MExFormer: Rethinking Early Channel Fusion for Multi-channel EEG Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

Learning informative EEG representations for identifying neurological states requires capturing channel-specific temporal dynamics and dependencies across spatially distinct electrodes. Existing multivariate time-series Transformers employ diverse strategies for channel interaction, including mixing information across channels during tokenization or embedding rather than retaining individual channels as explicit attention tokens, potentially limiting the direct modeling of inter-electrode interactions. We therefore introduce MExFormer, a Transformer that independently tokenizes each channel into temporally localized patches at multiple resolutions and defers channel fusion to explicit inter-channel self-attention, progressively integrating channel interactions, temporal structure, and multi-resolution information through a three-stage hierarchy. We evaluate the proposed approach on EEG classification across two public datasets under a subject-independent protocol. Beyond classification performance, we assess the quality of learned representations through robustness to corrupted EEG inputs and post-hoc evaluation of intermediate-layer representations using a weak classifier. Across the datasets and six evaluation metrics, MExFormer achieves the strongest overall performance on APAVA dataset and remains competitive on TDBRAIN dataset, while exhibiting greater robustness to channel corruption and more discriminative intermediate representations.

1 Introduction

Electroencephalography (EEG) provides a non-invasive measurement of brain electrical activity through multiple electrodes distributed across the scalp, with each channel recording a temporally evolving potential at a distinct spatial location (Roy et al., 2019). Analyzing EEG signals can provide useful information about a subject's cognitive and affective states, intentions, and neurological conditions (Badr et al., 2024; Roy et al., 2019). These states and conditions are reflected not only in characteristic temporal patterns within individual channels, but also in coordinated activity and dependencies across spatially distributed electrodes (Bazinet et al., 2023). Thus, an informative representation for classifying such states of interest must preserve both channel-specific temporal structure and inter-channel relationships. Given these distinctive features of EEG, Transformer-based models (Vaswani et al., 2017) provide a flexible framework for modeling the required dependencies: the way multichannel EEG is tokenized determines which temporal and cross-channel structures remain explicitly accessible, while the organization of attention determines how and when these structures interact.

Multivariate time-series Transformers expose these design choices through different tokenization and attention strategies, but they do not all preserve the same information for subsequent representation learning. When channel information is mixed during embedding, individual channels are no longer explicitly accessible as independent representation units during subsequent attention. Conversely, preserving channel identity alone does not ensure that inter-channel dependencies are explicitly modeled at the stage where channel-specific temporal patterns are still localized. In multichannel EEG, where coordinated activity across spatially distributed electrodes may occur at specific temporal locations and across multiple scales, this distinction is important: inter-channel interactions should be resolved while channel-specific temporal structure remains explicit, and be-

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.

fore subsequent temporal aggregation progressively removes this fine-grained correspondence. A representation-learning hierarchy should therefore retain individual channels as explicit units, first resolve their interactions at corresponding temporal positions, and only then integrate information across temporal positions and resolutions. This ordering provides a direct mechanism for progressively combining spatial, temporal, and multi-resolution structure while avoiding premature mixing or aggregation that can obscure the relationships between individual channels.

Taken together, these considerations motivate preserving localized temporal representations with explicit channel identity and resolving inter-channel interactions at the corresponding temporal positions to extract first the inter-electrode dynamics before temporal aggregation. MExFormer addresses this gap by independently patching each channel at multiple temporal granularities and deferring channel mixing to a dedicated inter-channel self-attention stage, making inter-channel interaction the first attention operation applied to the channel-specific temporal tokens. Its hierarchical attention encoder progressively integrates information in three stages: (i) channel-specific tokens at each temporal position first interact through explicit inter-channel self-attention, (ii) the resulting container representations extracted via self-attention are then integrated across temporal positions within each granularity, and (iii) the resulting router representations from different temporal granularities are finally combined. This ordering preserves channel identity during the initial interaction stage while progressively integrating cross-channel, temporal, and multi-resolution information.

The motivation for preserving channel-specific representations extends beyond classification accuracy. If explicit cross-channel interactions produce more informative representations, their benefits should also be reflected in how those representations respond to channel-level perturbations and how easily their information can be decoded by a weak downstream classifier. We therefore evaluate MExFormer not only under the standard subject-independent classification setting, but also through robustness to corrupted EEG inputs and post-hoc evaluation of intermediate-layer representations using a weak classifier. These complementary evaluations probe whether the proposed channel-preserving interaction strategy yields representations that remain informative under input corruption and become increasingly discriminative throughout the hierarchy.

2 Related Work

Tokenization and Cross-Channel Modeling in Multivariate Time Series. Multivariate time-series Transformers adopt substantially different tokenization strategies, which determine whether channel identity remains explicitly available to subsequent attention. A broad family follows the standard temporal-token formulation, in which the values of all channels at a given timestamp are jointly embedded into a single token. This formulation is used by widely adopted time-series Transformers such as Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), and the Non-stationary Transformer (Liu et al., 2022). Although these methods differ in how they model temporal dependencies and improve efficiency, the tokenization step mixes channel information before self-attention is applied. Consequently, subsequent attention cannot directly assign distinct representations or attention interactions to individual channels at a given temporal location.

A second family avoids this early fusion by processing channels independently. PatchTST (Nie et al., 2023) partitions each channel into temporal patches and applies shared self-attention independently to each channel, preserving channel identity and localized temporal structure but providing no explicit mechanism for learning inter-channel dependencies. MTST (Zhang et al., 2024) extends patch-based modeling to multiple temporal resolutions while retaining channel independence, whereas Pathformer (Chen et al., 2024) introduces adaptive multi-scale patch pathways and dual attention to model temporal structure across scales. These approaches therefore provide increasingly expressive temporal representations, including multi-resolution ones, but do not explicitly introduce inter-channel attention over the resulting channel-specific patch tokens. For EEG, this leaves the spatial dimension isolated from the localized temporal representations at precisely the stage where channel-specific temporal patterns are still explicitly identifiable.

