ETH zürich — ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

COPENHAGEN
47TH ANNUAL
International conference of the
IEEE Engineering in Medicine
And Biology Society

FEMBA: Efficient and Scalable EEG Analysis with a Bidirectional Mamba Foundation Model

Integrated Systems Laboratory (ETH Zürich)

Anna Tegon, Thorir Mar Ingolfsson, Xiaying Wang, Luca Benini, Yawei Li

ategon@iis.ee.ethz.ch

PULP Platform
Open Source Hardware, the way it should be!

@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

Why a foundational model for EEG

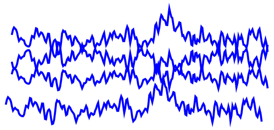
Challenges for EEG

Low signal-to-noise ratio

High subject-specific variability

Task-dependent annotation

$\mathbf{X}$: Raw Data **Cheap**



$\mathbf{Y}$: Labels **Expensive**

- **Supervised training**

$\mathbf{X}$ → **Training**

$\mathbf{Y}$ →

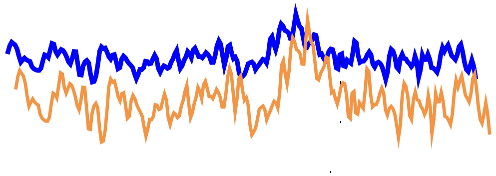
Foundation

- **Self supervised training**

$\mathbf{X}$ → **Training** →

→

Reconstructions



