

A DETAILS OF EXPERIMENTAL SETTINGS

A.1 DATASET DESCRIPTION

We describe the dataset we use for evaluation.

- **TUEV** (Obeid & Picone, 2016): A subset of the TUH EEG Corpus (Obeid & Picone, 2016), which is an abnormal-related event type dataset. This dataset was recorded using 23 electrodes at a sampling rate of 256 Hz. All data are categorized into one of six annotations: (1) spike and sharp wave (SPSW), (2) generalized periodic epileptiform discharges (GPED), (3) periodic lateralized epileptiform discharges (PLED), (4) eye movement (EYEM), (5) artifact (ARTF), and (6) background (BCKG).
- **DREAMER** (Katsigiannis & Ramzan, 2017): A multi-modal emotional dataset containing EEG (14 channels, 128 Hz) and ECG recordings evoked by audio-visual stimuli of 25 subjects. The recordings are labeled using the valence/arousal/dominance scale, as assessed through Self-Assessment Manikins (SAM).
- **Left/Right Hand Motor Imagery (LeftRight Hand)** (Zakrzewski et al., 2022): A dataset where EEG signals (64 channels, 512 Hz) were recorded by the Biosemi ActiveTwo system, including 52 participants. The EEG signals are classified into two motor imagery classes: left hand and right hand.
- **Crowdsourced** (Williams et al., 2023): The dataset includes EEG recordings (2048 Hz) from 60 participants engaged in a resting state task with eyes open and eyes closed. Only 13 participants who used the 14-channel EPOC+, EPOC X, and EPOC devices were included in the experiments.

A.2 DATASET PREPROCESSING

We follow the preprocessing steps of LaBraM (Jiang et al., 2024) as closely as possible to prevent distribution shifts caused by inconsistencies in preprocessing. First, to filter out noise, we apply a bandpass filter with a bandwidth between 0.1 Hz and 75 Hz. Then, we use a 50 Hz notch filter to remove power-line interference. Finally, all signals are resampled to 200 Hz. To ensure that neural networks process the signals stably, we scale the signal values to mainly lie between -1 and 1 by dividing them by 100. Each signal is segmented based on the dataset’s characteristics and references, with the following details.

- **TUEV** is segmented into 5-second, non-overlapping samples (Jiang et al., 2024).
- **DREAMER** includes 3 seconds of baseline data at the beginning of the signals. We only use the remaining 60 seconds and divide it into twenty 3-second samples. Valence is used as the label (Cui et al., 2020).
- **LeftRight Hand** is sampled for 2 seconds from each trial, specifically the part related to motor imagery (Zakrzewski et al., 2022).
- **Crowdsourced** is segmented into 4-second samples following Williams et al. (2023).

A.3 EVALUATION METRICS

We adapt the three metrics for evaluation.

- **Balanced accuracy (BACC)** is the average of recall across each class. BACC is particularly effective when class ratios are imbalanced. We use BACC for both binary and multi-class classification tasks.
- **AUROC** is the area under the receiver operating characteristic (ROC) curve. AUROC is used to determine how far the model’s predictions are from random predictions. Binary classification tasks are evaluated using AUROC.
- **Cohen’s Kappa** represents the measure of interrater agreement for qualitative items, commonly used to assess how much better the agreement is compared to random chance. We adopt Cohen’s Kappa for multi-class classification tasks.

A.4 BASELINE MODEL DESCRIPTION

We employ not only additive fine-tuning methods but also supervised and self-supervised modeling methods as baselines to confirm that the proposed method preserves the original superiority of EEG foundation models compared to other baselines.

Supervised modeling methods. The baselines of supervised modeling methods are from Yang et al. (2024) and Jiang et al. (2024). We use the open source code of Yang et al. (2024) and evaluate baseline models on datasets as mentioned above in a supervised learning manner. **SPaRCNet** (Jing et al., 2023) is a Dense-Net (Huang et al., 2017) type neural network, containing 1D-CNN based Dense blocks. **ContraWR** (Yang et al., 2023) is structured to apply a short-time Fourier transform (STFT) to the input signal and then pass the transformed 2D signal through a sequence of ResBlocks derived from ResNet (He et al., 2016). **CNN-Transformer** (Peh et al., 2022) is a neural network consisting of a CNN-based encoder and a transformer encoder designed to extract long-range signal patterns. **FFCL** (Li et al., 2022) is designed to fuse spatial and temporal features, which are extracted by the CNN network and the LSTM network, respectively. **ST-Transformer** (Song et al., 2021) has a hierarchical architecture that sequentially applies spatial and temporal attention to capture the global dependency in the signal.

