathcal{L}_{\text{prior}}. \tag{19}$$

During training, the teacher EFM $\mathcal{T}$ is frozen and provides only layer-wise representations for DLink. The student $\mathcal{S}$ and router $R_{\theta_R}$ are optimized jointly. Inference is performed only by the student.

## 4 Experiments

We evaluate DLink from four aspects: compact-student accuracy, learned layer aggregation, spectral transfer behavior, and inference efficiency.

### 4.1 Experimental Setup

**Datasets and metrics.** We evaluate six EEG benchmarks covering emotion recognition, mental-arithmetic decoding, depression diagnosis, motor-imagery decoding, and seizure prediction: FACED Chen et al. (2023), mental-arithmetic Zyma et al. (2019), Mumtaz2016 Mumtaz (2016), PhysioNet-MI Goldberger et al. (2000); Schalk et al. (2004), SHU-MI Ma et al. (2022), and Siena Detti et al. (2020). The main paper reports FACED, mental-arithmetic, Siena, and SHU-MI; Mumtaz2016 and PhysioNet-MI are reported in the Appendix. Following the common preprocessing protocol, all EEG signals are resampled to 200 Hz and represented as 1-second patches. We report balanced accuracy (ACC-B) for all tasks, together with task-appropriate secondary metrics. Dataset statistics and exact split rules are provided in Appendix A.2.

**Baselines.** We compare DLink with lightweight EEG models (EEGNet Lawhern et al. (2018), EEGConformer Song et al. (2023), and EEGDeformer Ding et al. (2025b)), direct feature/logit distillation methods (FitNets Romero et al. (2015) and Logit-std Sun et al. (2024)), and structured KD methods covering sample-wise multi-layer routing (SAKD Song et al. (2022)), contrastive representation matching (CRD Tian et al. (2020)), multi-layer feature aggregation (ReviewKD Chen et al. (2021)), relational distillation (RKD Park et al. (2019)), and frequency-aware distillation (FreeKD Zhang et al. (2024)). When required, lightweight feature-space adapters pool or map teacher hidden states to 200 dimensions before the method-specific loss is computed.

**Implementation.** The main DLink comparison uses the Stu-S student and a fine-tuned CBraMod teacher checkpoint Wang et al. (2025). The teacher is frozen during student training, supplies representations only, and is discarded at inference. The FACED scale comparison additionally evaluates DLink-M. For every dataset, teacher, student, and KD method, we perform a two-stage protocol. Stage 1 searches candidate classification heads, learning rates, and method-specific KD weights using only the training and validation splits, selecting the configuration with the best validation ACC-B. Candidate values and head definitions are provided in Appendix C. In Stage 2, the selected configuration is fixed, and seeds $\{0, 1, 2, 3, 4\}$ are run independently. Each seed selects the checkpoint with the best validation ACC-B before one evaluation on the held-out test split.

All runs use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.001, gradient clipping at maximum norm 1.0, cosine annealing to a minimum learning rate of $10^{-6}$, and at most 100 epochs. Early stopping is applied after 10 epochs without validation-ACC-B improvement. Further split, manifest, tensor-shape, and preprocessing details are provided in Appendix A.2 and Appendix C.

### 4.2 Main Results

Table 1 summarizes the performance of the Stu-S student across FACED, mental-arithmetic, Siena, and SHU-MI. DLink improves the matched supervised student under ACC-B and remains competitive with direct KD baselines and standalone lightweight models. Bold and underlined values denote the best and second-best results within each dataset and metric. The two student variants share the same projection-then-compression design but differ in capacity; their detailed configurations are given in Appendix B.1. Results on Mumtaz2016 and PhysioNet-MI are reported in Appendix A.4.

The main comparison uses a frozen fine-tuned CBraMod teacher. Table 2 reports the FACED teacher and student-scale comparison for DLink-S and DLink-M. DLink-M uses the same routed distillation

Table 1: Main results across four EEG benchmarks. Values are mean ± standard deviation over five seeds. The DLink rows use the fine-tuned CBraMod teacher. Bold and underlined values denote the best and second-best results within each dataset and metric.

| Category | Model | FACED (9-class) | | | mental-arithmetic (2-class) | | |
|---|---|---|---|---|---|---|---|
| | | ACC-B (%) | F1-W (%) | Kappa (%) | ACC-B (%) | AUROC (%) | AUC-PR (%) |
| Small Models | EEGNet | 22.81 ± 2.65 | 17.71 ± 3.53 | 12.94 ± 3.06 | 60.49 ± 2.79 | 71.48 ± 12.52 | 50.57 ± 16.12 |
| | EEGConformer | 43.24 ± 1.88 | **44.05 ± 1.88** | 37.72 ± 0.86 | 62.04 ± 11.86 | 70.22 ± 8.28 | **51.71 ± 12.01** |
| | EEGDeformer | 34.40 ± 1.26 | 34.08 ± 1.45 | 25.92 ± 1.23 | 61.35 ± 6.13 | 71.00 ± 4.67 | 50.11 ± 9.90 |
| | Stu(DLink) | 41.61 ± 0.76 | 41.29 ± 0.69 | 33.94 ± 0.85 | 60.90 ± 3.78 | 69.29 ± 4.16 | 43.67 ± 1.67 |
| KD | FitNets | 42.39 ± 0.68 | 42.17 ± 0.67 | 34.86 ± 0.83 | 60.49 ± 2.79 | 69.48 ± 10.23 | 42.93 ± 5.06 |
| | Logit-std | 41.25 ± 1.25 | 41.17 ± 1.27 | 33.67 ± 1.41 | 61.74 ± 3.17 | 71.32 ± 3.45 | 50.25 ± 21.79 |
| | S-A KD | 41.09 ± 0.76 | 40.61 ± 0.95 | 33.36 ± 0.91 | 59.10 ± 2.26 | 68.61 ± 4.43 | 42.95 ± 6.00 |
| | CRD | 42.59 ± 1.01 | 42.31 ± 0.89 | 35.01 ± 1.06 | 58.89 ± 4.48 | 68.56 ± 3.20 | 43.90 ± 6.40 |
| | ReviewKD | 41.22 ± 0.77 | 40.74 ± 0.98 | 33.54 ± 0.92 | 61.60 ± 2.89 | 67.25 ± 3.54 | 41.94 ± 5.29 |
| | RKD | 39.38 ± 1.92 | 39.21 ± 1.85 | 31.46 ± 2.10 | 57.15 ± 3.42 | 63.05 ± 7.34 | 36.91 ± 4.09 |
| | FreeKD | 41.29 ± 0.38 | 40.75 ± 0.57 | 33.56 ± 0.47 | 61.22 ± 2.02 | 70.78 ± 8.32 | 51.00 ± 12.93 |
| | **DLink** | **43.46 ± 0.80** | 43.06 ± 0.90 | **37.89 ± 0.95** | **62.21 ± 5.14** | **71.49 ± 4.60** | 44.59 ± 2.91 |
| **Category / Model** | | **Siena (2-class)** | | | **SHU (2-class)** | | |
| | | **ACC-B (%)** | **AUROC (%)** | **AUC-PR (%)** | **ACC-B (%)** | **AUROC (%)** | **AUC-PR (%)** |
| Small Models | EEGNet | 68.57 ± 6.09 | 90.90 ± 1.43 | 33.49 ± 4.98 | 57.14 ± 2.21 | 64.84 ± 1.25 | 63.05 ± 1.15 |
| | EEGConformer | 70.40 ± 4.75 | 90.23 ± 2.75 | 46.58 ± 1.90 | 60.14 ± 2.36 | 66.51 ± 2.22 | 64.63 ± 2.13 |
| | EEGDeformer | 72.72 ± 6.66 | 85.94 ± 3.74 | 40.15 ± 4.87 | 58.29 ± 1.85 | 66.76 ± 0.77 | 65.03 ± 0.70 |
| | Stu(DLink) | 69.55 ± 4.84 | 86.71 ± 3.88 | 40.75 ± 4.30 | 60.84 ± 0.92 | 67.07 ± 1.90 | 67.74 ± 1.84 |
| KD | FitNets | 69.40 ± 5.17 | 88.70 ± 3.12 | 38.19 ± 4.51 | 58.48 ± 1.97 | 62.10 ± 2.81 | 61.48 ± 3.03 |
| | Logit-std | 72.01 ± 7.26 | 87.94 ± 3.74 | 40.53 ± 6.83 | 56.23 ± 1.52 | 62.74 ± 1.63 | 62.95 ± 1.94 |
| | S-A KD | 67.14 ± 2.91 | 89.87 ± 7.91 | 38.99 ± 7.33 | 61.03 ± 0.68 | 66.47 ± 1.43 | 67.40 ± 1.20 |
| | CRD | 66.00 ± 3.12 | 88.24 ± 2.36 | 39.97 ± 5.04 | 61.88 ± 1.65 | 66.67 ± 2.19 | 67.56 ± 1.88 |
| | ReviewKD | 68.36 ± 4.11 | 89.91 ± 4.59 | 39.18 ± 6.06 | 61.15 ± 0.99 | 66.53 ± 1.46 | 66.22 ± 1.13 |
| | RKD | 68.36 ± 4.28 | 90.33 ± 1.13 | 39.47 ± 4.90 | 60.74 ± 1.46 | 66.51 ± 1.87 | 66.43 ± 1.59 |
| | FreeKD | 69.52 ± 2.87 | 89.03 ± 5.84 | 40.59 ± 9.10 | 61.26 ± 1.03 | 67.66 ± 1.08 | 67.98 ± 1.37 |
| | **DLink** | **78.79 ± 2.63** | **91.92 ± 2.15** | **48.73 ± 2.18** | **62.89 ± 1.17** | **68.43 ± 1.72** | **68.81 ± 1.47** |

