SUBJECT IDENTITY AS A SHORTCUT: RETHINKING SELF-SUPERVISED PRETRAINING FOR EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography foundation models (EEG-FMs) promise to learn transferable neural representations from unlabeled corpora, yet whether self-supervised pretraining achieves this remains unclear. We audit frozen representations of five released EEG-FMs across five canonical brain–computer interface (BCI) benchmarks using linear probing, variance decomposition, *t*-SNE, and representational similarity analysis. A consistent pattern emerges: pretraining strengthens subject identity information without comparable gains in task-related information, revealing *subject identity shortcuts* as a shared failure mode of prevailing EEG self-supervised paradigms. To identify the underlying mechanisms, we pretrain MAE-style and MoCo-style EEG-FMs on a multi-source corpus. Masked reconstruction preferentially exploits subject-related spatial structure over temporal context, whereas contrastive learning leverages subject differences for positive-pair consistency and separation from predominantly cross-subject negatives. Euclidean-aligned targets fail to suppress this shortcut in MAE, whereas same-subject negatives reduce it in MoCo without reliable task gains. These results motivate EEG-specific objectives that explicitly learn transferable task-related representations.

1 INTRODUCTION

Electroencephalography foundation models (EEG-FMs) aim to learn transferable, general-purpose representations through self-supervised pretraining on large-scale, heterogeneous, and unlabeled EEG corpora Kuruppu et al. (2026). Unlike EEG decoders trained from scratch for supervised, task-specific decoding on individual datasets Craik et al. (2019), EEG-FMs leverage pretraining followed by downstream adaptation, potentially reducing the demand for task-specific labels and improving robustness across subjects, electrode montages, and task settings. In recent years, EEG-FMs have proliferated, with many studies reporting competitive performance across diverse evaluation protocols and downstream datasets Yao et al. (2025).

Despite rapid progress, the practical value of EEG-FMs remains contested. Recent benchmarks suggest that prior estimates may have been inflated, as EEG-FMs often fail to consistently outperform compact supervised decoders Kastrati et al. (2025). Moreover, no clear scaling law has yet emerged: neither greater model capacity nor larger pretraining corpora reliably translate into better downstream results Xiong et al. (2026). Performance also depends substantially on evaluation protocols. Although full fine-tuning can yield moderately competitive results, strict linear probing often yields only marginal gains over random initialization, or even no reliable advantage over models pretrained on non-EEG modalities Li et al. (2026); Kontras et al. (2026). Feature-aggregation strategies and task-specific prediction heads can also affect downstream evaluation Yang et al. (2026), potentially confounding the assessment of pretrained representations.

These controversies appear to suggest that, despite substantial resources devoted to self-supervised pretraining, its benefits may lie more in providing a favorable parameter initialization for downstream fine-tuning than in yielding frozen representations that transfer directly to downstream tasks Wang et al. (2026). Relatedly, recent analyses identify an *identity trap* Lin et al. (2026) in EEG-FMs: task decoding may reflect subject identity features correlated with class labels, rather than

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.

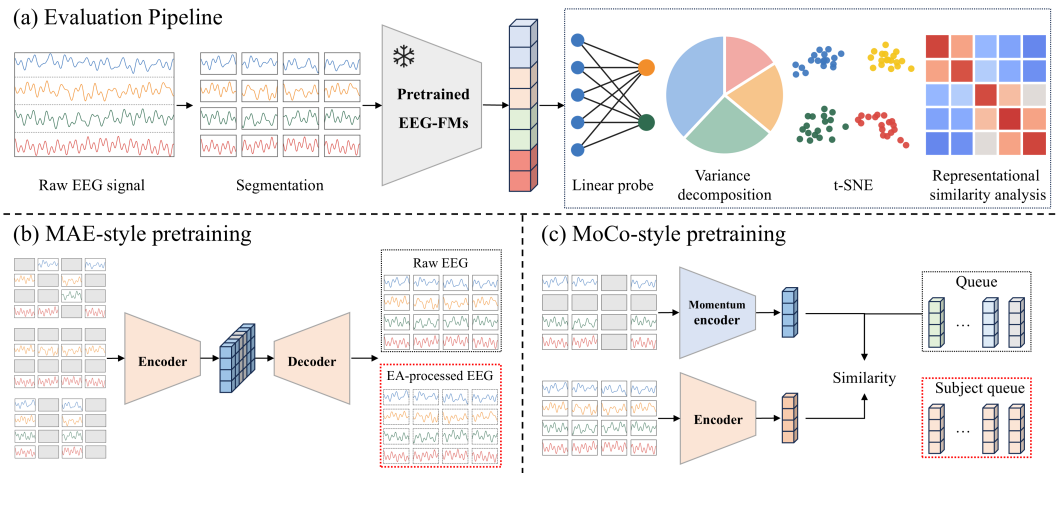


Figure 1: Overview of the study. (a) Segmented EEG is encoded by frozen EEG-FMs to obtain trial-level representations and evaluated using linear probing, variance decomposition, *t*-SNE, and RSA. (b) Schematic of MAE-style pretraining. (c) Schematic of MoCo-style pretraining.

transferable task-related information, potentially due to an aperiodic and low-frequency spectral bias induced during pretraining Kommineni et al. (2026).

Contributions: We evaluate representative EEG-FMs across canonical BCI tasks under population-decoding and strict linear-probing protocols, comparing frozen pretrained models with their frozen randomly initialized counterparts and combining variance decomposition, *t*-SNE Van der Maaten & Hinton (2008), and representational similarity analysis (RSA) Kriegeskorte et al. (2008) to characterize the encoding tendencies in frozen representations. We then investigate the mechanisms underlying these tendencies using masked autoencoder (MAE)-style He et al. (2022) and momentum contrast (MoCo)-style He et al. (2020) self-supervised models, examining how their optimization objectives may favor *subject identity shortcuts*. Finally, we attempt to constrain these shortcuts under both paradigms and examine whether doing so promotes more transferable task-related representations. An overview of the study is provided in Figure 1. Our key findings are as follows:

1. EEG-FMs preferentially encode subject identity. Across models, pretraining enhances the decodability and variance contribution of subject identity, without a comparable increase in task-related information.

2. Shortcut in MAE-style pretraining: spatiotemporal reconstruction asymmetry. Task and subject decodability are positively correlated across MAE-style models, reflecting an identity trap; structured masking shows that reconstruction mainly relies on subject-related spatial structure.

3. Shortcut in MoCo-style pretraining: individual differences. Contrastive pretraining likewise favors subject identity encoding over task-related encoding: within-subject similarity of positive pairs and cross-subject dissimilarity among negative pairs make subject identity an efficient solution.

4. Interventions on shortcuts do not promote task encoding. Neither MAE with Euclidean-aligned (EA) He & Wu (2019) targets nor MoCo restricted to same-subject negatives yields consistent task gains. The former fails to suppress subject identity, whereas the latter succeeds yet still does not reliably elicit task semantics.

Prevailing self-supervised paradigms favor subject identity over task-related information, yielding *subject identity shortcuts*. This calls for EEG-specific objectives that explicitly promote transferable task semantics rather than assuming they emerge from reconstruction or cross-view consistency.

2 Related Work

Recent EEG-FMs mainly follow two self-supervised paradigms: MAE-style masked modeling and contrastive learning. MAE-style models reconstruct masked continuous EEG signals, as exempli-

fied by CBraMod Wang et al. (2025) and REVE El Ouahidi et al. (2026), whereas LaBraM extends masked modeling to predict discretized neural tokens Jiang et al. (2024b). Contrastive models, such as BIOT Yang et al. (2023) and CoCoT-EEG Mahuas et al. (2026), learn invariant representations across augmented views. EEGPT Wang et al. (2024) further combines masked reconstruction with representation alignment. Despite differing objectives, these methods assume that self-supervised proxy tasks yield transferable representations that retain task-related neural information from large-scale unlabeled EEG corpora. However, recent benchmarks question this assumption, finding that EEG-FMs do not consistently surpass compact supervised decoders and that their outcomes are highly sensitive to evaluation protocols and downstream prediction heads Kastrati et al. (2025); Wang et al. (2026); Yang et al. (2026). A recent representation analysis further suggests that some EEG-FMs preferentially encode subject identity Kommineni et al. (2026). Building on these observations, we audit what information is retained in frozen EEG-FM representations and investigate how masked reconstruction and contrastive objectives may give rise to subject identity shortcuts.

3 METHODS

3.1 EVALUATION OF OFF-THE-SHELF EEG FOUNDATION MODELS

In this paper, we evaluate five representative EEG-FMs: BIOT, CBraMod, EEGPT, LaBraM, and REVE. Detailed descriptions of these models are provided in Appendix A. For all evaluations, we use the publicly available Braindecode[1] (v1.5 stable) Aristimunha et al. (2025) implementations with released Hugging Face checkpoints, verified against the official releases.

3.1.1 DOWNSTREAM DATASETS

Five canonical BCI datasets serve as downstream benchmarks: BCI-IV-2a Brunner et al. (2008), Kim2025 Kim et al. (2025), PhysioNet-P300 Citi et al. (2010), SEED-V Liu et al. (2021), and Tsinghua-RSVP Zheng et al. (2020), covering motor imagery, steady-state visually evoked potential (SSVEP), P300 speller, emotion recognition, and rapid serial visual presentation (RSVP), respectively. Details are provided in Appendix B.

3.1.2 EVALUATION METHODS

All analyses use trial-level representations extracted from frozen pretrained encoders. Following model-specific electrode adaptation, each trial is encoded by the frozen backbone and represented by its `[CLS]` token when available, or by average-pooling output tokens otherwise.

Linear probing (Appendix C.1). Linear probes compare balanced accuracy (BA) for task label and subject identity between each pretrained EEG-FM and its randomly initialized counterpart.

Variance decomposition (Appendix C.2). For each EEG-FM and its randomly initialized counterpart, we decompose the total variation in squared Euclidean distances across the full dataset into task, subject-identity, interaction, and residual components.

***t*-SNE visualization (Appendix C.3).** *t*-SNE is used to visualize frozen representations. For readability, at most 16 task classes and 16 subjects are retained in each visualization. Task- and subject-colored plots share the same coordinates.

Representational similarity analysis (Appendix C.4). Complementing *t*-SNE, RSA aggregates pairwise cosine similarities of representations into task- and subject-level similarity matrices.

3.2 CONSTRUCTION OF SELF-SUPERVISED PRETRAINING MODELS

To examine how MAE-style and MoCo-style pretraining shapes frozen EEG representations, we pretrain EEG-FMs on a multi-source unlabeled corpus and evaluate the resulting representations using the protocols in Section 3.1.2. All experiments are conducted using Python 3.12.13 and PyTorch 2.6.0 with CUDA 12.4 on four NVIDIA A800 GPUs (80 GB).

[1] `https://braindecode.org`

3.2.1 PRETRAINING DATASETS

Pretraining uses an unlabeled corpus comprising 1,583,853 four-second EEG segments from 2,156 subjects, aggregated from HBN Shirazi et al. (2024), M3CV Huang et al. (2022), Stieger21 Stieger et al. (2021), SEED Zheng & Lu (2015), SEED-VII Jiang et al. (2024a), TSU-Benchmark Wang et al. (2016), and XDU-RSVP Zhang et al. (2025); details are provided in Appendix D.

3.2.2 CONSTRUCTION OF MAE- AND MOCO-STYLE MODELS

To examine reconstruction-based self-supervised learning, we train a series of MAE-style EEG-FMs using a shared Vision Transformer (ViT) backbone Dosovitskiy et al. (2021) and a deliberately simple MAE objective He et al. (2022). Rather than varying architecture, capacity, or corpus scale, we vary the masking strategy and ratio: random token masking ($\rho \in [0.1, 0.995]$) versus spatial and temporal masking ($\rho \in [0.5, 0.95]$). Details are provided in Appendix E.1. We further examine whether MAE-style representational properties persist under contrastive pretraining, for which negative-sample selection is central. Among the publicly released EEG-FMs, BIOT is the only contrastive model, but it uses only mini-batch negatives. We therefore train a MoCo-style EEG-FM He et al. (2020), which aligns partially observed query views with full-signal key views using a momentum-updated encoder and a negative queue. Details are provided in Appendix E.2.

3.3 INTERVENTIONS ON SUBJECT IDENTITY SHORTCUTS

To constrain subject identity shortcuts, we modify reconstruction targets in MAE and negative-sample construction in MoCo. For MAE, Euclidean alignment (EA) He & Wu (2019) is applied to reconstruction targets, while the encoder receives the original EEG signal, thereby attenuating subject-specific channel covariance structure in the target. For MoCo, negative keys are drawn only from the queue associated with the query subject, preventing positive–negative separation by cross-subject differences. Details are provided in Appendices F.1 and F.2, respectively.

4 RESULTS

4.1 EEG-FMS PREFERENTIALLY ENCODE SUBJECT IDENTITY

Table 1 and Figure 2 present the quantitative results and representation visualizations on BCI-IV-2a. Results for all remaining datasets are provided in Appendix G for completeness.

Table 1: Linear-probe and variance decomposition of frozen representations on BCI-IV-2a.

Model	**Parameter**	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level		0.250	0.111	–	–	–	–
BIOT	RandomInit	**0.313**	0.733	**0.0045**	**0.4123**	0.0078	0.5754
	Pretrained	0.309	**0.770**	0.0033	0.4000	**0.0089**	**0.5878**
CBraMod	RandomInit	**0.283**	0.272	0.0010	0.0147	**0.0066**	**0.9777**
	Pretrained	0.265	**0.380**	**0.0030**	**0.2873**	0.0058	0.7039
EEGPT	RandomInit	0.285	0.502	0.0012	0.0417	0.0054	**0.9517**
	Pretrained	**0.287**	**0.633**	**0.0035**	**0.2109**	**0.0085**	0.7770
LaBraM	RandomInit	0.276	0.312	**0.0013**	0.0105	**0.0061**	**0.9821**
	Pretrained	**0.279**	**0.599**	0.0012	**0.1149**	0.0050	0.8790
REVE	RandomInit	0.309	0.356	**0.0033**	0.0276	**0.0094**	**0.9597**
	Pretrained	**0.372**	**0.844**	0.0024	**0.0616**	0.0062	0.9298

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively.

