EEG Foundation Model Representations Under Hierarchical Classification and Short Calibration

Eduarda Tessari Pereira[1], Felipe Correa Bitencourt[2], Claudio Salvalaio[2], Mina Jaberi[3], Tiago Henrique Falk[3], Mariana Recamonde-Mendoza[1], Alcyr Alves de Oliveira Jr[2]

[1] Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
[2] Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil
[3] Institut National de la Recherche Scientifique (INRS), Gatineau, QC, Canada

Key finding. Frozen and calibrated on ~58 epochs of the user, foundation-model representations do not beat classical baselines. Adapted on other subjects, they barely change detection but lift lateralization: **fine-tuned REVE reaches κ = 0.445 with 8 electrodes, against 0.283 for CSP.**

Background

Motor-imagery BCIs for rehabilitation need a **short calibration**. EEG **foundation models** promise that pretraining on large unlabeled corpora speeds up calibration when labels are scarce [2], but on motor imagery their performance remains limited [5] and they tend to need adaptation rather than serve as fixed feature extractors [6].

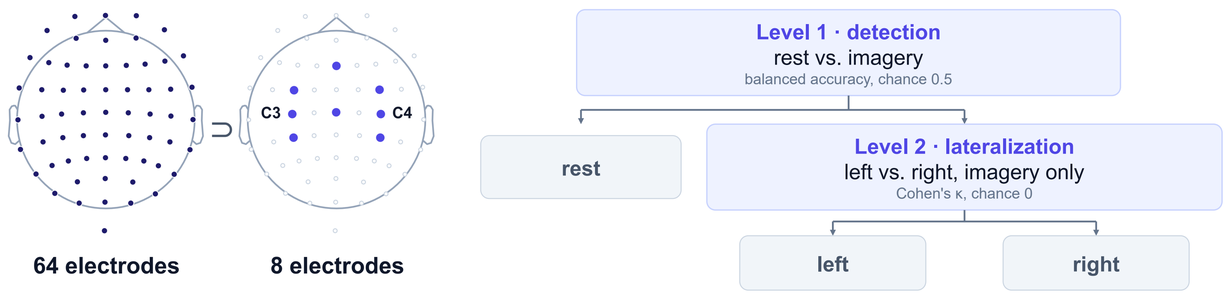
32–64 electrodes is what **128 motor-imagery studies** (2020–2025) predominantly use [1], far from a wearable montage.

Question
Does the pretrained representation help decode motor imagery split into **detection** and **lateralization** under short calibration, and from which adaptation regime onward does that help become measurable?

Methods

1 Data and hierarchical task

PhysioNet EEG Motor Movement/Imagery [7]: 109 subjects, 64 electrodes, 160 Hz. Subjects imagine opening and closing the left or right fist and rest in between. Each 0.5–2.5 s epoch is decoded at two levels, on all 64 electrodes and on **8 near C3/C4 nested in them**, so the comparison measures density, not region.



2 Models

Three encoders that accept **any montage by construction** are compared with two classical baselines, all read out through the same heads.

- **Band power** — log power, 5 bands, 4–40 Hz [9]
- **CSP** — 8–30 Hz, 3 filter pairs per level [8]
- **CBraMod** — criss-cross transformer, ICLR 2025 [3]
- **LUNA** — topology-agnostic, NeurIPS 2025 [4]
- **REVE** — pretrained on 25,000 subjects [2]

3 Adaptation regimes and evaluation

The four regimes differ in how much the task may change the model. In R1–R3 the encoder stays **frozen** and only the data seen by the head grows, from the subject's own epochs to those of 98 other subjects; R4 **fine-tunes** it on that pool, with each set of authors' recipe.

