Intelligent Automation & Soft Computing

Tech Science Press

https://doi.org/10.32604/iasc.2026.088039

ARTICLE

A Multi-Specialist Stacking Decoder of Cognitive Workload and a Decomposition of the Limits of Cross-Dataset Transfer in Electroencephalography

Sugeng Rifqi Mubaroq[1,*], **Rolly Maulana Awangga**[2], **Tegar Ditya Pragama**[1], **Sidiq Fathummubin**[3] **and Ali Yusuf Abdulhaq**[1]

[1]Digital Business Study Program, Akademi Digital Bandung, Jl. Rancamekar Blok Lio, Bandung, West Java, Indonesia
[2]Department of Informatics Engineering, Universitas Logistik dan Bisnis Internasional, Jl. Sariasih No. 54, Sarijadi, Sukasari, Bandung, West Java, Indonesia
[3]Software Engineering Technology Study Program, Akademi Digital Bandung, Jl. Rancamekar Blok Lio, Bandung, West Java, Indonesia

*Corresponding Author: Sugeng Rifqi Mubaroq. Email: mubaroq@digitalbdg.ac.id

Received: 27 June 2026; Accepted: 21 July 2026; Published: 11 August 2026

ABSTRACT: Decoding cognitive workload from electroencephalography (EEG) underpins passive brain–computer interfaces and adaptive learning technology, yet practical decoders share two weaknesses: they rely on a single family of features, and their accuracy collapses on recordings from an unfamiliar device, montage, or task. We address both. We first build a multi-specialist stacking decoder that fuses complementary spectral, Riemannian, and spatial views through a meta-learner. Evaluated across two public corpora under the leave-one-subject-out MOABB benchmarking protocol, it outperforms the best single specialist on the binary workload contrasts, and the strongest of these effects survives family-wide false-discovery-rate correction. The leading view switches between datasets, which is why single-view decoders are brittle. Effect-size profiling further shows that the decoders track the canonical neural signature of load: rising theta and falling alpha power. We then ask how workload models transfer between corpora. Naive pooling degrades accuracy, and we trace this negative transfer to two causes: the corpora label workload differently, and their signals differ at the device level, a component we corroborate with a third research-grade dataset that varies task and site while holding the device class fixed. Harmonizing the labels and applying class-conditional alignment then restores a small positive transfer, though only in one direction: COG-BCI benefits from Simultaneous Task EEG Workload (STEW), not the reverse. Rather than a headline accuracy figure, our contribution is an evidence-based account of when, and why, pooling EEG datasets is worthwhile.

KEYWORDS: Electroencephalography; cognitive workload; neural decoding; ensemble learning; cross-dataset transfer; domain adaptation; foundation models; Riemannian geometry

1 Introduction

In safety-critical and cognitively demanding settings, from aviation and industrial process control to driving and classroom learning, human performance falters once task demand drives an operator's mental workload beyond the level they can sustain [1,2]. Detecting objectively and in real time that an operator or learner is approaching this limit is a prerequisite for technology that can intervene before errors occur, such as passive brain–computer interfaces (BCIs) that offload an overloaded user or tutoring systems that adapt their pace to a learner's cognitive load [3]. Yet mental workload has resisted reliable, objective

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

measurement for decades and remains notoriously difficult to operationalize outside the laboratory [1]. Electroencephalography (EEG) offers the most practical window onto this state: portable, inexpensive relative to functional magnetic resonance imaging and magnetoencephalography, and directly sensitive to load-related cortical dynamics. Decades of cognitive neuroscience have mapped a robust spectral signature of workload: frontal-midline theta (4–8 Hz) power rises, driven by anterior-cingulate/medial-prefrontal circuits, while parieto-occipital alpha (8–13 Hz) desynchronizes as demand increases [4,5]. Deep-learning decoders now achieve strong within-dataset workload classification (around 84%–87% accuracy), and a recent systematic review documents how quickly this literature has grown [6]. Standardized public corpora such as COG-BCI [7] and STEW [8] have made the problem reproducible.

Two obstacles, however, limit translational impact. First, the dominant EEG feature families (spectral power, Riemannian covariance geometry, and spatial filters) capture complementary information, yet most studies commit to a single family. There is no principled, adaptive mechanism that exploits whichever view is most informative for a given recording, even though the optimal view plausibly varies with montage, noise, and task. Second, decoders rarely generalize across datasets. Differences in acquisition hardware (research amplifiers vs. consumer headsets), electrode montage and count, sampling rate, and the operational definition of "workload" itself mean that a model trained on one corpus typically fails on another, a data scarcity through non-transferability that is acute in applied neuroscience, where assembling large labeled cohorts is costly.

Two recent threads promise progress but remain underexploited for this problem. Multi-view ensemble learning, which combines complementary specialist models through a learned meta-classifier, is well established elsewhere but is rarely used for EEG workload decoding, where the choice of feature view is typically fixed in advance. Separately, EEG foundation models [9–11] and domain adaptation [12] promise cross-dataset generalization, yet a 2026 critical review [13] finds foundation-model generalizability largely undemonstrated, and prior work has not characterized cognitive-workload transfer between these heterogeneous datasets.

This paper makes three contributions. The individual techniques we use (stacking, foundation-model embeddings, class-conditional adaptation, and geometric alignment) are adopted, established methods; our contribution is diagnostic and empirical.

1. A multi-specialist stacking decoder for EEG cognitive-workload decoding that combines three complementary feature specialists through a meta-learner trained on out-of-fold predictions. We show it outperforms the best single specialist on the binary workload contrasts (significant, and surviving family-wide false-discovery-rate correction on COG-BCI binary workload), a target-dependent effect grounded in per-subject effect size rather than in a single threshold-crossing p-value, and that the dominant specialist differs by dataset, which explains why fixed single-view decoders are fragile.
2. A systematic, reproducible characterization of cross-dataset transfer for cognitive workload (COG-BCI↔STEW). We show that classical methods fail and then decompose the failure into label-construct mismatch and device/signal domain shift, measuring the contribution of each and corroborating the device component with a third research-grade dataset (ds003838) that varies task and site while holding device class fixed.
3. A method recipe that restores positive transfer: construct harmonization + class-conditional domain adaptation (with z-score input normalization for the foundation-model path), on both interpretable PSD features and frozen foundation-model embeddings. The restored transfer is asymmetric (COG-BCI(+STEW) only), and class-conditional alignment is the decisive ingredient: on the PSD feature space it improves over naive pooling by +0.047 macro-F1 (p = 0.0003, surviving family-wide false-discovery-rate (FDR) correction), and it is the only intervention that turns the frozen-embedding transfer positive,

albeit directionally (+0.015, $p = 0.10$); a second-order alignment control localizes the residual gap to domain shift.

We report modest effects. Our aim is a rigorous account of when and why multi-dataset EEG integration helps, not an inflated accuracy claim; we argue this characterization is itself the contribution.

2 Related Work

2.1 EEG Cognitive-Workload Decoding

Workload decoding traditionally relies on band-power features grounded in the theta/alpha signature, classified by support vector machines (SVMs) or random forests, or on covariance-based Riemannian pipelines that have proven strong for small-sample BCI. Deep models raise within-dataset accuracy but are data-hungry and prone to subject overfitting; these include convolutional neural networks (CNNs), long short-term memory networks, and transformers. Across studies, no single representation dominates, which motivates ensembles; prior work has largely explored these as fixed combinations, not learned stacking.

2.2 EEG Foundation Models

Self-supervised models pretrained on large heterogeneous EEG corpora (CBraMod [9], LaBraM [10], EEGPT [11]) are architecturally designed to ingest variable channel counts and lengths via patch tokenization and conditional positional encoding, directly targeting montage heterogeneity. However, a critical review [13] reports that many evaluations are in-sample and that larger models do not guarantee better generalization; none target cognitive-workload cross-montage transfer. A 2026 systematic channel-adaptation benchmark [14] further shows naive supervised fine-tuning causes negative transfer in 24.8% of cases (including a foundation-model + Riemannian collapse to chance), recommending frozen-encoder probing as the safe protocol, a recommendation we adopt.

2.3 Domain Adaptation and Geometric Alignment

A central insight, formalized by cdaDA [12], is that EEG domain alignment must be class-conditional: aligning only the global (marginal) distribution degrades workload-discriminative structure. Cross-subject methods such as CS-DASA [15] report sizeable gains within a single dataset. Geometric approaches (Riemannian Procrustes Analysis [16] and Euclidean Alignment [17]) align covariance distributions on the symmetric positive-definite (SPD) manifold, but their re-centering/translation step removes the covariance-magnitude information in which the workload signal partly resides, as we confirm. Note that these results are cross-subject within one dataset, not cross-dataset.

2.4 Gap

Prior work has not characterized COG-BCI↔STEW cognitive-workload transfer, decomposed the causes of transfer failure, or evaluated multi-specialist stacking across these heterogeneous corpora. We address all three.