Linear probing. Under random initialization and pretraining, subject identity is decodable well above chance from frozen representations, indicating readily distinguishable subject-specific information. Task decoding is more dataset-dependent: although some model–dataset pairs perform

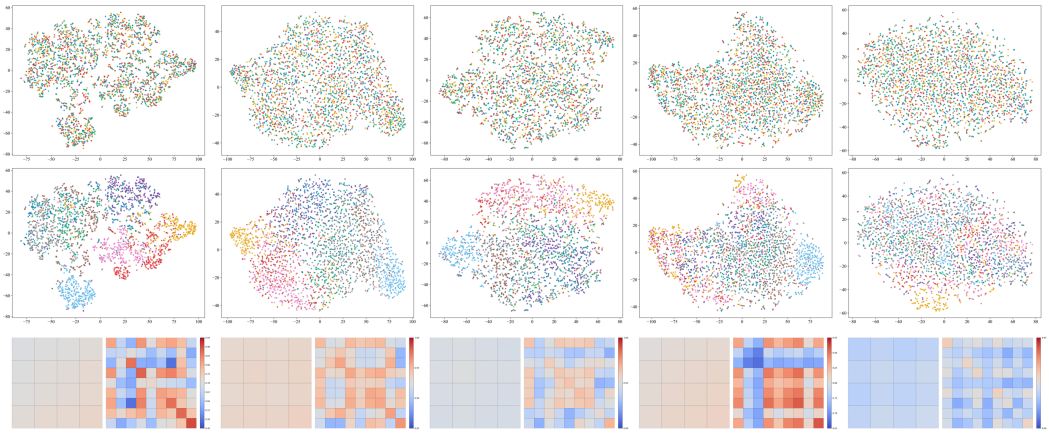


Figure 2: *t*-SNE and RSA of frozen EEG-FM representations on BCI-IV-2a. Columns show BIOT, CBraMod, EEGPT, LaBraM, and REVE from left to right. The top two rows show shared *t*-SNE coordinates colored by task labels and subject identity, respectively; the bottom row shows the corresponding task (left) and subject (right) RSA matrices. Best viewed via zoom-in.

modestly above chance, outcomes vary substantially across models and datasets. Because task and subject cardinalities vary across datasets, absolute BA values are not directly comparable. We therefore assess pretraining-induced changes relative to randomly initialized encoders.

With architecture and linear probe fixed, pretraining improves subject identity decoding in all 25 model–dataset pairs, but task decoding in only 15 pairs, generally marginally, and declines in the remaining 10 pairs. These differences are therefore attributable to pretrained parameters rather than backbone capacity or classifier complexity. Current EEG-FMs thus consistently increase the linear decodability of subject identity without a corresponding, consistent gain in task-related information.

Task and subject BA occasionally induce the same ranking of pretrained models. On BCI-IV-2a and SEED-V, the five EEG-FMs rank identically by task and subject BA, suggesting that these forms of decodability may co-vary across models. However, five models with differing corpus scales, capacities, and objectives are insufficient for robust statistical inference.

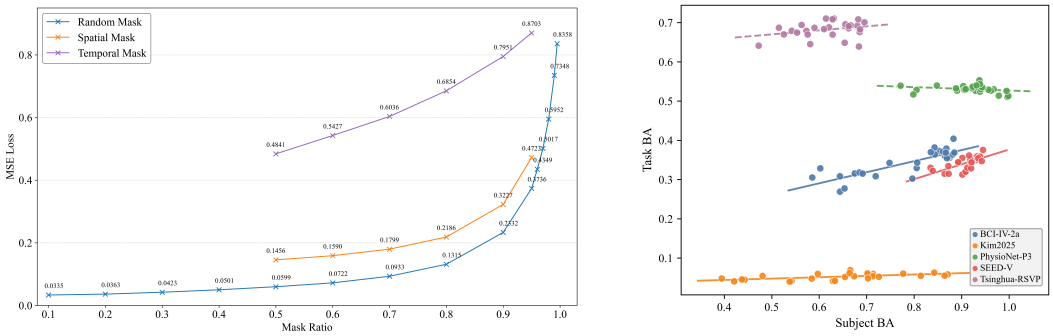
Variance decomposition. Variance decomposition complements linear probing by quantifying sources of representation space variation. In pretrained frozen representations, P_{task} is lower than both P_{subj} and P_{res} in all 25 model–dataset pairs, and lower than P_{inter} in 23 pairs. Thus, task-related variation is limited at the population level and is often expressed through subject–task interactions rather than consistently across subjects.

Randomly initialized representations are typically dominated by residual variation, with little variation associated with subject identity or task labels. P_{subj} increases in 23 of 25 pairs, alongside reductions in P_{res}. BIOT is the sole minor exception: on BCI-IV-2a and SEED-V, its already high P_{subj} decreases from 0.4123 to 0.4000 and from 0.4363 to 0.4038, respectively. In contrast, P_{task} increases in only 8 pairs and decreases in the remaining 17 pairs. Pretraining therefore reduces within-cell residual variation while increasing systematic subject-related variation, without a corresponding shift toward task-related variation.

***t*-SNE and RSA.** In the shared *t*-SNE embedding, trials colored by task are largely intermixed, whereas those colored by subject form more distinct local clusters. The corresponding RSA matrices exhibit limited task-wise diagonal structure but pronounced heterogeneity across subject pairs. Although *t*-SNE and RSA are descriptive analyses, these patterns align with linear probing and variance decomposition: the local geometry and pairwise similarity of frozen representations are more strongly structured by subject identity than by task labels.

4.2 SPATIOTEMPORAL RECONSTRUCTION ASYMMETRY IN MAE-STYLE PRETRAINING

To examine the origin of subject identity encoding in MAE-style models, we pretrain models with varied masking strategies and ratios. As shown in Figure 3a, validation reconstruction loss increases monotonically with the masking ratio under all strategies, consistent with standard MAE behavior in vision. At matched masking ratios, both spatial and temporal masking incur higher loss than random masking; temporal masking is markedly harder, with its loss at $\rho = 0.5$ comparable to that of random masking near $\rho = 0.97$. These results reveal a spatiotemporal reconstruction asymmetry.



(a) MSE loss under different configurations.

(b) Decodability across configurations.

Figure 3: Effects of masking configurations on MAE pretraining and linear-probe performance. In (b), solid and dashed lines denote significant and non-significant Pearson correlations ($p < 0.05$).

We evaluate frozen representations across masking configurations; detailed results are provided in Appendix H.

Linear probing shows that task and subject decoding generally rise together with the masking ratio, peak before near-complete masking, and decline at extreme ratios. Under random masking, task BA peaks at $\rho = 0.90$–0.97 on BCI-IV-2a, Kim2025, SEED-V, and Tsinghua-RSVP, whereas it fluctuates within a narrow range on PhysioNet-P300. These optima exceed the 0.75 ratio commonly used in visual MAE He et al. (2022), reflecting the high redundancy of EEG. In contrast, spatial and temporal masking typically reach peak task BA at slightly lower ratios.

Despite a substantially smaller corpus and model capacity than REVE, which is the strongest previously evaluated EEG-FM, our MAE-style models can match or exceed its task BA under suitable masking configurations. For example, random masking at $\rho = 0.9$ achieves task BA of 0.373 on BCI-IV-2a and 0.347 on SEED-V, exceeding REVE's 0.372 and 0.332, respectively. In the absence of a clear EEG-FM scaling law, these results indicate that masking configuration materially shapes frozen-representation decodability. Yet, relative to same-architecture random initialization, random-mask pretraining yields only marginal task-probing gains but substantial, consistent gains in subject decoding. Pretraining thus again preferentially enhances the decodability of subject identity.

Figure 3b plots task against subject BA across masking configurations. Pearson correlations are positive and significant for BCI-IV-2a ($r = 0.834$, $p < 0.001$), Kim2025 ($r = 0.594$, $p < 0.01$), and SEED-V ($r = 0.662$, $p < 0.001$); all three survive Bonferroni correction across five datasets. This extends the earlier conjecture that task and subject decodability co-vary to masking configurations.

Variance decomposition corroborates this pattern. Across configurations, P_{task} is substantially lower than P_{subj}, and is also below P_{inter} in roughly 80% of cases. Relative to random initialization, P_{subj} rises and P_{res} falls in every configuration, whereas P_{task} rises in fewer than half.

Figure 4 visualizes frozen representations from the MAE-style configuration achieving the highest task BA on each downstream dataset. Even under these task-optimal configurations, task-colored *t*-SNE embeddings remain mixed and task-wise RSA matrices lack consistent diagonal structure. In contrast, subject-colored *t*-SNE embeddings form clearer local clusters, while subject RSA matrices exhibit more heterogeneous subject-pair similarity structure.

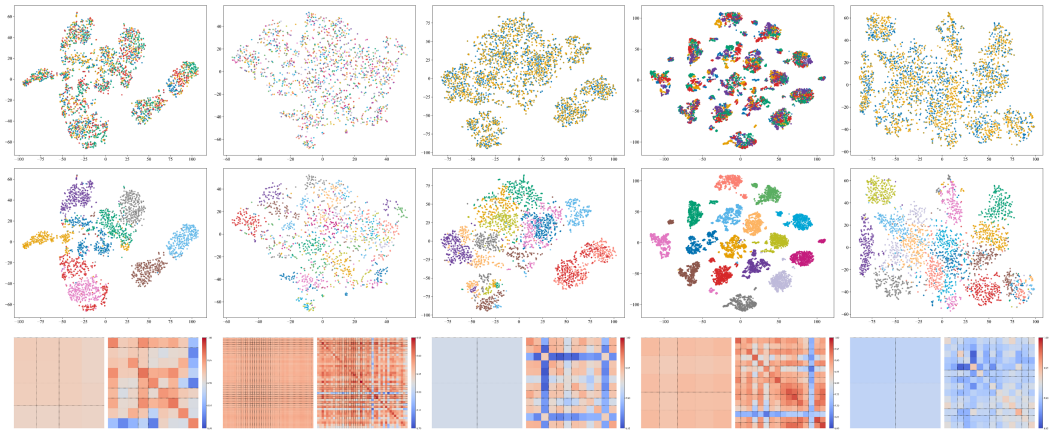
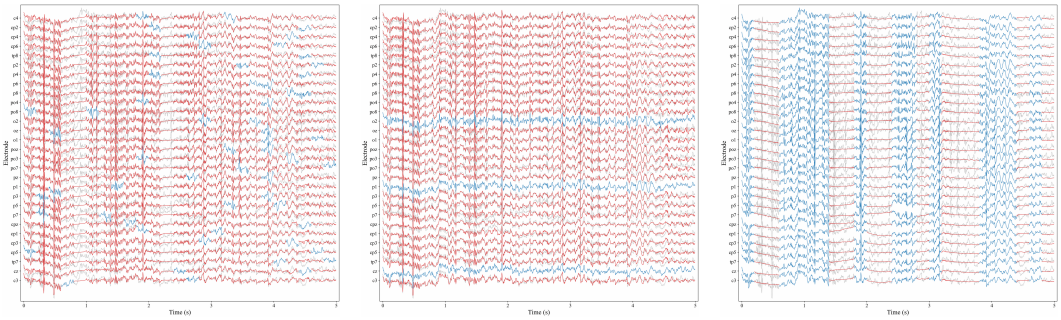


Figure 4: t-SNE and RSA of task-BA-optimal representations. From left to right: BCI-IV-2a (temporal masking, $\rho = 0.8$), Kim2025 (spatial masking, $\rho = 0.95$), PhysioNet-P300 (temporal masking, $\rho = 0.6$), SEED-V (spatial masking, $\rho = 0.5$), and Tsinghua-RSVP (temporal masking, $\rho = 0.6$).

Taken together, although some masking configurations yield state-of-the-art task-probe performance, these results suggest that such gains co-occur with subject-dominated representation reorganization rather than consistent increases in task-related variation shared across subjects.

Prior work attributes subject identity encoding in EEG-FMs to aperiodic low-frequency activity or a $1/f$ spectral bias Kommineni et al. (2026). Our reconstruction-loss results suggest this account may be incomplete. To characterize what different masking strategies preserve, we use the retained decoders of representative pretrained MAE-style models to visualize reconstructions under different masking strategies. Figure 5 shows an SSVEP example. Under random and spatial masking, temporal dynamics, including fine high-frequency waveform envelopes, are accurately recovered when sufficient electrodes remain visible. In contrast, temporal masking occludes concurrent spatial information, forcing inference from temporal context alone. This setting is markedly harder: even low-frequency components are visibly poorly recovered.

These observations suggest a plausible mechanism: EEG spatial redundancy and temporal nonstationarity make cross-channel information especially effective for minimizing reconstruction loss. Volume conduction induces correlations among scalp channels, allowing missing channels to be inferred from visible ones, whereas transient and non-stationary dynamics limit inference of missing segments from temporal context. Crucially, cross-channel covariance may encode individual differences in head anatomy, tissue conductivity, and electrode placement. MAE-style pretraining may therefore preferentially exploit subject-specific spatial cues, creating a potential subject identity shortcut because these cues are highly useful for reconstruction.



(a) Random masking ($\rho = 0.95$) (b) Spatial masking ($\rho = 0.9$) (c) Temporal masking ($\rho = 0.5$)

Figure 5: EEG reconstruction examples under three masking strategies. Gray, blue, and red traces indicate the original signal, visible segments, and masked-segment reconstructions, respectively.

4.3 Individual Differences in MoCo-Style Pretraining

Compared with masked reconstruction, subject identity encoding in contrastive pretraining arises more directly from positive–negative pair construction. Our variance decomposition shows that even randomly initialized models contain substantially more subject-related than task-related variation. Contrastive learning aligns two views of the same trial while separating them from negative trials. When negatives are sampled from a large multi-subject corpus, most negative pairs are cross-subject; subject differences can therefore facilitate both positive-pair consistency and negative-pair separation, furnishing an efficient shortcut for the contrastive objective.

BIOT exhibits this tendency with mini-batch negatives; Appendix I shows that it persists under queue-based MoCo. Relative to random initialization, MoCo increases subject BA and P_{subj} across all five downstream datasets, but yields only small and inconsistent changes in task BA and P_{task}. *t*-SNE and RSA likewise show representations more strongly structured by subject identity than task labels. This pattern is thus not tied to a particular architecture or negative-pool size, but reflects the utility of subject identity for positive-pair consistency and cross-subject negative separation.

4.4 Interventions on Shortcuts Do Not Promote Task Decoding

To test whether constraining subject identity shortcuts strengthens task-related information, we use MAE with EA reconstruction targets (EA-MAE) and MoCo restricted to same-subject negatives (SSN-MoCo). However, neither intervention promotes task decoding. Figure 6 summarizes changes in linear-probe BA, with detailed evaluations provided in Appendix J.

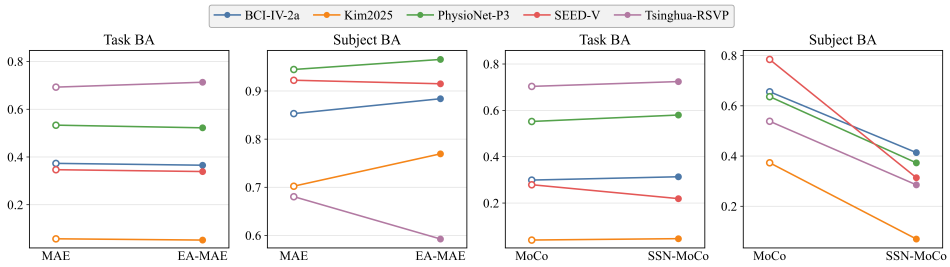


Figure 6: Effects of shortcut interventions on task and subject decoding.

For EA-MAE, subject identity remains decodable. EA aligns subject-level average second-order spatial covariance, but leaves trial-specific covariance deviations, higher-order statistics, and temporal dynamics intact. Reconstruction pretraining can therefore still exploit residual subject-specific regularities that reduce reconstruction loss.

For SSN-MoCo, both subject BA and P_{subj} decrease, confirming that the negative sample constraint weakens subject identity encoding. However, neither task BA nor P_{task} increases accordingly. Instead, reduction in P_{subj} is largely accompanied by an increase in P_{res} across datasets, which may reflect a shift toward trial-specific and view-consistent variation, such as waveform details and artifacts. As this intervention also changes the source, number, and diversity of negatives, the results show only that constraining subject identity alone does not automatically strengthen task decoding.

