Adapting Neural Audio Codecs to EEG

Ard Kastrati
ETH Zurich
akastrati@ethz.ch

Luca Lanzendörfer
ETH Zurich
lanzendoerfer@ethz.ch

Riccardo Rigoni
ETH Zurich
rrigoni@ethz.ch

John Staib Matilla
ETH Zurich
sjohn@ethz.ch

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

Abstract

EEG and audio are inherently distinct modalities, differing in sampling rate, channel structure, and scale. Yet, we show that pretrained neural audio codecs can serve as effective starting points for EEG compression, provided that the data are preprocessed to be suitable to the codec's input constraints. Using DAC, a state-of-the-art neural audio codec as our base, we demonstrate that raw EEG can be mapped into the codec's stride-based framing, enabling direct reuse of the audio-pretrained encoder-decoder. Even without modification, this setup yields stable EEG reconstructions, and fine-tuning on EEG data further improves fidelity and generalization compared to training from scratch. We systematically explore compression-quality trade-offs by varying residual codebook depth, codebook (vocabulary) size, and input sampling rate. To capture spatial dependencies across electrodes, we propose DAC-MC, a multi-channel extension with attention-based cross-channel aggregation and channel-specific decoding, while retaining the audio-pretrained initialization. Evaluations on the TUH Abnormal and Epilepsy datasets show that the adapted codecs preserve clinically relevant information, as reflected in spectrogram-based reconstruction loss and downstream classification accuracy.

1 Introduction and Related Work

Electroencephalography (EEG) plays a central role in clinical neurology and neuroscience, enabling the non-invasive monitoring of brain activity in applications such as epilepsy diagnosis, sleep staging, and cognitive assessment. As machine learning becomes increasingly integrated into healthcare [1, 2], there is growing interest in building large-scale models that can generalize across EEG datasets, subjects, and clinical conditions. Inspired by the success of foundation models in computer vision [3], language [4] and audio [5], early efforts are now being made to pretrain models on EEG at scale as well [6, 7, 8, 9]. A key enabler of such approaches is the ability to represent EEG signals compactly and discretely, making them easier to store, index, and model using architectures originally designed for token sequences. Neural codecs have recently shown promise in this direction for other modalities, particularly audio, where they compress raw signals into discrete tokens while preserving high reconstruction fidelity. This enables sequence models to process continuous signals as token streams, facilitating next-token prediction in time-series data using the same modeling strategies that have proven effective in natural language.

Recent advances in neural compression techniques [10, 11, 12, 13, 14] have revolutionized audio compression by employing residual vector quantization within adversarially trained autoencoders. These methods are usually trained with a large corpora of audio datasets. However, EEG is far less common as a data resource, as its collection is expensive, time-consuming, and subject to strict privacy and ethical constraints. As a result, publicly available EEG datasets are orders of magnitude

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

smaller than those for audio, making it considerably harder to train general-purpose codecs directly on EEG. With limited data, codecs struggle not only to generalize across different recording setups and subject populations, but even to avoid overfitting within a given dataset.

Despite the substantial differences between audio and EEG in sampling rates, channel structure, and signal characteristics, we find that pretrained audio codecs can serve as an unexpectedly strong starting point for EEG compression. In fact, we show that simply feeding EEG signals into an off-the-shelf audio codec already yields surprisingly reasonable reconstructions. Building on this insight, we fine-tune the neural audio codec DAC [11] on EEG recordings and show that this improves reconstruction fidelity and generalization compared to training a codec from scratch. To explore how DAC can be adapted to EEG, we vary key parameters that influence the compression rate, including the number of residual codebooks, the size of the vocabulary, and the internal sampling rate used to present EEG to the codec. In addition, we propose DAC-MC, an extension designed to handle the multi-channel nature of EEG. DAC-MC incorporates attention-based aggregation across channels and channel-specific conditioning in the decoder, enabling compression and reconstruction of multi-channel EEG within a unified framework. Importantly, all variants including the multi-channel extension, are initialized from the pretrained DAC checkpoint trained on 44.1 kHz audio, preserving the benefits of large-scale audio training throughout.

We evaluate our models on two largest EEG labeled datasets: Abnormal (TUAB) and Epilepsy (TUEP) dataset from the TUH EEG Corpus [15]. We use two types of evaluation: (i) reconstruction fidelity, measured using a spectrogram-based loss between the original and reconstructed signals, and (ii) downstream performance, where classification models are applied to reconstructed EEG to assess whether relevant information is preserved. While some performance degradation is observed after reconstruction, the results show that the trained codec retains the essential structure needed for meaningful downstream tasks in the clinical domain.