A third family exposes cross-channel dependencies by changing the semantic unit represented by a token. iTransformer (Liu et al., 2024) treats each entire channel as a single token and applies attention across them, directly modeling multivariate correlations but representing a complete temporal history within one token. This reverses the standard temporal-token formulation, but in doing so removes the patch-level temporal localization provided by patch-based representations. Cross-

former (Zhang & Yan, 2022) by using Dimension-Segment-Wise embedding, preserves channel- and segment-specific tokens and therefore retains both temporal segmentation and channel identity. Its Two-Stage Attention (TSA), however, first models dependencies across temporal segments and only subsequently performs cross-dimension attention. Moreover, the cross-dimension stage uses a small set of learned routers to gather and redistribute information across dimensions, providing a computationally efficient approximation to direct all-to-all cross-channel interaction. MExFormer differs in both respects. It resolves cross-channel dependencies first, at corresponding temporal positions, through direct self-attention among the channel-specific tokens, and only then efficiently aggregates information across temporal positions through its container representations.

Medical Time-Series Transformers. Medformer (Wang et al., 2024) introduces cross-channel patching, multi-granularity embeddings, and intra- and inter-granularity self-attention for medical time-series classification. Its cross-channel patching jointly incorporates the values of multiple channels within each temporal patch into a single embedded token before self-attention is applied. As a result, Medformer can exploit both temporal context and inter-channel information, while its subsequent intra- and inter-granularity attention models temporal structure within each granularity and interactions across them. However, the cross-channel dependencies at a given temporal position are introduced through the embedding operation itself rather than through learned attention between explicitly separated channel tokens. This distinction is important for EEG because the signals recorded by different electrodes can exhibit coordinated activity while simultaneously containing channel-specific temporal patterns; preserving those patterns as distinct representations allows their interactions to be learned explicitly rather than being resolved by a fixed projection during tokenization.

MExFormer retains the localized channel-specific temporal structure of patch-based models while providing explicit inter-channel attention, and performs this before temporal and multi-resolution aggregation. The resulting ordering is designed to preserve fine-grained channel correspondence when spatial dependencies are first learned, before temporal and multi-resolution aggregation progressively abstracts the representation. In this sense, MExFormer does not introduce multi-scale modeling or cross-channel attention in isolation; rather, it combines them with a specific ordering that keeps electrode identity and localized temporal structure explicit when inter-electrode dependencies are first extracted.

3 Proposed Method

Problem (Multichannel EEG Classification). Let an EEG sample be represented as $\boldsymbol{x} \in \mathbb{R}^{C\times T}$, where C denotes the number of EEG channels (electrodes) and T the number of temporal samples. The corresponding class label is $y \in \{1, \dots, K\}$, where K is the number of target classes. Our objective is to learn an encoder $\mathcal{F} : \mathbb{R}^{C\times T} \to \mathcal{H}$ that maps the multichannel EEG signal to a representation $\boldsymbol{h} \in \mathcal{H}$ that is informative for classification.

MExFormer consists of two conceptual components: a multi-resolution embedding module and a hierarchical attention encoder. The embedding module converts the raw EEG signal into a set of channel-preserving tokens at each temporal granularity. The encoder then progressively integrates information across channels and temporal positions before combining information across temporal resolutions.

3.1 Embedding

Temporal Patching. Let $\mathcal{L} = \{L_1, L_2, \dots, L_G\}$ denote the set of G temporal patch lengths, where L_g specifies the temporal extent (window) of the patches at granularity $g \in \{1, 2, \dots, G\}$. Each granularity therefore provides a distinct temporal view of the same multichannel EEG sample.

For each granularity g, the input signal is partitioned independently along the temporal dimension using non-overlapping patches of length L_g. The number of patches per channel is $N_g = \lceil T/L_g \rceil$. When T is not divisible by L_g, the signal is extended at the right boundary so that it can be partitioned into N_g complete patches. Denoting the resulting patch sequence by $\boldsymbol{x}_p^{(g)} \in \mathbb{R}^{C\times N_g\times L_g}$, the n-th patch of channel c derived from the input sample is:

$$\boldsymbol{x}_p^{(g)}[c, n] \in \mathbb{R}^{L_g}, \qquad c = 1, \dots, C \quad n = 1, \dots, N_g.$$

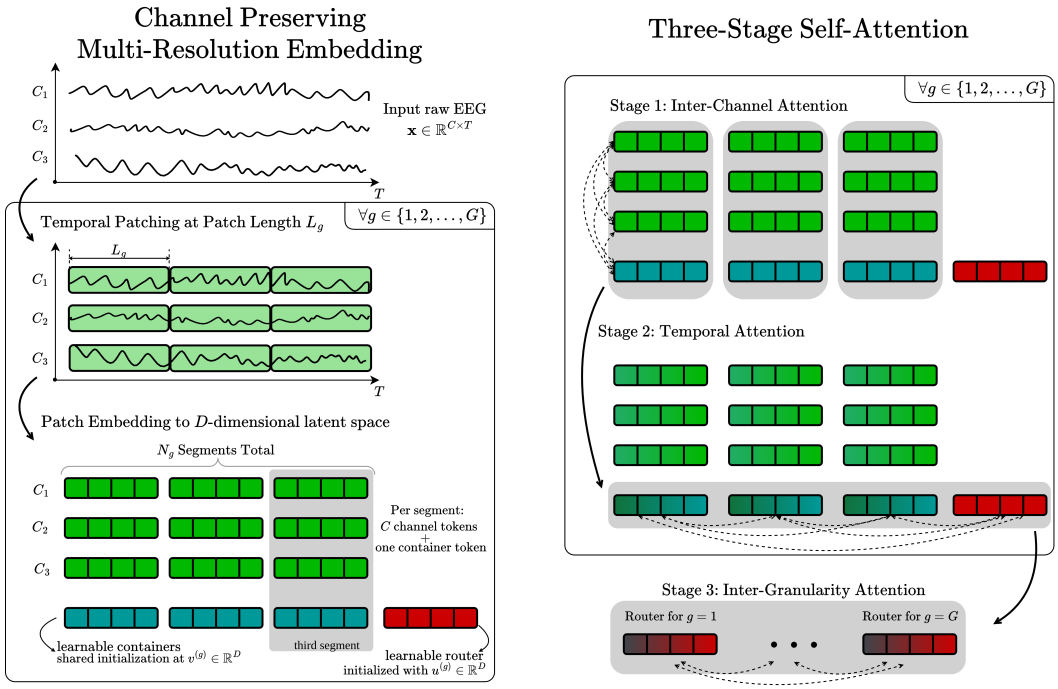

Figure 1: Embedding and Attention.