3 Materials and Methods

3.1 Datasets

COG-BCI [7]: 29 participants × 3 sessions of 63-channel EEG during the Multi-Attribute Task Battery (MATB), an N-Back working-memory task, a psychomotor vigilance task (PVT), a Flanker task, and resting state (eyes open/closed); 1044 recordings (~139,000 2-s epochs after preprocessing). STEW [8]: 48 participants, 14-channel Emotiv EPOC, recording rest ("low" workload) and a simultaneous-capacity

(SIMKAP) multitasking block ("high"); ~9500 epochs from 86 usable recordings. For the device de-confound analysis we add a third dataset, ds003838 [18]: an auditory digit-span working-memory study (rest vs. retention of a 13-digit load), 65 of the 86 subjects with usable EEG, 64-channel research-grade recording at a different laboratory (~36,000 harmonized epochs). ds003838 is included specifically because it shares the research-grade device class with COG-BCI but differs in task and recording site, which lets us vary the task/site confound while holding the device class fixed. All are de-identified public datasets used here for secondary analysis under their respective licenses.

3.2 Preprocessing and Quality Control

We used MNE-Python [19] with a config-driven, non-destructive pipeline. After a notch filter and 0.5–45 Hz band-pass, we ran extended-infomax independent component analysis (ICA) on a 1-Hz high-passed copy and rejected artifactual components automatically: via ICLabel [20] for COG-BCI (where the ≥32-channel montage makes ICLabel reliable), and via a frontal-correlation electro-oculogram (EOG) proxy heuristic for STEW (ICLabel is unreliable below ~20 channels). The dataset-dependent choice is itself a methodological point: consumer-grade STEW channels frequently saturate (e.g., T7 peak-to-peak > 9000 μV), so we added robust bad-channel detection (median-absolute-deviation z-score on log peak-to-peak, plus flat-channel detection) and spherical-spline interpolation before average referencing, following the Preprocessing Pipeline (PREP) robust-referencing standard [21]; otherwise a single bad channel contaminates all others through the common average. Recordings whose epoch retention fell below 30% were excluded as unsalvageable (10 of 96 STEW recordings). Median epoch retention was 97% (COG-BCI) and 77% (STEW). Non-EEG channels (e.g., electrocardiogram, ECG) and channels without montage coordinates were removed, as they otherwise propagate not-a-number (NaN) values through interpolation. Dataset checksums were verified against source records (we corrected two transcription errors in a download manifest), supporting reproducibility.

3.3 Cross-Dataset Feature Harmonization

To make features comparable across datasets (a prerequisite for transfer), both were projected onto the 14 Emotiv channels common to both montages (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) and resampled to a common 128 Hz (the CBraMod encoder resamples its input to 200 Hz internally). From the harmonized epochs we computed: (i) power spectral density (PSD) band-power: relative, log-transformed power in five bands ($\delta, \theta, \alpha, \beta, \gamma$) per channel (70 features); (ii) spatial covariance (14×14, oracle-approximating-shrinkage estimator); and (iii) Riemannian tangent-space projections (105 features). All three have identical dimensionality across datasets, enabling direct transfer. Projecting to 14 channels discards spatial detail. We found no detectable cost for the present task: per-subject macro-F1 from the full research-grade montage and from the 14 common channels is closely matched under the leave-one-subject-out (LOSO) protocol (0.770 vs. 0.768; difference not significant, paired Wilcoxon $p = 0.89$), even though the 14 channels carry only ~26% of the total band-power variance. Non-significance bounds rather than proves equivalence, so we claim only that no decoding-relevant loss is detectable at this sample size for this binary contrast; finer-grained targets may be more sensitive to the discarded detail.

3.4 Multi-Specialist Stacking Decoder

The decoder (Fig. 1) comprises three specialist modules, each consuming one feature view: spectral (radial-basis-function (RBF) SVM on PSD), Riemannian (tangent-space mapping [22] + logistic regression), and CSP (common spatial patterns + linear discriminant analysis). A meta-learner combines them via stacking [23]: a logistic regression trained on the specialists' out-of-fold class probabilities (we also report a

soft-voting variant). Out-of-fold training is essential to avoid leakage: the meta-learner never sees specialist predictions on data those specialists were trained on. Tangent-space and CSP transforms are fit within each training fold only. We refer to this trained combiner as the stacking meta-learner throughout; the decoder as a whole is a multi-specialist stacking decoder.

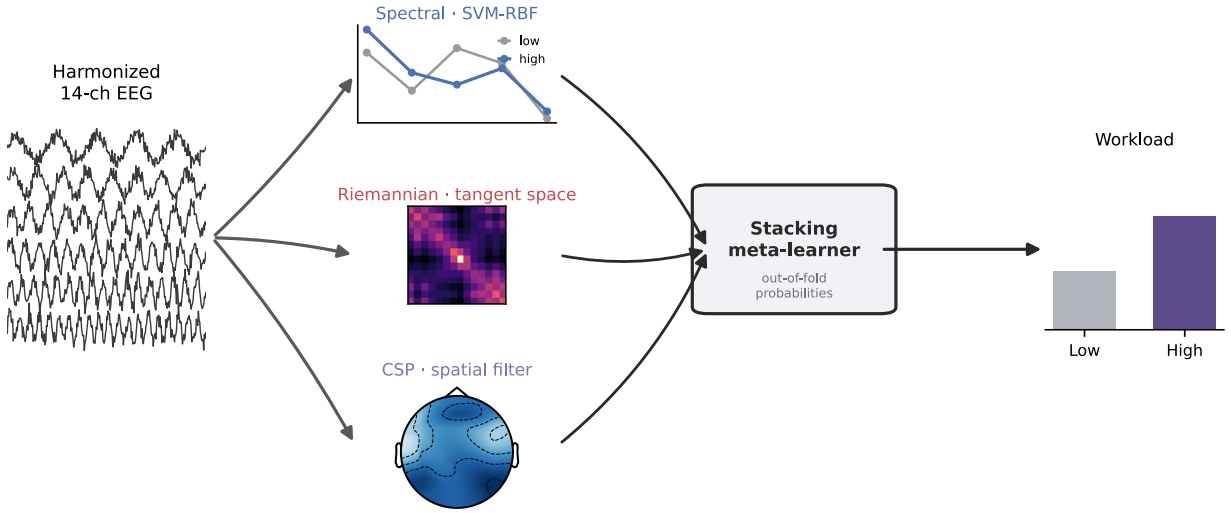

Figure 1: Multi-specialist stacking decoder: three complementary EEG feature views fused by a meta-learner.

3.5 Cross-Dataset Transfer Methods

We evaluated, in increasing sophistication: (a) naive pooling (concatenate source and target training data); (b) Riemannian re-centering (TLCenter, recentering each domain's covariances to identity); (c) multi-task shared-backbone, a neural network with a shared trunk and dataset-specific heads, so label semantics need not match; and (d) foundation-model frozen-encoder probing, using CBraMod [9] (encoder frozen) producing 200-dimensional embeddings (the native embedding width of the pretrained braindecode/cbramod-pretrained encoder, mean-pooled over channels and time patches, not a separate dimensionality reduction; mean-plus-std pooling would give 400), followed by a linear probe, optionally with class-conditional domain adaptation: per-class mean alignment, and a class-conditional correlation-alignment (CORAL) [24] variant aligning per-class mean and covariance (a second-order control). To isolate the role of label semantics we defined a construct-harmonized ("extreme") labeling: COG-BCI low = resting, high = MATBdiff + twoBACK (the most demanding, multitasking-like conditions), aligning it with STEW's rest-vs-SIMKAP contrast. Finally, to test whether transfer behaves differently for a learned, end-to-end representation rather than fixed or frozen features, we added (e) a learned domain-adversarial harmonization: a compact EEGNet [25] encoder (3764 parameters; temporal convolution → depthwise spatial filter → separable convolution) trained directly on the harmonized 14-channel epochs, with a workload-label head and a dataset-discriminator head connected through a gradient-reversal layer [26] (a domain-adversarial neural network, DANN). The adversary weight λ was swept over $\{0.1, 0.3, 1.0\}$; $\lambda = 0$ recovers naive pooling.