R1 NCM on target
pool size, for comparison
~58 epochs of the subject
head: **class mean (NCM)**
encoder frozen

R2 Linear on target
pool size, for comparison
~58 epochs of the subject
head: **logistic regression**
encoder frozen

R3 Linear on pool
~9,000 epochs of 98 other subjects
head: **logistic regression**
encoder frozen

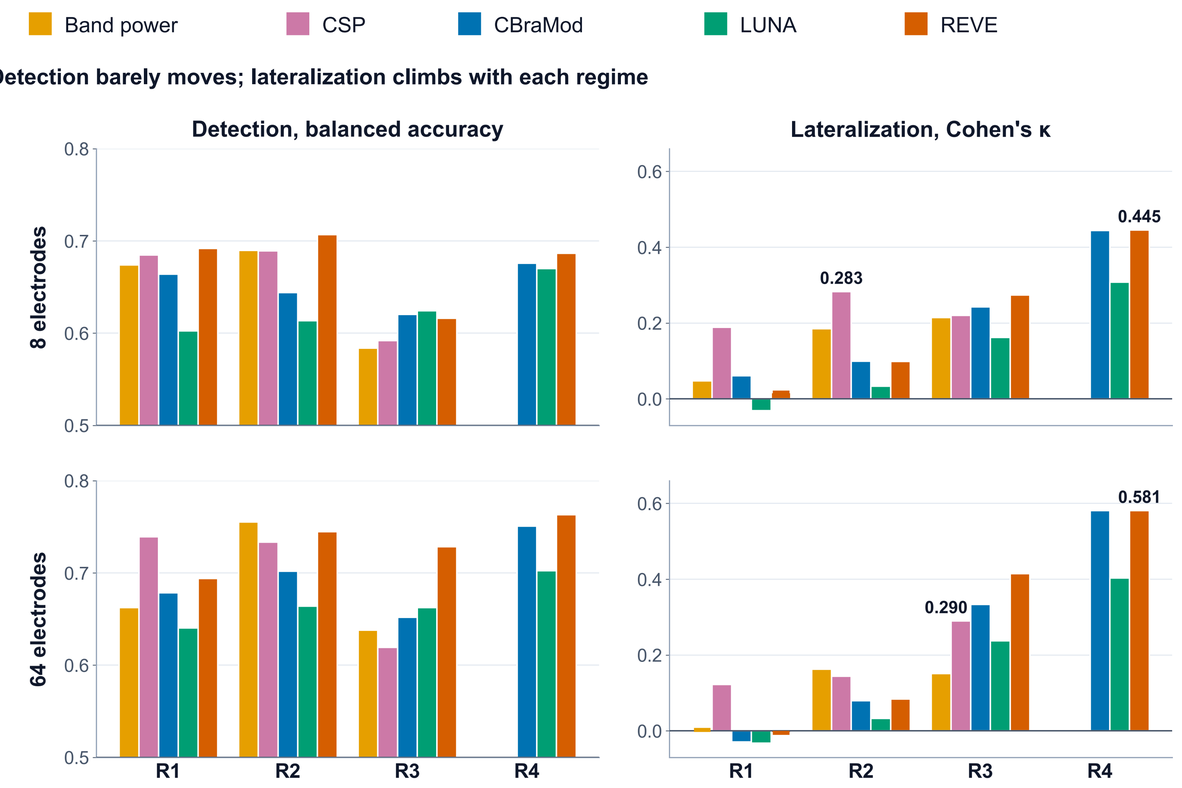
R4 Fine-tuned on pool
~9,000 epochs of 98 other subjects
head: **logistic regression**
encoder fine-tuned, 4–69 M weights