Patch Embedding. Each patch is linearly projected into a shared D-dimensional latent space:

$$\boldsymbol{z}^{(g)}[c, n] = \boldsymbol{W}^{(g)}\, \boldsymbol{x}_p^{(g)}[c, n] \in \mathbb{R}^D, \qquad c = 1, \ldots, C, \quad n = 1, \ldots, N_g,$$

where $\boldsymbol{W}^{(g)} \in \mathbb{R}^{D\times L_g}$ is a granularity-specific projection matrix, shared across all channels and all patch positions. Crucially, the projection acts independently on each channel: no mixing across the c axis occurs at embedding time, so the resulting tokens $\boldsymbol{z}^{(g)}[c, n]$ retain a one-to-one correspondence with individual electrodes.

Segment Container. To make channel interactions an explicit learnable computation rather than an artifact of embedding, we introduce for each granularity g a single learnable *container* vector $\boldsymbol{v}^{(g)} \in \mathbb{R}^D$. At every patch position n, the container is appended to the C channel tokens as an additional, $(C{+}1)$-th token:

$$\boldsymbol{z}^{(g)}[C + 1, n] := \boldsymbol{v}^{(g)}, \qquad n = 1, \ldots, N_g.$$

The container does not correspond to any physical electrode. It acts as a dedicated read-write slot that the inter-channel self-attention mechanism uses to aggregate and redistribute information across channels at a given temporal position (segment), while leaving the C channel tokens free for further processing at subsequent stages. The same parameter $\boldsymbol{v}^{(g)}$ is shared across all N_g segments of granularity g.

Positional Encoding. We inject two independent sources of positional information into every token, using the fixed sinusoidal encoding $\mathrm{PE} : \mathbb{N} \to \mathbb{R}^D$: a token-identity encoding $\mathrm{PE}_{\mathrm{tok}}(c)$, distinguishing the C channels and the container from one another, and a temporal encoding $\mathrm{PE}_{\mathrm{pos}}(n)$, encoding the patch position along the temporal axis. Writing $\bar{C} = C + 1$, the fully positioned token is:

$$\tilde{\boldsymbol{z}}^{(g)}[c, n] = \boldsymbol{z}^{(g)}[c, n] + \mathrm{PE}_{\mathrm{tok}}(c) + \mathrm{PE}_{\mathrm{pos}}(n), \qquad c = 1, \ldots, \bar{C}, \quad n = 1, \ldots, N_g.$$

Stacking all tokens of granularity g yields

$$\tilde{\boldsymbol{z}}^{(g)} \in \mathbb{R}^{\bar{C}\times N_g\times D}.$$

Granularity Router. Finally, to support communication across granularities during the last stage of the self-attention, we introduce for each granularity g a single learnable *router* vector $\boldsymbol{u}^{(g)} \in \mathbb{R}^D$, acting as a global summary token for that granularity. The router is placed at a temporal index immediately following the last segment position, $N_g + 1$:

$$\tilde{\boldsymbol{u}}^{(g)} = \boldsymbol{u}^{(g)} + \mathrm{PE}_{\mathrm{pos}}(N_g + 1) \in \mathbb{R}^D.$$

Positioning the router this way keeps it embedded in the same temporal coordinate system as the patch tokens, while ensuring it is never conflated with any of them.

Overall, the embedding module maps the raw sample $\boldsymbol{x}$ to G granularity-specific representations $\{(\tilde{\boldsymbol{z}}^{(g)}, \tilde{\boldsymbol{u}}^{(g)})\}_{g=1}^{G}$, each consisting of C channel-preserving tokens, one segment container per temporal position, and a single per-granularity router. These are passed jointly to the hierarchical attention encoder described next: channel representations are first contextualized via inter-channel self-attention over the $\bar{C}$ axis at each position n, temporal structure is then integrated along the N_g axis, and information across the G granularities is finally combined through the router tokens $\tilde{\boldsymbol{u}}^{(g)}$.

3.2 Self-Attention

All three stages below use the same primitive: standard scaled dot-product self-attention with H heads (Vaswani et al., 2017). No causal masking is applied at any stage.

Stage 1: Inter-Channel Attention. At every patch position n of every granularity g exist $\bar{C}$ tokens (C channel tokens together with the container), which are treated as a sequence and updated jointly via full self-attention, independently across n, with attention weights shared across positions but dedicated to each granularity:

$$\big[\boldsymbol{z}^{(g)}[1,n], \ldots, \boldsymbol{z}^{(g)}[\bar{C},n]\big] \;\leftarrow\; \mathrm{Attn}_{\mathrm{cc}}^{(g)}\Big(\big[\boldsymbol{z}^{(g)}[1,n], \ldots, \boldsymbol{z}^{(g)}[\bar{C},n]\big]\Big), \qquad n = 1, \ldots, N_g.$$

This is the mechanism that makes channel interactions explicit: since the container is a genuine query/key/value participant, it attends to, and is attended by, every channel token, so cross-channel dependencies are captured as attention weights rather than folded into a fixed linear combination at embedding time. The updated container $\boldsymbol{v}^{(g)}[n] := \boldsymbol{z}^{(g)}[\bar{C}, n]$ is carried forward to Stage 2.