3.6 Evaluation and Statistics

All decoding follows the Mother of All BCI Benchmarks (MOABB) cross-subject benchmarking protocol [27]: evaluation is leave-one-subject-out, so each subject is held out in turn and scored individually, yielding one score per subject (N = 29–47) rather than a handful of pooled folds. The primary metric is the

area under the receiver operating characteristic (ROC) curve (AUC; chance = 0.5 by construction, macro-averaged one-vs-rest for the 3-class targets); macro-F1 is reported alongside as a threshold-based secondary metric robust to class imbalance (e.g., COG-BCI workload is ~26% vs. 74%). For the stacking meta-learner we use a nested inner 10-fold subject-grouped cross-validation (CV) on the training subjects (out-of-fold specialist predictions for the meta-learner); the deep encoder's domain-adversarial weight λ is selected by a nested inner 5-fold subject-grouped CV. In either case no held-out subject informs model construction. For transfer, a model is trained on the source dataset together with the target's training subjects and scored per held-out target subject; "gain" is relative to a target-only model. Epochs are subsampled to a common budget of 100 per subject, stratified by class, so that every subject contributes comparably and no corpus dominates the pooled training (the two corpora otherwise differ by more than an order of magnitude, ~139,000 vs. ~9500 epochs); to confirm the results do not depend on a single draw, every analysis is repeated over three independent subsamples, each drawn with a distinct random seed that also sets the neural-encoder initialization, and we report the number of draws in which each effect keeps its direction and significance. Because LOSO test sets are disjoint across subjects, the anti-conservative fold-overlap correction of pooled cross-validation is unnecessary; we compare pipelines with the paired Wilcoxon signed-rank test across subjects and report the standardized mean difference (Cohen's d_z) as the effect size. Because many hypotheses are tested across specialists, datasets, and transfer conditions, we apply the Benjamini–Hochberg FDR correction across the full family of 16 tests and report adjusted q-values (Holm as a sensitivity check); every test, with both corrections, is listed in Table A2. Neurophysiological interpretability used Cohen's d on log band-power (averaged over the 14 common channels) between low and high workload. All model hyperparameters are listed in the Appendix A (Table A1), and the full pipeline configuration is released with the code.

4 Results

We first establish within-dataset decoding and ground it in known neurophysiology, then turn to cross-dataset transfer: why naive methods fail, how the failure decomposes into a construct-mismatch and a device component, and when integration can be made positive. Throughout, validation is leave-one-subject-out (no subject appears in both train and test), the most stringent regime for clinical/applied EEG, and every headline comparison is scored per subject and repeated over three subsamples.

4.1 Within-Dataset Decoding and the Stacking Decoder

Under the LOSO protocol the stacking decoder's advantage over the best single specialist is genuine but target-dependent (Table 1), and which specialist it must beat changes with the dataset (CSP for STEW, Riemannian for COG-BCI), confirming that no single view is universally best. The advantage is clearest on the binary workload contrasts. On COG-BCI binary workload the decoder reached AUC 0.696 ± 0.194 vs. the best specialist's 0.666 (Δ = +0.030, d_z = 0.62, Wilcoxon p = 0.0007), the only within-dataset comparison that survives the family-wide FDR correction on the primary AUC metric (q_{BH} = 0.004; the STEW stacking gain also survives on macro-F1, q_{BH} = 0.016), significant in 2/3 subsamples. On STEW workload it reached AUC 0.861 ± 0.118 vs. 0.840 (Δ = +0.021, p = 0.025) and macro-F1 0.716 ± 0.167 vs. 0.684 (Δ = +0.033, p = 0.006), positive and significant in all three subsamples. On the harder graded 3-class targets the margin shrinks to $\Delta \approx +0.01$ AUC and is not significant (MATB p = 0.31; N-back p = 0.41, where the ensemble neither helps nor hurts), consistent with these contrasts being intrinsically harder. We therefore report the stacking advantage as significant on the binary contrasts and directional-only on the graded ones, grounded in per-subject effect size and subsample reproducibility rather than in a single threshold-crossing.

Table 1: Within-dataset decoding: view-specialists vs. the stacking decoder (per-subject ROC-AUC, LOSO).

Target (AUC)	Spectral	Riemann.	CSP	Stacking	Δ	d_z	p	q
COG-BCI N-back, 3-class	0.651	0.633	0.634	0.650	−0.001	−0.02	0.41	0.466
COG-BCI MATB, 3-class	0.702	0.761	0.733	0.769	+0.008	0.13	0.31	0.454
COG-BCI workload, binary	0.656	0.666	0.593	0.696	+0.030	0.62	0.0007	0.004*
STEW workload, binary	0.808	0.766	0.840	0.861	+0.021	0.24	0.025	0.068

Note: Δ, stacking − best specialist; d_z, paired effect size; q, family-wide FDR. The Stacking column reports the proposed decoder, not the highest value in its row. *Survives the family-wide FDR correction on the primary AUC metric.

4.2 Ablation: Component Contribution and the "Dominant" Specialist

We define the dominant specialist by a single explicit criterion: the leave-one-view-out drop, i.e., the view whose removal most degrades the full stacked decoder (its marginal contribution to the ensemble), and apply it consistently (Table 2, per-subject AUC under the LOSO protocol). Under this protocol the criterion agrees with the highest-standalone view on both datasets: for STEW the dominant view is CSP (removal costs 0.030 AUC, the largest drop) and CSP also has the highest standalone AUC (0.840); for COG-BCI MATB the dominant view is Riemannian (drop 0.025) and Riemannian is likewise the strongest standalone view (0.761). No leave-one-out variant substantially exceeded the full fusion (the largest such difference, removing the redundant Riemannian view on STEW, was +0.002 AUC, within noise), and stacking and soft-voting performed comparably (soft-voting marginally higher on AUC). The stacking decoder's value is therefore its stability: it stays at or above the best single specialist on every dataset, whereas a single view collapses on the dataset outside its strength: CSP is strongest on STEW (0.840) but weakest on COG-BCI binary workload (0.593; Table 1).

Table 2: Ablation of the stacking decoder: single views, leave-one-out, and full fusion (ROC-AUC, LOSO).

Variant	COG-BCI MATB	STEW Workload
Only spectral	0.702	0.808
Only Riemannian	0.761	0.766
Only CSP	0.733	0.840
− spectral	0.766	0.837
− Riemannian	0.744	0.863
− CSP	0.762	0.831
Full, soft-vote	0.786†	0.863†
Full, stacking	0.769	0.861

Note: Single-view and full-stacking means match Table 1; the leave-one-out (−view) and soft-vote rows are specific to this ablation. A drop is full stacking minus the leave-one-out variant (e.g., Riemannian on COG-BCI, 0.769 − 0.744 = 0.025). †Highest AUC in the column.

4.3 Comparison with End-to-End Deep Architectures

To place the stacking decoder against modern end-to-end baselines under identical settings, we trained a compact EEGNet convolutional encoder and an EEG-Conformer (a convolutional transformer) [28] on the same harmonized 14-channel epochs, cross-subject LOSO, per-subject scored (Table 3). No single model dominates: on STEW the two end-to-end models modestly exceed the stacking decoder on AUC (0.884/0.875 vs. 0.861) and do not differ significantly on macro-F1 (Wilcoxon $p = 0.49$ and 0.39), whereas on the imbalanced COG-BCI binary workload the stacking decoder is clearly superior (AUC 0.696 vs. 0.602/0.581).

The transformer does not outperform the far smaller EEGNet on either dataset (Wilcoxon $p = 0.54$ and 0.25). The stacking decoder is thus competitive with, and on the harder research-grade target superior to, a modern transformer, while remaining a compact and interpretable pipeline of linear and kernel probes rather than a large neural network. This is consistent with the finding that robustness across heterogeneous corpora comes from combining complementary views rather than from any single architecture.

Table 3: End-to-end deep baselines vs. the stacking decoder (within-dataset, per-subject, LOSO; binary high-vs-low workload labels).

Model	Params	STEW (AUC)	COG-BCI (AUC)	STEW (F1)	COG-BCI (F1)
Stacking decoder	–	0.861	0.696†	0.716	0.614†
EEGNet	3.8k	0.884†	0.602	0.725	0.528
EEG-Conformer	225k	0.875	0.581	0.726†	0.549

Note: "Params" counts the trainable parameters of the neural encoders; it does not apply to the classical stacking pipeline of linear and kernel probes (dash). †Best value in the column.

4.4 Neurophysiological Interpretability

To check that the decoders exploit genuine physiology and not artifacts, we computed Cohen's d between low and high workload on log band-power (Fig. 2). The pattern matches the canonical workload signature: theta and delta power increase while alpha (and higher bands) decrease with load. Effect sizes are computed on band power averaged over the 14 common channels (Fig. 2a), with per-channel alpha-band d shown topographically (Fig. 2b). For COG-BCI the band-wise effects are large and ordered as theory predicts: alpha $d = -1.17$ (strong desynchronization), delta $d = +0.99$, theta $d = +0.80$, with weak beta/gamma effects ($|d| \approx 0.2$). The most discriminative channels in the alpha band are parieto-occipital (P8 −1.14, O2 −1.11, P7 −1.10, O1 −1.07), the expected generators of the alpha rhythm whose desynchronization indexes cognitive demand. STEW shows the same directionality with smaller, noisier effects (delta +0.63, theta +0.54, alpha −0.54), consistent with its consumer-grade signal quality. The largest and cleanest effects (COG-BCI alpha, $|d| > 1$) occur in the dataset on which the Riemannian specialist excels, whereas STEW's weaker, more diffuse effects are better exploited by the CSP spatial filter than by raw band power. This is consistent with the dataset-dependent specialist dominance in Table 2 (Riemannian for COG-BCI, CSP for STEW, both covariance-based views), and hence the value of stacking across views. These view preferences follow the datasets' covariance structure rather than dataset-specific artifacts.