5 Discussion

Across four complementary analyses, the evaluated EEG-FMs yield frozen representations in which subject identity is more readily decodable than task labels: relative to matched random initialization, pretraining consistently improves subject decoding and generally increases subject-related variation, without stable task-related gains. These findings caution against treating strong fine-tuned performance obtained with flattened token sequences and complex prediction heads as direct evidence of transferable task semantics. EEG-FM-Bench reports gradient conflicts between reconstruction objectives and downstream tasks Xiong et al. (2026), while another study shows that a purely multilayer perceptron achieves strong EEG decoding performance Zhang et al. (2026). Together, these

findings raise the possibility that high fine-tuned performance partly reflects task-specific adaptation capacity rather than task-related information acquired during pretraining.

MAE-style pretraining further shows that suitable masking configurations can yield representations in which subject identity is highly decodable, indicating that self-supervised EEG pretraining can reliably acquire this semantic factor. Such representations may merit evaluation for subject-centered applications, including biometric authentication, and motivate systematic tests of whether their decodability follows a scaling law. In contrast, task decoding does not improve reliably across the settings studied. Although some masking configurations achieve task-probe performance competitive with or exceeding current EEG-FMs, this alone does not establish stronger task-related representations. The correlation between task and subject identity decodability, together with typically larger subject–task interaction than task-attributable variance, is consistent with an identity trap: task gains on particular datasets tend to coincide with stronger subject identity encoding, whereas task-related variation may depend on subject rather than generalize consistently across subjects. This association is stronger for the motor imagery and emotion recognition datasets than for the two P300-based datasets. One possible explanation is that their class labels partly manifest through discriminative spatial patterns Ramoser et al. (2000); Wu et al. (2022), which may be more susceptible to entanglement with subject-specific cues. In contrast, P300 decoding relies on time- and phase-locked temporal responses Ming et al. (2010), potentially weakening this spatial entanglement.

MoCo-style pretraining exhibits the same overall tendency; however, despite sharing the MAE-style backbone, it shows a substantially weaker subject identity bias. Relative to mini-batch contrastive learning, MoCo-style pretraining maintains a larger negative queue, increasing the likelihood of encountering same-subject negatives. We speculate that these negatives become comparatively hard: although they constitute only a small fraction of the queue, their associated gradients may become increasingly influential as training progresses, while the marginal benefit of exploiting the subject identity shortcut to distinguish cross-subject negatives gradually diminishes.

Finally, the intervention results underscore the persistence of readily decodable subject identity and the absence of automatic task-decoding gains. The continued decodability of subject identity indicates that EA alone is insufficient to remove subject-specific information from MAE representations, likely because residual subject-related structure remains exploitable under the reconstruction objective. Same-subject negatives partially suppress the subject identity shortcut, but do not automatically induce task semantics. Future work should develop EEG-specific pretraining objectives that explicitly promote transferable task-related representations.

The current work has several limitations. The subject identity interventions are not strictly self-supervised under the conventional label-free sense, as they require subject-identity metadata. This motivates clearer terminology for metadata-assisted objectives. Moreover, this study is confined to BCI EEG; neither the pretraining corpus nor the downstream datasets include clinical EEG, and whether the reported findings extend to clinical settings remains to be established.

6 Conclusion

This study audits frozen representations of EEG foundation models across five models and five BCI benchmarks. Across complementary analyses, pretraining consistently strengthens subject identity information without comparable or reliable task-related gains. MAE- and MoCo-style pretraining experiments further suggest objective-specific mechanisms for these shortcuts: reconstruction may exploit subject-related cross-channel structure, whereas contrastive learning may benefit from individual differences for positive-pair consistency and cross-subject negative separation. Interventions aimed at these shortcuts do not automatically improve task decoding. Future EEG pretraining objectives should explicitly promote task semantics, rather than assuming they emerge from reconstruction or cross-view consistency.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62476205) and the Fundamental Research Funds for the Central Universities (Grant Nos. ZYTS25152 and ZYTS25136).

AI USE STATEMENT

In this work, we used generative AI tools for research assistance. Specifically, large language models (GPT-5.6 Terra and GPT-6 Sol) were used as auxiliary tools to refine hypotheses, assist in result interpretation, and support the implementation of parts of the code. We did not use generative AI to generate synthetic datasets, formulate mathematical claims, develop proofs, or conduct qualitative or quantitative analyses. Additionally, the models were used for manuscript drafting and editing, grammar checking, and code refinement. All AI-assisted work, including text and code, was reviewed and verified for correctness by two authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Reproducibility is supported by the Methods and Appendix, which detail the downstream and pretraining datasets, preprocessing, task splits, model architectures, training hyperparameters, linear-probing and variance-decomposition procedures, and intervention settings. **After the double-blind review period, we will release a complete GitHub repository**, which will contain the source code, configuration files, preprocessing and split scripts, evaluation pipelines, pretrained checkpoints, documentation, and usage examples. Public datasets can be accessed from their original providers in accordance with their respective licenses.

REFERENCES

Bruno Aristimunha, Pierre Guetschel, Martin Wimpff, Lukas Gemein, Cedric Rommel, Hubert Banville, Maciej Sliwowski, Daniel Wilson, Simon Brandt, Théo Gnassounou, et al. Braindecode: toolbox for decoding raw electrophysiological brain data with deep learning models. *Zenodo*, 2025. doi: 10.5281/zenodo.17699192. URL `https://github.com/braindecode/braindecode`.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. BCI competition 2008–graz data set A. *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, 16(1-6):1, 2008.

Luca Citi, Riccardo Poli, and Caterina Cinel. Documenting, modelling and exploiting P300 amplitude changes due to variable target delays in donchin's speller. *Journal of Neural Engineering*, 7 (5):056006, 2010.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16×16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, pp. 1–22. Curran Associates, Inc., 2021.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *International IEEE/EMBS Conference on Neural Engineering*, pp. 81–84. IEEE, 2013.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG-adapting to any setup with large-scale pretraining on 25,000 subjects. *Advances in Neural Information Processing Systems*, 38:22541–22577, 2026.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735. IEEE, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988. IEEE, 2022.

Gan Huang, Zhenxing Hu, Weize Chen, Shaorong Zhang, Zhen Liang, Linling Li, Li Zhang, and Zhiguo Zhang. M3CV: A multi-subject, multi-session, and multi-task database for EEG-based biometrics challenge. *NeuroImage*, 264:119666, 2022.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. SEED-VII: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2024a.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024b.

Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. EEG-Bench: A benchmark for EEG foundation models in clinical applications. *arXiv preprint arXiv:2512.08959*, 2025.

Heegyu Kim, Kyungho Won, Minkyu Ahn, and Sung Chan Jun. A 40-class ssvep speller dataset: Beta range stimulation for low-fatigue BCI applications. *Scientific Data*, 12(1):1751, 2025.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Simon Bock Segaard, Jeppe Roden Münster, Andreas Peter Juhl Hansen, Takfarinas Medani, Tiantian Feng, Richard Leahy, and Shrikanth Narayanan. Aperiodic and low-frequency spectral bias in reconstruction based EEG foundation models. *arXiv preprint arXiv:2605.26434*, 2026.

Konstantinos Kontras, Trui Osselaer, Stylianos G. Mouslech, Angeliki-Ilektra Karaiskou, Guido Gagliardi, Thomas Strypsteen, Mohammad Hossein Badiei, Anku Rani, Maarten Vanmarcke, Miguel Bhagubai, Chanakya Ekbote, Jaedong Hwang, Christos Chatzichristos, Paul Pu Liang, and Maarten De Vos. NeuroAtlas: Benchmarking foundation models for clinical EEG and brain-computer interfaces. In *Greeks in AI*, 2026.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:249, 2008.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. EEG foundation models: a critical review of current progress and future directions. *Journal of Neural Engineering*, 23(2):021001, 2026.

Ang Li, Zikai Wang, Liuyin Yang, Zhenyu Wang, Tianheng Xu, Honglin Hu, and Marc M Van Hulle. CoMET: A contrastive-masked brain foundation model for universal EEG representation. *arXiv preprint arXiv:2509.00314*, 2025.

Chenqi Li, Yu Liu, Shuo Zhang, Timothy Denison, and Tingting Zhu. Standing on the shoulders of giants: Rethinking EEG foundation model pretraining via multi-teacher distillation. *arXiv preprint arXiv:2603.04478*, 2026.

Jun-You Lin, Ying Choon Wu, and Tzyy-Ping Jung. The identity trap in EEG foundation models: A diagnostic audit. *arXiv preprint arXiv:2606.06647*, 2026.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

Gabriel Mahuas, Victoria Shevchenko, Ugo Tanielian, Yassir Bendou, and Richard Gao. CoCoT-EEG: Contrastive-pretrained multiscale convolutional transformer for EEG decoding. *arXiv preprint arXiv:2607.09543*, 2026.

Dong Ming, Xingwei An, Youyuan Xi, Yong Hu, Baikun Wan, Hongzhi Qi, Longlong Cheng, and Zhaojun Xue. Time-locked and phase-locked features of P300 event-related potentials (ERPs) for brain–computer interface speller. *Biomedical Signal Processing and Control*, 5(4):243–251, 2010.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00196.

Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B Esper, Dung Truong, Arnaud Delorme, Michael P Milham, and Scott Makeig. HBN-EEG: The FAIR implementation of the Healthy Brain Network (HBN) electroencephalography dataset. *bioRxiv*, pp. 2024–10, 2024.

James R Stieger, Stephen A Engel, and Bin He. Continuous sensorimotor rhythm based brain computer interface learning in a large population. *Scientific Data*, 8(1):98, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, volume 2025, pp. 75310–75346, 2025.

Xianheng Wang, Yige Yang, and Damien Coyle. EEG-FM-Audit: A systematic evaluation and analysis pipeline for EEG foundation models. *arXiv preprint arXiv:2605.26910*, 2026.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2016.

Xun Wu, Wei-Long Zheng, Ziyi Li, and Bao-Liang Lu. Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. *Journal of Neural Engineering*, 19(1): 016012, 2022.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. In *Forty-third International Conference on Machine Learning*, 2026.

Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuxuan Yao, Hongbo Wang, Li Chen, Yiheng Peng, and Jingjing Luo. Foundation models for EEG decoding: Current progress and prospective research. *Journal of Neural Engineering*, 22(6): 061002, 2025.

Ziyuan Zhang, Ziyu Wang, Kaitai Guo, Yang Zheng, Minghao Dong, and Jimin Liang. Boosting brain-computer interface performance through cognitive training: A brain-centric approach. *Journal of Information and Intelligence*, 3(1):19–35, 2025.

Ziyuan Zhang, Yang Zheng, Kaitai Guo, Jimin Liang, and Minghao Dong. A few-layer multilayer perceptron is worth attention for EEG classification in rapid serial visual presentation task. *International Journal of Neural Systems*, 36(09):2650030, 2026.

Li Zheng, Sen Sun, Hongze Zhao, Weihua Pei, Hongda Chen, Xiaorong Gao, Lijian Zhang, and Yijun Wang. A cross-session dataset for collaborative brain-computer interfaces based on rapid serial visual presentation. *Frontiers in neuroscience*, 14:579469, 2020.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

APPENDIX

A EEG FOUNDATION MODELS

We evaluate five publicly released pretrained checkpoints: BIOT (3.44M parameters), CBraMod (4.92M), EEGPT (25.29M), LaBraM-Base (5.82M), and REVE-Base (69.19M). We instantiate these models using the publicly available implementations provided by Braindecode Aristimunha et al. (2025) and obtain the corresponding pretrained weights from the Hugging Face Hub. The reported parameter counts were calculated from the downloaded pretrained checkpoints.

BIOT. BIOT Yang et al. (2023) was pretrained on 5 million 10-second resting-state EEG samples, either alone or jointly with 5 million sleep EEG samples. It splits each channel into fixed-length temporal segments, with optional overlap, and concatenates tokens from all channels into a single sequence. Each token contains a Fast Fourier Transform (FFT)-based spectral-energy feature, a channel embedding, and a relative temporal position embedding. During pretraining, parts of the input channels and temporal tokens are dropped. The model maps this perturbed signal toward the representation of the original signal through a two-layer predictor and a temperature-scaled contrastive cross-entropy loss.

CBraMod. CBraMod Wang et al. (2025) was pretrained on a cleaned subset of TUEGObeid & Picone (2016) containing 1,109,545 30-second samples, totaling more than 9,000 hours of EEG. It retains 19 common electrodes and divides each channel into one-second patches. A 30-second recording therefore yields 570 tokens. The patch encoder combines temporal convolutional features with FFT-derived frequency features, and the criss-cross Transformer models temporal and spatial dependencies in separate attention branches. Half of the patches are randomly masked during pretraining. A linear reconstruction head predicts only the masked raw patches under a mean squared error (MSE) loss.

EEGPT. EEGPT Wang et al. (2024) was pretrained on PhysioMI, HGD, TSU, SEED, and M3CV, covering motor imagery and execution, SSVEP, emotion recognition, and multi-paradigm EEG. EEGPT partitions each channel into non-overlapping 250-ms patches and adds a learned channel embedding to every patch token. Its objective combines two MSE losses: one reconstructs masked raw patches, while the other matches predictor outputs to representations generated by a momentum encoder from the complete input.

LaBraM. LaBraM Jiang et al. (2024b) was pretrained on 2,534.78 hours of EEG from roughly 20 public and self-collected datasets. It first divides each channel into non-overlapping temporal patches, then assigns each patch a discrete neural token using a vector-quantized tokenizer. This tokenizer is trained to reconstruct Fourier-spectrum amplitude and phase with MSE, together with the vector-quantization loss. LaBraM then masks a subset of patches and predicts their discrete token identities using cross-entropy loss. The authors also use symmetric masking, which applies complementary masks to the same tokenized sample.

REVE. REVE El Ouahidi et al. (2026) was pretrained on 61,415 hours of EEG from 92 datasets and 24,274 subjects. It divides each channel into overlapping one-second patches and encodes their electrode coordinates and temporal indices with a four-dimensional positional encoding. This allows the model to accommodate different montages without a fixed channel-index embedding. REVE uses spatio-temporal block masking and reconstructs masked raw patches with an L1 loss. Its auxiliary objective reconstructs the same patches from an attention-pooled global token, also with L1 loss and a weight of 0.1.

B DOWNSTREAM DATASETS

The five downstream benchmarks are summarized in Table 2. The detailed descriptions of each dataset are as follows:

BCI-IV-2a.[2] This dataset contains EEG recordings from nine subjects performing four motor-imagery tasks: left hand, right hand, feet, and tongue Brunner et al. (2008). Each subject completed two sessions of 288 trials, yielding 5,184 trials in total. Signals were recorded from 22 EEG

[2] https://www.bbci.de/competition/iv/

Table 2: Summary of downstream datasets.