Stage 2: Temporal Attention. Having summarized cross-channel information at each position into a single container token, temporal structure is integrated by letting the N_g containers of a granularity attend to one another, together with that granularity's router:

$$\big[\boldsymbol{v}^{(g)}[1], \ldots, \boldsymbol{v}^{(g)}[N_g], \boldsymbol{u}^{(g)}\big] \;\leftarrow\; \mathrm{Attn}_{\mathrm{temp}}^{(g)}\Big(\big[\boldsymbol{v}^{(g)}[1], \ldots, \boldsymbol{v}^{(g)}[N_g], \boldsymbol{u}^{(g)}\big]\Big).$$

This sequence has length $N_g + 1$: the router acts as a distinguished slot that exchanges information with every container in a single attention step. After this stage, the router carries a temporally-global summary of granularity g, while the containers are refined with context from other positions of the same granularity.

Stage 3: Inter-Granularity Attention. The routers of all G granularities, each now encoding a full-resolution, cross-channel summary from Stages 1-2, are collected into one sequence and jointly attended, using a single attention module:

$$\big[\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(G)}\big] \;\leftarrow\; \mathrm{Attn}_{\mathrm{res}}\Big(\big[\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(G)}\big]\Big).$$

This final self-attention operates purely over the G resolutions, letting the model combine coarse and fine temporal views of the same EEG sample.

Layer Composition and Depth. The three steps above, together with a position-wise feed-forward network, residual connections, and layer normalization (Post-LN), constitute a single *MExFormer layer*. Stacking E such layers, initialized from the embedding output $\{(\tilde{\boldsymbol{z}}^{(g)}, \tilde{\boldsymbol{u}}^{(g)})\}_{g=1}^{G}$, yields the full hierarchical attention encoder. After the final layer, the router tokens of all granularities are concatenated,

$$\boldsymbol{h} = \big[\boldsymbol{u}^{(1)}; \ldots; \boldsymbol{u}^{(G)}\big] \in \mathbb{R}^{G \cdot D},$$

giving the representation $\boldsymbol{h} \in \mathcal{H} = \mathbb{R}^{G \cdot D}$ passed to the task-specific head, and completing the encoder $\mathcal{F}$ introduced above.

4 EXPERIMENTS

4.1 DATASETS AND EXPERIMENTAL SETUP

We evaluate MExFormer on two public EEG datasets, APAVA (Escudero et al., 2006) and TDBRAIN (van Dijk et al., 2022), under a subject-independent classification protocol. APAVA contains EEG recordings from 23 subjects and is formulated as a binary classification task distinguishing subjects with Alzheimer's disease from healthy controls. TDBRAIN contains recordings from 72 subjects and similarly defines a binary classification task for Parkinson's disease. After preprocessing, each sample consists of a one-second EEG segment with 16 channels and 256 temporal samples for APAVA, and 33 channels and 256 temporal samples for TDBRAIN (both datasets are sampled at 256 Hz). The resulting datasets contain 5,967 and 8,860 samples for APAVA and TDBRAIN, respectively, as summarized in Table 1.

Table 1: Dataset summary

Dataset	Healthy	Non-healthy	Total
APAVA	2,367 samples from 11 subjects	3,600 samples from 12 subjects	5,967
TDBRAIN	5,620 samples from 47 subjects	3,240 samples from 25 subjects	8,860

For all experiments, the training, validation, and test sets are partitioned at the subject level using a fixed set of subject identifiers, such that samples from the same subject do not occur across different splits. The same partition is used for all models to ensure a controlled comparison. Each experiment is repeated over five random seeds, 41-45, (independent runs) varying the model initialization while keeping the data splits fixed. We report the mean and standard deviation across these five runs. We evaluate performance using accuracy, precision, recall, F1, AUROC, and AUPRC. These metrics provide complementary measures of classification quality, particularly in the presence of class imbalance.

4.2 SUBJECT-INDEPENDENT CLASSIFICATION

We compare MExFormer against eleven Transformer-based multivariate time-series baselines: Transformer (Vaswani et al., 2017), Informer (Zhou et al., 2021), Reformer (Kitaev et al., 2019), Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), Non-stationary Transformer (Liu et al., 2022), PatchTST (Nie et al., 2023), Crossformer (Zhang & Yan, 2022), iTransformer (Liu et al., 2024), MTST (Zhang et al., 2024), and Medformer (Wang et al., 2024). These baselines cover the principal tokenization and attention strategies discussed in Section 2, including early channel mixing, channel-independent patching, explicit channel-token modeling, and multi-granularity medical time-series modeling.

Table 2 presents the classification results under the subject-independent setting. On APAVA, MExFormer achieves the highest accuracy, recall, macro-F1, AUROC, and AUPRC among the evaluated models. On TDBRAIN, the overall classification task is stronger across several baselines, with MExFormer remaining highly competitive, achieving 90.23 ± 0.92 accuracy, 90.20 ± 0.92 macro-F1, 97.28 ± 0.81 AUROC, and 97.35 ± 0.88 AUPRC. The difference from the strongest clean baseline is relatively small, while the subsequent experiments show a clearer distinction under channel corruption and in intermediate-layer representation quality.

4.3 ROBUSTNESS TO CHANNEL EXCLUSION

To evaluate whether the learned representations remain informative when individual EEG channels are unavailable, we perform inference on test inputs with controlled channel exclusion. The models are trained only on the original clean data (whose results were presented above); channel exclusion is introduced at evaluation time without modifying the labels or sample ordering. We evaluate three exclusion levels, making the samples increasingly more corrupted, by randomly zeroing each input channel's whole values by 10%, 20%, and 30%. This setting directly probes the sensitivity of each architecture to the loss of spatially distributed EEG measurements (of some electrodes).