4.5 Cross-Dataset Transfer Baselines Fail

Naive pooling produced negative transfer in both directions: a model trained on pooled source + target data underperformed a target-only model on held-out target subjects. Riemannian re-centering was worse, collapsing COG-BCI workload toward chance, consistent with the 2026 benchmark in which a foundation model + Riemannian re-centering dropped to chance [14], because the workload signal resides in covariance magnitude (power), which re-centering removes. The multi-task shared-backbone, which avoids forcing identical label semantics, was the least harmful (gain ≈ −0.01 to −0.03 macro-F1) but still did not achieve positive transfer. The negative transfer of naive pooling is robust under the LOSO protocol: on per-subject PSD + logistic-regression transfer the multi-vs-single gain was negative for STEW (−0.033 macro-F1, paired Wilcoxon $p = 0.0009$) and, under the harmonized construct, for COG-BCI (−0.043, $p = 0.001$), both significant across all three subsamples and surviving the family-wide FDR correction, so the negative transfer of naive pooling is a consistently significant effect.

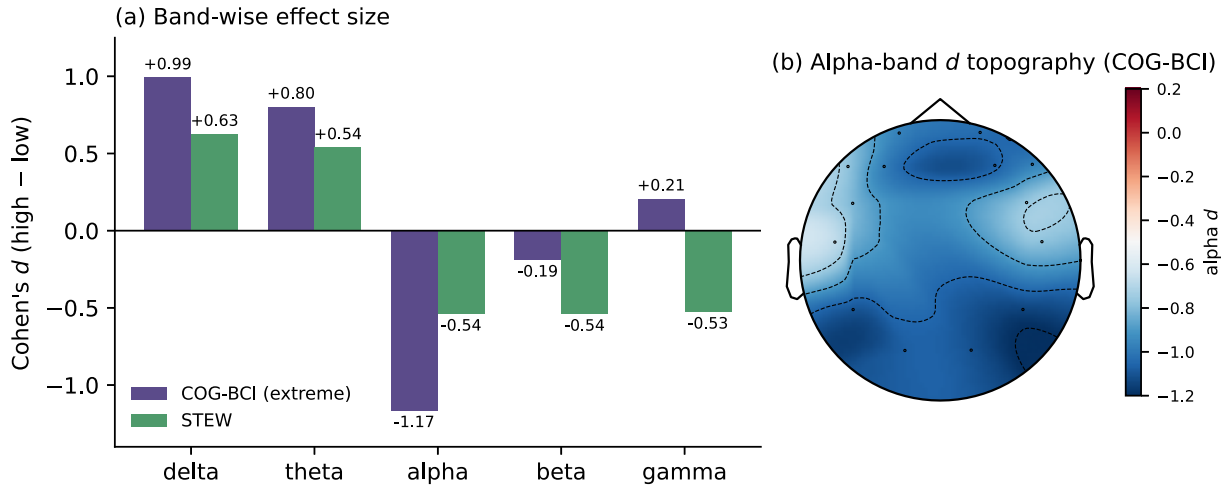

Figure 2: Neurophysiological signature of workload (Cohen's d, high − low).

4.6 Decomposing the Failure: Construct Mismatch vs. Domain Shift

Because COG-BCI and STEW operationalize "workload" differently (graded cognitive tasks vs. a single multitasking block), we tested whether harmonizing the label construct helps (Table 4). It did: for STEW the negative transfer shrank from −0.033 to −0.029, and within COG-BCI the harmonized contrast was far cleaner (single-dataset macro-F1 0.557 → 0.770). Label-construct mismatch is therefore a genuine, quantifiable contributor to transfer failure. (Note that harmonization redefines only the COG-BCI label mapping to match STEW's rest-vs-multitasking contrast; STEW's own task and labels are unchanged, so its single-dataset baseline in Table 4 is identical across the two schemes by construction.) Yet transfer remained non-positive, implicating a second, residual factor: device/signal domain shift, which we corroborate directly below.

Table 4: Construct harmonization: single-vs. multi-dataset transfer macro-F1 (PSD, LOSO).

Target	Label Scheme	Single	Multi (+Source)	Gain
STEW (+COG-BCI)	Binary (loose)	0.690	0.657	−0.033*
STEW (+COG-BCI)	Extreme	0.690	0.661	−0.029
COG-BCI (+STEW)	Binary	0.557	0.553	−0.004
COG-BCI (+STEW)	Extreme	0.770	0.726	−0.043*

Note: "Gain" is multi-dataset − single-dataset per-subject macro-F1; negative values indicate negative transfer. *Significant in all three subsamples and surviving the family-wide FDR correction.

4.7 Foundation Model: Normalization Study and a Recipe for Positive Transfer

Input normalization proved decisive for the foundation model. Under the LOSO protocol (Table 5) within-dataset STEW macro-F1 ranged from 0.649 (the common "μV/100" scaling) to 0.707 (μV scaling with mean pooling), with z-score per-epoch normalization close behind at 0.682, whereas a random-initialized encoder reached only 0.579. Measured consistently, the macro-F1 chance level for this binary target is 0.50 (a stratified-random predictor), so the random-initialized encoder sits just above chance; the majority-class accuracy sometimes quoted as "chance" is not the appropriate baseline for macro-F1. The gap between the random-initialized (0.579) and pretrained (0.682–0.707) encoders confirms the pretrained weights carry genuine signal. CBraMod frozen embeddings thus approach but do not surpass hand-crafted features here

(the PSD stacking decoder reaches 0.716 on STEW, Table 3), echoing the foundation-model generalizability critique [13].

Table 5: CBraMod frozen-probe normalization sweep (STEW workload, per-subject macro-F1, LOSO).

Normalization/Pooling	STEW Macro-F1
μV/100/mean (initial)	0.649
μV (scale 1)/mean	0.707
z-score/mean	0.682
z-score/mean + std	0.669
Random-init encoder (control)	0.579

Note: μV/mean scores highest within STEW, but the transfer recipe (Table 6) uses z-score/mean for its invariance to the amplitude-scale differences between devices.

Table 6: Foundation-model transfer recipe (CBraMod, harmonized construct; per-subject macro-F1, LOSO).

Target	Single	Multi (Naive)	+Class-Cond. (Mean)	+Class-Cond. (CORAL)
STEW (+COG-BCI)	0.682	0.652 (−0.030)	0.660 (−0.022)	0.661 (−0.022)
COG-BCI (+STEW)	0.714	0.707 (−0.007)	0.729 (+0.015)†	0.728 (+0.014)

Note: Gains are relative to the single-dataset probe; the COG-BCI class-conditional gain is directional (Wilcoxon p = 0.10). The significant, FDR-surviving version of this effect is on the PSD space (Table 7). †The only condition in this table with a positive transfer gain.

Table 7: Class-conditional alignment vs. naive pooling (PSD, per-subject macro-F1, LOSO).

Target (Scheme)	Naive	Class-Cond.	ΔF1	ΔAUC	Wilcoxon p	Sig.
STEW (+COG-BCI), binary	0.657	0.676	+0.020	+0.016	0.008	1/3
STEW (+COG-BCI), extreme	0.661	0.680	+0.019	−0.006	0.051	0/3
COG-BCI (+STEW), binary	0.553	0.550	−0.002	+0.026	0.70	0/3
COG-BCI (+STEW), extreme	0.726	0.773	+0.047	+0.043	0.0003*	3/3

Note: Naive and Class-cond. are per-subject macro-F1 means; ΔF1 and ΔAUC give the class-conditional advantage (class-conditional − naive) on each metric. Wilcoxon p and "sig." (subsamples of 3 with $p < 0.05$) are on macro-F1. On the primary AUC metric the headline COG-BCI (extreme) advantage is +0.043 (Wilcoxon p = 0.0004). On the two non-headline rows ΔF1 and ΔAUC differ in sign; we read neither as evidence, since both are non-significant on the primary metric. *Survives the family-wide FDR correction on both metrics (q_{BH} = 0.003).

Combining the z-score encoder, the harmonized construct, and class-conditional domain adaptation gave a small positive transfer on the frozen CBraMod embedding space under LOSO (Table 6): COG-BCI workload improved from single-dataset 0.714 to 0.729 (+0.015, directional, Wilcoxon p = 0.10) using STEW as auxiliary, whereas naive pooling gave −0.007, a +0.022 swing in favor of class-conditional alignment. Both CBraMod tables are now evaluated under the same LOSO protocol as the rest of the paper. The strong, FDR-surviving form of this effect lives on the interpretable PSD feature space (Fig. 3): there, class-conditional alignment improves per-subject COG-BCI macro-F1 over naive pooling by +0.047 (paired Wilcoxon p = 0.0003, d_z = 0.58), positive and significant in all three subsamples and surviving the family-wide FDR correction (q_{BH} = 0.003; Table 7). Second-order CORAL alignment (matching per-class covariance as well as mean) provided no further gain on the frozen embeddings (+0.014), indicating that first-order mean correction suffices and that the residual gap is fundamental domain shift, not insufficient alignment. Transfer

was asymmetric: STEW (single-dataset 0.682) did not benefit (−0.030 for naive pooling), consistent with its smaller headroom.