Table 2: FACED teacher and student-scale comparison. DLink-M denotes the larger student configuration; detailed model specifications are provided in Appendix B.1.

| DLink-S | | | | DLink-M | | | |
|---|---|---|---|---|---|---|---|
| Teacher | ACC-B (%) | F1-W (%) | Kappa (%) | Teacher | ACC-B (%) | F1-W (%) | Kappa (%) |
| Stu (supervised) | 41.61 ± 0.76 | 41.29 ± 0.69 | 33.94 ± 0.85 | Stu (supervised) | 47.35 ± 1.33 | 47.13 ± 1.56 | 40.36 ± 1.54 |
| CBraMod | 43.46 ± 0.80 | **43.06 ± 0.90** | **37.89 ± 0.95** | CBraMod | **49.81 ± 1.10** | **49.92 ± 0.81** | **43.21 ± 1.13** |
| CSBrain | **43.48 ± 0.66** | 42.91 ± 0.87 | 35.77 ± 1.08 | CSBrain | 49.62 ± 0.66 | 49.55 ± 0.43 | 42.97 ± 0.67 |
| LaBraM | 42.28 ± 1.14 | 41.66 ± 1.31 | 34.62 ± 1.23 | LaBraM | 49.20 ± 1.29 | 49.13 ± 1.31 | 42.56 ± 1.40 |

design as DLink-S with a slightly larger student, bringing performance into the fine-tuned-teacher range while retaining a substantially smaller parameter footprint (Fig. 5).

### 4.3 Dominant Knowledge Aggregation

#### 4.3.1 Ablation of Layer Routing

Table 3 reports routing ablations on FACED and SHU-MI. **Fixed Last** uses the final teacher representation, whereas **Fixed Avg** uniformly averages all teacher layers. Mean Power and Max Amp replace the spectral-concentration score without changing the router or distillation objective. **Routed L2** retains the same router and routed representation $f_T^*$ but uses pointwise squared-L2 instead of magnitude–phase matching. Freq-only and Time-only retain only the spectrum and pooled student-feature cues, respectively. DLink performs best across all reported metrics on both datasets. The prior and score ablations support spectral concentration as a soft routing cue, while Routed L2 supports representation-spectrum consistency over pointwise matching. The learned weights are neither near-one-hot nor exactly uniform; Appendix E reports detailed statistics.

Fig. 3 shows the FACED case, where the router progressively develops a stable top-1 preference toward $L9$, coinciding with the best linear-probe and student-matching layers in this run. The router

Table 3: Layer-routing and spectral-transfer ablations on FACED and SHU-MI.

| Variant | FACED (9-class) ACC-B (%) | F1-W (%) | Kappa (%) | Variant | SHU (2-class) ACC-B (%) | AUROC (%) | AUC-PR (%) |
|---|---|---|---|---|---|---|---|
| w/o $\mathcal{L}_{\text{prior}}$ | 39.68 ± 1.23 | 39.48 ± 1.19 | 31.86 ± 1.34 | w/o $\mathcal{L}_{\text{prior}}$ | 58.97 ± 1.20 | 65.15 ± 1.90 | 64.28 ± 2.00 |
| Fixed Last | 40.07 ± 0.69 | 39.95 ± 0.69 | 32.34 ± 0.80 | Fixed Last | 59.36 ± 1.00 | 65.62 ± 1.80 | 64.76 ± 2.00 |
| Fixed Avg | 42.48 ± 1.14 | 42.35 ± 0.95 | 35.02 ± 1.18 | Fixed Avg | 55.84 ± 0.98 | 55.62 ± 1.23 | 61.69 ± 1.87 |
| Fixed Best Val | 43.10 ± 0.97 | 42.04 ± 0.89 | 35.19 ± 0.92 | Fixed Best Val | 56.81 ± 0.93 | 55.98 ± 1.75 | 55.71 ± 2.31 |
| Mean Power | 42.73 ± 0.58 | 42.59 ± 0.42 | 35.33 ± 0.59 | Mean Power | 62.02 ± 1.20 | 68.26 ± 1.80 | 67.75 ± 2.00 |
| Max Amp | 42.12 ± 0.94 | 41.91 ± 1.01 | 34.36 ± 1.06 | Max Amp | 61.41 ± 1.20 | 67.58 ± 1.90 | 66.78 ± 2.10 |
| Routed L2 | 40.64 ± 0.40 | 39.98 ± 0.45 | 33.88 ± 0.43 | Routed L2 | 61.04 ± 2.34 | 61.56 ± 3.61 | 68.02 ± 3.21 |
| Freq-only router | 43.05 ± 0.42 | 42.71 ± 0.56 | 33.67 ± 0.62 | Freq-only router | 60.45 ± 5.21 | 60.12 ± 4.31 | 67.01 ± 2.71 |
| Time-only router | 42.98 ± 0.70 | 41.89 ± 0.87 | 31.96 ± 0.90 | Time-only router | 59.54 ± 2.84 | 59.82 ± 3.52 | 66.13 ± 4.38 |
| **DLink** | **43.46 ± 0.80** | **43.06 ± 0.90** | **37.89 ± 0.95** | **DLink** | **62.89 ± 1.17** | **68.43 ± 1.72** | **68.81 ± 1.47** |

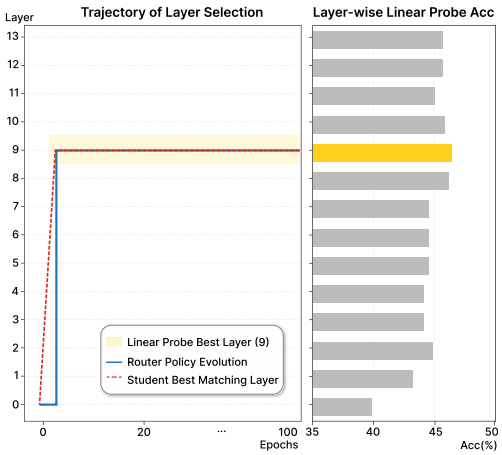


Figure 3: Layer routing behavior.

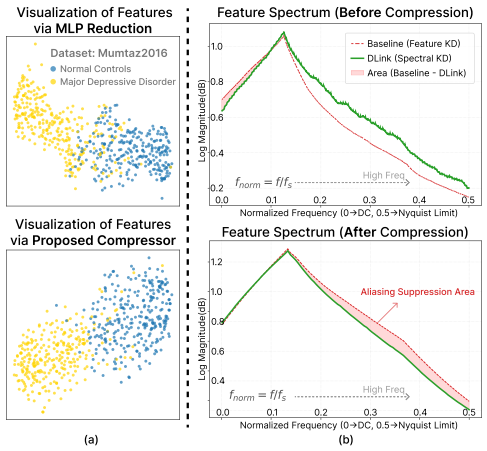


Figure 4: Feature and spectrum analysis.

nevertheless retains a high-entropy soft aggregation over teacher layers. The corresponding routing statistics are provided in Appendix E.

### 4.4 Spectral Distillation Analysis

On FACED, DLink improves ACC-B over the direct feature and logit KD baselines FitNets and Logit-std by $+1.07$ and $+2.21$ points, respectively (Table 1). Appendix F.6 provides focused objective comparisons and magnitude-only, phase-only, and shuffled-ordering controls.

Fig. 4 analyzes the learned representations before and after compression. The t-SNE plots show that compressed DLink features retain class separation comparable to the full-dimensional features. The spectrum plots further show a smaller high-frequency tail than Feature KD after compression, consistent with reduced representation-spectrum distortion.

### 4.5 Compact Internalization and Inference Cost

The router adds little training overhead because it operates on compact feature and representation-spectrum cues rather than the original high-dimensional feature. In CUDA profiling, it incurs $0.633$ ms forward latency per training batch, with $0.04$M parameters and $0.22$M FLOPs (Appendix Tables B3 and G1). CPU-only batch-1 profiling on Mumtaz2016 further shows that the compact student runs in $3.653$ ms, compared with $19.345$ ms for the fine-tuned CBraMod teacher, giving a $5.30\times$ speedup (Appendix G1).

Fig. 5 compares accuracy, FLOPs, and parameter volume on FACED. Under ACC-B, DLink-S is competitive with strong standalone lightweight EEG models. Table 2 further shows that the larger DLink-M configuration can substantially improve the compact student and reduce the remaining gap to teacher-guided results without requiring a teacher-scale inference model. The resulting model does not carry any teacher-side module into deployment.