Table 2: Test results on the APAVA and TDBRAIN datasets. All models are trained and evaluated using the same subject-independent data partitions and experimental protocol.

APAVA (2-Classes)

Models	Accuracy	Precision	Recall	F1 score	AUROC	AUPRC
Autoformer	$75.35_{\pm 8.17}$	$76.14_{\pm 5.60}$	$75.21_{\pm 6.18}$	$74.50_{\pm 7.96}$	$83.29_{\pm 5.15}$	$82.29_{\pm 6.76}$
Crossformer	$75.60_{\pm 2.00}$	$80.65_{\pm 3.74}$	$71.10_{\pm 2.09}$	$71.53_{\pm 2.36}$	$72.36_{\pm 4.36}$	$72.70_{\pm 4.06}$
FEDformer	$74.49_{\pm 2.35}$	$74.34_{\pm 1.48}$	$73.18_{\pm 4.03}$	$73.01_{\pm 3.88}$	$83.65_{\pm 2.19}$	$82.86_{\pm 2.56}$
Informer	$74.10_{\pm 2.87}$	$76.75_{\pm 3.09}$	$70.02_{\pm 3.54}$	$70.30_{\pm 4.27}$	$71.48_{\pm 2.70}$	$72.07_{\pm 2.60}$
MTST	$69.90_{\pm 2.31}$	$77.63_{\pm 1.90}$	$63.82_{\pm 2.98}$	$62.12_{\pm 4.12}$	$65.94_{\pm 1.03}$	$68.38_{\pm 1.63}$
Nonformer	$72.48_{\pm 4.84}$	$73.05_{\pm 6.03}$	$69.46_{\pm 4.91}$	$69.80_{\pm 5.20}$	$70.50_{\pm 6.32}$	$70.65_{\pm 7.79}$
PatchTST	$67.58_{\pm 1.61}$	$77.02_{\pm 1.93}$	$60.84_{\pm 2.24}$	$57.70_{\pm 3.73}$	$63.50_{\pm 3.79}$	$66.07_{\pm 3.18}$
iTransformer	$74.55_{\pm 1.85}$	$74.78_{\pm 2.34}$	$71.76_{\pm 1.92}$	$72.30_{\pm 2.00}$	$85.59_{\pm 1.73}$	$84.39_{\pm 1.76}$
Reformer	$78.14_{\pm 1.64}$	$81.93_{\pm 3.84}$	$74.41_{\pm 1.08}$	$75.26_{\pm 1.24}$	$73.91_{\pm 2.23}$	$76.23_{\pm 1.69}$
Transformer	$72.75_{\pm 4.10}$	$72.85_{\pm 3.75}$	$69.75_{\pm 5.26}$	$69.90_{\pm 5.94}$	$70.20_{\pm 5.70}$	$70.50_{\pm 7.28}$
Medformer	$76.58_{\pm 2.04}$	$78.08_{\pm 3.59}$	$73.68_{\pm 1.39}$	$74.30_{\pm 1.61}$	$82.30_{\pm 1.31}$	$83.08_{\pm 1.26}$
MExFormer	$\underline{79.71_{\pm 1.71}}$	$\underline{82.34_{\pm 1.58}}$	$\underline{76.42_{\pm 2.16}}$	$\underline{77.36_{\pm 2.22}}$	$\underline{89.79_{\pm 2.83}}$	$\underline{89.36_{\pm 3.06}}$

TDBRAIN (2-Classes)

Models	Accuracy	Precision	Recall	F1 score	AUROC	AUPRC
Autoformer	$86.52_{\pm 7.08}$	$87.22_{\pm 6.02}$	$86.52_{\pm 7.08}$	$86.39_{\pm 7.31}$	$92.75_{\pm 4.73}$	$92.30_{\pm 5.02}$
Crossformer	$78.08_{\pm 4.30}$	$78.43_{\pm 4.13}$	$78.08_{\pm 4.30}$	$78.00_{\pm 4.37}$	$88.21_{\pm 3.14}$	$88.61_{\pm 3.08}$
FEDformer	$72.96_{\pm 0.72}$	$73.25_{\pm 0.82}$	$72.96_{\pm 0.72}$	$72.87_{\pm 0.70}$	$81.66_{\pm 1.44}$	$81.96_{\pm 1.64}$
Informer	$\underline{90.50_{\pm 0.34}}$	$\underline{90.95_{\pm 0.36}}$	$\underline{90.50_{\pm 0.34}}$	$\underline{90.47_{\pm 0.35}}$	$\underline{97.38_{\pm 0.32}}$	$\underline{97.42_{\pm 0.27}}$
MTST	$76.96_{\pm 3.31}$	$77.45_{\pm 3.73}$	$76.96_{\pm 3.31}$	$76.87_{\pm 3.26}$	$85.91_{\pm 3.91}$	$84.15_{\pm 5.20}$
Nonformer	$88.50_{\pm 1.53}$	$89.13_{\pm 1.30}$	$88.50_{\pm 1.53}$	$88.45_{\pm 1.56}$	$96.62_{\pm 0.76}$	$96.56_{\pm 0.80}$
PatchTST	$77.88_{\pm 2.19}$	$78.16_{\pm 2.46}$	$77.88_{\pm 2.19}$	$77.83_{\pm 2.15}$	$86.87_{\pm 2.93}$	$84.70_{\pm 4.12}$
iTransformer	$71.46_{\pm 0.61}$	$71.57_{\pm 0.68}$	$71.46_{\pm 0.61}$	$71.42_{\pm 0.59}$	$80.10_{\pm 0.65}$	$80.08_{\pm 0.59}$
Reformer	$90.22_{\pm 1.10}$	$90.59_{\pm 1.21}$	$90.22_{\pm 1.09}$	$90.20_{\pm 1.09}$	$96.91_{\pm 0.90}$	$97.01_{\pm 0.84}$
Transformer	$87.94_{\pm 1.89}$	$89.19_{\pm 1.50}$	$87.94_{\pm 1.89}$	$87.83_{\pm 1.95}$	$96.78_{\pm 0.57}$	$96.81_{\pm 0.51}$
Medformer	$87.71_{\pm 0.99}$	$87.79_{\pm 1.07}$	$87.71_{\pm 0.99}$	$87.70_{\pm 0.99}$	$95.45_{\pm 0.40}$	$95.53_{\pm 0.38}$
MExFormer	$90.23_{\pm 0.92}$	$90.64_{\pm 0.86}$	$90.23_{\pm 0.92}$	$90.20_{\pm 0.92}$	$97.28_{\pm 0.81}$	$97.35_{\pm 0.88}$