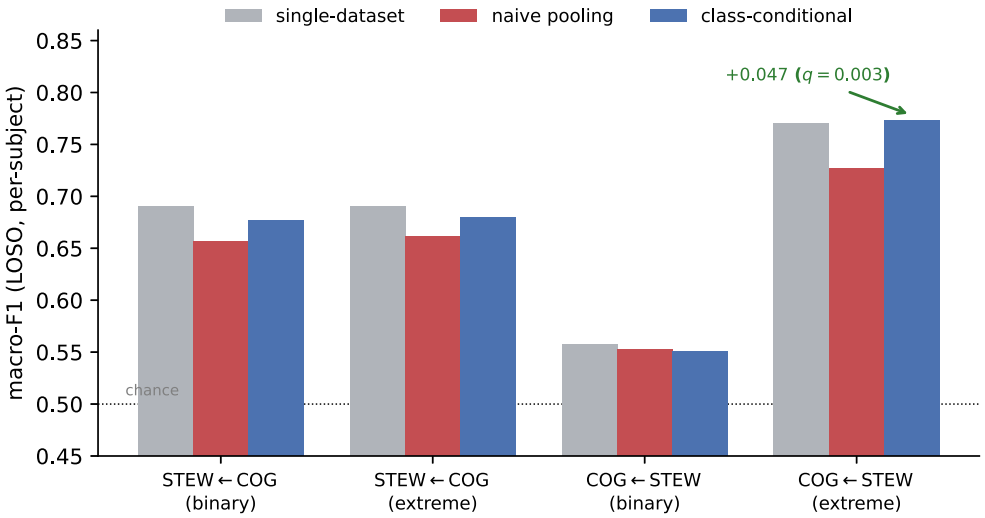

Figure 3: Cross-dataset transfer under LOSO: naive pooling degrades, class-conditional alignment recovers (PSD, macro-F1).

4.8 Transfer Is Model-Class-Dependent: A Learned Deep Encoder

The transfer behavior above was obtained with fixed (PSD) or frozen (foundation-model) features. To test whether a learned, end-to-end representation behaves differently, we trained the compact EEGNet encoder directly on the harmonized epochs and compared three regimes per target: target-only training (single), naive pooling of source and target (pool, i.e., $\lambda = 0$), and learned domain-adversarial harmonization (DANN, gradient-reversal with λ selected by nested cross-validation). The linear-regime conclusion is qualified rather than reversed. Under the LOSO protocol the learned deep encoder shows no pooling benefit on either target (Table 8): pooling is neutral for STEW (macro-F1 $0.727 \rightarrow 0.737$, +0.010; AUC $0.883 \rightarrow 0.897$; both n.s.) and mildly negative for COG-BCI (macro-F1 $0.758 \rightarrow 0.741$, −0.017, n.s.; AUC $0.887 \rightarrow 0.859$). Per-subject evaluation thus yields no evidence of a pooling benefit for the deep encoder. We report the deep-encoder analysis as exploratory: naive pooling never helps a data-hungry deep encoder either, echoing the negative transfer it causes for fixed PSD features. For the domain adversary we selected λ per outer fold by nested inner 5-fold cross-validation (the grid $\{0.1, 0.3, 1.0\}$ was chosen mostly at $\lambda = 0.1$); even so it recovered nothing over pooling (+0.004 and −0.002 macro-F1; Wilcoxon $p = 0.52$, 0.70), so we rest no claim on it. The encoder is extremely compact (3764 parameters; ~0.02 ms per epoch on GPU), supporting edge-deployment feasibility.

Table 8: Deep-encoder transfer under LOSO: single, naive pooling, and nested-λ domain adversary (EEGNet, per-subject macro-F1, harmonized (extreme) labels).

Target	Single	Pool ($\lambda = 0$)	DANN (Nested-λ)	Pool − Single	DANN − Pool
STEW (+COG-BCI)	0.727	0.737	0.741	+0.010	+0.004
COG-BCI (+STEW)	0.758	0.741	0.739	−0.017	−0.002

Note: Per-subject macro-F1. Wilcoxon (one-sided) pool > single $p = 0.45$ (STEW), 0.89 (COG-BCI); DANN > pool $p = 0.52$, 0.70. The domain-adversary weight λ was chosen per outer fold by nested inner 5-fold CV; the grid $\{0.1, 0.3, 1.0\}$ was picked mostly at $\lambda = 0.1$.

Synthesizing across all three representation regimes (Table 9), what helps cross-dataset transfer depends on the model class. With fixed linear features, naive pooling hurts robustly (surviving FDR) and class-conditional alignment is required to recover. This recovery is significant and reproducible across all three subsamples. With frozen foundation-model embeddings, pooling is mildly negative and class-conditional adaptation yields a small, directional positive transfer. With an end-to-end deep encoder, pooling is neutral for STEW and mildly negative for COG-BCI, never positive, and a nested-λ domain adversary recovers nothing. The unifying principle is that transfer is governed by the conditional (class-wise) distribution: naive marginal pooling never improves any of the fixed, frozen, or learned representations, whereas explicitly correcting the class-conditional shift is what recovers positive transfer.

Table 9: Transfer behavior by representation regime (LOSO).

Representation Regime	Naive Pooling	Class-Conditional vs. Pooling
Fixed linear (PSD + LogReg)	Negative (survives FDR)	Recovers, +0.047, $p = 0.0003$, survives FDR
Frozen foundation model (CBraMod)	Mildly negative	Small, directional (+0.015, $p = 0.10$)
Learned deep encoder (EEGNet)	Neutral (STEW) to negative (COG-BCI)	Adversarial adds nothing (nested-λ)

4.9 De-Confounding the Device Attribution with a Third Dataset

The decomposition above attributes the residual, construct-invariant transfer gap to device/signal domain shift, but with only COG-BCI (research-grade) and STEW (consumer) that attribution is confounded: the two corpora also differ in task, site, channel count and preprocessing. To separate the device factor from the task/site confound we added a third dataset, ds003838 (an auditory digit-span working-memory study; 64-channel research-grade EEG recorded at a different laboratory), which shares the research-grade device class with COG-BCI but differs in task and site. If the residual gap were merely a task/site artifact, matching the task should matter more than matching the device; if it is device-driven, matching the device class should transfer better even across tasks. The latter holds (Table 10): under the LOSO protocol, scoring each held-out target subject on the harmonized PSD space, device-matched transfer (research↔research, i.e., COG-BCI↔ds003838, despite the working-memory-vs-MATB task change) averaged AUC 0.816 and macro-F1 0.571, whereas cross-device transfer to or from the consumer headset averaged AUC 0.738 and macro-F1 0.446. This gap of +0.078 AUC (+0.125 macro-F1) is highly significant (Mann–Whitney $p = 2.8 \times 10^{-4}$ for AUC, $p = 3.2 \times 10^{-7}$ for macro-F1), it is stable across all three subsamples, and it survives the family-wide FDR correction. Matching the device class thus yields more transferable workload signal than matching the task. We read this as a strengthening check rather than a proof: ds003838 still differs from COG-BCI in channel layout and preprocessing, so it does not isolate device perfectly, and we keep those residual confounds explicit.

Table 10: Device de-confound via ds003838: directional transfer per target subject (LOSO; AUC, macro-F1).

Transfer $T \leftarrow S$	Relation	AUC	Macro-F1
COG-BCI ← ds003838	Device-matched	0.814	0.723
ds003838 ← COG-BCI	Device-matched	0.818	0.504
STEW ← ds003838	Cross-device	0.681	0.477
ds003838 ← STEW	Cross-device	0.700	0.335
STEW ← COG-BCI	Cross-device	0.782	0.512
COG-BCI ← STEW	Cross-device	0.837	0.539
Device-matched (mean)		0.816	0.571
Cross-device (mean)		0.738	0.446

Note: $T \leftarrow S$, train on S, test per subject of T. Group means are pooled over every held-out subject in the category (subject-weighted, N = 94 device-matched, N = 188 cross-device), so they need not equal the arithmetic average of the rows. The tendency is not a law: one cross-device pair (COG-BCI ← STEW, AUC 0.837) transfers as well as the device-matched pairs, so device class is a strong but not exclusive factor.

5 Discussion

This study set out to answer three questions. Does multi-specialist stacking improve workload decoding over a single feature view, and why? Why does cross-dataset transfer between heterogeneous corpora fail, and can that failure be decomposed? And can multi-dataset integration be made to help rather than hurt? We take each in turn and position our answers against recent work.