Dataset	Type	Subjects	Samples	Electrodes	Length (s)	Classes
BCI-IV-2a	Motor imagery	9	5,184	22	4	4
Kim2025	SSVEP	40	9,600	31	5	40
PhysioNet-P300	P300 speller	12	28,463	64	2	2
SEED-V	Emotion recognition	16	29,168	62	4	5
Tsinghua-RSVP	RSVP	14	117,600	62	1	2

electrodes at 250 Hz. We band-pass filtered the continuous recordings between 0.5 and 40 Hz and resampled them to 200 Hz. EEG segments from 2 to 6 s after trial onset were then extracted. For population decoding, samples from the original training sessions were split into training and validation sets at a 3:1 ratio within each subject–class stratum, while all samples from the original evaluation sessions were used for testing. This produced 1,944, 648, and 2,592 trials for training, validation, and testing, respectively.

Kim2025.[3] This dataset contains EEG recordings from 40 subjects performing a 40-class SSVEP task Kim et al. (2025). Each subject completed six experimental blocks, yielding 9,600 trials in total. Signals were recorded from 31 scalp EEG electrodes and two mastoid reference electrodes at 1,024 Hz. We removed the mastoid reference electrodes, applied a 0.3–89 Hz band-pass filter and a 60 Hz notch filter, and resampled the data to 200 Hz. We then extracted EEG segments from 0 to 5 s after stimulus onset. For each subject and stimulus target, the six blocks were partitioned into training, validation, and test sets in a 2:2:2 ratio, yielding 3,200 trials in each split.

PhysioNet-P300.[4] This dataset contains EEG recordings from 12 subjects performing a P300 speller task, comprising 28,463 samples Citi et al. (2010). Signals were recorded from 64 scalp EEG electrodes at 2,048 Hz. For each row or column flash, a trial was labeled as a target when the flashed row or column contained the target character, and as a non-target otherwise. We applied a 0.3–75 Hz band-pass filter and a 50 Hz notch filter to the continuous recordings, resampled them to 200 Hz, and extracted epochs from 0.7 s before to 1.3 s after each flash event. For population decoding, samples were split into training, validation, and test sets at a 4:1:3 ratio within each subject–class stratum, resulting in 14,228, 3,550, and 10,685 samples, respectively.

SEED-V.[5] This dataset contains EEG recordings from 16 subjects under five emotion-elicitation conditions Liu et al. (2021). Each subject completed three sessions, each containing 15 emotion-elicitation trials. Recordings were acquired from 62 EEG electrodes and yielded 29,168 four-second samples after segmentation. We removed mastoid reference electrodes and electrooculography channels, applied a 0.3–75 Hz band-pass filter and a 50 Hz notch filter, and resampled the data to 200 Hz. Emotion-elicitation segments were extracted according to the experimental annotations and divided into non-overlapping four-second windows. For population decoding, samples were split into training, validation, and test sets at a 4:1:3 ratio within each subject–class stratum, resulting in 14,560, 3,600, and 11,008 samples, respectively.

Tsinghua-RSVP.[6] This dataset contains EEG recordings from 14 subjects performing an RSVP task, comprising 117,600 target and non-target trials Zheng et al. (2020). Signals were recorded from 62 EEG electrodes at 1,000 Hz. We applied a 0.3–75 Hz band-pass filter and a 50 Hz notch filter, resampled the recordings to 200 Hz, and extracted epochs from 0.2 s before to 0.8 s after each stimulus trigger. For population decoding, samples were split into training, validation, and test sets at a 4:1:3 ratio within each subject–class stratum, yielding 58,800, 14,700, and 44,100 samples, respectively.

[3] https://doi.org/10.6084/m9.figshare.28806815.v2

[4] https://physionet.org/content/erpbci/1.0.0/

[5] https://bcmi.sjtu.edu.cn/home/seed/seed-v.html

[6] https://bci.med.tsinghua.edu.cn/download.html

C Evaluation Methods

C.1 Linear Probing

Linear probing is performed separately for task-label and subject-identity decoding. For each EEG-FM, the pretrained backbone and its randomly initialized counterpart are frozen throughout training, including all normalization and dropout layers. Trial-level representations are extracted, and a single linear classifier is trained on top of each representation. Thus, no nonlinear prediction head or backbone adaptation is introduced during probing.

All runs use the predefined population-protocol splits, with samples stratified by subject identity and task labels. Each EEG trial is independently z-score normalized along the temporal dimension within each channel before model-specific electrode alignment and sampling-rate adaptation. The linear classifier is optimized for 500 epochs using AdamW with a learning rate of 1×10^{-3}, weight decay of 0.01, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-8}$.

The objective is inverse-frequency-weighted cross-entropy:

$$\mathcal{L}_{\mathrm{WCE}} = -\frac{\sum_{n=1}^{N}\sum_{c=1}^{C} w_c y_{n,c} \log\left(p_{n,c}\right)}{\sum_{n=1}^{N}\sum_{c=1}^{C} w_c y_{n,c}}, \tag{1}$$

where N is the batch size, C is the number of classes, $y_{n,c}$ is the one-hot label of trial n, and $p_{n,c}$ is the predicted probability of class c. The weight for class c is computed from the training split as

$$w_c = \frac{N_{\mathrm{train}}}{C N_c}, \tag{2}$$

where N_{train} is the number of training trials and N_c is the number of training trials in class c. Class weights are computed separately for task-label and subject identity decoding.

Training uses a batch size of 256 per GPU (1,024 in total over four GPUs). The checkpoint with the highest validation balanced accuracy is selected and evaluated once on the test split. Randomly initialized baselines use the same architectures, data splits, optimization settings, and frozen-backbone setting, with checkpoint loading disabled.

C.2 Variance Decomposition

To quantify the relative proportions of variation in frozen representations associated with subject identity, task labels, and other factors, we decompose the total sum of squared deviations of trial-level representations over the complete downstream dataset. Variance decomposition is performed separately for representations extracted from frozen pretrained EEG-FMs and their frozen randomly initialized counterparts. This analysis quantifies the proportions of total representation variation associated with subject identity, task labels, their interaction, and trial-level variation within the same subject–task cell.

For each trial, let $\mathbf{z}_{str} \in \mathbb{R}^D$ denote its frozen representation, where $s = 1, \ldots, S$, $t = 1, \ldots, T$, and $r = 1, \ldots, n_{st}$ index subject identity, task label, and trial, respectively. A subject–task cell (s, t) contains all trials sharing subject s and task label t, and n_{st} denotes its number of trials. All representations are L2-normalized before decomposition to remove scale differences across models.

Since the number of trials may differ across cells, each trial in cell (s, t) is assigned the weight

$$w_{str} = \frac{1}{n_{st}}, \qquad \sum_{r=1}^{n_{st}} w_{str} = 1. \tag{3}$$

Thus, every subject–task cell has equal total weight. The mean representation of cell (s, t) is

$$\bar{\mathbf{z}}_{st\cdot} = \sum_{r=1}^{n_{st}} w_{str} \mathbf{z}_{str}. \tag{4}$$

The grand mean, subject mean, and task mean are defined as

$$\bar{\mathbf{z}}_{\cdots} = \frac{1}{ST} \sum_{s=1}^{S}\sum_{t=1}^{T} \bar{\mathbf{z}}_{st\cdot}, \tag{5}$$

$$\bar{\mathbf{z}}_{s\cdot\cdot} = \frac{1}{T}\sum_{t=1}^{T}\bar{\mathbf{z}}_{st\cdot}, \tag{6}$$

and

$$\bar{\mathbf{z}}_{\cdot t\cdot} = \frac{1}{S}\sum_{s=1}^{S}\bar{\mathbf{z}}_{st\cdot}. \tag{7}$$

Using squared Euclidean distance in representation space, the total sum of squared deviations is

$$SS_{\text{total}} = \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{r=1}^{n_{st}} w_{str}\,\|\mathbf{z}_{str} - \bar{\mathbf{z}}_{\cdots}\|_2^2\,. \tag{8}$$

The subject and task components are respectively defined as

$$SS_{\text{subj}} = T\sum_{s=1}^{S}\|\bar{\mathbf{z}}_{s\cdot\cdot} - \bar{\mathbf{z}}_{\cdots}\|_2^2\,, \tag{9}$$

$$SS_{\text{task}} = S\sum_{t=1}^{T}\|\bar{\mathbf{z}}_{\cdot t\cdot} - \bar{\mathbf{z}}_{\cdots}\|_2^2\,. \tag{10}$$

The subject–task interaction component is

$$SS_{\text{inter}} = \sum_{s=1}^{S}\sum_{t=1}^{T}\|\bar{\mathbf{z}}_{st\cdot} - \bar{\mathbf{z}}_{s\cdot\cdot} - \bar{\mathbf{z}}_{\cdot t\cdot} + \bar{\mathbf{z}}_{\cdots}\|_2^2\,, \tag{11}$$

and the within-cell residual component is

$$SS_{\text{res}} = \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{r=1}^{n_{st}} w_{str}\,\|\mathbf{z}_{str} - \bar{\mathbf{z}}_{st\cdot}\|_2^2\,. \tag{12}$$

The four components satisfy

$$SS_{\text{total}} = SS_{\text{subj}} + SS_{\text{task}} + SS_{\text{inter}} + SS_{\text{res}}. \tag{13}$$

Here, SS_{inter} measures whether the effect of task labels on representations varies across subjects, whereas SS_{res} captures within-cell trial-level variation not accounted for by subject identity, task labels, or their interaction. For each component c, its relative contribution to total representation variation is calculated as

$$P_c = \frac{SS_c}{SS_{\text{total}}}, \qquad c \in \{\text{subj}, \text{task}, \text{inter}, \text{res}\}\,. \tag{14}$$

C.3 t-SNE VISUALIZATION

t-SNE is used to visualize frozen representations. To control visualization scale, we retain at most 16 task classes and 16 subjects, with at most 400 trials per subject–task cell. PhysioNet-P300 and Tsinghua-RSVP are additionally downsampled to balance task classes.

Each point represents one trial. For each model–dataset pair, we compute one t-SNE embedding and render it twice, colored by task label and anonymized subject identity, respectively; the paired plots share the same coordinates. Task colors denote task labels, whereas subject colors denote anonymized subject identity. Within each task- or subject-colored panel, colors are assigned in the order listed below.

C.3.1 BCI-IV-2A

Task colors:

● left hand; ● right hand; ● feet; ● tongue.

Subject colors:

● S_1; ● S_2; ● S_3; ● S_4; ● S_5; ● S_6; ● S_7; ● S_8; ● S_9.

C.3.2 KIM2025

Task colors:

14.0 Hz; 15.0 Hz; 16.0 Hz; 17.0 Hz; 18.0 Hz; 19.0 Hz; 20.0 Hz; 21.0 Hz;
14.2 Hz; 15.2 Hz; 16.2 Hz; 17.2 Hz; 18.2 Hz; 19.2 Hz; 20.2 Hz; 21.2 Hz.

Subject colors:

S_1; S_2; S_3; S_4; S_5; S_6; S_7; S_8;
S_9; S_{10}; S_{11}; S_{12}; S_{13}; S_{14}; S_{15}; S_{16}.

C.3.3 PHYSIONET-P300

Task colors:

non-target; target.

Subject colors:

S_1; S_2; S_3; S_4; S_5; S_6;
S_7; S_8; S_9; S_{10}; S_{11}; S_{12}.

C.3.4 SEED-V

Task colors:

disgust; fear; sad; neutral; happy.

Subject colors:

S_1; S_2; S_3; S_4; S_5; S_6; S_7; S_8;
S_9; S_{10}; S_{11}; S_{12}; S_{13}; S_{14}; S_{15}; S_{16}.

C.3.5 TSINGHUA-RSVP

Task colors:

non-target; target.

Subject colors:

S_1; S_2; S_3; S_4; S_5; S_6; S_7;
S_8; S_9; S_{10}; S_{11}; S_{12}; S_{13}; S_{14}.

C.4 REPRESENTATIONAL SIMILARITY ANALYSIS

Representational similarity analysis is performed on trial-level representations extracted from frozen pretrained encoders. For each model–dataset pair, all trials from the training, validation, and test splits are concatenated. All trial representations are L2-normalized before similarity computation.

Let $\mathbf{z}_i \in \mathbb{R}^D$ denote the normalized representation of trial i, and let t_i and s_i denote its task label and subject identity, respectively. The task similarity matrix is defined as

$$M_{ab}^{\text{task}} = \frac{1}{N_{ab}^{\text{task}}} \sum_{\substack{i \neq j \\ t_i = a,\ t_j = b}} \mathbf{z}_i^\top \mathbf{z}_j, \tag{15}$$

where N_{ab}^{task} is the number of ordered trial pairs satisfying $t_i = a$ and $t_j = b$. Similarly, the subject similarity matrix is

$$M_{ab}^{\text{subj}} = \frac{1}{N_{ab}^{\text{subj}}} \sum_{\substack{i \neq j \\ s_i = a,\ s_j = b}} \mathbf{z}_i^\top \mathbf{z}_j, \tag{16}$$

where N_{ab}^{subj} is the number of ordered trial pairs satisfying $s_i = a$ and $s_j = b$.

Thus, each matrix entry is the mean raw cosine similarity between trials from the corresponding task or subject groups. The condition $i \neq j$ excludes self-pairs from diagonal entries. Task and subject matrices are displayed as paired heatmaps using the same cosine-similarity scale from -1 to 1.

D PRETRAINING DATASETS

The pretraining corpus comprises seven EEG datasets, summarized in Table 3. Unless otherwise noted, recordings are band-pass filtered between 0.3 and 75 Hz, notch-filtered according to the acquisition environment, and resampled to 200 Hz. Electrodes are aligned to a common 60-channel montage, with missing channels completed by spatial interpolation when necessary Li et al. (2025). Signals are divided into four-second segments and independently z-score normalized along the temporal dimension within each channel, yielding an input of size 60×800. Approximately 5% of samples from each subject are randomly assigned to validation, resulting in 1,495,149 pretraining and 88,704 validation samples.

Table 3: Summary of pretraining datasets.

Dataset	Type	Subjects	Samples	Electrodes	Length (s)
HBN	Multi-task	1,895	1,062,438	128	4
M3CV	Multi-task	95	96,228	64	4
Stieger21	Motor imagery	62	269,099	62	4
SEED-VII	Emotion recognition	20	68,758	62	4
SEED	Emotion recognition	15	37,890	62	4
TSU-Benchmark	SSVEP	35	16,800	64	4
XDU-RSVP	RSVP	34	32,640	63	4

The detailed descriptions of each dataset are as follows:

HBN.[7] HBN-EEG is a large-scale high-density EEG dataset for studying brain development in children and adolescents Shirazi et al. (2024). It includes resting-state, surround-suppression, movie-watching, sequence-learning, contrast-change-detection, and symbol-search recordings. We use available recordings from 1,895 subjects in Releases 1–11, all acquired with a 128-channel HydroCel system. Four-second segments are extracted from continuous recordings with a five-second stride, yielding 1,062,438 samples.

M3CV.[8] M3CV is a multi-subject, multi-session, and multi-task dataset designed to study individual variability in EEG Huang et al. (2022). It covers resting-state, transient sensory evoked, steady-state sensory evoked, cognitive oddball, motor-execution, and selective-attention SSVEP paradigms, comprising 14 tasks. Of the 106 recruited participants, 95 completed two sessions on separate days; their recordings are used in this study. The publicly available pre-segmented 64-channel trials are filtered, aligned to the common montage, and retained as four-second segments, yielding 96,228 samples.