Table 3 reports accuracy and macro-F1 under the three channel exclusion levels. MExFormer achieves the highest accuracy at all channel exclusion levels on both datasets. On APAVA, MExFormer also obtains the highest macro-F1 at 10% channel exclusion, while at 20% and 30% channel exclusion Autoformer attains a higher macro-F1. On TDBRAIN, MExFormer achieves the highest accuracy and macro-F1 at all three channel exclusion levels.

The absolute advantage of MExFormer becomes particularly pronounced on TDBRAIN as the amount of channel exclusion increases: its accuracy decreases from 88.90 ± 1.32 at 10% channel exclusion to 86.60 ± 1.39 at 30%, whereas the corresponding results of the other models decline more substantially. These results suggest that retaining channel-specific representations while explicitly learning inter-channel interactions provides a useful basis for prediction when a subset of spatially distributed measurements is unavailable.

4.4 Intermediate Representation Analysis

Classification accuracy alone does not reveal whether the proposed hierarchy produces increasingly informative representations throughout the encoder. We therefore perform a post-hoc evaluation of intermediate-layer representations using a weak downstream classifier (logistic regression). We extract the representation produced at encoder layers 1, 3, and 6 while keeping the encoder fixed, and fit the same downstream evaluation procedure to the resulting representations. This evaluation isolates the discriminative information available in the learned representation at different depths, independently of the original task-specific prediction head.

The core reason for performing this analysis is to observe how the proposed three-step attention mechanism differs from other tokenization and attention approaches in terms of channel fusion. These approaches include having an exclusive token for each channel and performing cross-channel

Table 3: Results under channel corruption at different corruption levels. The training, validation, and test sets are distributed based on subjects according to a predetermined ratio/IDs. Results are reported for the APAVA and TDBRAIN datasets.

APAVA (2-Classes)

Models	10% Channel Corruption		20% Channel Corruption		30% Channel Corruption	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Autoformer	$73.81_{\pm 4.79}$	$72.53_{\pm 4.39}$	$72.48_{\pm 3.26}$	$\underline{70.61_{\pm 3.11}}$	$70.17_{\pm 2.00}$	$\underline{67.26_{\pm 2.47}}$
Crossformer	$70.76_{\pm 1.11}$	$67.46_{\pm 1.58}$	$66.57_{\pm 1.10}$	$64.27_{\pm 1.49}$	$63.69_{\pm 0.48}$	$62.08_{\pm 0.95}$
FEDformer	$73.42_{\pm 3.03}$	$70.70_{\pm 5.42}$	$71.80_{\pm 3.31}$	$67.44_{\pm 6.46}$	$69.35_{\pm 3.42}$	$62.74_{\pm 7.36}$
Informer	$69.29_{\pm 1.87}$	$63.32_{\pm 2.98}$	$66.93_{\pm 1.23}$	$58.31_{\pm 2.40}$	$63.44_{\pm 1.00}$	$51.64_{\pm 3.08}$
MTST	$68.32_{\pm 2.63}$	$61.27_{\pm 4.06}$	$66.74_{\pm 2.54}$	$60.39_{\pm 3.22}$	$65.72_{\pm 2.28}$	$60.48_{\pm 2.18}$
Nonformer	$68.46_{\pm 5.54}$	$65.48_{\pm 5.43}$	$65.58_{\pm 6.62}$	$62.35_{\pm 5.50}$	$62.75_{\pm 5.83}$	$58.62_{\pm 3.77}$
PatchTST	$64.61_{\pm 1.89}$	$57.19_{\pm 2.86}$	$61.93_{\pm 2.05}$	$56.26_{\pm 1.78}$	$60.50_{\pm 1.56}$	$56.01_{\pm 0.73}$
iTransformer	$71.50_{\pm 1.82}$	$67.29_{\pm 2.40}$	$69.07_{\pm 1.67}$	$62.24_{\pm 3.13}$	$66.57_{\pm 2.09}$	$56.85_{\pm 4.47}$
Reformer	$73.11_{\pm 1.39}$	$67.69_{\pm 2.27}$	$68.92_{\pm 1.13}$	$60.49_{\pm 2.73}$	$64.81_{\pm 0.88}$	$52.99_{\pm 2.95}$
Transformer	$69.45_{\pm 4.65}$	$66.90_{\pm 5.01}$	$66.37_{\pm 6.67}$	$63.47_{\pm 5.53}$	$62.99_{\pm 5.92}$	$58.94_{\pm 3.70}$
Medformer	$73.26_{\pm 2.43}$	$71.43_{\pm 1.64}$	$69.60_{\pm 3.82}$	$68.04_{\pm 2.97}$	$66.33_{\pm 4.72}$	$64.87_{\pm 3.93}$
MExFormer	$\underline{77.30_{\pm 1.99}}$	$\underline{73.79_{\pm 2.80}}$	$\underline{74.35_{\pm 2.52}}$	$69.21_{\pm 3.81}$	$\underline{70.61_{\pm 2.46}}$	$62.84_{\pm 4.31}$

TDBRAIN (2-Classes)