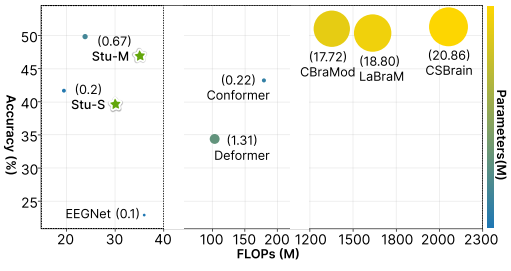


Figure 5: Accuracy–efficiency trade-off

## 5 Related Work

### 5.1 EEG Foundation Models

Recent EEG foundation models (EFMs) have advanced neural representation learning with large-scale and heterogeneous EEG corpora Ma et al. (2026); Ouahidi et al. (2026). Representative models such as EEGPT Wang et al. (2024), LaBraM Jiang et al. (2024), CBraMod Wang et al. (2025), and CSBrain Zhou et al. (2025b) learn transferable representations for downstream decoding, while benchmarks such as AdaBrainBench Wu et al. (2025) highlight their cross-subject and cross-task potential. However, EFMs usually rely on high-capacity backbones, making it costly to retain them directly for inference. DLink therefore treats EFMs as training-time teachers and focuses on transferring their layer-wise knowledge into a compact inference model.

### 5.2 Knowledge Distillation and Spectral Motivation

Knowledge distillation transfers information from a high-capacity teacher to a compact student through logit-level or feature-level supervision Park et al. (2021); Gou et al. (2021); Ko et al. (2024). Classical feature distillation encourages students to mimic teacher representations Romero et al. (2015); Gu et al. (2024), while logit distillation matches the output distribution Sun et al. (2024); related ideas have also been studied in physiological and EEG settings Liu et al. (2023); Kan et al. (2025). For hierarchical EFMs, however, useful decoding information may not be concentrated in the final representation. Prior layer-wise transfer and probing studies show that different pretrained layers can encode different task-relevant information Liang et al. (2023); Ramtoula et al. (2026). DLink follows this motivation but avoids manually fixing a teacher layer by using a lightweight router to automatically aggregate dominant intermediate layers.

Frequency-domain modeling is central to EEG analysis, and recent EEG representation learning methods also exploit spectral or spectral-spatial-temporal structures Li et al. (2022); Ding et al. (2023; 2024); Wang et al. (2024). Frequency-domain objectives have further been used for representation transfer and distillation beyond EEG Zhang et al. (2024). In convolutional networks, anti-aliased designs show that frequency-aware regularization can be useful on ordered learned feature maps before resolution reduction Zhang (2019); Ribeiro & Schön (2021). DLink follows this representation-level motivation by aligning magnitude and phase spectra along the temporal-latent axis as a regularizer for compact transfer. This objective is applied to learned features rather than raw EEG reconstruction, so it is used to mitigate representation-spectrum distortion during teacher–student transfer rather than to claim recovery of physiological rhythms.

## 6 Conclusion

This paper presents DLink, a compact EFM distillation framework that transfers dominant intermediate teacher knowledge through input-conditioned layer routing and representation-spectrum consistency. A project-then-compress student internalizes the routed knowledge, and both the teacher and router are discarded after training. Across six EEG benchmarks, DLink performs competitively against diverse EEG baselines, providing a practical route for distilling hierarchical EFM representations into efficient inference models.

## AI USE STATEMENT

During preparation of this manuscript, generative AI tools were used to assist with LaTeX formatting and manuscript editing. The authors reviewed and verified the resulting text, equations, claims, and code-related content, and take responsibility for the final content of this work.

## REFERENCES

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10(1):740, oct 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02650-w. URL https://doi.org/10.1038/s41597-023-02650-w.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5006–5015, 2021. doi: 10.1109/CVPR46437.2021.00497.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. Eeg synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7), 2020. ISSN 2227-9717. doi: 10.3390/pr8070846. URL https://www.mdpi.com/2227-9717/8/7/846.

Yi Ding and Cuntai Guan. Gign: Learning graph-in-graph representations of eeg signals for continuous emotion recognition. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–5, 2023. doi: 10.1109/EMBC40787.2023.10340644.

Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2023. doi: 10.1109/TAFFC.2022.3169001.

Yi Ding, Chengxuan Tong, Shuailei Zhang, Muyun Jiang, Yong Li, Kevin Lim Jun Liang, and Cuntai Guan. EmT: A novel transformer for generalized cross-subject EEG emotion recognition, June 2024.

Yi Ding, Joon Hei Lee, Shuailei Zhang, Tianze Luo, and Cuntai Guan. Decoding human attentive states from spatial-temporal eeg patches using transformers. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–5, 2025a. doi: 10.1109/EMBC58623.2025.11254148.

Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. EEG-Deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2025b. doi: 10.1109/JBHI.2024.3504604.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025. URL https://openreview.net/forum?id=TWaw5qtKQf.

Yann Dubois, Douwe Kiela, David J. Schwab, and Ramakrishna Vedantam. Learning optimal representations with the decodable information bottleneck. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Ary L. Goldberger, Luis A. N. 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. 101(23), 2000. ISSN 0009-7322, 1524-4539. doi: 10.1161/01.CIR.101.23.e215. URL https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *Int. J. Comput. Vision*, 129(6):1789–1819, June 2021. ISSN 0920-5691. doi: 10.1007/s11263-021-01453-z. URL https://doi.org/10.1007/s11263-021-01453-z.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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*, MM '25, pp. 5509–5518, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754775. URL https://doi.org/10.1145/3746027.3754775.

Wei-Bang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Representation Learning*, volume 2024, pp. 16405–16426, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/47393e8594c82ce8fd83adc672cf9872-Paper-Conference.pdf.

Siyuan Kan, Huanyu Wu, Zhenyao Cui, Fan Huang, Xiaolong Xu, and Dongrui Wu. CMCRD: cross-modal contrastive representation distillation for emotion recognition. *CoRR*, abs/2504.09221, 2025. doi: 10.48550/ARXIV.2504.09221. URL https://doi.org/10.48550/arXiv.2504.09221.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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, jul 2018. doi: 10.1088/1741-2552/aace8c. URL https://dx.doi.org/10.1088/1741-2552/aace8c.

Na Lee, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Are large brainwave foundation models capable yet ? insights from fine-tuning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=J5SbLoq7Uv.

Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu. A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 6–14, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548243. URL https://doi.org/10.1145/3503161.3548243.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: task-aware layer-wise distillation for language model compression. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG-FM-Compass: Progress, Benchmarking, and Future Directions for EEG Foundation Models. *National Science Review*, pp. nwag466, 08 2026. ISSN 2095-5138. doi: 10.1093/nsr/nwag466. URL https://doi.org/10.1093/nsr/nwag466.

Yucheng Liu, Ziyu Jia, and Haichao Wang. EmotionKD: A cross-modal knowledge distillation framework for emotion recognition based on physiological signals. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 6122–6131, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612277. URL https://doi.org/10.1145/3581783.3612277.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=msJgEkjwh5.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, September 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01647-1.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Alexandre Gramfort, and Denis A. Engemann. Geodesic optimization for predictive shift adaptation on EEG data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 32828–32855. Curran Associates, Inc., 2024. doi: 10.52202/079017-1033.

Wajid Mumtaz. MDD patients and healthy controls EEG data (new). 11 2016. doi: 10.6084/m9.figshare.4244171.v2. URL https://figshare.com/articles/dataset/EEG_Data_New/4244171.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG - adapting to any setup with large-scale pretraining on 25,000 subjects. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=ZeFMtRBy4Z.

Bei Pan, Kaoru Hirota, Zhiyang Jia, and Yaping Dai. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561:126866, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126866. URL https://www.sciencedirect.com/science/article/pii/S092523122300989X.

Dae Young Park, Moon-Hyun Cha, changwook jeong, Daesin Kim, and Bohyung Han. Learning student-friendly teacher networks for knowledge distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13292–13303. Curran Associates, Inc., 2021.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, 2019. doi: 10.1109/CVPR.2019.00409.

Benjamin Ramtoula, Pierre-Yves Lajoie, Paul Newman, and Daniele De Martini. Fantastic features and where to find them: A probing method to combine features from multiple foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=PxgIElCohI.

Antônio H. Ribeiro and Thomas B. Schön. How convolutional neural networks deal with aliasing. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2755–2759, 2021. doi: 10.1109/ICASSP39728.2021.9414627.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6550.

G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on Biomedical Engineering*, 51 (6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072.

Jie Song, Ying Chen, Jingwen Ye, and Mingli Song. Spot-adaptive knowledge distillation. *IEEE Transactions on Image Processing*, 31:3359–3370, 2022. doi: 10.1109/TIP.2022.3170728.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740, 2024. doi: 10.1109/CVPR52733.2024.01489.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgpBJrtvS.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=lvS2b8CjG5.

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 Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 75310–75346, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bbbd6d915cb90be21c1254a82d45cedd-Paper-Conference.pdf.

Jiamin Wu, Zichen Ren, Junyu Wang, Pengyu Zhu, Yonghao Song, Mianxin Liu, Qihao Zheng, Lei Bai, Wanli Ouyang, and Chunfeng Song. Adabrain-bench: Benchmarking brain foundation models for brain-computer interface applications. *CoRR*, abs/2507.09882, 2025. doi: 10.48550/ARXIV.2507.09882. URL https://doi.org/10.48550/arXiv.2507.09882.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh.

Richard Zhang. Making convolutional networks Shift-Invariant again. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7324–7334. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19a.html.

Yuan Zhang, Tao Huang, Jiaming Liu, Tao Jiang, Kuan Cheng, and Shanghang Zhang. FreeKD: Knowledge distillation via semantic frequency prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15931–15940, June 2024.

Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *IEEE Signal Processing Magazine*, 42(5):22–35, 2025a. doi: 10.1109/MSP.2025.3592356.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 87150–87195. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2915. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7e199ad8ae40eb19b2980f61f659cb07-Paper-Conference.pdf.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1), 2019. ISSN 2306-5729. doi: 10.3390/data4010014. URL https://www.mdpi.com/2306-5729/4/1/14.

# SUPPLEMENTARY MATERIAL

## A DATASETS, METRICS, AND PREPROCESSING

### A.1 EVALUATION METRICS

We report task metrics according to the label structure of each benchmark. For all datasets, balanced accuracy (ACC-B) is used to reduce the effect of class imbalance. For multi-class benchmarks, we additionally report weighted F1 (F1-W) and Cohen's Kappa. For binary benchmarks, we additionally report AUROC and AUC-PR.

Let $n_c$ be the number of samples in class $c$, and let $\mathrm{TP}_c$ be the number of correctly predicted samples in class $c$. Balanced accuracy is

$$\text{ACC-B} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \frac{\mathrm{TP}_c}{n_c}. \tag{20}$$

Weighted F1 averages class-wise F1 scores using class support as weights. Cohen's Kappa measures agreement between prediction and label after correcting for chance agreement. AUROC is the area under the receiver operating characteristic curve, and AUC-PR is the area under the precision–recall curve.

Table A1: Evaluation metrics reported for each benchmark according to task label structure.

| **Dataset** | **Task** | **Main metric** | **Additional metrics** |
|---|---|---|---|
| FACED | 9-class emotion recognition | ACC-B | F1-W, Kappa |
| mental-arithmetic | Mental arithmetic | ACC-B | AUROC, AUC-PR |
| Mumtaz2016 | Binary depression diagnosis | ACC-B | AUROC, AUC-PR |
| PhysioNet-MI | 4-class motor imagery | ACC-B | F1-W, Kappa |
| SHU-MI | Binary motor imagery | ACC-B | AUROC, AUC-PR |
| Siena | Binary seizure prediction | ACC-B | AUROC, AUC-PR |

### A.2 DATASET DETAILS AND PREPROCESSING

Table A2 summarizes the downstream benchmark sample counts, and Table A3 reports the exact train/validation/test partitions. We follow the common preprocessing setting: EEG signals are resampled to 200 Hz and represented as 1-second patches with 200 time points per patch. For benchmark comparability, Mumtaz2016 follows the subject-independent protocol reported by CBraMod Wang et al. (2025), and Siena follows the cross-subject test boundary of AdaBrain-Bench Wu et al. (2025). All partitions are defined before patching or windowing using the dataset-specific grouping unit stated in Table A3. The preprocessing scale factor is 0.01.

Table A2: Dataset statistics and post-split sample counts for the six downstream benchmarks.

| **Dataset** | **Subjects /files** | **Task** | **Train** | **Val** | **Test** | **Total** |
|---|---|---|---|---|---|---|
| FACED | 123 subjects | 9-class emotion | 6,720 | 1,680 | 1,932 | 10,332 |
| mental-arithmetic | 36 subjects | Mental arithmetic | 1,343 | 172 | 192 | 1,707 |
| Mumtaz2016 | 34 MDD / 30 NC | Binary depression | 4,891 | 1,041 | 1,211 | 7,143 |
| PhysioNet-MI | 109 subjects | 4-class motor imagery | 6,300 | 1,734 | 1,803 | 9,837 |
| SHU-MI | 25 subjects | Binary motor imagery | 7,210 | 2,431 | 2,347 | 11,988 |
| Siena | 14 subjects | Binary seizure prediction | 33,479 | 14,149 | 3,679 | 51,307 |

Table A3: Train/validation/test partitions for each downstream benchmark.

| Dataset | Train/Val/Test split |
|---|---|
| FACED | 80/20/23 subjects: `sub000-079 / sub080-099 / sub100-122` |
| mental-arithmetic | 28/4/4 subjects: `1-28 / 29-32 / 33-36` |
| Mumtaz2016 | Subject-independent split following CBraMod: MDD/NC participants are 24/19 for training, 5/4 for validation, and 5/5 for testing |
| PhysioNet-MI | 70/19/20 subjects: `S001-070 / S071-089 / S090-109` |
| SHU-MI | 15/5/5 subjects: `1-15 / 16-20 / 21-25` |
| Siena | AdaBrain-Bench cross-subject test boundary: `PN16, PN17` are held out entirely for testing. The remaining 12 patients form the development pool; training/validation use a label-constrained, source-recording-grouped 80/20 split before windowing |

### A.3 Dataset Sources and Licenses

We use existing datasets, teacher EFMs, and baseline methods only under their original access terms and cite the corresponding papers in the main text. We do not redistribute restricted raw EEG data. Users of the released DLink code should obtain each dataset from its official source and follow the corresponding license, data-use agreement, and citation requirements.

Table A4: Dataset sources and license or usage terms.

| Dataset | Source | License / usage terms |
|---|---|---|
| FACED | FACED dataset, Synapse project `syn50614194`; `https://www.nature.com/articles/s41597-023-02650-w` | CC-BY metadata; the Scientific Data article is CC BY 4.0, with download terms governed by the Synapse page. |
| SHU-MI | figshare "`shu_dataset`", DOI 10.6084/m9.figshare.19228725.v1; `https://doi.org/10.6084/m9.figshare.19228725.v1` | CC BY 4.0. |
| PhysioNet-MI | PhysioNet EEG Motor Movement/Imagery Dataset; `https://www.physionet.org/content/eegmmidb/` | Open Data Commons Attribution License v1.0. |
| mental-arithmetic | EEGMAT, *EEG During Mental Arithmetic Tasks*, version 1.0.0, DOI 10.13026/C2JQ1P; original paper: Zyma et al., *Data* 2019, 4(1), 14 Zyma et al. (2019); `https://www.physionet.org/content/eegmat/1.0.0/` | Open Data Commons Attribution License v1.0. Users should download the dataset from PhysioNet and follow the citation and attribution requirements. |
| Siena | *Siena Scalp EEG Database*, version 1.0.0, DOI 10.13026/5d4a-j060; original paper: Detti et al., *Processes* 2020, 8(7), 846 Detti et al. (2020); `https://www.physionet.org/content/siena-scalp-eeg/1.0.0/` | PhysioNet Open Data Commons Attribution License v1.0. Users should obtain the data from PhysioNet and cite Detti (2020) and Detti et al. (2020). |
| Mumtaz2016 (MDD Patients and Healthy Controls EEG Data (New)) | figshare, DOI 10.6084/m9.figshare.4244171.v2; `https://figshare.com/articles/dataset/EEG_Data_New/4244171` | CC BY 4.0. |

### A.4 Additional Benchmark Results

The main text reports four benchmarks; the two remaining benchmarks are included here for completeness. The additional benchmarks and ConformerKD comparisons are arranged as paired tables to match the layout of the main results.

The additional-dataset tables follow the same three-metric layout as the main table; only the benchmark-specific metric names differ for binary tasks.

## B Architecture and Algorithm Details

This section provides the implementation-level details needed to reproduce the compact student, the router, and the training/inference procedures. The emphasis is on the components that determine the distillation pathway, not on introducing an additional architectural contribution.

Table A5: Results on the two additional benchmarks.

| Category | Model | Mumtaz2016 (2-class) | | | PhysioNet-MI (4-class) | | |
|---|---|---|---|---|---|---|---|
| | | ACC-B (%) | AUROC (%) | AUC-PR (%) | ACC-B (%) | F1-W (%) | Kappa (%) |
| Small Models | EEGNet | 89.58 ± 0.56 | 96.16 ± 1.28 | 96.78 ± 0.73 | 58.90 ± 0.62 | 59.68 ± 1.33 | 48.53 ± 0.76 |
| | EEGConformer | 89.23 ± 0.75 | 95.02 ± 0.60 | 95.26 ± 0.80 | 58.79 ± 0.57 | 58.78 ± 0.56 | 45.04 ± 0.77 |
| | EEGDeformer | 90.03 ± 2.47 | **97.60 ± 1.37** | **98.06 ± 1.56** | 60.56 ± 0.42 | 60.65 ± 0.52 | 47.68 ± 1.06 |
| | Stu(DLink) | 88.97 ± 1.67 | 95.47 ± 1.31 | 96.29 ± 1.00 | 58.60 ± 0.58 | 58.85 ± 0.74 | 44.93 ± 0.53 |
| KD | FitNets | 88.01 ± 0.42 | 96.31 ± 0.69 | 96.97 ± 0.56 | 59.34 ± 0.66 | 59.42 ± 0.94 | 45.79 ± 0.89 |
| | Logit-std | 84.91 ± 2.34 | 93.25 ± 1.61 | 94.49 ± 1.30 | 56.09 ± 2.32 | 56.66 ± 2.38 | 41.45 ± 3.10 |
| | S-A KD | 89.37 ± 1.74 | 96.48 ± 0.73 | 97.10 ± 0.61 | 57.84 ± 0.95 | 58.00 ± 0.99 | 43.79 ± 1.27 |
| | CRD | 88.60 ± 1.56 | 95.31 ± 0.36 | 96.16 ± 0.24 | 57.71 ± 1.46 | 57.80 ± 1.47 | 43.61 ± 1.95 |
| | ReviewKD | 89.85 ± 1.31 | 96.30 ± 1.17 | 96.93 ± 0.95 | 58.14 ± 0.55 | 58.14 ± 0.97 | 44.19 ± 0.73 |
| | RKD | 90.07 ± 0.99 | 95.85 ± 0.53 | 96.65 ± 0.38 | 57.47 ± 1.33 | 57.52 ± 1.25 | 43.29 ± 1.77 |
| | FreeKD | 88.95 ± 1.35 | 96.09 ± 1.09 | 96.62 ± 1.01 | 58.02 ± 0.52 | 57.99 ± 0.91 | 44.03 ± 0.70 |
| | **DLink** | **90.17 ± 1.53** | 95.69 ± 1.31 | 96.47 ± 1.22 | **60.83 ± 0.62** | **60.89 ± 0.91** | **48.94 ± 0.82** |