Mamba for an efficient foundational model

Transformer

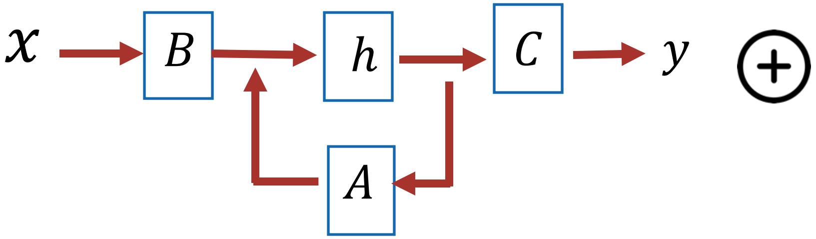
✓ Strong performances: SoTA models

✗ $O(n^2)$ **computation** and **memory** complexity

↓

Mamba

State space models



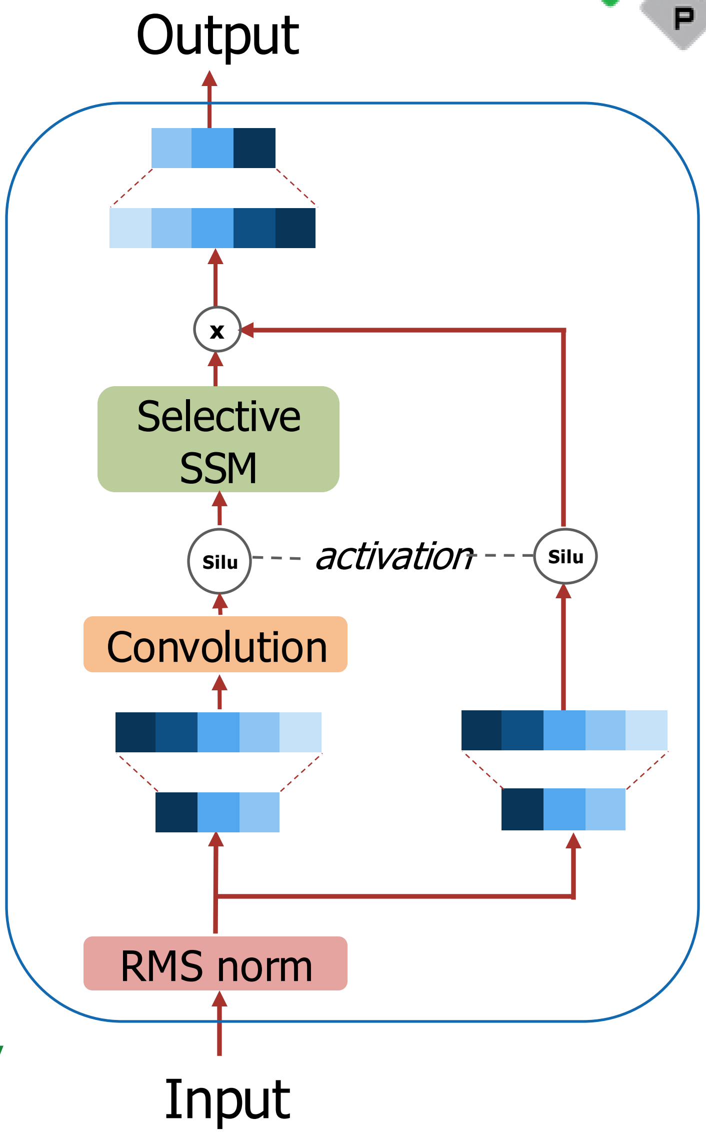
⊕

Selective scan operation

Kernel fusion

Recomputation

⟶ $O(n)$ **computation** and **memory** complexity



Steps

1. Pretraining description:

Pretraining data and setup to reach a broad and comprehensive knowledge

2. Model description:

Definition of the model components:

- Backbone: tokenizer
- Central: encoder
- Head: decoder

3. Finetuning evaluation

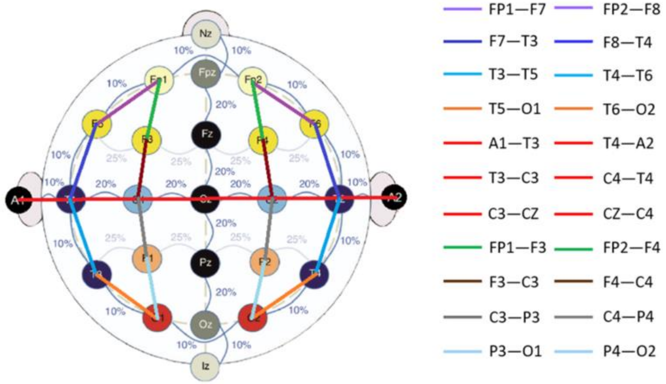
Evaluate transferability in different downstream tasks

Pretraining dataset

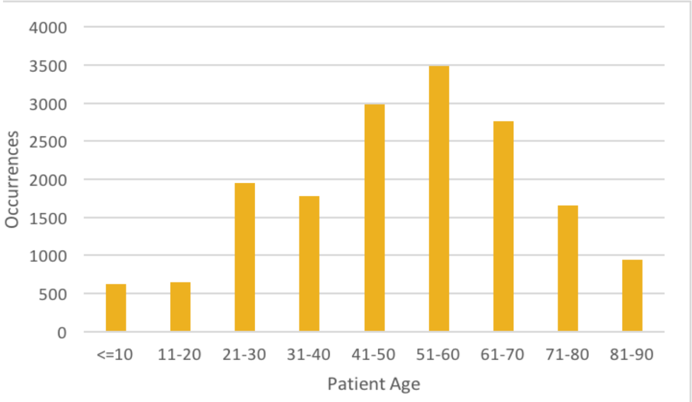
TUEG: *Temple University Hospital EEG:*

- Over **25,000** EEG recordings
- Around **21,000** hours of EEG recordings
- Over **10,000** patients:
    - Range from 1 year to 90 years
    - Average of 51 years
    - 51% female