Self-supervised modeling methods. The baselines for self-supervised modeling methods include those that first learn semantic representations through a pre-training process and then perform the downstream task. **BIOT** (Yang et al., 2024) is pre-trained with contrastive learning, which aims to align the embeddings of the original signal and the perturbed signal. BIOT’s architecture consists of a biosignal tokenization module and multiple transformer blocks. BIOT uses six EEG datasets recorded using the same channel configuration for pre-training. **EEG2Rep** (Foumani et al., 2024), a transformer-based neural network, performs a self-prediction pretext task that predicts masked patches in the latent space. EEG2Rep is pre-trained on two types of datasets with the same channel configuration. **LaBraM** (Jiang et al., 2024) is an EEG foundation model pre-trained on a large EEG dataset, regardless of configurations such as channel order and signal length. LaBraM uses masked patch prediction as a pretext task and is validated on multiple out-of-source downstream tasks.

A.5 HYPERPARAMETER SETTINGS

We describe the hyperparameter settings we use for training here.

Table 4: Hyperparameters for TaKF⁺ training on LaBraM.

	Hyperparameters	Values
TaKF	Learnable latent query vector number	5
	Cross-attention blocks	6
	latent query vector dimension	32
	Cross-attention head number	4
Adapter Modules	Adapter layers	6
	Down-projection dimension	25
LaBraM	Transformer encoder layers	12
	Token dimension	200
	MLP size	800
	Attention head number	10
	Batch size	64
	Learning rate	5e-4
	Minimal learning rate	1e-6
	Optimizer	AdamW
	Weight decay	0.05
	Total epochs	50
	Early stop patience	5

Table 5: Hyperparameters for TaKF⁺ training on BIOT.

	Hyperparameters	Values
TaKF	Learnable latent query vector number	5
	Cross-attention blocks	2
	latent query vector dimension	24
	Cross-attention head number	4
Adapter Modules	Adapter layers	2
	Down-projection dimension	64
BIOT	Transformer encoder layers	4
	Token dimension	256
	MLP size	1024
	Attention head number	8
	Batch size	64
	Learning rate	1e-3
	Optimizer	Adam
	Weight decay	1e-5
	Total epochs	100
	Early stop patience	5

Table 6: Hyperparameters for TaKF(FF) training on LaBraM.

	Hyperparameters	Values
TaKF	Learnable latent query vector number	5
	Cross-attention blocks	8
	latent query vector dimension	32
	Cross-attention head number	4
	MLP size	128
LaBraM	Transformer encoder layers	12
	Token dimension	200
	MLP size	800
	Attention head number	10
	Batch size	64
	Learning rate	5e-4
	Minimal learning rate	1e-6
	Optimizer	AdamW
	Weight decay	0.05
	Total epochs	50
	Early stop patience	5

B DETAIL EXPERIMENTS ON BCI TASK

The details of our evaluation results are described as follows. The results are divided by the categories of baseline models, with sections separated by horizontal lines in the table. Our primary comparison focuses on additive fine-tuning methods. Therefore, for both LaBraM and BIOT, we mark the best (**Bold**) and second-best (underline) results only within the additive fine-tuning methods.

Table 7: Performance comparison to the competitors in emotion recognition. FT and LP denote fine-tuning and linear probing, respectively, while PT represents Prefix-Tuning.