5.1 Multi-Specialist Stacking Is a Dependable Framing for Neural Decoding

Because the most informative feature view differs by dataset (Tables 1 and 2), a stacking meta-learner that weights specialists tracks the best representation per dataset and does not fall below the best single view on any target we evaluated, unlike any fixed single-view decoder. The gain over the best specialist is target-dependent: it is significant on the binary workload contrasts, surviving family-wide FDR correction on COG-BCI binary workload (q_{BH} = 0.004; the STEW gain also survives on macro-F1, q_{BH} = 0.016) and significant on STEW (Δ = +0.033 macro-F1, p = 0.006), and directional only on the harder graded three-class targets. We are explicit about this asymmetry rather than claiming a uniform advantage. The ablation localizes the advantage to complementarity (by leave-one-view-out, CSP dominates STEW and Riemannian dominates COG-BCI) and not to any single component. This dataset-dependence is corroborated by the recent systematic review of Pušica et al. [29], which finds that no single model or representation dominates EEG workload classification across task types and that the absence of a standardized benchmark frustrates direct comparison. A stacked multi-specialist decoder meets this fragility directly, because it does not commit to any single view. Our approach also connects to an independent, concurrent trend: the strongest recent EEG foundation models route inputs to specialized sub-networks via mixture-of-experts (MoE) layers, e.g., Uni-NTFM [30], a 1.9-billion-parameter MoE transformer whose dynamic routing assigns signal patterns to specialized experts to prevent task interference. Our contribution is the parsimonious, interpretable counterpart of that idea: a learned combination of three transparent, neurophysiologically meaningful specialists, achieved without large-scale pretraining and with each specialist's contribution auditable by ablation. We credit the underlying mechanism (stacking [23]) rather than claim architectural novelty. The contribution is the systematic evaluation of heterogeneous-input multi-specialist stacking for workload decoding and the empirical demonstration that it confers robustness.

5.2 Why Cross-Dataset Transfer Is Hard, and What the Decomposition Reveals

The transfer failure is real and decomposable into two distinct, separately measurable causes. (i) Construct mismatch: harmonizing the label definition substantially reduced the negative transfer, so part of the failure is simply that the two corpora label different mental states "high workload." (ii) Domain shift: the residual deficit persisted under harmonization and was unchanged by second-order (covariance) alignment.

That CORAL did not improve over first-order mean alignment is the most informative negative result here, and it speaks directly to an active domain-adaptation debate. A foundational principle, formalized for adversarial adaptation by Long et al. [31] (conditioning the alignment on classifier predictions through multilinear and entropy conditioning) and for EEG specifically by cdaDA [12], holds that aligning the marginal distribution is insufficient and that the class-conditional distribution must be matched. Our results refine this principle empirically: for EEG-embedding workload transfer the transferable discriminative structure is carried by a class-conditional mean shift, while conditional covariance carries no additional recoverable signal. Cheap first-order class-conditional correction was therefore both necessary and sufficient; the more expensive second-order alignment (CORAL [24]) and adversarial feature matching [26] added nothing. The benefit of class-conditional alignment over naive pooling is a significant, FDR-surviving effect (+0.047 macro-F1, $p = 0.0003$, on the interpretable PSD space), so the recipe's gain is not noise. It is quantified evidence that escalating to higher-order or adversarial objectives is not always warranted.

The decomposition also reframes the wider negative-transfer literature. Multi-source EEG methods increasingly mitigate negative transfer by selecting or re-weighting source domains, e.g., the Multi-source Selective Graph Domain Adaptation Network [32], which down-weights dissimilar sources to avoid harmful transfer. Our finding is complementary and, we argue, more diagnostic: rather than asking which source to admit, we ask why a given source hurts, and separate a fixable cause (construct) from a residual physical cause (device). We test the device attribution directly with a third research-grade dataset: because ds003838 shares COG-BCI's research-grade device class but differs in task and site, it lets us vary the task/site confound while holding device fixed, and device-matched transfer nonetheless exceeds cross-device transfer by +0.125 macro-F1 ($p = 3.2 \times 10^{-7}$) and +0.078 AUC ($p = 2.8 \times 10^{-4}$). This is a strengthening check rather than a proof (ds003838 still differs in channel layout and preprocessing), so we describe the device component as an empirically supported interpretation, not an isolated bottleneck, and enumerate the remaining confounds in the Threats. The residual barrier is also physically consistent with documented differences between consumer- and research-grade EEG: recent device comparisons report systematically narrower effective bandwidth and distorted spectral characteristics for consumer dry-electrode systems relative to research amplifiers [33]. Pairing a 63-channel amplifier with a 14-channel Emotiv introduces front-end discrepancies (impedance, reference scheme, signal-to-noise) that label- and covariance-alignment cannot undo. The transfer asymmetry (COG-BCI benefits from STEW, not the reverse) follows from headroom: STEW's within-dataset model is already strong, leaving little for an auxiliary source to add. This is not an artifact of the harmonized contrast, since single-dataset baselines accompany every transfer gain.

5.3 Positive Transfer Is Achievable, but Model-Class-Dependent

Multi-dataset integration can be turned positive, but the outcome is governed by the interaction of three factors: the label construct, the conditional (class-wise) distribution, and the model class. The same naive pooling that caused robust negative transfer for fixed linear features was, for an end-to-end deep encoder, merely neutral on STEW (+0.010, n.s.) and mildly negative on COG-BCI (−0.017, n.s.). A data-hungry network absorbs the extra examples without the full distribution-shift penalty that degrades a fixed linear probe, but it does not convert them into gain. How much pooling costs is therefore a property of the learner, not of the datasets alone. The more sophisticated intervention, learned domain-adversarial harmonization, did not beat simple pooling for the deep encoder even with λ selected per fold by nested cross-validation

(which favored a small adversary weight), consistent with the well-documented training instability of adversarial domain adaptation [26]. This suggests a practical ordering: respect the conditional distribution (class-conditional alignment) when the representation is fixed, and reserve adversarial or higher-order machinery for cases where the cheaper levers are exhausted. This ordering also clarifies why source-selection strategies [32] and class-conditional alignment can both "work" in the literature while appearing to compete: they operate on different terms of the same conditional-distribution objective. We report the deep-encoder results as exploratory, given the single compact architecture; its domain-adversarial variant, with λ selected by nested cross-validation, added nothing over pooling.

5.4 Foundation Models in Practice

Frozen CBraMod embeddings approached but did not surpass hand-crafted PSD features (best CBraMod STEW macro-F1 0.707 under μV scaling; 0.682 under the z-score normalization we adopt for transfer, for its invariance to device amplitude scale, vs. the PSD stacking decoder's 0.716). This result is in line with the evaluation literature that has matured alongside these models. The critical review of Kuruppu et al. [13] reports that foundation-model generalizability is largely undemonstrated. The channel-adaptation benchmark of Kokate et al. [14] shows that external channel-adaptation methods can cause severe negative transfer when flexible foundation models are fine-tuned. And EEG-FM-Bench [34], which standardizes evaluation across fourteen datasets and ten paradigms, finds that inconsistent protocols make cross-model comparisons unreliable and that fair, like-for-like comparison deflates headline claims. For a 14-channel workload task, the expensive foundation model offered no accuracy advantage over interpretable band-power features. The constructive corollary is that its value here was not raw accuracy but the embedding space in which class-conditional adaptation produced the frozen-embedding route to positive transfer, i.e., foundation models as a substrate for adaptation, not as drop-in classifiers. The frozen-probe protocol mattered, because the same benchmark reports that fine-tuning flexible foundation models can trigger severe negative transfer, which frozen-encoder probing avoids [14].

5.5 Positioning Relative to Reported State-of-the-Art Accuracy

Absolute accuracies in the applied workload literature can appear far higher than ours: a hybrid autoencoder–CNN–gradient-boosting model reports average accuracies around 90% on STEW (92.1% on the SIMKAP block, 89.9% on the no-task condition) [35]. These are, however, predominantly within-subject or within-dataset, the regime that the systematic review of Pušica et al. [29] identifies as the most optimistic and the least comparable across studies. Two regime differences explain and, we argue, justify the gap. First, our validation is cross-subject and, for the transfer experiments, cross-dataset: no subject, and in transfer no device or task-construct, is shared between train and test. Second, the same review documents a marked accuracy drop specifically for multitasking studies with quantitative task-load labels, the MATB/SIMKAP family we study, relative to single-task or subjectively rated paradigms. Our modest numbers are therefore not evidence of a weak decoder but of a stringent, ecologically meaningful evaluation. The contribution is a characterization of when integration helps, which a headline within-subject accuracy cannot provide.