2 Methods

2.1 From Audio Codec to EEG Codec

We use the Improved RVQGAN (DAC) [11], a residual vector-quantized encoder–decoder. Residual quantization builds a stack of codebooks: shallow codebooks capture coarse structure; deeper codebooks add detail. We refer to [11] for details and focus here on how we can use this model for EEG compression.

Dealing with sampling rate. Our main idea behind adapting an audio codec at 44.1 kHz to EEG at 512 Hz is simple: we directly feed raw EEG segments (clipped to $\pm 200\mu$V and normalized to the [-1, 1] range similar to audio scale) into the pretrained model, even though the *temporal meaning* of each segment differs. In DAC [11], a fixed window determines how many input samples correspond to one output token, i.e., 512 samples per token. While 512 samples corresponds to $\sim$13 ms in 44.1 kHz audio, it spans 1 s at 512 Hz EEG. Nevertheless, we observe that the pretrained model can still produce stable and interpretable outputs, allowing us to treat EEG compression as a repurposing of an off-the-shelf audio codec.

Dealing with multi channels. The simplest way to apply DAC to EEG is to compress each channel independently, using the original audio-pretrained encoder and decoder without modification. We refer to this approach as `DAC-SC` (DAC Single-Channel). While straightforward, it ignores spatial correlations across electrodes and treats each channel as an isolated 1D signal. On the other hand, to exploit cross-channel dependencies and further increase compression, we introduce `DAC-MC`, a multi-channel extension that encodes multiple channels jointly. Each EEG channel is first processed by the `DAC-SC` encoder to produce a latent sequence. These latents are concatenated across channels (dimension: `hidden_dim`×`num_channels`), with zero-padding applied when needed to support variable-length inputs. We then apply self-attention over the concatenated latents to capture both temporal and inter-channel structure, followed by a projection back to the original hidden dimension for quantization and decoding. To enable channel-specific decoding, the decoder is modulated by learned style vectors (one per channel) which apply affine transformations (scale and bias) to the latent space, inspired by StyleGAN [16]. Fixed channel ordering and positional encodings are used to preserve spatial context. Because full attention across many channels can be computationally expensive (due to DAC's 1024-dimensional latent space), we group EEG channels into smaller subsets (variable size, at most 5) and process each group independently. We evaluate two grouping strategies: *Random Groups*, which vary per training step to improve generalization, and *Manual Groups*, which reflect anatomical or montage-based spatial structures (see Appendix A for detailed description). Importantly, `DAC-MC` adds lightweight adapters around the pretrained DAC model, so all components are compatible with audio-based initialization (see Figure 1).

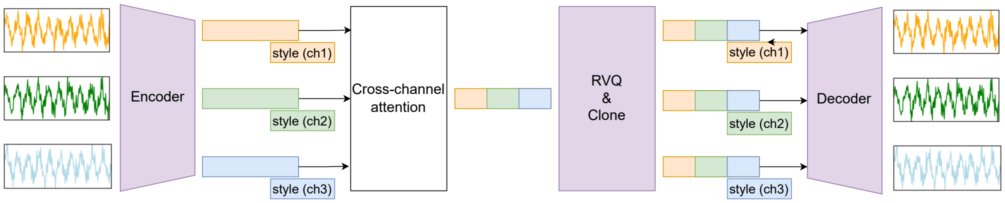


Figure 1: DAC-MC. Purple modules form the pretrained DAC backbone.

Configuration Options. Several modeling decisions remain flexible and influence the trade-off between compression and reconstruction fidelity. These choices let us use the same model in different ways, without modifying its architecture: *(i) Upsampling.* By feeding the model EEG at its native rate (e.g., 256 Hz), a token corresponds to a longer time span, resulting in stronger compression. Upsampling to 512 Hz produces more tokens per second, effectively using the model with a finer temporal resolution at the cost of a higher bitrate: 90 bts (resp. 180 bts) at 512Hz (resp. 256Hz). *(ii) Vocabulary size.* We also can reduce the vocabulary size of each residual codebook (e.g. from 1024 to 512) by merging similar entries in the codebook during finetuning. This limits the model's expressive capacity, but also reduces bitrate. Importantly, this variant still uses the pretrained weights as a starting point and is adapted to EEG through fine-tuning. *(iii) Residual depth.* We can also reduce the number of residual codebooks (e.g., from 9 to 3). This controls how many tokens are combined to represent each input. Here we distignuish two types: *(i) post fine-tune pruning:* fine-tune the model with all 9 codebooks, but during evaluation reconstruct signals using only the first $k < 9$ codebooks, *(ii) pre fine-tune pruning:* train a separate model for each residual depth (from 3 to 9 codebooks), yielding 7 distinct fine-tuned models.