Models	10% Channel Corruption		20% Channel Corruption		30% Channel Corruption	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Autoformer	$81.12_{\pm 4.69}$	$81.03_{\pm 4.80}$	$76.25_{\pm 3.44}$	$76.05_{\pm 3.43}$	$70.52_{\pm 3.02}$	$70.11_{\pm 2.82}$
Crossformer	$76.25_{\pm 2.59}$	$76.00_{\pm 2.51}$	$72.12_{\pm 2.14}$	$71.12_{\pm 2.39}$	$65.88_{\pm 2.54}$	$62.86_{\pm 3.57}$
FEDformer	$71.08_{\pm 0.78}$	$70.87_{\pm 0.74}$	$68.00_{\pm 1.63}$	$67.54_{\pm 1.72}$	$65.21_{\pm 1.97}$	$64.41_{\pm 2.17}$
Informer	$85.37_{\pm 1.22}$	$85.19_{\pm 1.32}$	$81.06_{\pm 2.05}$	$80.53_{\pm 2.32}$	$74.79_{\pm 2.45}$	$73.39_{\pm 2.95}$
MTST	$74.29_{\pm 3.38}$	$74.20_{\pm 3.36}$	$71.46_{\pm 3.69}$	$71.22_{\pm 3.85}$	$68.23_{\pm 3.48}$	$67.68_{\pm 3.99}$
Nonformer	$84.17_{\pm 0.88}$	$83.97_{\pm 0.93}$	$80.58_{\pm 1.23}$	$80.12_{\pm 1.37}$	$74.48_{\pm 2.03}$	$73.25_{\pm 2.62}$
PatchTST	$74.46_{\pm 1.87}$	$74.29_{\pm 1.77}$	$71.27_{\pm 2.49}$	$70.69_{\pm 2.77}$	$67.81_{\pm 2.87}$	$66.42_{\pm 4.08}$
iTransformer	$71.71_{\pm 0.87}$	$71.61_{\pm 0.85}$	$71.12_{\pm 0.46}$	$70.84_{\pm 0.50}$	$69.71_{\pm 1.29}$	$68.93_{\pm 1.69}$
Reformer	$87.04_{\pm 1.57}$	$86.95_{\pm 1.61}$	$83.71_{\pm 1.55}$	$83.45_{\pm 1.67}$	$78.35_{\pm 1.61}$	$77.62_{\pm 1.91}$
Transformer	$83.62_{\pm 1.46}$	$83.36_{\pm 1.56}$	$79.10_{\pm 1.01}$	$78.45_{\pm 1.12}$	$73.48_{\pm 1.11}$	$72.00_{\pm 1.42}$
Medformer	$85.60_{\pm 1.27}$	$85.60_{\pm 1.27}$	$83.04_{\pm 0.93}$	$83.02_{\pm 0.94}$	$80.15_{\pm 1.14}$	$80.11_{\pm 1.18}$
MExFormer	$\underline{88.90_{\pm 1.32}}$	$\underline{88.87_{\pm 1.34}}$	$\underline{88.27_{\pm 1.31}}$	$\underline{88.25_{\pm 1.32}}$	$\underline{86.60_{\pm 1.39}}$	$\underline{86.59_{\pm 1.38}}$

attention (iTransformer); representing values of each timestamp as a single token, as incorporated by the Transformer model in our expirements; and performing early cross-channel mixing through patches at different granularities, as in Medformer. Comparing intermediate representations against this spectrum isolates the effect of our staged fusion hierarchy from each alternative fusion timing/granularity, letting us attribute differences in layer-wise representation quality to the proposed mechanism itself (instead of efficiency derived inductive bias) and pinpoint at which depth each design's representations diverge from ours or converge.

Table 4 summarizes the results. The progressive improvement across layers is consistent with the hierarchical design of MExFormer, in which channel interactions are first established at localized temporal positions and are subsequently integrated across temporal positions and temporal resolutions. Taken together, these results indicate that the proposed encoder does not only produce competitive final predictions, but also develops increasingly discriminative intermediate representations as depth increases.

5 Discussion and Conclusion

Taken together, the experiments indicate that the stage at which channel information is integrated affects not only final classification performance, but also robustness and the quality of intermediate representations. MExFormer was designed to preserve channel-specific, temporally localized representations until explicit inter-channel self-attention is applied, before progressively integrating information across temporal positions and resolutions. This design yields the strongest overall

Table 4: Results of the post-hoc weak-classifier evaluation of intermediate-layer representations. The training, validation, and test sets are distributed based on subjects according to a predetermined ratio/IDs. Accuracy and F1 score are reported for representations extracted from encoder layers 1, 3, and 6.

APAVA (2-Classes)

Models	Layer 1		Layer 3		Layer 6	
	Accuracy	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
MExFormer	$78.71_{\pm 1.62}$	$76.44_{\pm 2.12}$	$79.50_{\pm 1.53}$	$77.03_{\pm 2.21}$	$79.59_{\pm 1.95}$	$77.05_{\pm 2.69}$
Medformer	$77.48_{\pm 2.41}$	$76.04_{\pm 2.52}$	$77.06_{\pm 1.96}$	$75.74_{\pm 1.96}$	$77.33_{\pm 1.04}$	$75.48_{\pm 1.06}$
iTransformer	$67.77_{\pm 1.35}$	$65.82_{\pm 1.32}$	$72.13_{\pm 2.10}$	$70.49_{\pm 2.11}$	$72.75_{\pm 2.61}$	$71.20_{\pm 2.76}$
Transformer	$64.37_{\pm 4.09}$	$63.18_{\pm 4.84}$	$67.21_{\pm 3.26}$	$65.95_{\pm 3.69}$	$68.69_{\pm 2.85}$	$67.31_{\pm 3.12}$

TDBRAIN (2-Classes)