5.6 Threats to Validity

We used linear probes and avoided deep fine-tuning, and the transfer analysis rests on three public datasets. Adopting the leave-one-subject-out MOABB protocol places the statistical evidence on the per-subject distribution (N = 29–47 subjects) rather than a handful of pooled folds, and materially strengthens it: under the family-wide false-discovery-rate correction, five of sixteen tests survive at $q < 0.05$ on the primary AUC metric (seven on macro-F1), including the device de-confound, the class-conditional advantage on COG-BCI, the negative transfer of naive pooling, and the within-dataset stacking advantage on COG-BCI

binary workload. Each surviving effect is also stable across all three independent subsamples. Where an effect is significant only in direction (the graded three-class within-dataset targets and the STEW class-conditional advantage), we say so explicitly and rest no confirmatory claim on it. The learned deep-encoder experiments are exploratory: a single compact architecture, a pool-vs-single effect ranging from neutral (STEW) to mildly negative (COG-BCI) under per-subject evaluation, and a domain-adversarial weight selected by nested cross-validation that recovered nothing, so we advance no confirmatory deep-encoder claim. The device de-confound is a strengthening check rather than a proof: ds003838 still differs from COG-BCI in channel layout and preprocessing, so the device attribution remains an interpretation with those residual confounds acknowledged. The construct-harmonization comparison changes the task difficulty (cleaner contrast), which we control by reporting single-dataset baselines alongside transfer gains. Absolute targets common in the applied literature (>85% accuracy, +40% over single-dataset) were not met for transfer; we report these unmet targets plainly, since our contribution is characterizing why transfer is limited rather than reporting a headline number.

5.7 Deployment and Generalization

The decoders are cheap enough for edge use: the classical specialists are linear probes over low-dimensional features, and the deep encoder has only 3764 parameters and runs at ~0.02 ms per epoch on a GPU, so a 2-s decision costs a negligible fraction of its window and real-time, per-epoch inference on modest hardware is realistic. Two caveats bound generalization, however. First, all three datasets are non-clinical research/consumer recordings from healthy adults; clinical-grade EEG (pathological rhythms, montage and artifact profiles unlike ours, medication effects) would very likely need recalibration, and our cross-device results suggest that moving to yet another acquisition device is precisely where accuracy is most at risk. Second, our absolute macro-F1 values are modest by within-subject standards because the evaluation is cross-subject and, for transfer, cross-dataset; a deployed system would benefit from a short per-user or per-device calibration, which our decomposition predicts should mostly remove the construct and device gaps. Methodologically, because the residual barrier is device/signal domain shift, collecting primary single-device data (e.g., a collaborative-team study on one headset) should remove inter-device shift and make transfer substantially easier. This is the focus of the project's second phase. Continuous workload-index regression could bridge label constructs more finely than a binary contrast, and partial fine-tuning under the safe protocols of [14] may extend the positive-transfer regime. Having benchmarked a modern transformer (EEG-Conformer) here, extending the comparison to state-space (Mamba-style) EEG encoders is a natural next step.

6 Conclusion

We presented a multi-specialist stacking EEG decoder that combines three complementary feature specialists through a meta-learner. Under the leave-one-subject-out MOABB protocol it outperforms single-view specialists on the binary workload contrasts, a target-dependent gain that is significant and, on COG-BCI binary workload, survives family-wide FDR correction. Its performance also tracks the canonical neurophysiological signature of workload. We then provided a systematic account of cross-dataset cognitive-workload transfer between COG-BCI and STEW. Multi-dataset integration can help, though only asymmetrically: class-conditional alignment, combined with construct harmonization, turns transfer positive on the interpretable PSD features, where its advantage over naive pooling of +0.047 macro-F1 survives family-wide FDR correction, and, directionally, on frozen foundation-model embeddings. At the same time we quantify the two barriers to transfer, corroborate the device component with a third dataset, and localize the remaining limit to domain shift. The pipeline is fully reproducible, and the analysis offers a principled guide to when multi-dataset EEG integration is worthwhile.

Acknowledgement: None.

Funding Statement: This research was funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kementerian Pendidikan Tinggi, Sains, dan Teknologi, Kemendiktisaintek) under the Penelitian Fundamental Reguler scheme, Fiscal Year 2026, grant numbers 283/C3/DT.05.00/PL-BARU/2026, 1529/LL4/PG/2026, and 400A/ADB.WD2/SU/KU/2026.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Sugeng Rifqi Mubaroq and Rolly Maulana Awangga; methodology, Sugeng Rifqi Mubaroq; software, Sugeng Rifqi Mubaroq and Tegar Ditya Pragama; validation, Sugeng Rifqi Mubaroq, Tegar Ditya Pragama and Sidiq Fathummubin; formal analysis, Sugeng Rifqi Mubaroq and Sidiq Fathummubin; data curation, Sugeng Rifqi Mubaroq and Tegar Ditya Pragama; writing, original draft preparation, Sugeng Rifqi Mubaroq and Ali Yusuf Abdulhaq; writing, review and editing, Rolly Maulana Awangga and Ali Yusuf Abdulhaq; supervision and funding acquisition, Rolly Maulana Awangga. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed in this study are publicly available: COG-BCI (CC-BY-4.0) at https://doi.org/10.5281/zenodo.6874129, STEW (IEEE DataPort) at https://doi.org/10.21227/44r8-ya50, and ds003838 (OpenNeuro, CC0) at https://doi.org/10.18112/openneuro.ds003838.v1.0.2. The analysis code is available at https://github.com/AgentiAI-neural-decoding/agentic-eeg-decoder and will be released publicly upon publication.

Ethics Approval: Not applicable. This study is a secondary analysis of publicly available, fully de-identified EEG datasets (COG-BCI, STEW, and ds003838), whose original collection was approved by the respective providers' ethics committees and conducted with participant informed consent. No new data were collected and no identifiable human data were processed.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Hyperparameters

Table A1: Model hyperparameters (library defaults stated explicitly).

Component	Setting
Spectral specialist	RBF-SVM, $C = 1.0$ (default), $\gamma =$ scale (default), probability outputs, class-weight balanced; input: standardized 70-d relative log band-power
Riemannian specialist	Tangent-space map (105-d) → standardize → logistic regression ($C = 1.0$, ℓ_2, lbfgs, max_iter 2000, balanced)
CSP specialist	6 CSP components (log-variance) → linear discriminant analysis (svd)
Covariance estimator	Oracle-approximating shrinkage (OAS), 14×14 per epoch
Meta-learner	Logistic regression (balanced) over out-of-fold specialist probabilities; nested inner 10-fold subject-grouped
Deep encoder (EEGNet)	$F_1 = 8$, $D = 2$, $F_2 = 16$, temporal kernel 64, dropout 0.3 (3764 params); Adam, lr 10^{-3}, weight decay 10^{-4}, 25 epochs, batch 128
EEG-Conformer	Convolutional transformer [28], default configuration (224,594 params); Adam, lr 10^{-3}, 30 epochs, batch 128, same LOSO split
Domain-adversarial (DANN)	Gradient-reversal; λ grid $\{0.1, 0.3, 1.0\}$, sigmoid ramp, selected per outer fold by nested inner 5-fold subject-grouped CV
CBraMod probe	Frozen encoder, mean pooling (200-d), per-epoch z-score, logistic-regression probe
Evaluation	Leave-one-subject-out (MOABB CrossSubject); nested inner 10-fold; 100 epochs/subject; 3 independent subsamples; AUC primary, macro-F1 secondary

Appendix B Family-Wide Multiple-Comparison Correction

Table A2 lists every hypothesis test reported in the paper with its raw p value and two corrected values: the Benjamini–Hochberg false-discovery-rate q and the Holm-adjusted p. Tests are pooled into one family per metric (16 tests each). On the primary AUC metric, five tests survive Benjamini–Hochberg at $q < 0.05$ and four survive the stricter Holm correction; on macro-F1, seven survive Benjamini–Hochberg and five survive Holm. Every claim we label as surviving family-wide correction is drawn from these survivors.

Table A2: All hypothesis tests with Benjamini–Hochberg (q_{BH}) and Holm-adjusted values, per metric family.

#	Test	Side	AUC p	AUC q_{BH}	AUC q_{Holm}	F1 p	F1 q_{BH}	F1 q_{Holm}
1	E9 device vs. cross-device	>	2.8×10^{-4}	0.003	0.004	3.2×10^{-7}	5.0×10^{-6}	5.0×10^{-6}
2	E3 ccda COG ext←STEW	>	3.7×10^{-4}	0.003	0.006	3.3×10^{-4}	0.003	0.005
3	E1 stack COG bin	>	7.4×10^{-4}	0.004	0.010	0.335	0.487	1.000
4	E3 pool COG ext←STEW	≠	0.001	0.005	0.015	9.7×10^{-4}	0.004	0.013
5	E3 pool STEW bin←COG	≠	0.004	0.014	0.051	9.1×10^{-4}	0.004	0.013
6	E1 stack STEW bin	>	0.025	0.068	0.280	0.006	0.016	0.065
7	E3 pool COG bin←STEW	≠	0.048	0.108	0.480	0.865	0.885	1.000
8	E3 ccda STEW bin←COG	>	0.054	0.108	0.484	0.008	0.017	0.075
9	E3 ccda COG bin←STEW	>	0.063	0.112	0.504	0.696	0.856	1.000
10	E3 pool STEW ext←COG	≠	0.277	0.443	1.000	0.003	0.009	0.033
11	E1 stack COG MATB	>	0.312	0.454	1.000	0.196	0.314	1.000
12	E3 ccda STEW ext←COG	>	0.373	0.466	1.000	0.051	0.091	0.410
13	E11 deep-pool STEW←COG	>	0.389	0.466	1.000	0.453	0.604	1.000
14	E1 stack COG n-back	>	0.407	0.466	1.000	0.038	0.076	0.342
15	E6 full vs. 14-ch montage	≠	0.550	0.587	1.000	0.885	0.885	1.000
16	E11 deep-pool COG←STEW	>	0.997	0.997	1.000	0.886	0.886	1.000

Note: >, one-sided; ≠, two-sided. E1 stacking gain, E3 transfer (pool and class-conditional), E11 deep-encoder pooling, E6 channel-reduction, E9 device de-confound. Tests ordered by AUC p.