2.2 Datasets and Preprocessing

Datasets. We used EEG recordings from the publically-available dataset: the Temple University Hospital EEG Corpus (TUH EEG) [15]. TUH EEG served as the primary dataset for model fine-tuning. Benchmark evaluation used (labeled) TUAB subset (abnormal EEG detection) and TUEP subset (epilepsy detection).

Preprocessing. Both datasets underwent identical preprocessing: noisy initial segments were removed, empty channels were excluded, signals were resampled to 512 Hz, and a 0.1 Hz high-pass filter was applied. Amplitude values were clipped to ±200μV, normalized to the [-1, 1] range, and finally segmented into non-overlapping 30-second windows.

2.3 Training and Optimization

We considered three approaches for employing `DAC-SC` to EEG signals, and namely: training the model from scratch using only EEG data (`Scratch`), use the pre-trained `DAC-SC` model checkpoint on 44.1 kHz audio (`Pre-trained`), and fine-tune the pre-trained checkpoint on EEG data (`Fine-tuned`). All models were trained using the Adam optimizer with a learning rate of `1e-5` and β parameters set to $(0.8, 0.999)$, using 4 RTX3090 GPUs.

We employed the composite loss function from the original DAC paper [11], including waveform, multi-scale STFT, Mel-spectrogram, adversarial (GAN), commitment, and codebook losses. To better match EEG characteristics, the Mel-spectrogram loss was replaced with a standard spectrogram loss. The waveform loss was initially weighted more heavily to guide adaptation to EEG, then gradually reduced to prevent smoothing effects that diminished peak amplitudes and biased outputs toward the signal mean. Adversarial losses proved unstable on EEG data, often diverging due to domain mismatch. To mitigate this, we adopted a two-phase training strategy: GAN losses were included with reduced weight in the first phase, then removed entirely once divergence was observed.

2.4 Evaluation

Our evaluation framework measures EEG reconstruction quality with a focus on clinical indistinguishability, analogous to perceptual indistinguishability in audio codecs. The goal is to ensure that reconstructed EEG signals preserve key clinical features relevant for neurological diagnosis and classification. We adopt two complementary evaluation strategies: *(i) Reconstruction loss.* We compute spectrogram-based loss between original and reconstructed signals to quantify how well the model preserves spectral structure over time. *(ii) Downstream classification.* We assess whether clinically relevant information is retained in the reconstructed signals by training classifiers on frequency-domain features extracted from reconstructions alone. We use standard models such as Random Forests and Decision Trees, trained with features from the Brainfeatures[1] library. Benchmarks include two widely used clinical classification tasks: (i) anomaly detection using the TUH Abnormal subset (TUAB), and (ii) epilepsy diagnosis using the TUH Epilepsy subset (TUEP).

3 Results

We begin by examining whether audio-pretrained codecs can reconstruct EEG signals. Figure 2 shows a representative example from the test set of single-channel EEG reconstruction using `DAC-SC`, where the reconstructed signal closely follows the morphology of the original. In Appendix B we provide also examples that extend this to multi-channel EEG using `DAC-MC`, illustrating that spatial and temporal structure is preserved across several electrodes. These examples demonstrate that even though the codec was pretrained on audio, it can produce stable and realistic EEG reconstructions when adapted appropriately.