	LeftRight Hand		DREAMER	
	BACC	AUROC	BACC	AUROC
SPaRCNet (Jing et al., 2023)	69.06 \pm 5.16	77.75 \pm 6.02	53.61 \pm 2.05	54.86 \pm 3.82
ContraWR (Yang et al., 2023)	59.20 \pm 5.47	66.25 \pm 8.31	55.67 \pm 4.96	58.85 \pm 5.23
CNN-Transformer (Peh et al., 2022)	50.28 \pm 0.56	51.94 \pm 1.17	50.37 \pm 0.64	49.23 \pm 4.13
FFCL (Li et al., 2022)	60.79 \pm 5.80	66.58 \pm 8.33	56.25 \pm 4.10	59.04 \pm 7.17
ST-Transformer (Song et al., 2021)	62.70 \pm 5.87	70.62 \pm 8.47	49.93 \pm 0.93	49.51 \pm 2.03
EEG2Rep (Foumani et al., 2024)	60.26 \pm 5.79	66.34 \pm 8.18	55.63 \pm 2.85	58.02 \pm 2.74
LaBraM-FT (Jiang et al., 2024)	60.74 \pm 3.51	41.44 \pm 5.32	55.67 \pm 3.64	59.60 \pm 4.79
BIOT-FT (Jiang et al., 2024)	49.32 \pm 0.70	49.55 \pm 0.91	49.04 \pm 1.94	48.88 \pm 3.13
LaBraM-LP	54.95 \pm 3.23	68.11 \pm 7.67	34.61 \pm 2.25	39.68 \pm 3.29
LaBraM-Adapter (Houlsby et al., 2019)	<u>65.37 \pm 11.17</u>	<u>74.75 \pm 5.89</u>	59.86 \pm 0.98	56.88 \pm 1.52
LaBraM-PT (Li & Liang, 2021)	63.18 \pm 10.45	72.06 \pm 14.11	55.56 \pm 1.94	52.44 \pm 3.48
LaBraM-MAM Adapter (He et al., 2021)	64.55 \pm 9.87	71.93 \pm 14.79	57.73 \pm 0.59	51.72 \pm 2.05
(Ours) LaBraM-TaKF ⁺	67.04 \pm 14.20	75.46 \pm 12.74	56.17 \pm 1.45	<u>54.27 \pm 1.14</u>
BIOT-LP	50.11 \pm 0.61	50.53 \pm 1.09	49.78 \pm 0.95	49.78 \pm 2.66
BIOT-Adapter	49.85 \pm 0.90	49.81 \pm 1.46	50.35 \pm 1.73	49.32 \pm 2.35
BIOT-PT	49.89 \pm 0.43	<u>51.53 \pm 0.69</u>	50.40 \pm 1.31	49.57 \pm 1.69
BIOT-MAM Adapter	50.15 \pm 0.42	51.81 \pm 1.51	50.23 \pm 1.38	51.49 \pm 3.51
(Ours) BIOT-TaKF ⁺	50.49 \pm 0.75	50.61 \pm 0.87	50.43 \pm 1.55	<u>49.92 \pm 1.20</u>

Table 8: Performance comparison to the competitors in motor-imagary classification. FT and LP denote fine-tuning and linear probing, respectively, while PT represents Prefix-Tuning.