References

1. Young MS, Brookhuis KA, Wickens CD, Hancock PA. State of science: mental workload in ergonomics. Ergonomics. 2015;58(1):1–17. doi:10.1080/00140139.2014.956151.
2. Dehais F, Lafont A, Roy RN, Fairclough S. A neuroergonomics approach to mental workload, engagement and human performance. Front Neurosci. 2020;14:268. doi:10.3389/fnins.2020.00268.
3. Beauchemin N, Charland P, Karran A, Boasen J, Tadson B, Sénécal S, et al. Enhancing learning experiences: EEG-based passive BCI system adapts learning speed to cognitive load in real-time. Front Hum Neurosci. 2024;18:1416683. doi:10.3389/fnhum.2024.1416683.
4. Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Res Rev. 1999;29(2–3):169–95. doi:10.1016/S0165-0173(98)00056-3.
5. Gevins A, Smith ME. Neurophysiological measures of cognitive workload during human–computer interaction. Theor Issues Ergon Sci. 2003;4(1–2):113–31. doi:10.1080/14639220210159717.
6. Kingphai K, Moshfeghi Y. Mental workload assessment using deep learning models from EEG signals: a systematic review. IEEE Trans Cogn Dev Syst. 2025;17(1):40–60. doi:10.1109/tcds.2024.3460750.
7. Hinss MF, Jahanpour ES, Somon B, Pluchon L, Dehais F, Roy RN. Open multi-session and multi-task EEG cognitive dataset for passive brain–computer interface applications. Sci Data. 2023;10(1):85. doi:10.1038/s41597-022-01898-y.
8. Lim WL, Sourina O, Wang L. STEW: simultaneous task EEG workload dataset. IEEE Trans Neural Syst Rehabil Eng. 2018;26(11):2106–14. doi:10.1109/TNSRE.2018.2872924.

9. Wang J, Zhao S, Luo Z, Zhou Y, Jiang H, Li S, et al. CBraMod: a criss-cross brain foundation model for EEG decoding. In: Proceedings of the International Conference on Learning Representations (ICLR); 2025 Apr 24–28; Singapore. p. 62056–92. doi:10.48550/arXiv.2412.07236.
10. Jiang WB, Zhao LM, Lu BL. Large brain model for learning generic representations with tremendous EEG data in BCI. In: Proceedings of the International Conference on Learning Representations (ICLR); 2024 May 7–11; Vienna, Austria. doi:10.48550/arXiv.2405.18765.
11. Wang G, Liu W, He Y, Xu C, Ma L, Li H. EEGPT: pretrained transformer for universal and reliable representation of EEG signals. In: Proceedings of the 38th International Conference on Neural Information Processing Systems; 2024 Dec 10–15; Vancouver, BC, Canada. Vol. 37, p. 39249–80. doi:10.52202/079017-1239.
12. Zhou Y, Wang P, Gong P, Wan P, Wen X, Zhang D. Cross-subject mental workload recognition using bi-classifier domain adversarial learning. Cogn Neurodyn. 2025;19(1):16. doi:10.1007/s11571-024-10215-9.
13. Kuruppu G, Wagh N, Kremen V, Varatharajah Y. EEG foundation models: a critical review of current progress and future directions. J Neural Eng. 2026;23(2):021001. doi:10.1088/1741-2552/ae4455.
14. Kokate K, Aristimunha B, Truong D, Delorme A. Channel adaptation for EEG foundation models: a systematic benchmark across architectures, tasks, and training regimes. arXiv:2604.23091. 2026. doi:10.48550/arXiv.2604.23091.
15. Chen J, Li S, Pi D. Cross-subject domain adaptation for classifying working memory load with multi-frame EEG images. J Supercomput. 2025;81(13):1272. doi:10.1007/s11227-025-07748-z.
16. Rodrigues PLC, Jutten C, Congedo M. Riemannian Procrustes analysis: transfer learning for BCIs. IEEE Trans Biomed Eng. 2019;66(8):2390–401. doi:10.1109/TBME.2018.2889705.
17. He H, Wu D. Transfer learning for brain–computer interfaces: a Euclidean space data alignment approach. IEEE Trans Biomed Eng. 2020;67(2):399–410. doi:10.1109/TBME.2019.2913914.
18. Pavlov YG, Kasanov D, Kosachenko AI, Kotyusov AI, Busch NA. Pupillometry and electroencephalography in the digit span task. Sci Data. 2022;9(1):325. doi:10.1038/s41597-022-01414-2.
19. Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MNE software for processing MEG and EEG data. NeuroImage. 2014;86(2):446–60. doi:10.1016/j.neuroimage.2013.10.027.
20. Pion-Tonachini L, Kreutz-Delgado K, Makeig S. ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. NeuroImage. 2019;198(5):181–97. doi:10.1016/j.neuroimage.2019.05.026.
21. Bigdely-Shamlo N, Mullen T, Kothe C, Su KM, Robbins KA. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. Front Neuroinform. 2015;9:16. doi:10.3389/fninf.2015.00016.
22. Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass brain–computer interface classification by Riemannian geometry. IEEE Trans Biomed Eng. 2012;59(4):920–8. doi:10.1109/TBME.2011.2172210.
23. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59. doi:10.1016/S0893-6080(05)80023-1.
24. Sun B, Saenko K. Deep CORAL: correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops. Lecture Notes in Computer Science. Vol. 9915. Berlin/Heidelberg, Germany: Springer; 2016. p. 443–50. doi:10.1007/978-3-319-49409-8_35.
25. Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. J Neural Eng. 2018;15(5):056013. doi:10.1088/1741-2552/aace8c.
26. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. J Mach Learn Res. 2016;17(59):1–35. doi:10.1007/978-3-319-58347-1_10.
27. Jayaram V, Barachant A. MOABB: trustworthy algorithm benchmarking for BCIs. J Neural Eng. 2018;15(6):066011. doi:10.1088/1741-2552/aadea0.
28. Song Y, Zheng Q, Liu B, Gao X. EEG Conformer: convolutional transformer for EEG decoding and visualization. IEEE Trans Neural Syst Rehabil Eng. 2023;31:710–9. doi:10.1109/TNSRE.2022.3230250.
29. Pušica M, Mijović B, Leva MC, Gligorijević I. Machine learning performance in EEG-based mental workload classification across task types: a systematic review. Front Neuroergonomics. 2025;6:1621309. doi:10.3389/fnrgo.2025.1621309.

30. Chen Z, Zhang Y, Lan Q, Liu T, Wang H, Ding Y, et al. Uni-NTFM: a unified foundation model for EEG signal representation learning. In: Proceedings of the International Conference on Learning Representations (ICLR); 2026 Apr 23–27; Rio de Janeiro, Brazil. doi:10.48550/arXiv.2509.24222.
31. Long M, Cao Z, Wang J, Jordan MI. Conditional adversarial domain adaptation. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, QC, Canada. Vol. 31. doi:10.48550/arXiv.1705.10667.
32. Wang J, Ning X, Xu W, Li Y, Jia Z, Lin Y. Multi-source selective graph domain adaptation network for cross-subject EEG emotion recognition. Neural Netw. 2024;180(1):106742. doi:10.1016/j.neunet.2024.106742.
33. Mikhaylov D, Saeed M, Alhosani MH, Al Wahedi YF. Comparison of EEG signal spectral characteristics obtained with consumer- and research-grade devices. Sensors. 2024;24(24):8108. doi:10.3390/s24248108.
34. Xiong W, Li J, Li J, Zhu K, Jiang C. EEG-FM-Bench: a comprehensive benchmark for the systematic evaluation of EEG foundation models. arXiv:2508.17742. 2025. doi:10.48550/arXiv.2508.17742.
35. Abinaya G, Dinakaran K. ACXNet hybrid deep learning model for cross-task mental workload estimation using EEG neural manifolds. Sci Rep. 2025;15(1):35178. doi:10.1038/s41598-025-19144-x.