Table A6: ConformerKD results on FACED (9-class).

| Model | ACC-B (%) | F1-W (%) | Kappa (%) |
|---|---|---|---|
| EEGConformer | 43.24 ± 1.88 | 44.05 ± 1.88 | 37.72 ± 0.86 |
| **DLink (Conformer)** | **46.58 ± 0.53** | **46.30 ± 1.71** | **40.72 ± 1.80** |

Table A7: ConformerKD results on SHU-MI (2-class).

| Model | ACC-B (%) | AUROC (%) | AUC-PR (%) |
|---|---|---|---|
| EEGConformer | 60.14 ± 2.36 | 66.51 ± 2.22 | 64.63 ± 2.13 |
| **DLink (Conformer)** | **61.47 ± 1.18** | **68.68 ± 2.27** | **67.81 ± 2.62** |

### B.1 Compact Project-then-Compress Student

We instantiate two compact student variants, Stu-S and Stu-M, which share the same projection-then-compression architecture but differ in capacity. Each student first uses a depthwise-separable temporal convolutional stem to extract local EEG patterns. In parallel, a lightweight Transformer branch models dependencies among the ordered latent tokens. A learnable scalar gate adaptively fuses the convolutional and Transformer representations, producing the student projection features used for distillation. The fused features are then passed through a compact compression block before classification. Detailed implementation stages are summarized in Table B1.

Table B1: Project-then-compress student architectures used for DLink. Stu-S and Stu-M share the projection and fusion design but use different Transformer depth, channel width, and latent-resolution compression.

| Stage | Operation | Stu-S | Stu-M | Notes |
|---|---|---|---|---|
| Projection | Depthwise-separable temporal CNN stem | kernel 35 | kernel 35 | Local temporal feature extraction |
| Projection | Transformer encoder over latent units | 1 block | 3 blocks | Long-range context modeling |
| Projection | Learnable CNN/Transformer fusion | scalar gate $\alpha$ | scalar gate $\alpha$ | Produces feature $f_S^p$ for distillation |
| Compression | Temporal compressor | 15 channels; segment reduction | 30 channels; segment resolution preserved | Compact transfer |
| Prediction | MLP classifier | hidden size 100 | hidden size 100 | Operates on compressed representation |

Stu-S is the default student used in the multi-dataset experiments. For the FACED scale comparison, we additionally instantiate Stu-M to examine whether a modest increase in student capacity can improve performance under the same routed distillation design. The teacher, router, loss functions, and training protocol remain unchanged. Stu-M uses three Transformer blocks instead of one, a 30-channel temporal compressor instead of 15 channels, and preserves the latent segment resolution, whereas Stu-S uses reduced segment resolution. Thus, the comparison isolates the effect of the student configuration while keeping the DLink distillation mechanism fixed. DLink-M is evaluated only on FACED and is not intended as a replacement for the default Stu-S configuration across all benchmarks. In both variants, the final classifier operates only on the compressed student representation, so the teacher and layer router are discarded at inference.

We ablate the CNN and Transformer branches to check whether both projection components contribute to compact transfer. Table B2 reports the FACED results for Stu-S.

The full projection block performs best for Stu-S, indicating that local temporal extraction and latent-unit context modeling provide complementary support for distillation.

Table B2: Ablation of the student projection components on FACED.

| Variant | Stu-S | | |
|---|---|---|---|
| | ACC-B (%) | F1-W (%) | Kappa (%) |
| w/o CNN | 39.33 ± 0.49 | 38.97 ± 0.42 | 31.34 ± 0.53 |
| w/o Trans | 42.24 ± 2.26 | <u>41.66 ± 2.32</u> | 34.59 ± 2.50 |
| **DLink (Ours)** | **43.46 ± 0.80** | **43.06 ± 0.90** | **37.89 ± 0.95** |

### B.2 Layer Router Architecture

The router predicts input-conditioned soft weights over teacher layers. Its input is constructed from compact student-feature cues and a representation-spectrum cue, rather than from the original high-dimensional feature map. In the profiled setting, the router input shape is $(64, 38, 5)$.

Table B3: Layer router architecture. The router has $0.04$M parameters and $0.22$M FLOPs in our profiling.

| Component | Operation | Configuration | Notes |
|---|---|---|---|
| Input normalization | LayerNorm | compact feature/spectrum cue | Applied before routing blocks |
| Channel fuser | Conv1D + BatchNorm + GELU | router hidden dim 64 | Fuses channel-wise cues |
| Sequence processor | Transformer encoder + LayerNorm | 2 attention heads | Models dependencies in routing input |
| Pooling | Adaptive average pooling | global pooling | Produces global routing summary |
| Routing head | Linear projection | output dimension $L$ | Layer-weight logits |
| Layer weights | Temperature softmax | temperature $\tau$ | $\mathrm{softmax}(a/\tau)$ |

The router is lightweight because it processes compact cues rather than the teacher's full feature map. It is therefore suitable as a training-time module that can be discarded after the student has internalized the routed supervision.

### B.3 DLink Training Algorithm

Algorithm B1 summarizes one DLink training step. The teacher provides frozen layer-wise representations, while the student and router are optimized jointly.

**Algorithm B1** DLink training procedure

**Require:** Frozen teacher EFM $\mathcal{T}$, compact student $\mathcal{S}$, layer router $R$, training set $\mathcal{D}$
1: **for** each mini-batch $(X, y)$ **do**
2: Extract the routed teacher hidden-state list $\{f_T^{(l)}\}_{l=1}^{L}$ using frozen $\mathcal{T}$
3: Compute student feature $f_S^p$, compressed feature $f_S^c$, and prediction $\hat{y}_S$
4: Construct router input $z_{\text{in}}$ from pooled student features and representation-spectrum cues
5: Predict layer weights $w = \mathrm{softmax}(R(z_{\text{in}})/\tau)$
6: Aggregate routed teacher knowledge $f_T^* = \sum_{l=1}^{L} w_l f_T^{(l)}$
7: Compute classification loss $\mathcal{L}_{\text{cls}}$
8: Compute spectral distillation loss $\mathcal{L}_{\text{kd}}$
9: Compute spectral-concentration routing prior loss $\mathcal{L}_{\text{prior}}$
10: Update $\mathcal{S}$ and $R$ with $\mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{kd}} + \mathcal{L}_{\text{prior}}$
11: **end for**
12: Discard $\mathcal{T}$ and $R$ after training
13: **return** Compact student $\mathcal{S}$

After training, both the teacher and the router are removed. The returned model is therefore the compact student alone.

### B.4 DLink Inference Algorithm

Algorithm B2 shows the resulting inference path. No teacher forward pass, routing computation, or distillation loss is used during deployment.

**Algorithm B2** DLink inference procedure

**Require:** Trained compact student $\mathcal{S}$, EEG sample $X$
1: Compute prediction $\hat{y} = \mathcal{S}(X)$
2: **return** $\hat{y}$

Thus, the inference cost is identical to a standalone compact student with the same architecture.

## C Implementation Details

All models are implemented in PyTorch and optimized with AdamW. We use a two-stage head and hyperparameter protocol for every dataset, teacher, student, and KD method. Stage 1 uses only the train/validation split to compare candidate heads, learning rates, and method-specific KD weights, and selects the winner by validation ACC-B. Stage 2 fixes the selected configuration, trains five independent seeds (0–4), selects the checkpoint with the best validation ACC-B for each seed, and evaluates that checkpoint once on the held-out test split.

Each run uses one teacher EFM at a time. The fine-tuned teacher checkpoint is frozen during student training and supplies representations only; the teacher-side classifier is not used. The main multi-dataset comparison uses Stu-S, while the FACED scale comparison additionally reports DLink-M. The teacher and router are discarded after training.

Table C1: Shared head-search and optimization protocol. All methods receive the same validation-search budget. AdamW uses $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

| Parameter | Stage-1 candidates | Stage-2 and shared setting |
|---|---|---|
| Classification head | `full_mlp`; `avgpool_linear` | Freeze the validation winner for five seeds |
| Learning rate | 0.005, 0.002, 0.0008 | Freeze the validation winner |
| KD weight $\lambda$ | 0.3, 0.5, 0.7, 1.0 | Freeze the validation winner; not used by non-KD baselines |
| Checkpoint | Validation-ACC-B sweep | Best validation ACC-B per seed |
| Optimizer | AdamW, $(\beta_1, \beta_2) = (0.9, 0.999)$ | Weight decay 0.001; gradient norm $\leq$ 1.0 |
| Schedule | Cosine annealing | Minimum learning rate $10^{-6}$; up to 100 epochs |
| Stopping | Validation-ACC-B monitoring | Patience 10 epochs; CUDA mixed precision enabled |
| Seeds | – | $0, 1, 2, 3, 4$ |

The selected dataset-specific configurations used in the main evaluation are summarized in Table C2.