Evaluation protocol

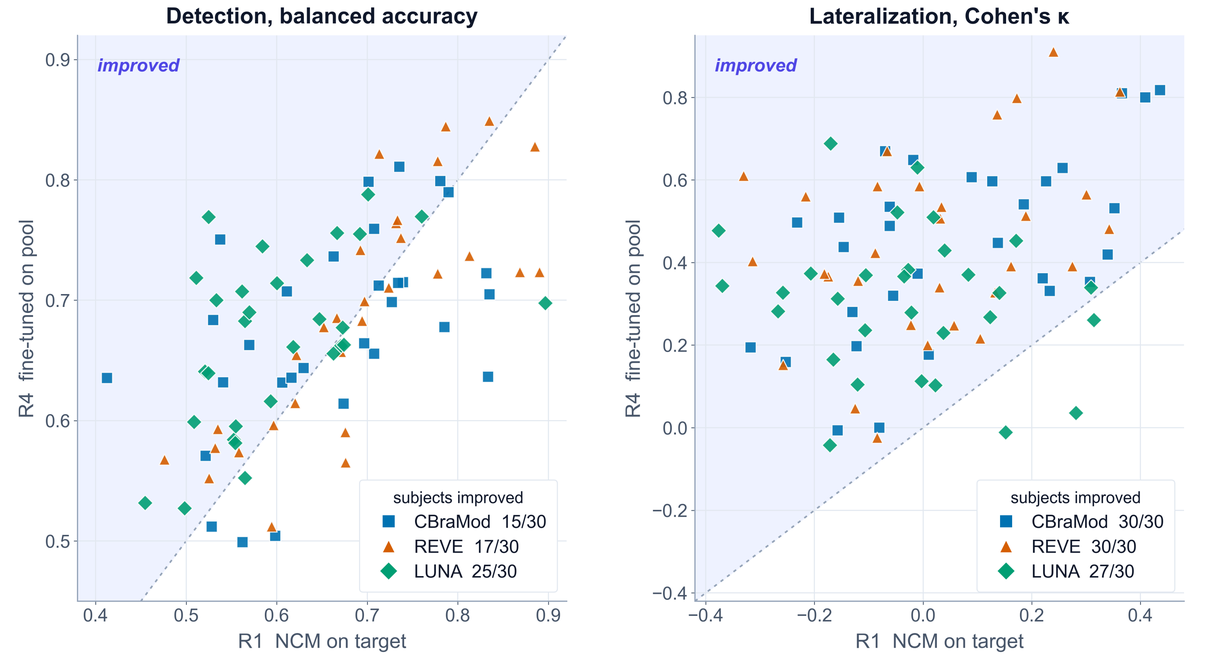
- 30 targets in 6 groups of 5
- each group's pool excludes its own targets
- 3-class balanced accuracy (chance 0.333)
- paired tests (n = 30, Holm) over 3 partitions

Results

With the encoder frozen and the head calibrated on the target alone (R1–R2), no encoder separates from band power. Separation starts when the head is trained on the 98 other subjects (R3) and consolidates with fine-tuning (R4). The two levels respond differently: **detection barely moves, lateralization climbs at every regime**, consistently across subjects.



(A) Mean over the 30 targets and the three partitions; bars start at chance (0.5 for detection, 0 for κ). Band power and CSP have no encoder to fine-tune, hence no R4.



(B) Each marker is one target subject, mean of the three partitions; the shaded area above the diagonal means the subject improved from R1 to R4.

0.445 vs 0.283
lateralization κ with 8 electrodes: fine-tuned REVE vs CSP, the best baseline

0.581 vs 0.290
the same comparison with 64 electrodes, where the gap is larger

+0.062 / +0.063
3-class balanced accuracy of fine-tuned REVE / CBraMod over band power in its best regime (8 electrodes)

Conclusions

Frozen, the pretrained representation is not enough. Calibrated on the subject's ~58 epochs, it matches log band power and falls below CSP in lateralization with 8 electrodes.

Adaptation converts into lateralization, not detection, consistently across subjects.

Pretraining does not remove the need for the user's data in a clinical setting with short calibration, in line with [10].

Limitations: one database and paradigm; two montages that fix the extremes, not the middle; offline evaluation; errors correlated within target groups (effective n < 30); R4 does not separate pretraining from end-to-end training on the pool.

Acknowledgements

To the authors of CBraMod, LUNA and REVE for the public weights, and to PhysioNet for the data.

NERV UFCSPA
MuSAE Lab INRS

References, data and code
scan or visit **eramia2026.cienciaberta.com**

Contact
Eduarda Tessari Pereira
eduarda.tpereira@inf.ufrgs.br

ERAMIA-RS II ESCOLA REGIONAL DE APRENDIZADO DE MÁQUINA E INTELIGÊNCIA ARTIFICIAL DA REGIÃO SUL 2026