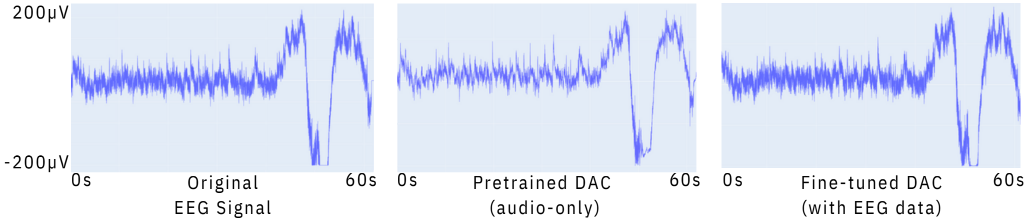


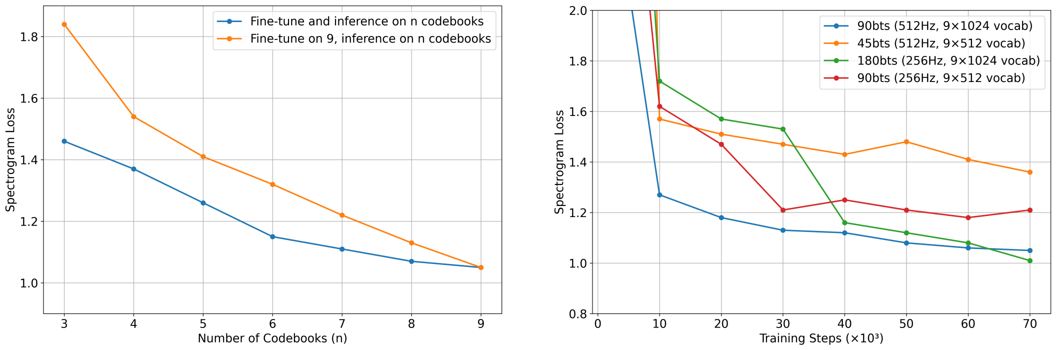
Figure 2: Example reconstruction with audio-pretrained codec and fine-tuned codec with EEG data.

Using pretrained audio codecs is better than training from scratch. Table 1a demonstrates the effectiveness of fine-tuning the DAC model, originally pretrained on audio data, for EEG signal reconstruction. The `Audio-to-EEG Fine-tuned` model achieved significantly lower spectrogram reconstruction loss (approximately 1.05) compared to both the `Scratch` (1.46) and `Audio-Pretrained` models without fine-tuning (2.5). Remarkably, the fine-tuned EEG model nearly matched the original DAC model's performance on audio data (1.09, see [11]), highlighting DAC's robust generalization capabilities.

Impact of codebook size. Figure 3a explores the impact of reducing residual codebooks for higher compression. A reduction from 9 to 6 codebooks resulted in less than a 10% increase in spectrogram loss, representing a favorable trade-off between computational efficiency and reconstruction accuracy. Conversely, reducing codebooks directly from a fully fine-tuned model consistently produced higher spectrogram losses than fine-tuning with fewer codebooks. Notably, reducing from 9 to 3 codebooks doubled the spectrogram loss for the `Audio-to-EEG Fine-tuned` model.

Further hyperparameter experiments, depicted in Figure 3b, examined the combined effects of increasing the internal sampling rate (upsampling from 256 Hz to 512 Hz) and reducing the codebook alphabet size (from 1024 to 512 entries). Upsampling offered marginal improvements in

[1] `https://github.com/TNTLFreiburg/brainfeatures`



(a) Effect of pruning residual codebooks (last-to-first) from the vector quantizer before and after fine-tuning the `Audio-to-EEG` model. *After*: the model is fine-tuned with all nine codebooks and, at inference, only the first n are kept. *Before*: the model is fine-tuned from scratch using only the first n codebooks.

(b) Effect of upsampling and alphabet size reduction on reconstruction fidelity.

Figure 3: Comparison of EEG codec adaptations. (a) Losses for different training strategies. (b) Trade-offs with residual codebooks. (c) Sampling rate and alphabet size effects.

temporal granularity and reconstruction fidelity, which were negligible relative to the substantial increase in computational requirements. Additionally, alphabet size reduction consistently degraded reconstruction quality unless combined with iterative pruning and further fine-tuning.

Model	**Loss**
`Audio-Pretrained`	2.50
`Scratch`	1.46
`Audio-to-EEG`	1.05

(a) Comparison of the spectrogram reconstruction losses from test set showing that fine-tuning an audio-pretrained codec for EEG yields the best performance.