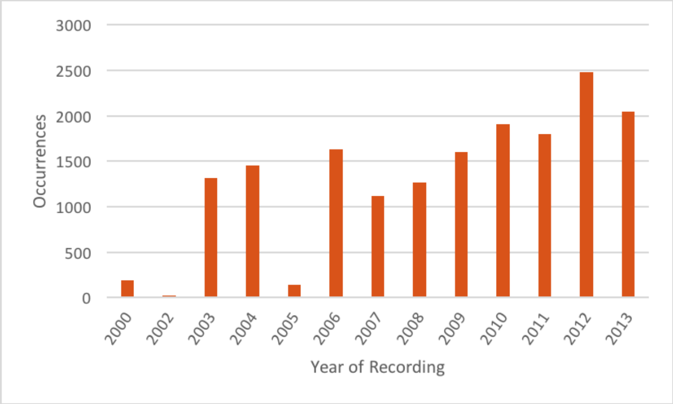
Bipolar TCP montage



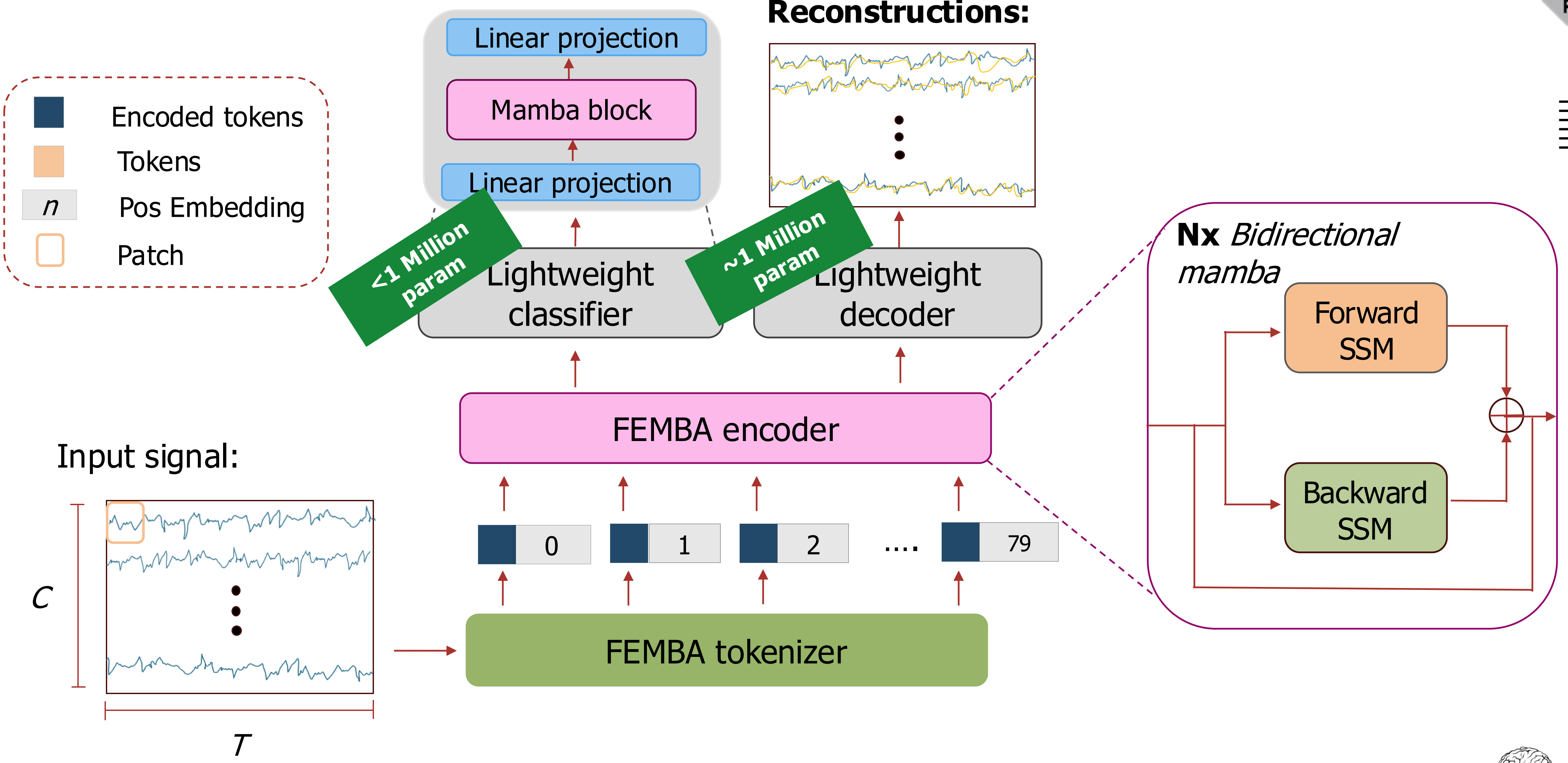
Age distribution



Year of recording



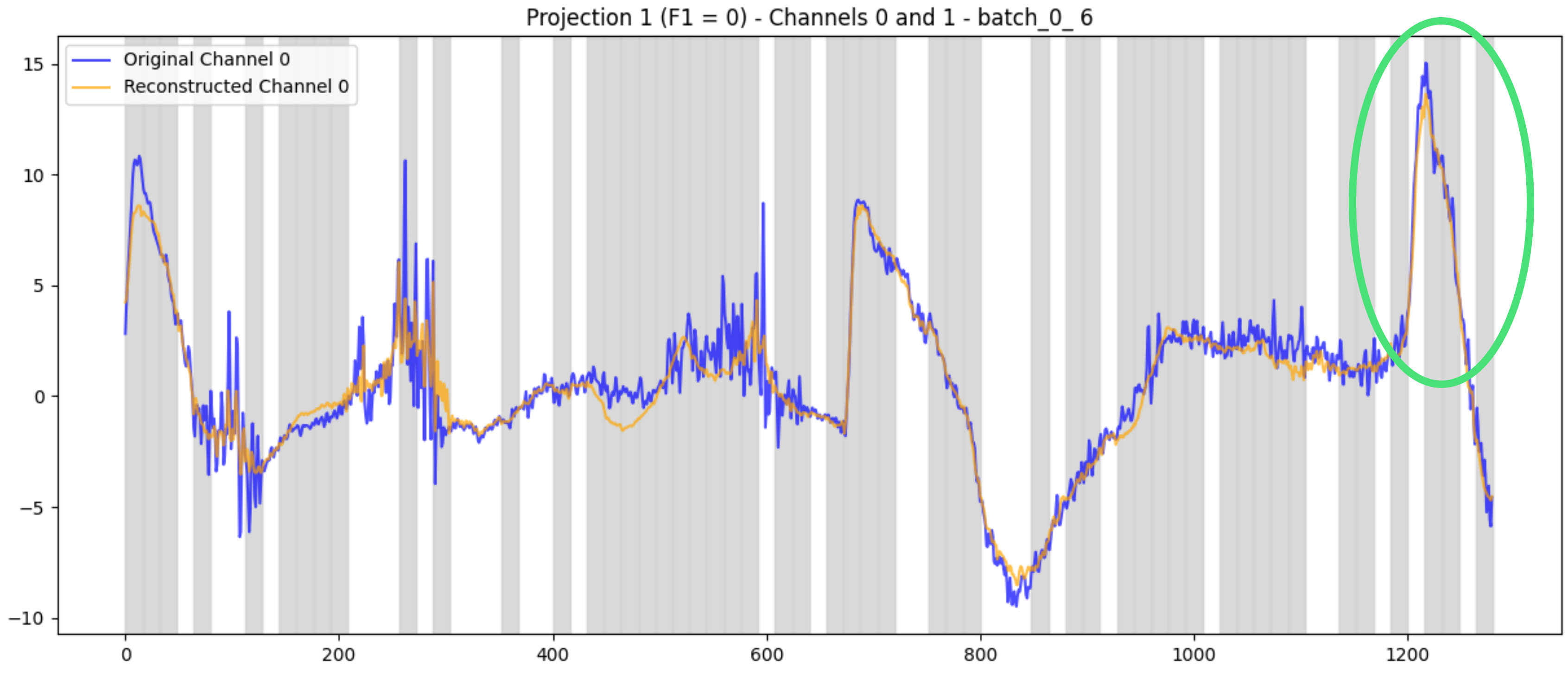
FEMBA: model architecture



Mamba: Robust Signal Reconstruction

Results for a **Base model**:
Encoder: 12 Bidirectional Mamba blocks
Embedded dimension:35
Total number of parameters: 47.7 Million

Impressive signal match!



Benchmarking Foundation Models: The Datasets

Seizure detection

TUAB: *Temple University Hospital Abnormal*

- 2,329 patients
- 2 Classes
- Balanced dataset:

  1.521 Normal
  1.472 Abnormal

Artifact detection

TUAR: *Temple University Hospital Artifact*

- 213 patients
- 14 artifacts → 5 groups
- Unbalanced dataset

Slowing events detection

TUSL: *Temple University Slowing Events*

- 1000 patients
- 4 Classes

SOTA results for 2 out of 3 datasets:

Seizure detection

Previous SOTA model **LaBraM**

Accuracy : 82.52% vs 81.72%

↓0.7% accuracy loss

Artifact detection

Previous SOTA model **EEGFormer**

AUROC : 85.2% vs **91.80%**

6% gain

Slowing events detection

Previous SOTA model **EEGFormer**

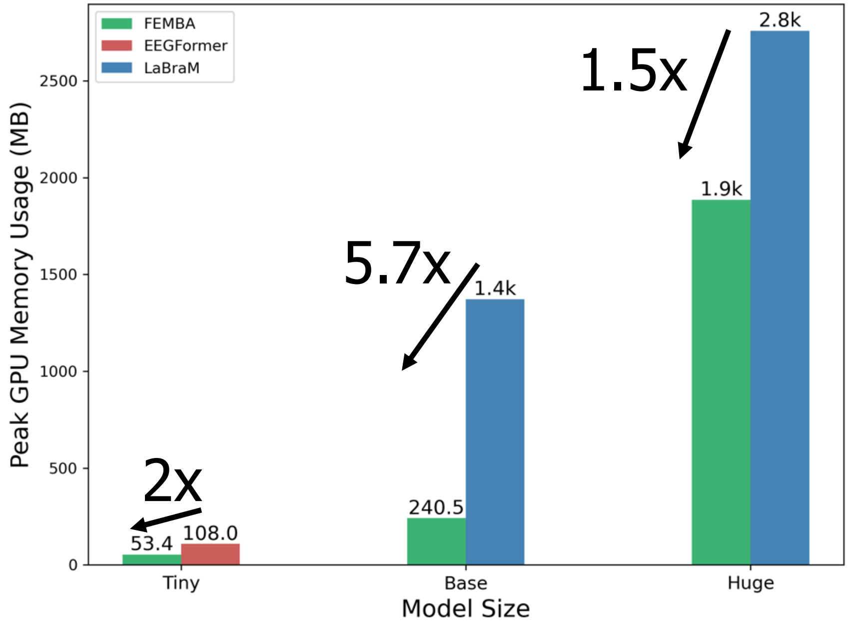
AUROC : 71.3% vs **73.1%**

2% gain

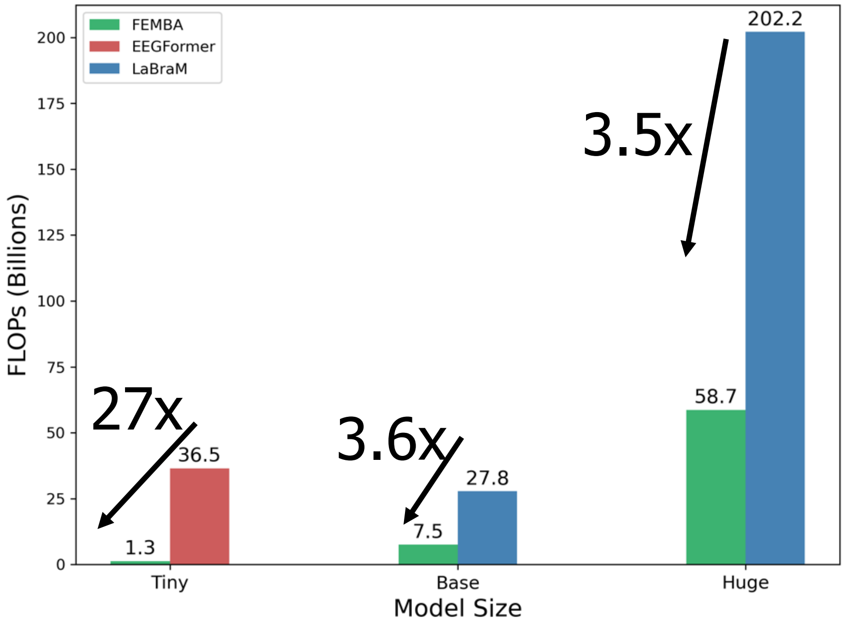
[1] W. Jiang et al., "Large brain model for learning generic representa- tions with tremendous eeg data in bci," in The Twelfth International Conference on Learning Representations, 2024
[2] Y. Chen et al., "Eegformer: Towards transferable and interpretable large-scale eeg foundation model," in AAAI 2024 Spring Symposium on Clinical Foundation Models

Light on memory, **high** on speed!

Peak of GPU memory usage(MB)



FLOPs (Billion)



up to **5.7x less** memory than LaBraM

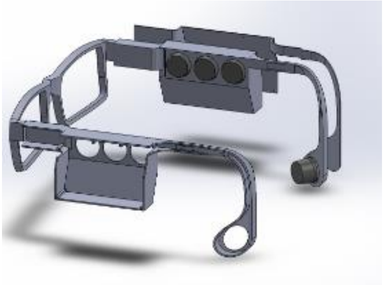
2x less memory than EEGFormer Tiny

~ **3.5x faster** than LaBraM

27x faster than EEGFormer

FEMBA: **high** performance at a **cheaper** price

- **SOTA results for artifact detection**
- **SOTA results for slowing events detection**
- **Close to SOTA for seizure detection**



Up to 27x speedup

Up to 5.7x less memory

FEMBA- Tiny (<8 million parameters) Indicates potential for future use in **battery-constrained wearable devices**