Table C2: Selected dataset-specific hyperparameters for the main evaluation.

| Dataset | Learning rate | KD weight $\lambda$ | Classification head |
|---|---|---|---|
| FACED | 0.002 | 1.0 | `full_mlp` |
| mental-arithmetic | 0.005 | 0.7 | `full_mlp` |
| Mumtaz2016 | 0.005 | 0.5 | `avgpool_linear` |
| PhysioNet-MI | 0.005 | 0.7 | `full_mlp` |
| SHU-MI | 0.005 | 0.7 | `full_mlp` |
| Siena | 0.0008 | 0.7 | `avgpool_linear` |

The two candidate heads are defined as follows. `full_mlp` flattens all compressed spatio-temporal positions, applies a 100-unit hidden layer, and uses dropout 0.3. `avgpool_linear` applies adaptive global average pooling over the compressed feature and then a linear classifier, with no hidden layer and no dropout. Thus, the classification head is a dataset-, teacher-, and method-dependent training hyperparameter rather than a fixed component shared across all runs. The router temperature $\tau$ follows the standard temperature-scaled softmax formulation used in knowledge distillation Hinton et al. (2015); we fix $\tau = 2.0$ and $\tau_q = 2.0$ for all datasets. These temperatures are used only during training because the router is discarded after distillation.

### C.1 Compute Resources

All experiments were conducted on a server equipped with four NVIDIA A100-SXM4-80GB GPUs, providing 320 GB of total GPU memory. The server uses dual Intel Xeon Platinum 8452Y CPUs with 72 physical cores and 144 logical threads, and has approximately 512 GB of system memory. The software environment uses Python 3.9.23 and PyTorch 2.5.1+cu121; PyTorch was built against CUDA 12.1 with cuDNN 9.1.0.

The model-scale compute cost is reported through the parameter and FLOP comparisons in Fig. 5. The training-time router cost and CPU-only inference latency are separately profiled in Appendix G. These measurements are intended to support the reported efficiency claim and do not include EEG acquisition, online preprocessing, communication, or operating-system scheduling overhead.

## D Teacher Robustness and Hyperparameter Analyses

This section reports experiments that vary the teacher or training hyperparameters while keeping the compact student design unchanged. These analyses clarify how teacher selection, teacher-side tuning, and validation-based hyperparameter choices affect DLink training rather than the inference architecture.

### D.1 Robustness to Different Teacher EFMs

We first test whether DLink depends on a specific EFM teacher. For each run, the teacher is fixed to one EFM, and the routed hidden-state list is constructed within that teacher. Table D1 reports the three-teacher comparison for the retained Stu-S student across all six datasets under the same five-seed protocol. We fix the student to Stu-S in the teacher comparison, while Stu-M is evaluated separately in the FACED scale comparison.

These results use the same validation-ACC-B selection rule and five-seed evaluation protocol as the main comparison. The FACED DLink-M scale comparison is reported in the main text because the corresponding M-scale runs are currently available only for that benchmark.

### D.2 Hyperparameter Sensitivity

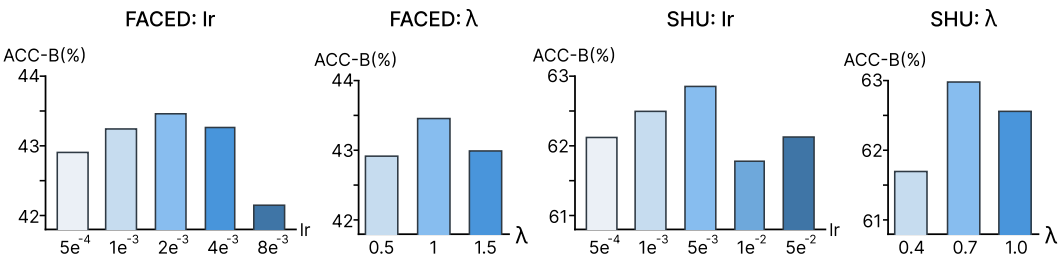


Figure D1: Post-selection, one-at-a-time sensitivity analysis around the validation-selected learning rate and spectral distillation weight $\lambda$ on FACED and SHU-MI.

The Stage-1 candidate grid in Table C1 is used only to select the final dataset-specific configurations. After selection, we perform a separate local sensitivity analysis by varying either the learning rate or

Table D1: Teacher-robustness comparison for Stu-S across six benchmarks. All teacher checkpoints are frozen and provide representations only.

| Dataset | Teacher | ACC-B (%) | F1-W/AUROC (%) | Kappa/AUC-PR (%) |
|---|---|---|---|---|
| FACED | CBraMod | 43.46 ± 0.80 | **43.06 ± 0.90** | **37.89 ± 0.95** |
| | CSBrain | **43.48 ± 0.66** | 42.91 ± 0.87 | 35.77 ± 1.08 |
| | LaBraM | 42.28 ± 1.14 | 41.66 ± 1.31 | 34.62 ± 1.23 |
| Mumtaz2016 | CBraMod | **90.17 ± 1.53** | 95.69 ± 1.31 | 96.47 ± 1.22 |
| | CSBrain | 90.14 ± 0.35 | 96.17 ± 0.51 | **97.24 ± 0.50** |
| | LaBraM | 89.35 ± 0.80 | **96.27 ± 1.14** | 96.93 ± 0.80 |
| PhysioNet-MI | CBraMod | **60.83 ± 0.62** | 60.89 ± 0.91 | **48.94 ± 0.82** |
| | CSBrain | 60.81 ± 1.00 | **61.01 ± 0.88** | 47.15 ± 0.81 |
| | LaBraM | 59.44 ± 1.00 | 59.70 ± 0.89 | 45.92 ± 1.34 |
| SHU-MI | CBraMod | 62.89 ± 1.17 | 68.43 ± 1.72 | 68.81 ± 1.47 |
| | CSBrain | 62.77 ± 1.08 | 68.58 ± 1.43 | 68.19 ± 1.76 |
| | LaBraM | **63.09 ± 0.83** | **69.10 ± 1.71** | **69.19 ± 1.64** |
| mental-arithmetic | CBraMod | 62.21 ± 5.14 | 71.49 ± 4.60 | 44.59 ± 2.91 |
| | CSBrain | 62.28 ± 6.00 | **72.68 ± 5.06** | **52.05 ± 7.50** |
| | LaBraM | **63.82 ± 6.31** | 72.42 ± 5.02 | 49.57 ± 5.69 |
| Siena | CBraMod | **78.79 ± 2.63** | **91.92 ± 2.15** | **48.73 ± 2.18** |
| | CSBrain | 74.55 ± 2.54 | 90.93 ± 4.13 | 44.18 ± 3.71 |
| | LaBraM | 71.26 ± 3.59 | 88.78 ± 4.39 | 35.13 ± 6.26 |

$\lambda$ around its validation-selected value while holding all other hyperparameters fixed. These auxiliary perturbation points are not part of the selection grid and do not affect the configurations used for the main evaluation.

As shown in Fig. D1, the selected settings—$0.002$ and $1.0$ for FACED, and $0.005$ and $0.7$ for SHU-MI—lie near the local optima of the corresponding one-at-a-time sweeps. Performance changes smoothly under nearby perturbations, indicating that the reported configurations are not isolated optima.

# E Additional Routing Dynamics

This section reports the layer-wise routing statistics behind Fig. 3. SHU-MI provides the clearest per-sample evidence of a non-uniform layer preference. At the saved best checkpoint, $L_7$ is top-1 for 7,207 of 7,210 training samples (99.96%). Its margin over the next-highest layer exceeds 0.001 for 98.93% of samples and 0.002 for 56.75%. The maximum layer-weight spread exceeds 0.01 for every sample and 0.02 for 99.97% of samples.

This preference nevertheless remains a soft aggregation. On SHU-MI, the median normalized entropy is 0.99472 and the median effective layer count is 13.806. FACED is closer to uniform, with mean normalized entropy 0.9991, an effective layer count of 13.968, and a mean $L_9$ margin of $0.00163 \pm 0.00035$. Thus, the learned distributions are neither near-one-hot hard routing nor exact uniform averaging. Their degree of layer preference is dataset-dependent, with SHU-MI showing clearer separation than FACED.

Here, "dominant" denotes relatively higher adaptive importance during student-oriented transfer, not a uniquely optimal or independently most discriminative teacher layer. The router is optimized through the distillation and routing-prior objectives rather than direct task-label supervision. Its student-derived inputs are jointly shaped by task supervision. The learned weights should therefore be interpreted as student-conditioned transfer preferences rather than direct estimates of layer-wise task discriminability.

For FACED, the top-1 preference shifts from $L_0$ early in training to $L_9$ at the best epoch and remains stable thereafter. Tables E1 and E2 report the corresponding online and post-hoc statistics. Their

batch-frequency and mean-weight values can differ slightly because they use different evaluation passes.