	Crowdsourced		TUEV	
	BACC	AUROC	BACC	Cohen’s κ
SPaRCNet (Jing et al., 2023)	60.93 \pm 15.24	70.11 \pm 14.28	41.61 \pm 2.62	42.33 \pm 1.81
ContraWR (Yang et al., 2023)	56.85 \pm 15.59	68.15 \pm 23.67	43.84 \pm 3.49	39.12 \pm 2.37
CNN-Tran. (Peh et al., 2022)	52.91 \pm 5.41	60.29 \pm 14.88	40.87 \pm 1.61	38.15 \pm 1.34
FFCL (Li et al., 2022)	60.94 \pm 8.91	69.26 \pm 18.27	39.79 \pm 1.04	37.32 \pm 1.88
ST-Tran. (Song et al., 2021)	60.71 \pm 11.99	74.99 \pm 9.32	39.84 \pm 2.28	37.65 \pm 3.06
EEG2Rep (Foumani et al., 2024)	69.27 \pm 3.08	76.22 \pm 4.40	23.47 \pm 0.35	12.81 \pm 2.10
LaBraM-FT (Jiang et al., 2024)	62.04 \pm 10.51	65.30 \pm 11.15	64.09 \pm 0.65	66.37 \pm 0.93
BIOT-FT (Yang et al., 2024)	57.71 \pm 8.11	69.42 \pm 9.39	52.81 \pm 2.25	52.73 \pm 2.49
LaBraM-LP	54.95 \pm 3.23	68.11 \pm 7.67	34.61 \pm 2.25	39.68 \pm 3.29
LaBraM-Adapter (Houlsby et al., 2019)	<u>65.37 \pm 11.17</u>	<u>74.75 \pm 5.89</u>	59.86 \pm 0.98	56.88 \pm 1.52
LaBraM-PT (Li & Liang, 2021)	63.18 \pm 10.45	72.06 \pm 14.11	55.56 \pm 1.94	52.44 \pm 3.48
LaBraM-MAM Adapter (He et al., 2021)	64.55 \pm 9.87	71.93 \pm 14.79	57.73 \pm 0.59	51.72 \pm 2.05
(Ours) LaBraM-TaKF ⁺	67.04 \pm 14.20	75.46 \pm 12.74	56.17 \pm 1.45	<u>54.27 \pm 1.14</u>
BIOT-LP	61.87 \pm 8.16	68.62 \pm 9.17	37.47 \pm 1.25	46.66 \pm 2.48
BIOT-Adapter	58.16 \pm 8.24	65.44 \pm 5.70	45.54 \pm 2.77	51.12 \pm 4.33
BIOT-PT	59.98 \pm 7.56	72.55 \pm 8.30	36.01 \pm 1.45	35.09 \pm 2.67
BIOT-MAM Adapter	58.36 \pm 7.76	70.08 \pm 6.90	48.49 \pm 1.83	47.58 \pm 3.46
(Ours) BIOT-TaKF ⁺	63.83 \pm 7.06	<u>70.44 \pm 4.70</u>	<u>46.66 \pm 1.22</u>	51.56 \pm 2.03

C ABLATION ON FEATURE DISTILLATION POINTS IN TRANSFORMER BLOCKS

To ensure that TaKF functions as intended, we conduct an ablation study to determine the optimal point for TaKF to extract representation features from the transformer block of the EEG foundation model. We adopt the TaKF(+FF) introduced in Section 6.3 for the ablation study, using LaBraM as the EEG foundation model. We elaborate on three cases. **Case 1:** extraction before the projection layer in the attention layer. **TaKF(+FF):** extraction before the residual connection in the

attention layer. **Case 2:** extraction before the feed-forward layer. **Case 3** extraction after the feed-forward layer. We use three datasets to highlight the differences in extraction positions and report the results in Table 9. It is noteworthy that whether the representation features pass through the feed-forward layer is a key factor in determining the characteristics of TaKF. It can be observed that on the DREAMER and Crowdsourced datasets, the results of TaKF, Case 1, and Case 2 are quite similar, while in Case 3, there is a slight degradation. In contrast, in TUEV, Case 3 achieves the best performance. We designed each component of TaKF⁺ with two intentions. First, TaKF functions solely to make the model more expressive by learning new task-relevant patterns in a low-dimensional space. Second, the adapter modules transform the prior knowledge of the EEG foundation model into a task-specific form. Therefore, extracting the representation features obtained before the feed-forward layer, specifically before the residual connection in the attention layer, aligns more closely with the intended purpose compared to those obtained after the feed-forward layer.

Table 9: Ablation study to validate the effectiveness of where the representation features are extracted from within the transformer block. The lowest performance values are underlined.

	TUEV		DREAMER		Crowdsourced	
	BACC	Cohen’s κ	BACC	AUROC	BACC	AUROC
LaBram-TaKF(+FF)	<u>53.21 \pm 2.44</u>	51.90 \pm 1.72	56.85 \pm 4.29	60.01 \pm 5.52	62.30 \pm 8.22	68.76 \pm 6.87
Case 1	53.31 \pm 1.53	<u>48.86 \pm 2.99</u>	55.65 \pm 2.79	60.70 \pm 6.07	62.01 \pm 11.45	68.00 \pm 18.83
Case 2	57.25 \pm 3.89	53.48 \pm 3.30	55.66 \pm 1.87	60.25 \pm 5.28	<u>62.61 \pm 9.98</u>	67.49 \pm 12.22
Case 3	57.76 \pm 2.90	50.97 \pm 3.89	<u>54.65 \pm 2.90</u>	<u>59.04 \pm 8.43</u>	<u>61.87 \pm 5.50</u>	<u>65.13 \pm 18.11</u>