Mode	**Epilepsy**	**Abnormal**
Single-Channel (`DAC-SC`)	80 %	83 %
Single-Channel (`DAC-MC`)	82 %	81 %
Random-Groups (`DAC-MC`)	85 %	78 %
Manual-Groups (`DAC-MC`)	85 %	78 %
Baseline	84 %	82 %

(b) Benchmark accuracy for Epilepsy and Abnormal EEG datasets. `DAC-SC` is a single-channel model. `DAC-MC` is a multi-channel model that can be used either per-channel decoding ("Single-Channel (DAC-MC)") or jointly on multiple channels with groups chosen at random or manually ("Random-Groups"/"Manual-Groups" (DAC-MC)). The Baseline is the best accuracy obtained when training and testing on the original signals.

Downstream Tasks Table 1b reports classification accuracies for epilepsy and abnormal EEG detection using `DAC-SC` and `DAC-MC` under different decoding setups. For epilepsy, grouped `DAC-MC` (Random/Manual Groups) reaches 85% vs. 80% for `DAC-SC`, indicating gains from modeling spatial dependencies across channels. For abnormal EEG, however, grouped `DAC-MC` underperforms at 78%, while running the same multi-channel model in a *single-channel* mode (i.e., decoding one channel at a time) attains 81%—closer to the baseline trained and tested on original signals (82%) and to `DAC-SC` (83%). This suggests that, for abnormality detection, per-channel decoding preserves the relevant features better than cross-channel grouping.

4 Conclusions

We show that pretrained neural audio codecs can be repurposed for EEG compression with minimal modifications. By fine-tuning DAC on EEG data and introducing a multi-channel extension (DAC-MC), we achieve high-fidelity reconstructions that preserve clinically relevant structure across datasets. Our analysis spans compression-quality trade-offs via sampling rate, codebook size, and depth.

References

[1] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[2] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025.

[3] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook, 2023.

[4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[5] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen, 2023.

[6] Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15, 2021.

[7] Di Wu, Siyuan Li, Jie Yang, and Mohamad Sawan. neuro2vec: Masked fourier spectrum prediction for neurophysiological representation learning, 2022.

[8] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-gpt: Towards a foundation model for eeg, 2024.

[9] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci, 2024.

[10] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[11] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

[12] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[13] Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. Snac: Multi-scale neural audio codec. *arXiv preprint arXiv:2410.14411*, 2024.

[14] Luca A Lanzendörfer, Florian Grötschla, Amir Dellali, and Roger Wattenhofer. Neural audio codec for latent music representations. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

[15] Iyad Obeid and Joseph Picone. The temple university hospital eeg corpus: Electrophysiological data for neurological research. *Frontiers in Neuroscience*, 10:196, 2016.

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[17] Asanagi. Electrode locations of international 10–20 system for eeg. Wikimedia Commons, 2010. Public domain.

A Multi-channel grouping

A.1 Random grouping

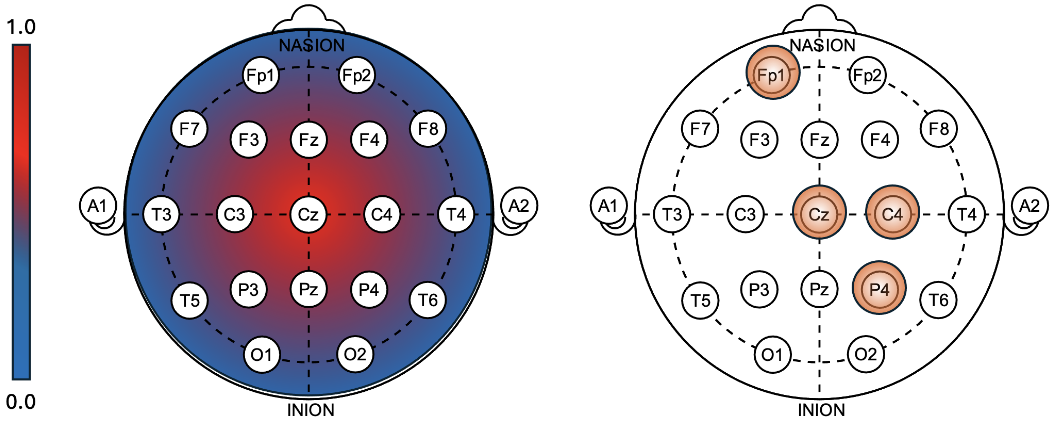


Figure 4: Random subset of 4 channels with pivot Cz on the 10-20 System. On the left, the induced probability distribution from pivot Cz. On the right, the sampled channels.. *Adapted from* [17]. Public domain.