Models	Layer 1		Layer 3		Layer 6	
	Accuracy	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
MExFormer	$80.83_{\pm 6.56}$	$80.79_{\pm 6.62}$	$85.75_{\pm 5.27}$	$85.73_{\pm 5.28}$	$89.40_{\pm 1.11}$	$89.38_{\pm 1.10}$
Medformer	$81.40_{\pm 1.06}$	$81.39_{\pm 1.06}$	$85.08_{\pm 0.82}$	$85.08_{\pm 0.82}$	$87.56_{\pm 1.22}$	$87.56_{\pm 1.22}$
iTransformer	$61.63_{\pm 1.40}$	$61.54_{\pm 1.39}$	$68.50_{\pm 0.39}$	$68.30_{\pm 0.41}$	$71.10_{\pm 1.11}$	$71.01_{\pm 1.10}$
Transformer	$81.04_{\pm 1.14}$	$80.86_{\pm 1.17}$	$85.19_{\pm 1.42}$	$85.08_{\pm 1.44}$	$88.21_{\pm 1.10}$	$88.16_{\pm 1.11}$

performance on APAVA, while remaining competitive on the clean TDBRAIN task and showing substantially clearer advantages under channel exclusion and at deeper representation levels. These results suggest that the ordering of channel, temporal, and multi-resolution integration is an important design consideration for multichannel EEG representation learning.

The stronger relative gains of MExFormer on APAVA may be consistent with the electrophysiological profile of Alzheimer's disease (AD), where disrupted large-scale functional connectivity is a repeatedly reported EEG finding. Systematic reviews have identified reduced functional connectivity and coherence, particularly in the alpha band, as common characteristics of AD and mild cognitive impairment, supporting the view that altered inter-regional synchronization is an important component of the disease phenotype (Paitel et al., 2025; Fischer et al., 2023). Since MExFormer preserves electrode-specific representations and explicitly models their interactions before temporal aggregation, this inductive bias may be especially useful when discriminative information is expressed through altered coordination across spatially distributed channels.

In Parkinson's disease (PD), by contrast, the most consistently reported scalp-EEG abnormalities are spectral slowing, reduced dominant frequency, increased theta activity, and reduced faster-frequency activity, although altered functional connectivity has also been reported (Geraedts et al., 2018; Shirahige et al., 2020). This may partly explain why the advantage of MExFormer is less pronounced on the clean TDBRAIN classification task, while becoming more evident under channel corruption and in the quality of deeper intermediate representations. Rather than implying that cross-channel interactions are unimportant in PD, these results suggest that the benefit of explicit channel modeling may depend on the dominant electrophysiological characteristics of the target condition.

Overall, the results support preserving channels as explicit representation units and resolving their interactions before temporal aggregation as a promising inductive bias for multichannel EEG. More broadly, they highlight the importance of **when** channel information is integrated, rather than only **whether** cross-channel information is modeled.

The present study is limited to two binary neurological classification tasks, and further evidence is needed to determine how broadly the proposed representation strategy generalizes across EEG paradigms, recording configurations, and clinical conditions. Future work should therefore investigate additional EEG datasets, directly isolate the contribution of channel-first interaction through controlled architectural ablations, and examine whether the learned cross-channel representations correspond to established electrophysiological relationships. Extending the approach to other multichannel physiological signals may also help determine whether the same inductive bias is useful beyond EEG.

AI USE STATEMENT

Generative AI tools were used for language polishing and manuscript review. The authors reviewed and approved the final content.

REPRODUCIBILITY STATEMENT

The source code and implementation details will be made publicly available upon publication.

REFERENCES

Yara Badr, Usman Tariq, Fares Al-Shargie, Fabio Babiloni, Fadwa Al Mughairbi, and Hasan Al-Nashash. A review on evaluating mental stress by deep learning using eeg signals. *Neural Computing and Applications*, pp. 1–26, 2024.

Vincent Bazinet, Justine Y Hansen, and Bratislav Misic. Towards a biologically annotated brain connectome. *Nature reviews neuroscience*, 24(12):747–760, 2023.

Peng Chen, Yingying ZHANG, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lJkOCMP2aW.

J Escudero, Daniel Abásolo, Roberto Hornero, Pedro Espino, and Miguel López. Analysis of electroencephalograms in alzheimer's disease patients with multiscale entropy. *Physiological measurement*, 27(11):1091, 2006.

Michael Hén Forbord Fischer, Ivan Chrilles Zibrandtsen, Peter Høgh, and Christian Sandøe Musaeus. Systematic review of EEG coherence in Alzheimer's disease. *Journal of Alzheimer's Disease*, 91(4):1261–1272, 2023.

Victor J. Geraedts, Lennard I. Boon, Johan Marinus, Alida A. Gouw, Jacobus J. van Hilten, Cornelis J. Stam, Martijn R. Tannemaat, and Maria Fiorella Contarino. Clinical correlates of quantitative EEG in Parkinson disease: A systematic review. *Neurology*, 91(19):871–883, 2018.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2019.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35:9881–9893, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *International Conference on Learning Representations*, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *ICLR*, 2023.

Elizabeth R. Paitel, Christian B. D. Otteman, Mary C. Polking, Henry J. Licht, and Kristy A. Nielson. Functional and effective EEG connectivity patterns in Alzheimer's disease and mild cognitive impairment: a systematic review. *Frontiers in Aging Neuroscience*, 17, 2025.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019. doi: 10.1088/1741-2552/ab260c.

Lívia Shirahige, Marina Berenguer-Rocha, Sarah Mendonça, Sérgio Rocha, Marcelo Cairrão Rodrigues, and Kátia Monte-Silva. Quantitative electroencephalography characteristics for Parkinson's disease: A systematic review. *Journal of Parkinson's Disease*, 10(2):455–470, 2020.

Hanneke van Dijk, Guido van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (tdbrain) database. *Scientific data*, 9(1):333, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *International Conference on Artificial Intelligence and Statistics*, pp. 4222–4230. PMLR, 2024.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pp. 11106–11115. AAAI Press, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML 2022)*, 2022.