Case 1: Extraction from before the projection layer in the attention layer

Case 2: Extraction from before the feed-forward layer

Case 3: Extraction from after the feed-forward layer

D ABLATION ON DATA SCALE

We present the detailed results of the ablation study on data scale in Tables 10 and 11. To verify the strength of our proposed methods in a data scarcity scenario, we selected two datasets, Crowdsourced and LeftRight Hand. Bold values represent the best results. Notably, TaKF⁺ demonstrates high data efficiency in low-data scenarios.

Table 10: Ablation about data scale on Crowdsourced. BACC and AUROC are used as evaluation metric. FT denotes fine-tuning, and PT represents Prefix-Tuning.

	Crowdsourced (BACC)		
	4-shot	8-shot	12-shot
LaBraM-FT (Jiang et al., 2024)	70.29 \pm 14.00	75.04 \pm 15.80	79.54 \pm 14.34
LaBraM-Adapter (Houlsby et al., 2019)	60.47 \pm 9.31	67.44 \pm 14.11	71.59 \pm 12.45
LaBraM-PT (Li & Liang, 2021)	53.67 \pm 7.96	60.57 \pm 14.25	65.59 \pm 15.01
LaBraM-MAM Adapter (He et al., 2021)	55.95 \pm 8.43	58.31 \pm 7.84	57.81 \pm 8.79
(Ours) LaBraM-TaKF ⁺	73.42 \pm 8.43	78.63 \pm 12.88	81.46 \pm 11.16
	Crowdsourced (AUROC)		
	4-shot	8-shot	12-shot
LaBraM-FT	76.62 \pm 15.80	81.16 \pm 16.21	87.14 \pm 11.55
LaBraM-Adapter	67.81 \pm 14.53	74.29 \pm 16.15	79.11 \pm 15.58
LaBraM-PT	55.72 \pm 11.83	63.94 \pm 17.61	71.03 \pm 18.75
LaBraM-MAM Adapter	58.81 \pm 10.62	62.08 \pm 9.29	60.33 \pm 11.24
(Ours) LaBraM-TaKF ⁺	80.03 \pm 14.52	84.71 \pm 13.86	88.99 \pm 10.82

Table 11: Ablation about data scale on LeftRight Hand. BACC and AUROC are used as evaluation metric. FT denotes fine-tuning, and PT represents Prefix-Tuning.

	LeftRight Hand (BACC)		
	4-shot	8-shot	12-shot
LaBraM-FT (Jiang et al., 2024)	53.52 \pm 8.31	53.22 \pm 11.05	54.49 \pm 16.14
LaBraM-Adapter (Houlsby et al., 2019)	50.55 \pm 3.89	51.72 \pm 6.07	53.05 \pm 8.23
LaBraM-PT (Li & Liang, 2021)	51.30 \pm 7.47	50.75 \pm 6.02	55.53 \pm 14.48
LaBraM-MAM Adapter (He et al., 2021)	51.00 \pm 4.88	50.30 \pm 6.44	50.27 \pm 6.30
(Ours) LaBraM-TaKF ⁺	53.35 \pm 10.10	53.80 \pm 12.13	55.46 \pm 13.33
	LeftRight Hand (AUROC)		
	4-shot	8-shot	12-shot
LaBraM-FT	53.97 \pm 10.64	54.01 \pm 13.22	55.00 \pm 17.44
LaBraM-Adapter	52.09 \pm 7.00	54.16 \pm 10.46	54.45 \pm 11.15
LaBraM-PT	52.20 \pm 8.43	50.41 \pm 7.86	55.63 \pm 16.48
LaBraM-MAM Adapter	52.23 \pm 6.22	53.80 \pm 8.13	49.91 \pm 8.08
(Ours) LaBraM-TaKF ⁺	53.40 \pm 13.29	55.50 \pm 14.33	57.12 \pm 15.04