Stieger21.[9] Stieger21 contains EEG recordings from 62 subjects across 598 online sensorimotor-rhythm BCI training sessions Stieger et al. (2021). Subjects completed 7–11 sessions and controlled a two-dimensional cursor using left-hand, right-hand, bimanual, or rest motor-imagery strategies. Each session typically contains 450 trials recorded from 62 channels at 1,000 Hz. The final four seconds of every trial are retained, resulting in 269,099 samples.

SEED.[10] SEED is an emotion-recognition dataset in which film clips induce positive, negative, and neutral emotional states Zheng & Lu (2015); Duan et al. (2013). Fifteen subjects completed three sessions approximately one week apart, each containing 15 film clips of about four minutes. EEG was recorded using a 62-channel ESI NeuroScan system at 1,000 Hz. Mastoid-reference, electrooculography, and electrocardiography channels are removed. Valid intervals associated with each film stimulus are segmented into non-overlapping four-second windows, yielding 37,890 samples.

[7] https://nemar.org/dataset/on005505

[8] https://www.kaggle.com/competitions/eeg-biometric-competition/data

[9] https://doi.org/10.6084/m9.figshare.13123148

[10] https://bcmi.sjtu.edu.cn/home/seed/seed.html

SEED-VII.[11] SEED-VII is a fine-grained emotion-recognition dataset with seven emotional states: happiness, sadness, fear, disgust, neutrality, anger, and surprise Jiang et al. (2024a). Twenty subjects completed four sessions, each containing 20 video clips. After removing mastoid-reference, electrooculography, and electrocardiography channels, valid intervals determined by video-event boundaries are segmented into non-overlapping four-second windows, yielding 68,758 samples.

TSU-Benchmark.[12] TSU-Benchmark contains 40-class SSVEP spelling recordings from 35 subjects, including eight experienced and 27 inexperienced SSVEP-BCI users Wang et al. (2016). Each subject completed six blocks with 40 stimulation targets per block. Recordings were acquired from 64 channels at 1,000 Hz; each six-second trial contains a five-second visual-stimulation period. Two four-second segments, spanning 0–4 s and 1–5 s after stimulus onset, are extracted from each trial, yielding 16,800 samples.

XDU-RSVP.[13] XDU-RSVP dataset contains recordings from 34 subjects performing an RSVP task, acquired using 63 EEG electrodes Zhang et al. (2025). Two four-second segments, spanning 1–5 s and 2–6 s after sequence onset, are extracted from every recording, yielding 32,640 samples.

E Construction of Self-Supervised Pretraining Models

E.1 MAE-Style Model

To examine the representational properties of reconstruction-based self-supervised learning, we train a series of MAE-style EEG-FMs that vary in masking strategy and ratio. Each model contains 4.76M parameters. The architecture, masking schemes, training objective, and optimization settings are described below; hyperparameters are listed in Table 4.

Table 4: Hyperparameters for MAE-style reconstruction pretraining.

Category	Hyperparameter	Value
Encoder	Token embedding dimension	256
	Channel positional encoding	Learned
	Temporal positional encoding	Sinusoidal
	Transformer blocks	6
	Attention heads	4
	MLP ratio	4
	Dropout	0.05
Decoder	Transformer blocks	2
	Attention heads	4
	MLP ratio	4
	Dropout	0.05
Training	Epochs	100
	Batch size per GPU	256
	Optimizer	AdamW
	Learning-rate range	$[2 \times 10^{-5}, 5 \times 10^{-4}]$
	Learning-rate schedule	Warmup–plateau–cosine decay
	Warmup / decay fractions	0.02 / 0.80
	Weight decay	0.01
	Gradient clipping threshold	3.0
	Training precision	bfloat16

Each EEG sample is a 60×800 channel–time matrix, corresponding to 4 s sampled at 200 Hz. Within each channel, the signal is partitioned into 20 non-overlapping temporal segments of 40 samples, each treated as an EEG token, yielding 1,200 channel–time tokens per sample. Tokens are

[11] https://bcmi.sjtu.edu.cn/home/seed/seed-vii.html
[12] https://bci.med.tsinghua.edu.cn/download.html
[13] https://doi.org/10.6084/m9.figshare.33404251

linearly projected into the embedding space, after which learned channel positional encodings and sinusoidal temporal positional encodings are added before masking.

For random masking, all channel–time tokens in a sample are flattened, and a specified proportion is randomly selected for masking. We train one model at each of 15 masking ratios:

$$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.96, 0.97, 0.98, 0.99, 0.995\}.$$

We further consider structured spatial and temporal masking to probe the model's reliance on spatial and temporal information. Under spatial masking, randomly selected channels have all their temporal tokens masked; under temporal masking, randomly selected temporal segments are masked across all channels. Each structured masking scheme is evaluated at six masking ratios:

$$\{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}.$$

Spatial masking preserves complete temporal dynamics in visible channels, whereas temporal masking preserves complete spatial topographies in visible time segments.

Only visible tokens are passed to the Transformer encoder. During decoding, encoded visible tokens are returned to their original positions in the full sequence, while learned mask tokens fill the masked positions. Channel and temporal positional encodings are added again before a Transformer decoder and linear prediction layer reconstruct the original signal segment of each token. The reconstruction objective is

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x}_i \in \mathcal{B}} \frac{\sum_{i=1}^{N} m_i \left\| \hat{\mathbf{x}}_i - \mathbf{x}_i \right\|_2^2}{\sum_{i=1}^{N} m_i}. \tag{17}$$

where $\mathcal{B}$ is the batch size, N is the number of tokens per trial, and $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ denote the original and reconstructed EEG segments for token i, respectively. The mask indicator $m_i \in \{0, 1\}$ equals one when the token is masked and zero otherwise. Thus, the reconstruction loss is computed only over masked tokens. Except for the masking strategy and ratio, all training configurations are shared across MAE variants. After pretraining, the decoder is discarded, and only the 4.75M-parameter encoder (slightly smaller than representative EEG-FMs such as CBraMod) is retained for frozen-representation evaluation.

E.2 MoCo-style Model

To examine whether MAE-style representational properties persist under contrastive learning, we train a MoCo-style EEG-FM He et al. (2020). The model comprises a query encoder, a momentum key encoder, two projection heads, and a first-in, first-out queue of negative keys. The query and key encoders share the same architecture. The query encoder and its projection head are updated by backpropagation, whereas the corresponding key-branch parameters are updated by exponential moving average:

$$\boldsymbol{\theta}_k \leftarrow \mu \boldsymbol{\theta}_k + (1 - \mu) \boldsymbol{\theta}_q, \tag{18}$$

where $\boldsymbol{\theta}_q$ and $\boldsymbol{\theta}_k$ denote the query- and key-branch parameters, respectively, and μ is the momentum coefficient. Each EEG trial is a 60×800 channel–time matrix, corresponding to 4 s sampled at 200 Hz. For each trial, the key branch receives the complete signal, whereas the query branch receives a partially observed view formed by randomly dropping 40% of channels and 15% of temporal segments. This requires the query encoder to produce a representation consistent with that of the momentum key encoder despite missing spatial and temporal information.

Following the MAE-style tokenization scheme, each encoder partitions an EEG trial into 1,200 channel–time tokens, with 20 non-overlapping temporal segments per channel. Tokens are linearly projected to 256 dimensions and combined with learned channel positional encodings and sinusoidal temporal positional encodings. The resulting token sequence is processed by a six-layer, four-head Transformer encoder, and mean pooling over its output tokens yields a trial-level representation.

Query and key representations are independently passed through two-layer projection heads with GELU activation and are L2-normalized. Let $\mathbf{q}_i$, $\mathbf{k}_i^+$, and $\mathbf{k}_j^-$ denote the normalized query representation for trial i, its corresponding positive key representation, and the j-th negative key in the

Table 5: Hyperparameters for MoCo-style contrastive pretraining.

Category	Hyperparameter	Value
Encoder	Token embedding dimension	256
	Channel positional encoding	Learned
	Temporal positional encoding	Sinusoidal
	Transformer blocks	6
	Attention heads	4
	MLP ratio	4
	Dropout	0.05
Contrastive learning	Projection head	$256 \to 128$ (GELU)
	Momentum coefficient	0.999
	Queue size	65,536
	Temperature	0.07
Training	Epochs	100
	Batch size per GPU	256
	Optimizer	AdamW
	Learning-rate range	$[2 \times 10^{-5}, 5 \times 10^{-4}]$
	Learning-rate schedule	Warmup–plateau–cosine decay
	Warmup / decay fractions	0.02 / 0.80
	Weight decay	0.01
	Gradient clipping threshold	3.0
	Training precision	bfloat16

queue, respectively. The contrastive objective is

$$\mathcal{L}_{\mathrm{MoCo}} = -\frac{1}{|\mathcal{B}|} \sum_{\mathbf{q}_i \in \mathcal{B}} \log \frac{\exp\left(\mathbf{q}_i^{\mathsf{T}} \mathbf{k}_i^{+} / \tau\right)}{\exp\left(\mathbf{q}_i^{\mathsf{T}} \mathbf{k}_i^{+} / \tau\right) + \sum_{j=1}^{K} \exp\left(\mathbf{q}_i^{\mathsf{T}} \mathbf{k}_j^{-} / \tau\right)}, \tag{19}$$

where $\mathcal{B}$ is the batch size, K is the queue size, and τ is the temperature. At each iteration, the current batch of key representations is appended to the queue, while the oldest entries are removed to maintain a fixed queue size.

The model shares the remaining optimization settings with MAE-style pretraining. After pretraining, the projection head and momentum key encoder are discarded, and only the query encoder is retained for frozen-representation evaluation. Table 5 summarizes the hyperparameters.

F Subject Identity Shortcut Interventions

F.1 MAE Pretraining with Euclidean-Aligned Targets

Reconstruction-based self-supervised learning may exploit subject-specific spatial covariance to predict masked EEG segments. We therefore apply Euclidean alignment (EA) He & Wu (2019) only to construct reconstruction targets, while the encoder receives the original EEG signal.

For subject s, let $\mathbf{X}_{s,n} \in \mathbb{R}^{C \times T}$ denote trial n, where C and T are the numbers of channels and time samples, respectively. The reference covariance matrix is computed from all N trials of subject s:

$$\bar{\mathbf{\Sigma}}_s = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{T-1} \mathbf{X}_{s,n} \mathbf{X}_{s,n}^{\mathsf{T}}. \tag{20}$$

The Euclidean-aligned reconstruction target for each trial is

$$\widetilde{\mathbf{X}}_{s,n} = \bar{\mathbf{\Sigma}}_s^{-\frac{1}{2}} \mathbf{X}_{s,n}. \tag{21}$$

This transformation makes the average channel covariance of each subject approach the identity matrix, reducing subject-related information carried by channel correlations and spatial covariance in the target signal.

During pretraining, the encoder receives visible tokens from the original signal $\mathbf{X}_{s,n}$, whereas the decoder reconstructs the corresponding Euclidean-aligned target $\widetilde{\mathbf{X}}_{s,n}$ at masked positions. Thus, the model predicts an EA target from partially observed original EEG rather than reconstructing its input directly. The architecture and optimization settings follow the standard MAE configuration, with global random masking at a ratio of 0.9.

F.2 MoCo Pretraining restricted to Same-Subject Negatives

In standard MoCo, representations of other trials in the dynamic queue serve as negative keys. Since EEG differences across subjects can be substantial, this construction allows subject identity to reduce the contrastive objective. To suppress this shortcut, a separate key queue is maintained for each subject, and only keys from the query subject are used as negatives.

Let $\mathbf{q}_i$ and $\mathbf{k}_i^+$ denote the normalized query and positive-key representations of trial i, whose subject identity is s_i. Let $\mathcal{Q}_{s_i}$ denote the key queue associated with subject s_i. The same-subject-negative contrastive loss is

$$\mathcal{L}_{\text{SSN-MoCo}} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\mathbf{q}_i^\top\mathbf{k}_i^+/\tau\right)}{\exp\left(\mathbf{q}_i^\top\mathbf{k}_i^+/\tau\right)+\sum_{\mathbf{k}_j^-\in\mathcal{Q}_{s_i}}\exp\left(\mathbf{q}_i^\top\mathbf{k}_j^-/\tau\right)}. \tag{22}$$

If the queue contains the same trial as the current query, that key is excluded from the negative set. Each subject queue has a capacity of 64. After each iteration, current key representations are added to their corresponding subject queues and the oldest keys are removed.

Because both the positive key and all negative keys belong to the same subject, subject identity cannot directly distinguish positive from negative pairs. Apart from subject-specific negative queues and their capacity, all model and optimization settings are identical to those of standard MoCo-style pretraining.

G Evaluation of Off-the-Shelf EEG Foundation Models

G.1 Kim2025

On Kim2025 (Table 6), pretraining improves subject BA and increases P_{subj} for all five EEG-FMs, with particularly large gains for REVE. In contrast, task BA changes inconsistently, improving for CBraMod, EEGPT, and REVE but declining for BIOT and LaBraM; P_{task} increases only for BIOT and CBraMod. In every pretrained model, P_{task} remains far below both P_{subj} and P_{inter}, while residual variation decreases. Thus, on this dataset, pretraining consistently reorganizes frozen representations around subject identity without a corresponding increase in shared task-related variation.

Figure 7 provides convergent qualitative evidence. Task-colored *t*-SNE embeddings remain largely mixed, whereas subject-colored embeddings show clearer local clustering. RSA likewise exhibits limited task-wise diagonal structure but pronounced heterogeneity across subject pairs, consistent with the linear-probing and variance-decomposition results.

Table 6: Linear-probe and variance decomposition of frozen representations on Kim2025.

Model	**Parameter**	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level		0.025	0.025	–	–	–	–
BIOT	RandomInit	**0.191**	0.472	0.0101	0.3046	**0.1023**	**0.5830**
	Pretrained	0.087	**0.610**	**0.0115**	**0.3843**	0.0986	0.5056
CBraMod	RandomInit	0.025	0.069	0.0067	0.0158	**0.1631**	**0.8143**
	Pretrained	**0.049**	**0.183**	**0.0078**	**0.2453**	0.1197	0.6272
EEGPT	RandomInit	0.052	0.185	**0.0119**	0.0452	**0.1647**	**0.7782**
	Pretrained	**0.066**	**0.347**	0.0103	**0.2038**	0.1372	0.6487
LaBraM	RandomInit	**0.033**	0.099	**0.0062**	0.0461	**0.1543**	**0.7934**
	Pretrained	0.024	**0.279**	0.0046	**0.0884**	0.1443	0.7627
REVE	RandomInit	0.060	0.062	**0.0171**	0.0216	**0.1645**	**0.7968**
	Pretrained	**0.063**	**0.815**	0.0129	**0.1227**	0.1431	0.7213

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively.