Table E1: Online epoch-level routing dynamics and intra-batch routing statistics on FACED with DLink-S and a fine-tuned CBraMod teacher. "#Top-1" denotes the average number of distinct top-1 layers per batch, and "Multi" denotes the fraction of batches with more than one top-1 layer.

| Stage | Average router-weight top-3 | Lowest average-MSE top-3 | Top-1 router distribution | #Top-1 | Multi |
|---|---|---|---|---|---|
| Early (epoch 1) | $L_0 = 0.077112,\ L_9 = 0.072327,\ L_8 = 0.071976$ | $L_0,\ L_9,\ L_8$ | $L_0$ 83.51%, $L_9$ 3.97%, $L_3$ 1.88% | 5.27 | 95.24% |
| Best epoch (24) | $L_9 = 0.079403,\ L_8 = 0.077879,\ L_7 = 0.075705$ | $L_9,\ L_8,\ L_7$ | $L_9$ 97.83%, $L_0$ 0.76%, $L_{12}$ 0.52% | 2.12 | 74.29% |
| Stop epoch (34) | $L_9 = 0.079600,\ L_8 = 0.078038,\ L_7 = 0.075893$ | $L_9,\ L_8,\ L_7$ | $L_9$ 97.84%, $L_0$ 0.65%, $L_{12}$ 0.52% | 2.17 | 81.90% |

Table E2: Post-hoc dataset-level routing statistics at the saved best checkpoint on FACED with DLink-S. Statistics are recomputed over all 6,720 training samples. $H(w)$ denotes routing entropy and $\bar{w}$ denotes the mean routing distribution.

| Statistic at best checkpoint | Value |
|---|---|
| Mean normalized entropy $H(w)/\log 14$ | 0.9991 |
| Effective number of layers $\exp(H(w))$ | 13.968 |
| $w_9$, mean $\pm$ std | $0.07967 \pm 0.00176$ |
| $w_9 - w_8$, mean $\pm$ std | $0.00163 \pm 0.00035$ |
| Mean $\mathrm{KL}(w_i \,\|\, \bar{w})$ | $1.09 \times 10^{-4}$ |
| Samples with top-1 $\neq L_9$ | 2.26% |
| Batches with multiple top-1 layers | 76.19% |

The SHU-MI epoch dynamics provide a second case. The best standalone linear-probe layer varies across the five seeds and therefore does not provide a stable reference. The router nevertheless develops a stable $L_7$ preference, supporting dataset-dependent transfer preferences rather than a universal rule that routing must recover the probe-optimal layer.

Table E3: Online epoch-level routing dynamics and intra-batch routing statistics on SHU-MI with DLink-S and a fine-tuned CBraMod teacher. "#Top-1" denotes the average number of distinct top-1 layers per batch, and "Multi" denotes the fraction of batches with more than one top-1 layer.

| Stage | Average router-weight top-3 | Lowest average-MSE top-3 | Top-1 router distribution | #Top-1 | Multi |
|---|---|---|---|---|---|
| Early (epoch 1) | $L_5 = 0.077966,\ L_4 = 0.077687,\ L_6 = 0.076266$ | $L_5,\ L_4,\ L_6$ | $L_5$ 29.22%, $L_4$ 23.38%, $L_3$ 13.59% | 5.50 | 100.00% |
| Best epoch (3) | $L_7 = 0.084873,\ L_6 = 0.083917,\ L_5 = 0.083060$ | $L_7,\ L_6,\ L_5$ | $L_7$ 69.13%, $L_6$ 17.03%, $L_5$ 13.51% | 2.25 | 76.11% |
| Stop epoch (13) | $L_7 = 0.087429,\ L_6 = 0.086328,\ L_5 = 0.084876$ | $L_7,\ L_6,\ L_5$ | $L_7$ 97.14%, $L_6$ 2.47%, $L_5$ 0.29% | 1.23 | 18.58% |

Table E4: Post-hoc dataset-level routing statistics at the saved best checkpoint on SHU-MI with DLink-S. Statistics are recomputed over all 7,210 training samples. $H(w)$ denotes routing entropy and $\bar{w}$ denotes the mean routing distribution.

| Statistic at best checkpoint | Value |
|---|---|
| Mean normalized entropy $H(w)/\log 14$ | 0.9950 |
| Effective number of layers $\exp(H(w))$ | 13.816 |
| $w_7$, mean $\pm$ std | $0.08543 \pm 0.00134$ |
| $w_7 - w_6$, mean $\pm$ std | $0.00210 \pm 0.00051$ |
| Mean $\mathrm{KL}(w_i \,\|\, \bar{w})$ | $1.00 \times 10^{-4}$ |
| Samples with top-1 $\neq L_7$ | 0.04% |
| Batches with multiple top-1 layers | 2.65% |

A representative SHU-MI sample assigns weights $w_7 = 0.08722$, $w_6 = 0.08353$, $w_8 = 0.08285$, $w_9 = 0.08104$, $w_{13} = 0.05646$, and $w_0 = 0.04673$. Its $L_7$ margin is 0.003689 and its maximum spread is 0.040486, while its normalized entropy remains 0.994613. This example illustrates a measurable layer preference within a broad soft distribution.

# F Spectral Distillation and Representation-Spectrum Consistency

This section clarifies the representation-level spectral objective used by DLink. It separates the mathematical formulation from the interpretation boundary, because the objective operates on learned temporal-latent representations rather than on raw EEG frequency bands.

## F.1 Spectral Distillation Formulation

Spectral distillation is applied to learned feature maps exposed to DLink. We use $f \in \mathbb{R}^{N\times d}$ as a notational view: $N$ denotes the ordered temporal-latent axis used for spectral alignment, and $d$ denotes all remaining feature indices. The Fourier transform is a one-dimensional transform $\mathcal{F}_N$ along the $N$ axis, computed independently over the remaining feature indices. It is not applied to raw EEG signals, but to learned representations before compression. Given student projection feature $f_S^p$ and routed teacher representation $f_T^*$, we compute

$$\mathcal{F}_N(f_S^p) = M_S \odot e^{j\Phi_S}, \qquad \mathcal{F}_N(f_T^*) = M_T \odot e^{j\Phi_T}. \tag{21}$$

The magnitude consistency term is

$$\mathcal{L}_{\text{mag}} = \|M_S - M_T\|_F^2. \tag{22}$$

To avoid phase discontinuity from $2\pi$ periodicity, phase is represented by sine and cosine components:

$$\text{Enc}(\Phi) = [\cos\Phi, \sin\Phi], \qquad \mathcal{L}_{\text{phase}} = \|\text{Enc}(\Phi_S) - \text{Enc}(\Phi_T)\|_F^2. \tag{23}$$

The spectral distillation loss is $\mathcal{L}_{\text{kd}} = \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{phase}}$.

## F.2 Why Representation-Spectrum Consistency Helps Compact Transfer

During compact transfer, projection and downsampling can alter learned feature spectra. Representation-spectrum consistency provides an auxiliary constraint that encourages the compact student feature to follow the spectral structure of the routed teacher representation before compression. This complements pointwise feature matching, which constrains Euclidean discrepancy but does not explicitly regularize the temporal-latent spectrum entering the compression block.

## F.3 Representation-Spectrum View

The spectral bins used here describe variation across adjacent temporal latent units in a learned representation, not physiological EEG bands in Hz. This is meaningful because the temporal-latent axis preserves the order used by the feature layout entering compression, and the compression block reduces this latent resolution before classification.

Our motivation is representation-level rather than a strict signal-reconstruction claim. Anti-aliased CNNs show that ordered learned feature maps can benefit from frequency-aware treatment before resolution reduction Zhang (2019); Ribeiro & Schön (2021). Analogously, DLink constrains the pre-compression temporal-latent spectrum to follow the routed teacher spectrum. A smaller high-frequency tail after compression indicates smoother compact features and reduced representation-spectrum distortion.

## F.4 What We Do Not Claim

We do not claim that spectral distillation recovers physiological rhythms, identifies causal EEG frequency bands, or preserves task-relevant oscillatory activity in a neurophysiological sense. Our claim is limited to regularization for compact transfer: spectral distillation improves compact transfer and reduces distortion observed in learned representation spectra.

## F.5 Layer Score Definitions and Routed L2 Control

The routing ablation compares alternative layer scores computed from a teacher feature $f_T^{(l)}$. Mean Power and Max Amp replace the spectral-concentration score in the routing prior distribution $q_l$; the

router architecture, spectral distillation loss, and training protocol remain unchanged. Routed L2 is a separate objective-level control: it retains the same learned router, routed teacher representation $f_T^*$, student representation $f_S^p$, and training protocol as DLink, but replaces $\mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{phase}}$ with pointwise squared-L2 feature matching $\|f_S^p - f_T^*\|_F^2$. $A_l(\omega) = \|\mathcal{F}_N(f_T^{(l)})(\omega)\|_2$ denote the feature-index-aggregated spectral amplitude at temporal-latent representation-frequency bin $\omega \in \Omega$, and let $E_l(\omega) = A_l(\omega)^2$.

$$\text{MeanPower}_l = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} E_l(\omega), \qquad \text{MaxAmp}_l = \max_{\omega \in \Omega} A_l(\omega). \tag{24}$$

The spectral-concentration score first forms a normalized energy distribution over representation-frequency bins:

$$p_l(\omega) = \frac{E_l(\omega)}{\sum_{\omega' \in \Omega} E_l(\omega') + \epsilon}. \tag{25}$$

We then reduce the distribution to a scalar layer score using spectral concentration:

$$\text{SpecConc}_l = \sum_{\omega \in \Omega} p_l(\omega)^2. \tag{26}$$

In our experiments, this spectral-concentration score provides a more stable cue for routing than mean power or maximum amplitude.