We outline here the *random groups* sampling procedure employed during fine-tuning. For a given EEG recording, we start with all EEG channels $\mathcal{C} = \{1, \ldots, N\}$, each with a 3D scalp location $\mathbf{x}_c \in \mathbb{R}^3$ on the unit sphere ($\|\mathbf{x}_c\|_2 = 1$). We keep a pool $\mathcal{R}$ of channels not yet assigned, and build groups $G_1, G_2, \ldots$ until $\mathcal{R}$ is empty.

1) Pick a group size (1–5). For each group, draw $X \sim \mathrm{Exp}(\lambda = 3)$, round up, and clip to the range $[1, 5]$:

$$s = \min\{5, \max\{1, \lceil X \rceil\}\}.$$

Finally, make sure we don't ask for more channels than remain: $s \leftarrow \min\{s, |\mathcal{R}|\}$. (If you want larger or smaller typical groups, scale X before rounding.)

A.2 Manual grouping

We outline here the *manual groups* crafted for the multi-channel EEG reconstruction for the epilepsy and anomaly detection tasks.

Group	Channels
1	F3, F4, F7, F8
2	FP1, FP2, P3, P4
3	T3, T4, T5, T6
4	C3, C4, CZ
5	O1, O2

Table 2: Manual channel groups for the epilepsy task.

2) Choose a pivot. Pick one pivot channel p uniformly at random from $\mathcal{R}$ and put it in the group ($G \leftarrow \{p\}$).

3) Fill the rest by proximity. For each candidate $c \in \mathcal{R} \setminus G$, compute its distance to the pivot (using chord distance on the unit sphere, which has the property of being monotonic with the geodesic angle):

$$d(p, c) = \|\mathbf{x}_p - \mathbf{x}_c\|_2.$$

Group	Channels
1	EEG 26-REF, EEG 27-REF, EEG 28-REF, EEG 29-REF
2	EEG 30-REF, EEG 31-REF, EEG 32-REF
3	EEG C3-REF, EEG C3P-REF, EEG C4-REF, EEG C4P-REF, EEG CZ-REF
4	EEG FP1-REF, EEG F3-REF, EEG F7-REF, EEG FZ-REF
5	EEG F4-REF, EEG FP2-REF, EEG F8-REF
6	EEG T1-REF, EEG T2-REF, EEG T3-REF, EEG T4-REF, EEG T5-REF
7	EEG O1-REF, EEG O2-REF, EEG OZ-REF, EEG T6-REF
8	EEG P3-REF, EEG P4-REF, EEG PG1-REF, EEG PG2-REF, EEG PZ-REF
9	EEG EKG1-REF, EEG LOC-REF, EEG ROC-REF
10	EEG A1-REF, EEG A2-REF, EEG SP1-REF, EEG SP2-REF

Table 3: Manual channel groups for the TUAB (Abnormal EEG) task.

Turn distances into sampling weights so nearer channels are more likely:

$$w_c = \exp\left(-\frac{d(p,c)}{\tau}\right), \qquad \pi(c \mid p) = \frac{w_c}{\sum_{c' \in \mathcal{R} \setminus G} w_{c'}}.$$

Sample $s-1$ distinct neighbors without replacement from $\mathcal{R} \setminus G$ using $\pi(\cdot \mid p)$, add them to G, remove G from $\mathcal{R}$, and repeat. We set the temperature $\tau = 1$.

B DAC-DC Reconstructions

Below we give example reconstructions of 5 neighboring channels with DAC-MC.

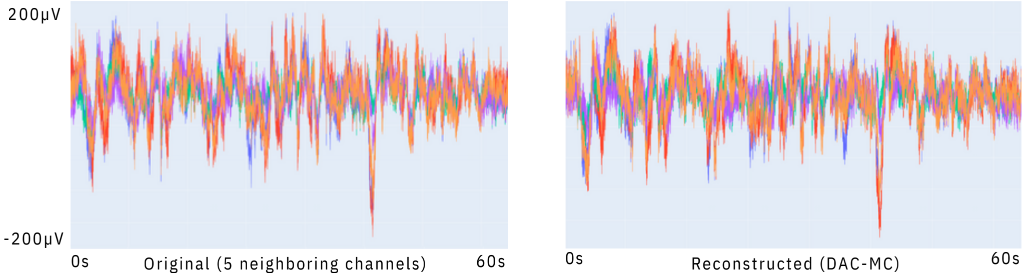


Figure 5: Example reconstruction with fine-tuned codec with multi-channels.