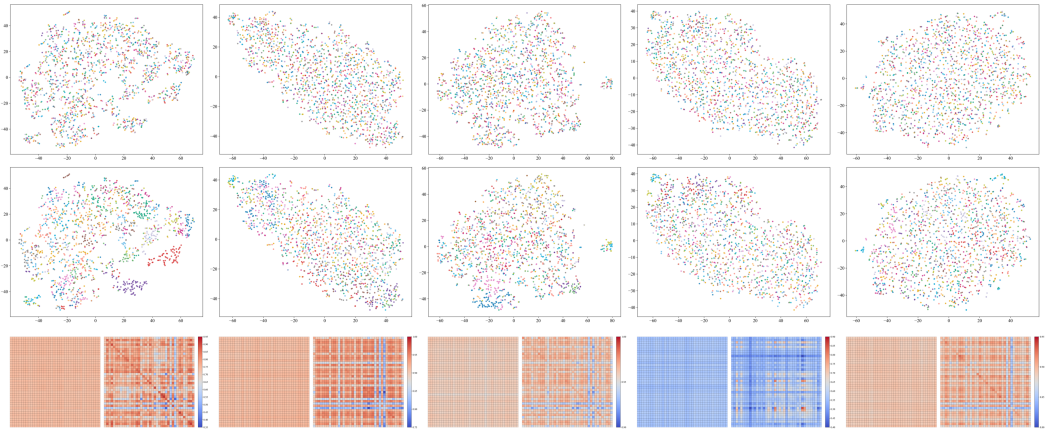


Figure 7: *t*-SNE and RSA of frozen representations on Kim2025. Columns show BIOT, CBraMod, EEGPT, LaBraM, and REVE from left to right. The top two rows show shared *t*-SNE coordinates colored by task labels and subject identity, respectively; the bottom row shows the corresponding task (left) and subject (right) RSA matrices. Best viewed via zoom-in.

G.2 PHYSIONET-P300

On PhysioNet-P300 (Table 7), task BA remains close to the 0.5 chance level and improves after pretraining only for EEGPT and LaBraM. In contrast, subject BA and P_{subj} increase for all five EEG-FMs. The task component remains negligible ($P_{\text{task}} \leq 0.0008$) and is lower than the subject, interaction, and residual components for every pretrained model, whereas residual variation decreases. Thus, pretraining again strengthens subject-related structure without consistently improving task decoding or task-related variation.

Figure 8 mirrors this pattern qualitatively. Task-colored *t*-SNE plots show little separation between the two classes, and task-wise RSA matrices are nearly uniform. Subject-colored embeddings, however, reveal clearer local structure, while subject-wise RSA matrices show substantial between-subject heterogeneity.

Table 7: Linear-probe and variance decomposition of frozen representations on PhysioNet-P300.

Model	**Parameter**	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level		0.500	0.083	–	–	–	–
BIOT	RandomInit	**0.522**	0.496	**0.0002**	0.2889	0.0003	**0.7106**
	Pretrained	0.506	**0.624**	<0.0001	**0.4024**	**0.0004**	0.5972
CBraMod	RandomInit	**0.551**	0.244	**0.0005**	0.0101	**0.0017**	**0.9877**
	Pretrained	0.532	**0.397**	0.0002	**0.2084**	0.0008	0.7906
EEGPT	RandomInit	0.501	0.499	**<0.0001**	0.0372	**0.0014**	**0.9613**
	Pretrained	**0.519**	**0.637**	<0.0001	**0.1951**	0.0012	0.8036
LaBraM	RandomInit	0.538	0.253	**0.0006**	0.0046	**0.0015**	**0.9933**
	Pretrained	**0.542**	**0.541**	0.0002	**0.0545**	0.0008	0.9445
REVE	RandomInit	**0.591**	0.288	**0.0009**	0.0090	0.0019	**0.9881**
	Pretrained	0.590	**0.970**	0.0008	**0.0910**	**0.0021**	0.9062

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively.

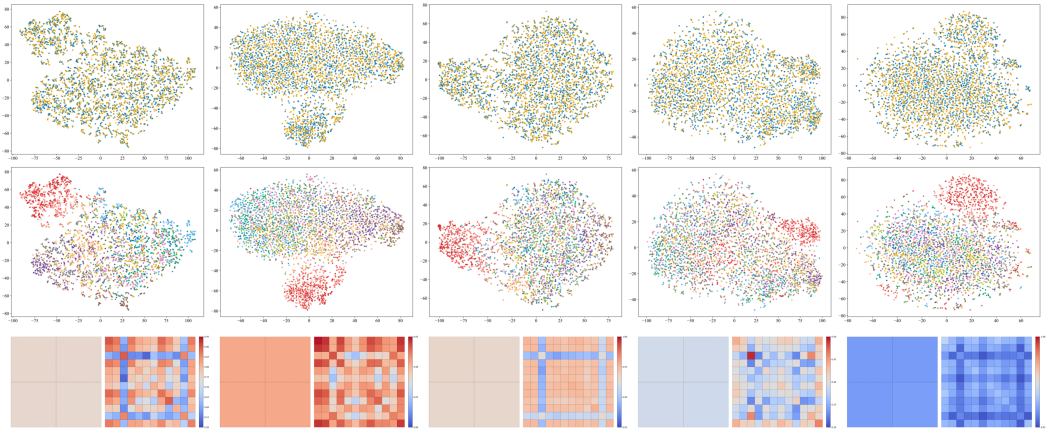


Figure 8: *t*-SNE and RSA of frozen representations on PhysioNet-P300. Columns show BIOT, CBraMod, EEGPT, LaBraM, and REVE from left to right. The top two rows show shared *t*-SNE coordinates colored by task labels and subject identity, respectively; the bottom row shows the corresponding task (left) and subject (right) RSA matrices. Best viewed via zoom-in.

G.3 SEED-V

On SEED-V (Table 8), pretraining improves both task and subject BA for all five EEG-FMs, with the largest task gain for REVE. However, the variance decomposition remains dominated by subject-related structure: P_{task} is lower than P_{subj}, P_{inter}, and P_{res} for every pretrained model. Although P_{task} increases in three models, it remains small and does not provide a consistent account of the task-BA gains. Meanwhile, P_{subj} increases in four models; BIOT is the exception, although its subject component remains high after pretraining. Thus, improved task decoding coincides with, rather than clearly reflects, stronger shared task-related variation.

Figure 9 further illustrates this dissociation. Task labels remain mixed within the local *t*-SNE clusters and show little diagonal structure in the task-wise RSA matrices. In contrast, subject coloring separates these clusters more clearly, and subject-wise RSA reveals pronounced between-subject heterogeneity.

Table 8: Linear-probe and variance decomposition of frozen representations on SEED-V.

Model	**Parameter**	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level		0.200	0.063	–	–	–	–
BIOT	RandomInit	0.268	0.738	0.0080	**0.4363**	0.0247	0.5310
	Pretrained	**0.287**	**0.777**	**0.0141**	0.4038	**0.0279**	**0.5542**
CBraMod	RandomInit	0.221	0.274	**0.0320**	0.0008	0.0040	**0.9633**
	Pretrained	**0.258**	**0.475**	0.0091	**0.3594**	**0.0208**	0.6107
EEGPT	RandomInit	0.263	0.601	**0.0526**	0.0016	0.0061	**0.9398**
	Pretrained	**0.288**	**0.785**	0.0054	**0.2557**	**0.0195**	0.7195
LaBraM	RandomInit	0.221	0.307	0.0009	0.0386	0.0045	**0.9560**
	Pretrained	**0.266**	**0.684**	**0.0053**	**0.1639**	**0.0120**	0.8188
REVE	RandomInit	0.214	0.230	0.0003	0.0053	0.0027	**0.9916**
	Pretrained	**0.332**	**0.922**	**0.0030**	**0.1502**	**0.0095**	0.8373

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively.

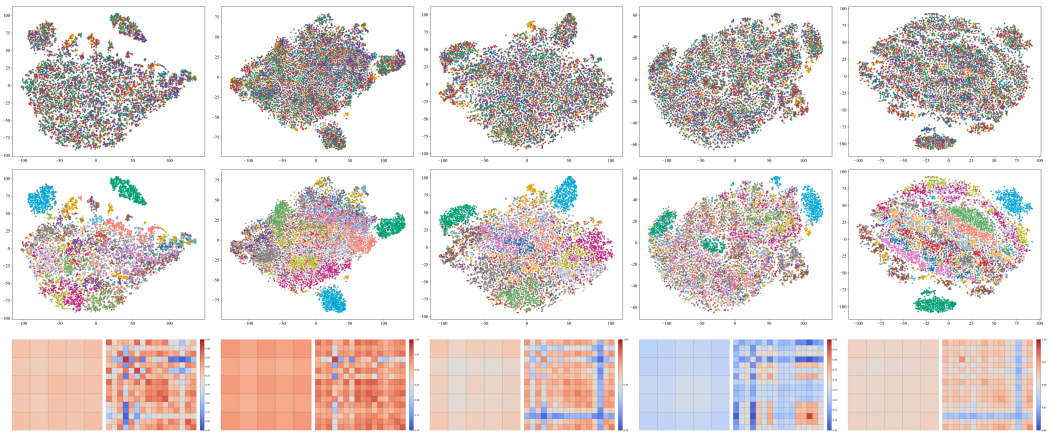


Figure 9: *t*-SNE and RSA of frozen representations on SEED-V. Columns show BIOT, CBraMod, EEGPT, LaBraM, and REVE from left to right. The top two rows show shared *t*-SNE coordinates colored by task labels and subject identity, respectively; the bottom row shows the corresponding task (left) and subject (right) RSA matrices. Best viewed via zoom-in.

G.4 TSINGHUA-RSVP

On Tsinghua-RSVP (Table 9), task BA is already above chance under random initialization, but pretraining improves it only slightly for EEGPT and REVE and reduces it for the other three models. In contrast, subject BA and P_{subj} increase for all five models, accompanied by lower residual variation. P_{task} increases only for EEGPT and declines otherwise. Thus, pretraining consistently amplifies subject-related structure without reliably improving task decoding or task-related variation.

Figure 10 provides corresponding qualitative evidence. Task-colored *t*-SNE embeddings show limited class separation, and the task-wise RSA matrices lack a consistent diagonal pattern. Subject-colored embeddings instead reveal clearer local clustering, with marked heterogeneity in subject-wise RSA matrices.

Table 9: Linear-probe and variance decomposition of frozen representations on Tsinghua-RSVP.

Model	**Parameter**	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level		0.500	0.071	–	–	–	–
BIOT	RandomInit	**0.582**	0.428	**0.0028**	0.1962	0.0015	**0.7995**
	Pretrained	0.572	**0.457**	0.0017	**0.2864**	**0.0018**	0.7101
CBraMod	RandomInit	**0.696**	0.161	**0.0106**	0.0133	**0.0064**	**0.9696**
	Pretrained	0.624	**0.330**	0.0032	**0.2093**	0.0029	0.7845
EEGPT	RandomInit	0.577	0.390	0.0008	0.0108	**0.0022**	**0.9863**
	Pretrained	**0.582**	**0.425**	**0.0019**	**0.0872**	0.0017	0.9092
LaBraM	RandomInit	**0.720**	0.161	**0.0136**	0.0135	**0.0091**	**0.9637**
	Pretrained	0.644	**0.397**	0.0013	**0.0582**	0.0020	0.9385
REVE	RandomInit	0.743	0.226	**0.0090**	0.0150	**0.0125**	**0.9636**
	Pretrained	**0.746**	**0.514**	0.0056	**0.0566**	0.0055	0.9323

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively.

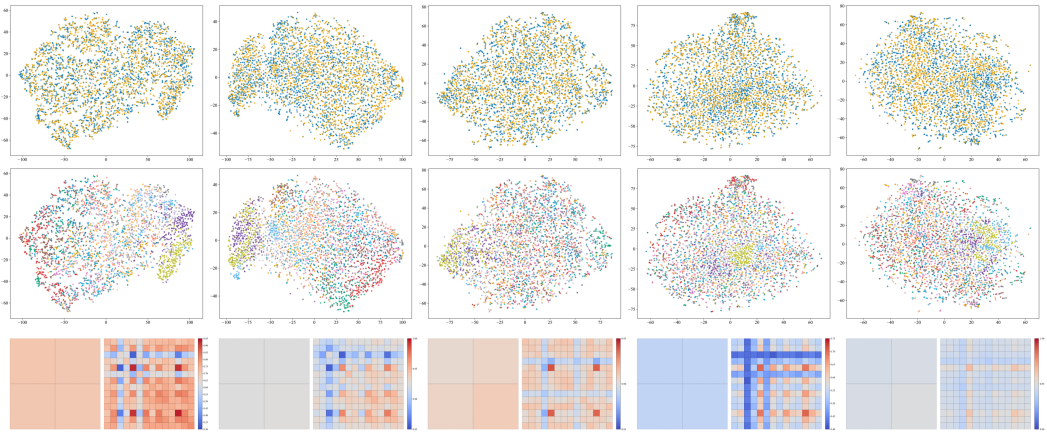


Figure 10: *t*-SNE and RSA of frozen representations on Tsinghua-RSVP. Columns show BIOT, CBraMod, EEGPT, LaBraM, and REVE from left to right. The top two rows show shared *t*-SNE coordinates colored by task labels and subject identity, respectively; the bottom row shows the corresponding task (left) and subject (right) RSA matrices. Best viewed via zoom-in.

H EVALUATION OF MAE-STYLE MODELS

All 27 MAE-style models use the same pretraining corpus and settings, differing only in masking configuration; linear probes use a common seed. Across datasets, pretraining consistently improves subject-identity decoding and subject-related variation over random initialization. Task gains vary by dataset and configuration and, even when large, are not accompanied by comparably large task-attributable variation. Figure 4 shows *t*-SNE and RSA for the task-BA-optimal configuration on each dataset; other configurations exhibit similar task–subject patterns and are omitted. Detailed linear-probe and variance-decomposition results for all masking configurations are reported in the following tables.

H.1 BCI-IV-2A

On BCI-IV-2a (Table 10), all masking configurations substantially improve subject BA over random initialization, while 25 of 27 configurations also improve task BA. The highest task BA (0.404) is achieved by temporal masking at $\rho = 0.8$. However, variance decomposition consistently remains dominated by non-task components: P_{task} is lower than P_{subj}, P_{inter}, and P_{res} under every masking configuration. Relative to random initialization, P_{subj} increases substantially and P_{res} decreases throughout, whereas P_{task} remains small even at its maximum (0.0109). Thus, although suitable masking configurations substantially improve task decoding on BCI-IV-2a, these gains coincide with strong subject-related and subject-dependent representational variation rather than a comparably large increase in shared task-related variation.

Table 10: Linear-probe and variance decomposition of frozen representations on BCI-IV-2a.