The input ablations test the two cues used by the router. Freq-only router retains only the pooled representation-spectrum cue $\text{Pool}(|\mathcal{F}_N(f_S^p)|)$, whereas Time-only router retains only the pooled student-feature cue $\text{Pool}(f_S^p)$. These variants use the same router, layer aggregation, distillation losses, and optimization protocol as DLink.

### F.6 Focused Distillation Objective Comparison

We compare DLink with direct feature-level and logit-level distillation under the same student backbones. This focused comparison isolates the effect of routed representation-spectrum transfer from the use of a compact student alone.

Table F1: Comparison of distillation objectives on FACED using the same student backbones.

| Objective | Stu-S | | |
|---|---|---|---|
| | ACC-B (%) | F1-W (%) | Kappa (%) |
| Feature KD | 42.39 ± 0.68 | <u>42.17 ± 0.67</u> | <u>34.86 ± 0.83</u> |
| Logit KD | 41.25 ± 1.25 | 41.17 ± 1.27 | 33.67 ± 1.41 |
| **DLink (Ours)** | **43.46 ± 0.80** | **43.06 ± 0.90** | **37.89 ± 0.95** |

DLink gives the best FACED performance for Stu-S, supporting the combination of layer routing and representation-spectrum consistency over direct feature or logit matching.

To further isolate the two components of the representation-spectrum objective, we run three controls on FACED using the same Stu-S backbone, routed teacher representation $f_T^*$, student projection feature $f_S^p$, and training protocol. Magnitude-only retains $\mathcal{L}_{\text{mag}}$ and removes phase matching; phase-only retains $\mathcal{L}_{\text{phase}}$ and removes magnitude matching. Shuffled ordering disrupts the ordered temporal-latent structure before spectral comparison. These controls probe whether the gain depends on the joint magnitude–phase objective and on the ordering used by the learned temporal-latent spectrum.

We also compare two routing controls. Global Learnable replaces the input-conditioned router with one trainable layer-weight vector shared by all samples, while keeping the remaining losses and training protocol unchanged. Uniform Prior retains the input-conditioned router but replaces the spectral-concentration guidance distribution with a uniform target. This provides a strict comparison between the learned spectral prior and a uniform prior while preserving the router, student, teacher, and distillation objective.

Table F2: Additional routing and spectral-objective controls on FACED using the Stu-S student. All controls use the same teacher, student, and training protocol as DLink.

| Variant | ACC-B (%) | F1-W (%) | Kappa (%) |
|---|---|---|---|
| Magnitude-only | $42.34 \pm 0.80$ | $42.24 \pm 0.60$ | $34.88 \pm 0.86$ |
| Phase-only | $42.39 \pm 0.57$ | $42.15 \pm 0.44$ | $34.82 \pm 0.58$ |
| Shuffled ordering | $\underline{42.86 \pm 0.85}$ | $\underline{42.85 \pm 0.64}$ | $35.43 \pm 0.90$ |
| Global Learnable | $41.88 \pm 1.13$ | $41.95 \pm 0.88$ | $34.35 \pm 1.20$ |
| Uniform Prior | $42.83 \pm 0.81$ | $42.36 \pm 0.46$ | $\underline{35.83 \pm 0.74}$ |
| **DLink (full)** | $\mathbf{43.46 \pm 0.80}$ | $\mathbf{43.06 \pm 0.90}$ | $\mathbf{37.89 \pm 0.95}$ |

DLink outperforms all five controls. The magnitude-only and phase-only results indicate that both components contribute to the full objective, while the lower shuffled-ordering result is consistent with the usefulness of ordered temporal-latent structure during compact transfer. Global Learnable is also below DLink, providing evidence that the gain is not explained by a single shared layer mixture. Uniform Prior remains below DLink, supporting the contribution of the spectral-concentration guidance beyond a uniform target. These results support the representation-spectrum objective and input-conditioned aggregation as transfer components, but do not imply recovery of physiological rhythms or causal EEG frequency bands.

The Global Learnable model produces the following shared layer-weight distribution:

Table F3: Shared layer weights learned by the Global Learnable routing control on FACED.

| Layer | Weight | Layer | Weight |
|---|---|---|---|
| $L_0$ | 0.068105 | $L_7$ | 0.075370 |
| $L_1$ | 0.068296 | $L_8$ | 0.077311 |
| $L_2$ | 0.071174 | $L_9$ | 0.078703 |
| $L_3$ | 0.073200 | $L_{10}$ | 0.074141 |
| $L_4$ | 0.071055 | $L_{11}$ | 0.068181 |
| $L_5$ | 0.072313 | $L_{12}$ | 0.063820 |
| $L_6$ | 0.074511 | $L_{13}$ | 0.063819 |

Table F4: Five-seed mean layer weights produced by the input-conditioned router under Uniform Prior on FACED.

| Layer | Weight | Layer | Weight |
|---|---|---|---|
| $L_0$ | 0.071271 | $L_7$ | 0.076130 |
| $L_1$ | 0.069251 | $L_8$ | 0.079338 |
| $L_2$ | 0.071165 | $L_9$ | 0.082276 |
| $L_3$ | 0.073030 | $L_{10}$ | 0.076572 |
| $L_4$ | 0.070852 | $L_{11}$ | 0.067991 |
| $L_5$ | 0.072304 | $L_{12}$ | 0.057416 |
| $L_6$ | 0.074990 | $L_{13}$ | 0.057415 |

The shared weights have normalized entropy $0.99928 \pm 0.00004$ and an effective layer count of $13.973 \pm 0.001$, indicating a mild global preference rather than a concentrated single-layer solution. Under Uniform Prior, the mean routing distribution has normalized entropy $0.99821 \pm 0.00006$ and an effective layer count of $13.934 \pm 0.002$. The router still produces $1.96 \pm 0.17$ distinct top-1 layers per batch, with $64.19 \pm 10.75\%$ of batches containing multiple top-1 layers, and selects $L_9$ as the top-1 layer in 98.13% of cases on average. Thus, Uniform Prior does not equal Fixed Avg: the uniform target constrains only the prior term, while the input-conditioned router remains free to learn non-uniform preferences from the distillation objective and task-shaped student features.

# G Deployment-oriented Profiling Protocol

## G.1 CPU-only Profiling Setup

We evaluate pure forward latency under a conservative CPU-only setting: one CPU thread, batch size 1, no CUDA acceleration, 10 warm-up runs, and 100 timed runs. The reported values measure model forward computation only.

## G.2 Teacher, Student, and Router Timing

Table G1 reports the timing measurements used to support the deployment-oriented efficiency claim. Teacher and student latency are measured under the same CPU-only batch-1 setting, whereas the router is profiled separately as a training-time CUDA module.

Table G1: Forward-latency profiling for the fine-tuned CBraMod teacher, compact student, and training-time router. Teacher and student timings use CPU-only batch-1 inference; the router timing reports its separate training-time CUDA forward cost.

| Model/Module | Device | Batch | Latency |
|---|---|---|---|
| CBraMod teacher checkpoint | CPU | 1 | 19.345 ms/sample |
| Compact student | CPU | 1 | 3.653 ms/sample |
| Layer router | CUDA | 64 | 0.633 ms/batch |

The CPU timings use 10 warm-up and 100 timed runs; the router uses 10 warm-up and 1,000 timed runs. Because the router is discarded after training, deployment uses only the compact student timing in this table.

## G.3 Boundary of the Profiling Claim

This profiling evaluates pure forward latency under controlled CPU-only or module-level CUDA settings. It is not validation on commercial BCI hardware, and it does not include EEG acquisition, preprocessing, communication, operating-system scheduling, power measurement, or end-to-end online streaming overhead.

# H Additional Discussion and Limitations

**Scope of deployment claims.** DLink is evaluated through deployment-oriented profiling rather than real-device deployment. The profiling results support the efficiency of the compact inference path, but they should not be interpreted as validation on commercial BCI hardware.

**Scope of spectral claims.** The spectral analysis supports regularization for compact transfer and reduced distortion in learned representation spectra. It does not establish a causal physiological interpretation of EEG frequency bands.

**Future work.** Future work includes validation on real BCI hardware, online EEG streaming experiments, broader teacher EFMs and downstream tasks, larger subject populations, and energy or memory profiling.

# I Broader Impacts

DLink may help reduce the inference cost of EEG foundation-model adaptation by transferring teacher knowledge into compact students. This can support more efficient research prototypes for EEG decoding and may lower the computational barrier for studying compact models in BCI, affective computing, and neurological assessment settings.

The main risk is reliability under realistic acquisition conditions. The experiments use existing benchmark datasets rather than online clinical or commercial-device recordings, and real-world EEG data may contain stronger sensor noise, subject variability, artifacts, and distribution shifts

than those represented in the benchmarks. DLink should therefore not be used as a clinical decision tool or deployed in safety-critical BCI settings without additional validation on the target hardware, acquisition protocol, and user population.