Parameter	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level	0.250	0.111	–	–	–	–
RandomInit	0.297	0.289	0.0013	0.0150	0.0059	**0.9779**
Random mask ($\rho = 0.1$)	0.306	0.585	0.0028	0.1345	0.0061	0.8566
Random mask ($\rho = 0.2$)	0.309	0.644	0.0028	0.1709	0.0059	0.8204
Random mask ($\rho = 0.3$)	0.329	0.602	0.0020	0.3053	0.0045	0.6882
Random mask ($\rho = 0.4$)	0.269	0.644	0.0025	0.1952	0.0058	0.7965
Random mask ($\rho = 0.5$)	0.277	0.653	0.0022	0.3217	0.0050	0.6710
Random mask ($\rho = 0.6$)	0.309	0.719	0.0030	0.1988	0.0065	0.7918
Random mask ($\rho = 0.7$)	0.316	0.675	0.0025	0.2828	0.0056	0.7091
Random mask ($\rho = 0.8$)	0.302	0.796	0.0031	0.2895	0.0066	0.7009
Random mask ($\rho = 0.9$)	0.373	0.853	0.0050	0.3350	0.0090	0.6510
Random mask ($\rho = 0.95$)	0.356	0.875	0.0063	0.3792	0.0091	0.6054
Random mask ($\rho = 0.96$)	0.359	0.873	0.0037	0.4137	0.0074	0.5751
Random mask ($\rho = 0.97$)	0.360	0.865	0.0040	0.4198	0.0080	0.5682
Random mask ($\rho = 0.98$)	0.329	0.805	0.0035	0.4210	0.0073	0.5682
Random mask ($\rho = 0.99$)	0.343	0.748	0.0045	0.3416	0.0067	0.6472
Random mask ($\rho = 0.995$)	0.318	0.685	0.0036	0.3617	0.0055	0.6292
Spatial mask ($\rho = 0.5$)	0.371	0.861	0.0030	0.3582	0.0071	0.6317
Spatial mask ($\rho = 0.6$)	0.364	0.881	0.0046	0.3653	0.0084	0.6217
Spatial mask ($\rho = 0.7$)	0.379	0.868	0.0039	0.3925	0.0074	0.5961
Spatial mask ($\rho = 0.8$)	0.364	0.843	0.0039	0.3766	0.0073	0.6122
Spatial mask ($\rho = 0.9$)	0.355	0.869	0.0042	0.4008	0.0086	0.5863
Spatial mask ($\rho = 0.95$)	0.343	0.806	0.0030	0.3739	0.0069	0.6162
Temporal mask ($\rho = 0.5$)	0.369	**0.885**	0.0092	0.4179	0.0176	0.5553
Temporal mask ($\rho = 0.6$)	0.382	0.843	0.0091	0.4199	**0.0190**	0.5521
Temporal mask ($\rho = 0.7$)	0.369	0.867	0.0058	0.4740	0.0121	0.5081
Temporal mask ($\rho = 0.8$)	**0.404**	0.883	0.0109	0.4475	0.0176	0.5240
Temporal mask ($\rho = 0.9$)	0.370	0.835	**0.0109**	0.4454	0.0149	0.5288
Temporal mask ($\rho = 0.95$)	0.315	0.691	0.0048	0.3945	0.0074	0.5933

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the maximum value within each column.

H.2 KIM2025

On Kim2025 (Table 11), all masking configurations substantially improve subject BA over random initialization, reaching a maximum of 0.871 under spatial masking at $\rho = 0.5$. Task BA improves in 20 of 27 configurations, but remains modest overall; its maximum (0.069) is obtained with spatial masking at $\rho = 0.95$. Variance decomposition again shows that pretrained representations are dominated by subject-related structure: under every masking configuration, P_{task} is lower than P_{subj}, P_{inter}, and P_{res}. Moreover, P_{subj} increases in all configurations, whereas P_{task} exceeds its random-initialization value in only three. Thus, the masking configurations that yield the strongest task decoding on Kim2025 still primarily reorganize representation-space variation around subject identity rather than shared task-related structure.

Table 11: Linear-probe and variance decomposition of frozen representations on Kim2025.

Parameter	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level	0.025	0.025	–	–	–	–
RandomInit	0.046	0.059	0.0097	0.0183	**0.1731**	**0.7989**
Random mask ($\rho = 0.1$)	0.045	0.443	0.0061	0.1962	0.1282	0.6694
Random mask ($\rho = 0.2$)	0.041	0.421	0.0061	0.2161	0.1263	0.6515
Random mask ($\rho = 0.3$)	0.042	0.542	0.0055	0.2479	0.1170	0.6296
Random mask ($\rho = 0.4$)	0.040	0.538	0.0053	0.2383	0.1215	0.6349
Random mask ($\rho = 0.5$)	0.048	0.584	0.0058	0.2300	0.1218	0.6425
Random mask ($\rho = 0.6$)	0.042	0.629	0.0055	0.2558	0.1176	0.6211
Random mask ($\rho = 0.7$)	0.042	0.633	0.0058	0.2258	0.1246	0.6438
Random mask ($\rho = 0.8$)	0.052	0.654	0.0062	0.2359	0.1235	0.6343
Random mask ($\rho = 0.9$)	0.057	0.702	0.0068	0.3173	0.1069	0.5690
Random mask ($\rho = 0.95$)	0.060	0.713	0.0074	0.3091	0.1084	0.5751
Random mask ($\rho = 0.96$)	0.056	0.708	0.0074	0.3075	0.1091	0.5760
Random mask ($\rho = 0.97$)	0.061	0.702	0.0072	0.2901	0.1124	0.5903
Random mask ($\rho = 0.98$)	0.054	0.674	0.0071	0.3045	0.1112	0.5772
Random mask ($\rho = 0.99$)	0.060	0.597	0.0090	0.2611	0.1177	0.6122
Random mask ($\rho = 0.995$)	0.054	0.481	0.0073	0.2639	0.1152	0.6136
Spatial mask ($\rho = 0.5$)	0.058	**0.871**	0.0048	**0.4615**	0.0862	0.4475
Spatial mask ($\rho = 0.6$)	0.055	0.864	0.0046	0.4406	0.0880	0.4667
Spatial mask ($\rho = 0.7$)	0.055	0.814	0.0062	0.3634	0.1005	0.5299
Spatial mask ($\rho = 0.8$)	0.063	0.842	0.0071	0.3575	0.1046	0.5308
Spatial mask ($\rho = 0.9$)	0.061	0.777	0.0083	0.3290	0.1091	0.5536
Spatial mask ($\rho = 0.95$)	**0.069**	0.666	0.0104	0.3103	0.1121	0.5672
Temporal mask ($\rho = 0.5$)	0.062	0.664	**0.0127**	0.3088	0.1097	0.5688
Temporal mask ($\rho = 0.6$)	0.048	0.394	0.0102	0.1847	0.1312	0.6740
Temporal mask ($\rho = 0.7$)	0.054	0.713	0.0083	0.3139	0.1061	0.5716
Temporal mask ($\rho = 0.8$)	0.052	0.726	0.0070	0.3305	0.1011	0.5614
Temporal mask ($\rho = 0.9$)	0.048	0.703	0.0057	0.3396	0.0980	0.5567
Temporal mask ($\rho = 0.95$)	0.046	0.437	0.0063	0.3084	0.1019	0.5834

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the maximum value within each column.

H.3 PHYSIONET-P300

On PhysioNet-P300 (Table 12), task BA remains close to its random-initialization value (0.537) across masking configurations. Only 6 of 27 configurations improve task BA, with the maximum (0.553) obtained under temporal masking at $\rho = 0.6$. In contrast, every configuration substantially improves subject BA, from 0.294 at random initialization to 0.771–0.998 after pretraining. Variance decomposition shows the same dissociation: P_{subj} increases in every configuration and P_{res} decreases, whereas P_{task} remains extremely small (≤ 0.0003) and is lower than P_{subj}, P_{inter}, and P_{res} throughout. Thus, on PhysioNet-P300, masking-based pretraining strongly increases subject-identity decodability without a corresponding, reliable improvement in task decoding or shared task-related variation.

Table 12: Linear-probe and variance decomposition of frozen representations on PhysioNet-P300.

Parameter	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level	0.500	0.083	–	–	–	–
RandomInit	0.537	0.294	0.0002	0.0100	**0.0013**	**0.9886**
Random mask ($\rho = 0.1$)	0.528	0.805	0.0002	0.1112	0.0010	0.8875
Random mask ($\rho = 0.2$)	0.538	0.904	0.0002	0.1892	0.0009	0.8097
Random mask ($\rho = 0.3$)	0.533	0.927	0.0002	0.2294	0.0008	0.7697
Random mask ($\rho = 0.4$)	0.526	0.929	0.0002	0.2149	0.0008	0.7841
Random mask ($\rho = 0.5$)	0.524	0.939	0.0002	0.2262	0.0008	0.7728
Random mask ($\rho = 0.6$)	0.530	0.967	0.0001	0.2808	0.0007	0.7184
Random mask ($\rho = 0.7$)	0.526	0.962	0.0001	0.2357	0.0007	0.7634
Random mask ($\rho = 0.8$)	0.529	0.957	0.0001	0.2249	0.0007	0.7742
Random mask ($\rho = 0.9$)	0.533	0.945	0.0002	0.2694	0.0006	0.7298
Random mask ($\rho = 0.95$)	0.533	0.922	0.0001	0.2845	0.0006	0.7148
Random mask ($\rho = 0.96$)	0.529	0.906	0.0001	0.2684	0.0005	0.7310
Random mask ($\rho = 0.97$)	0.531	0.909	<0.0001	0.2368	0.0005	0.7626
Random mask ($\rho = 0.98$)	0.527	0.890	<0.0001	0.2529	0.0005	0.7465
Random mask ($\rho = 0.99$)	0.540	0.848	**0.0003**	0.2546	0.0007	0.7443
Random mask ($\rho = 0.995$)	0.539	0.771	0.0003	0.2560	0.0007	0.7431
Spatial mask ($\rho = 0.5$)	0.511	0.996	<0.0001	0.3397	0.0006	0.6596
Spatial mask ($\rho = 0.6$)	0.513	**0.998**	<0.0001	0.3835	0.0005	0.6159
Spatial mask ($\rho = 0.7$)	0.526	0.995	<0.0001	0.3600	0.0005	0.6393
Spatial mask ($\rho = 0.8$)	0.514	0.978	<0.0001	0.2998	0.0006	0.6995
Spatial mask ($\rho = 0.9$)	0.530	0.941	<0.0001	0.2734	0.0005	0.7260
Spatial mask ($\rho = 0.95$)	0.533	0.888	<0.0001	0.2161	0.0005	0.7833
Temporal mask ($\rho = 0.5$)	0.537	0.923	0.0002	0.3236	0.0005	0.6758
Temporal mask ($\rho = 0.6$)	**0.553**	0.937	0.0001	0.3721	0.0004	0.6274
Temporal mask ($\rho = 0.7$)	0.535	0.943	0.0001	0.3816	0.0004	0.6178
Temporal mask ($\rho = 0.8$)	0.544	0.939	0.0001	0.4191	0.0004	0.5804
Temporal mask ($\rho = 0.9$)	0.540	0.931	0.0001	0.4016	0.0004	0.5979
Temporal mask ($\rho = 0.95$)	0.517	0.798	<0.0001	0.3842	0.0003	0.6154

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the maximum value within each column.

H.4 SEED-V

On SEED-V (Table 13), all masking configurations improve both task and subject BA over random initialization. The highest task BA (0.375) is achieved by spatial masking at $\rho = 0.5$, which also yields the highest subject BA (0.945) and the largest interaction component ($P_{\text{inter}} = 0.0283$). Although P_{task} increases under every configuration, it remains small (0.0063–0.0108) and lower than P_{subj}, P_{inter}, and P_{res} throughout. Meanwhile, P_{subj} increases substantially relative to random initialization, accompanied by lower P_{res}. Thus, on SEED-V, strong task-decoding gains co-occur with strong subject-identity decoding and subject–task interaction, rather than a comparably large increase in shared task-related variation.

Table 13: Linear-probe and variance decomposition of frozen representations on SEED-V.

Parameter	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level	0.200	0.062	–	–	–	–
RandomInit	0.222	0.279	0.0005	0.0259	0.0040	**0.9695**
Random mask ($\rho = 0.1$)	0.315	0.864	0.0090	0.4116	0.0203	0.5592
Random mask ($\rho = 0.2$)	0.315	0.873	0.0087	0.4226	0.0207	0.5480
Random mask ($\rho = 0.3$)	0.319	0.907	0.0097	0.4455	0.0216	0.5232
Random mask ($\rho = 0.4$)	0.313	0.901	0.0087	0.4296	0.0214	0.5403
Random mask ($\rho = 0.5$)	0.320	0.908	0.0104	0.4381	0.0236	0.5279
Random mask ($\rho = 0.6$)	0.330	0.912	0.0087	0.4350	0.0230	0.5332
Random mask ($\rho = 0.7$)	0.329	0.920	0.0080	0.4305	0.0216	0.5399
Random mask ($\rho = 0.8$)	0.349	0.928	0.0091	0.4308	0.0226	0.5375
Random mask ($\rho = 0.9$)	0.347	0.922	0.0076	0.4676	0.0233	0.5016
Random mask ($\rho = 0.95$)	0.355	0.928	0.0086	0.4305	0.0251	0.5359
Random mask ($\rho = 0.96$)	0.348	0.918	0.0084	0.3997	0.0260	0.5659
Random mask ($\rho = 0.97$)	0.350	0.921	0.0089	0.3949	0.0272	0.5690
Random mask ($\rho = 0.98$)	0.345	0.893	0.0093	0.3783	0.0261	0.5863
Random mask ($\rho = 0.99$)	0.335	0.872	0.0070	0.3764	0.0221	0.5945
Random mask ($\rho = 0.995$)	0.330	0.834	0.0064	0.4046	0.0240	0.5650
Spatial mask ($\rho = 0.5$)	**0.375**	**0.945**	0.0082	0.4571	**0.0283**	0.5065
Spatial mask ($\rho = 0.6$)	0.352	0.918	0.0089	0.4422	0.0282	0.5207
Spatial mask ($\rho = 0.7$)	0.360	0.933	0.0084	0.4720	0.0279	0.4917
Spatial mask ($\rho = 0.8$)	0.358	0.938	0.0085	0.4628	0.0267	0.5020
Spatial mask ($\rho = 0.9$)	0.359	0.939	0.0089	0.4786	0.0267	0.4858
Spatial mask ($\rho = 0.95$)	0.355	0.935	0.0087	**0.4822**	0.0264	0.4827
Temporal mask ($\rho = 0.5$)	0.347	0.942	0.0072	0.4253	0.0228	0.5448
Temporal mask ($\rho = 0.6$)	0.344	0.921	0.0069	0.4188	0.0232	0.5511
Temporal mask ($\rho = 0.7$)	0.361	0.916	0.0065	0.4181	0.0217	0.5537
Temporal mask ($\rho = 0.8$)	0.355	0.902	0.0063	0.4098	0.0229	0.5610
Temporal mask ($\rho = 0.9$)	0.344	0.893	0.0070	0.3932	0.0239	0.5759
Temporal mask ($\rho = 0.95$)	0.323	0.839	**0.0108**	0.4280	0.0278	0.5334

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the maximum value within each column.

H.5 TSINGHUA-RSVP

On Tsinghua-RSVP (Table 14), all masking configurations improve both task and subject BA over random initialization. Task BA reaches 0.711 under temporal masking at $\rho = 0.6$, while subject BA reaches 0.695 under random masking at $\rho = 0.97$. However, variance decomposition shows no corresponding increase in shared task-related variation: P_{task} is lower than its random-initialization value (0.0123) under every masking configuration. P_{inter} likewise decreases throughout, whereas P_{subj} increases substantially and P_{res} decreases. Thus, although masking-based pretraining consistently improves task decoding on Tsinghua-RSVP, these gains do not reflect greater global task-attributable variation in the frozen representation space.

Table 14: Linear-probe and variance decomposition of frozen representations on Tsinghua-RSVP.

Parameter	**Task BA**	**Subject BA**	P_{task}	P_{subj}	P_{inter}	P_{res}
Chance level	0.500	0.071	–	–	–	–
RandomInit	0.634	0.248	**0.0123**	0.0253	**0.0062**	**0.9562**
Random mask ($\rho = 0.1$)	0.641	0.472	0.0074	0.1077	0.0043	0.8806
Random mask ($\rho = 0.2$)	0.670	0.525	0.0056	0.2073	0.0043	0.7828
Random mask ($\rho = 0.3$)	0.687	0.515	0.0063	0.2164	0.0037	0.7736
Random mask ($\rho = 0.4$)	0.679	0.542	0.0053	0.2287	0.0034	0.7627
Random mask ($\rho = 0.5$)	0.679	0.573	0.0052	0.2512	0.0036	0.7400
Random mask ($\rho = 0.6$)	0.674	0.553	0.0049	0.2439	0.0031	0.7481
Random mask ($\rho = 0.7$)	0.685	0.662	0.0045	0.2136	0.0038	0.7781
Random mask ($\rho = 0.8$)	0.688	0.621	0.0047	0.2320	0.0046	0.7587
Random mask ($\rho = 0.9$)	0.692	0.680	0.0044	0.2703	0.0030	0.7222
Random mask ($\rho = 0.95$)	0.708	0.682	0.0042	0.2520	0.0029	0.7409
Random mask ($\rho = 0.96$)	0.696	0.655	0.0042	0.2441	0.0029	0.7487
Random mask ($\rho = 0.97$)	0.701	**0.695**	0.0040	0.2186	0.0030	0.7744
Random mask ($\rho = 0.98$)	0.693	0.666	0.0041	0.2182	0.0029	0.7748
Random mask ($\rho = 0.99$)	0.683	0.610	0.0050	0.2245	0.0037	0.7668
Random mask ($\rho = 0.995$)	0.670	0.576	0.0043	0.2155	0.0042	0.7760
Spatial mask ($\rho = 0.5$)	0.649	0.654	0.0025	**0.3642**	0.0018	0.6315
Spatial mask ($\rho = 0.6$)	0.639	0.684	0.0031	0.2866	0.0024	0.7078
Spatial mask ($\rho = 0.7$)	0.669	0.628	0.0030	0.3398	0.0019	0.6552
Spatial mask ($\rho = 0.8$)	0.676	0.685	0.0030	0.3152	0.0020	0.6797
Spatial mask ($\rho = 0.9$)	0.683	0.686	0.0031	0.3050	0.0021	0.6897
Spatial mask ($\rho = 0.95$)	0.691	0.662	0.0028	0.2083	0.0027	0.7862
Temporal mask ($\rho = 0.5$)	0.710	0.614	0.0031	0.2511	0.0026	0.7432
Temporal mask ($\rho = 0.6$)	**0.711**	0.631	0.0035	0.2487	0.0027	0.7451
Temporal mask ($\rho = 0.7$)	0.708	0.628	0.0032	0.2664	0.0027	0.7277
Temporal mask ($\rho = 0.8$)	0.687	0.590	0.0034	0.2641	0.0026	0.7299
Temporal mask ($\rho = 0.9$)	0.694	0.563	0.0029	0.2837	0.0026	0.7109
Temporal mask ($\rho = 0.95$)	0.645	0.581	0.0032	0.2943	0.0027	0.6999

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the maximum value within each column.

I EVALUATION OF MOCO-STYLE MODELS

Table 15 shows that MoCo-style pretraining increases subject BA and P_{subj} across all five downstream datasets, accompanied by lower P_{res}. Task BA improves on four datasets, but the gains are dataset-dependent and negligible on BCI-IV-2a; it decreases on Kim2025. Likewise, P_{task} increases on only three datasets and decreases on Kim2025 and Tsinghua-RSVP. In every pretrained model–dataset pair, P_{task} remains lower than P_{subj}, P_{inter}, and P_{res}.

The *t*-SNE and RSA visualizations in Figure 11 provide a qualitative complement, showing weaker task-wise organization than subject-wise structure. Thus, queue-based contrastive pretraining reproduces the preferential strengthening of subject identity, without a consistent increase in shared task-related variation.

Table 15: Linear-probe and variance decomposition of frozen representations from MoCo-style pretraining across downstream datasets.

Dataset	Parameter	Task BA	Subject BA	P_{task}	P_{subj}	P_{inter}	P_{res}
BCI-IV-2a	RandomInit	0.297	0.289	0.0013	0.0150	0.0059	**0.9779**
	Pretrained	**0.299**	**0.656**	**0.0021**	**0.0628**	**0.0075**	0.9276
Kim2025	RandomInit	**0.046**	0.059	**0.0097**	0.0183	**0.1731**	**0.7989**
	Pretrained	0.040	**0.373**	0.0080	**0.0666**	0.1526	0.7727
PhysioNet-P300	RandomInit	0.537	0.294	0.0002	0.0100	**0.0013**	**0.9886**
	Pretrained	**0.552**	**0.636**	**0.0004**	**0.0867**	0.0010	0.9118
SEED-V	RandomInit	0.222	0.279	0.0005	0.0259	0.0040	**0.9695**
	Pretrained	**0.279**	**0.784**	**0.0029**	**0.1074**	**0.0089**	0.8809
Tsinghua-RSVP	RandomInit	0.634	0.248	**0.0123**	0.0253	0.0062	**0.9562**
	Pretrained	**0.703**	**0.538**	0.0053	**0.0954**	**0.0064**	0.8929

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the larger value between random initialization and pretraining within each dataset and metric.

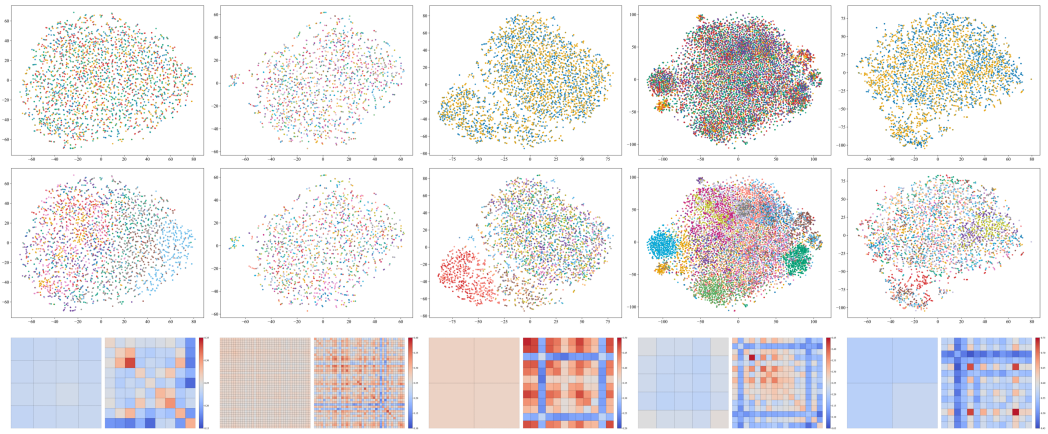


Figure 11: *t*-SNE and RSA of frozen MoCo-style representations. Columns correspond, from left to right, to BCI-IV-2a, Kim2025, PhysioNet-P300, SEED-V, and Tsinghua-RSVP datasets.

J Interventions on Subject Identity Shortcuts

J.1 MAE Pretraining with Euclidean-Aligned Targets

Table 16 shows that Euclidean alignment of MAE reconstruction targets does not consistently suppress subject-identity information. Relative to MAE, EA-MAE increases subject BA on BCI-IV-2a, Kim2025, and PhysioNet-P300, while P_{subj} increases on four of five datasets. Subject identity therefore remains strongly decodable after alignment. In contrast, task BA decreases on four datasets and improves only on Tsinghua-RSVP; its associated P_{task} is likewise inconsistent, increasing only on Kim2025 and SEED-V.

Figure 12 provides a qualitative complement, with *t*-SNE and RSA continuing to show stronger subject-wise than task-wise structure. Thus, aligning subject-level covariance in reconstruction targets does not reliably remove subject-specific information or improve task decoding.

Table 16: Linear-probe and variance decomposition of frozen representations before and after MAE pretraining with Euclidean-aligned targets across downstream datasets.

Dataset	Configuration	Task BA	Subject BA	P_{task}	P_{subj}	P_{inter}	P_{res}
BCI-IV-2a	MAE	**0.373**	0.853	**0.0050**	0.3350	**0.0090**	**0.6510**
	EA-MAE	0.365	**0.884**	0.0040	**0.4158**	0.0075	0.5727
Kim2025	MAE	**0.057**	0.702	0.0068	0.3173	**0.1069**	**0.5690**
	EA-MAE	0.052	**0.769**	**0.0070**	**0.3530**	0.0997	0.5402
PhysioNet-P300	MAE	**0.533**	0.945	**0.0002**	0.2694	**0.0006**	**0.7298**
	EA-MAE	0.522	**0.966**	<0.0001	**0.3848**	0.0004	0.6147
SEED-V	MAE	**0.347**	**0.922**	0.0076	**0.4676**	0.0233	0.5016
	EA-MAE	0.339	0.915	**0.0089**	0.4543	**0.0246**	**0.5121**
Tsinghua-RSVP	MAE	0.692	**0.680**	**0.0044**	0.2703	**0.0030**	**0.7222**
	EA-MAE	**0.713**	0.592	0.0027	**0.3607**	0.0023	0.6343

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the larger value between MAE and EA-MAE within each dataset and metric.

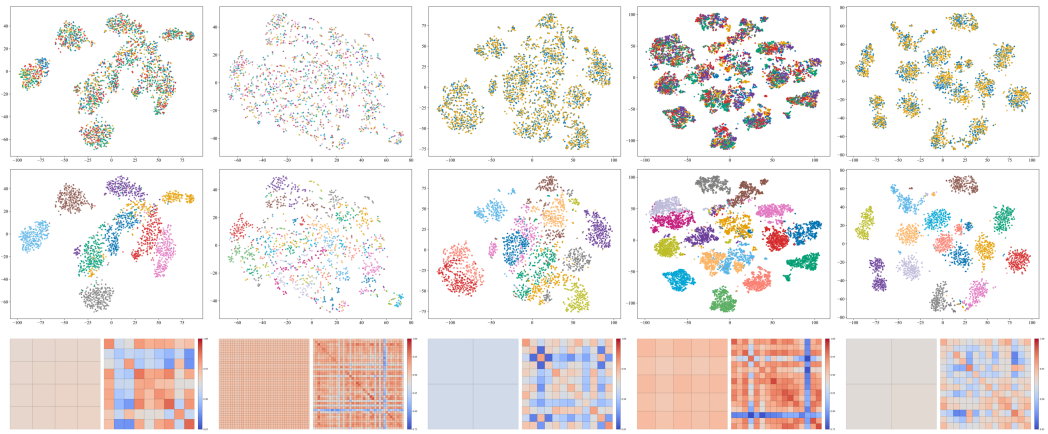


Figure 12: *t*-SNE and RSA of frozen representations from EA-MAE. Columns correspond, from left to right, to BCI-IV-2a, Kim2025, PhysioNet-P300, SEED-V, and Tsinghua-RSVP.

J.2 MoCo Pretraining restricted to Same-Subject Negatives

Table 17 shows that restricting MoCo negatives to the query subject consistently suppresses subject-identity information. Relative to standard MoCo, SSN-MoCo decreases both subject BA and P_{subj} across all five datasets, accompanied by higher P_{res}. Task BA increases on BCI-IV-2a, Kim2025, PhysioNet-P300, and Tsinghua-RSVP, but decreases on SEED-V; P_{task} increases on only three datasets and decreases on the other two. Thus, reduced subject-identity encoding does not yield a consistent increase in shared task-related variation.

Figure 13 visualizes the corresponding frozen representations using *t*-SNE and RSA. task-colored *t*-SNE embeddings remain mixed, and task RSA matrices show no clearer diagonal structure. In contrast, subject-colored local clusters and the heterogeneity of subject RSA matrices are markedly attenuated, consistent with reduced subject-identity encoding.

Table 17: Linear-probe and variance decomposition of frozen representations from standard MoCo and MoCo restricted to same-subject negatives across downstream datasets.

Dataset	Configuration	Task BA	Subject BA	P_{task}	P_{subj}	P_{inter}	P_{res}
BCI-IV-2a	MoCo	0.299	**0.656**	**0.0021**	**0.0628**	**0.0075**	0.9276
	SSN-MoCo	**0.313**	0.414	0.0014	0.0125	0.0068	**0.9793**
Kim2025	MoCo	0.040	**0.373**	0.0080	**0.0666**	0.1526	0.7727
	SSN-MoCo	**0.046**	0.070	**0.0084**	0.0101	**0.1726**	**0.8089**
PhysioNet-P300	MoCo	0.552	**0.636**	0.0004	**0.0867**	0.0010	0.9118
	SSN-MoCo	**0.580**	0.373	**0.0008**	0.0116	**0.0018**	**0.9858**
SEED-V	MoCo	**0.279**	**0.784**	**0.0029**	**0.1074**	**0.0089**	0.8809
	SSN-MoCo	0.219	0.314	0.0003	0.0066	0.0025	**0.9907**
Tsinghua-RSVP	MoCo	0.703	**0.538**	0.0053	**0.0954**	**0.0064**	0.8929
	SSN-MoCo	**0.724**	0.285	**0.0068**	0.0186	0.0063	**0.9683**

Note: BA denotes balanced accuracy. P_{task}, P_{subj}, P_{inter}, and P_{res} denote the proportions of variation associated with task labels, subject identity, their interaction, and residual variation, respectively. Boldface denotes the larger value between MoCo and SSN-MoCo within each dataset and metric.

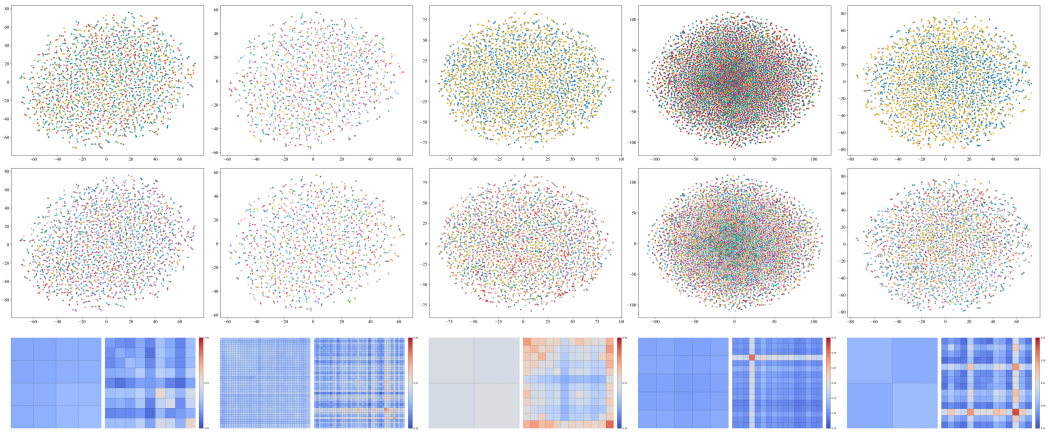


Figure 13: *t*-SNE and RSA of frozen representations from MoCo restricted to same-subject negatives. Columns correspond, from left to right, to BCI-IV-2a, Kim2025, PhysioNet-P300, SEED-V, and Tsinghua-RSVP.