STRUM: A DYADIC MULTIMODAL BENCHMARK FOR COMPARISON OF PHYSIOLOGICAL MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation learning for physiological signals has focused on scale, prioritizing longer recordings and larger cohorts. How the signals compare to one another is harder to establish: cross-modality comparisons are typically drawn from separate studies, each with its own cohort and instrumentation, so any apparent difference carries everything else that varied along with it. Isolating it requires recording every candidate modality from the same participants simultaneously, which no EEG benchmark provides. We present STRUM (Small Team Reconnaissance Urban Missions), a corpus of 66 people recorded as 33 pairs playing a cooperative mission game, with both partners instrumented at once across eleven stream families on a single clock: 205-channel scalp EEG each, muscle, cardiac, ocular, electrodermal and respiratory channels, eye tracking, motion capture, the in-game avatar, a force plate and the controller, over roughly 87 h, with the two participants' neural recordings aligned to a median of 0.208 ms. We define two continuous, event-free regression tasks on the corpus, predicting a person's phasic electrodermal arousal and mean heart rate from the remaining simultaneously acquired channels, evaluated under leave-one-dyad-out cross-validation, released with fixed splits and a reference loader. Across both targets, modality choice affects decoding more than model choice: the neck muscle band alone reaches $r = 0.4449$ on heart rate, ahead of every other modality in the same comparison, and pretrained encoders on EEG do not surpass a small network trained directly on muscle on matched windows. Multimodal fusion improves further even with encoders reused off the shelf: the full suite reaches 0.5072 against the best single modality's 0.4449, a gain of 0.062. Simultaneous acquisition separates an apparent neural result into the muscular, ocular and cardiac signal inside it, and recording both partners makes pair-level targets measurable. For continuous autonomic decoding, the instrumentation chosen at acquisition constrains performance more than the encoder chosen at training.

1 INTRODUCTION

Physiological foundation models have scaled along two axes. Neural models pretrain on ever larger corpora of a single modality, acquired one person at a time (Jiang et al., 2024; Wang et al., 2025; 2023), while wearable models push the other direction, across far more people and far more hours, with no neural signal at all (Narayanswamy et al., 2025). Which signals are worth recording is a harder question, and answering it requires something else: every candidate signal recorded from the same people at the same moments, so that one can be withheld with the others held fixed. Existing multimodal corpora approach that requirement from different directions, and none of them meets it. Some record rich peripheral channels and no neural signal (Meier et al., 2024; Chaptoukaev et al., 2023), others pair a neural recording with a single behavioral stream (Lin et al., 2026b; Kastrati et al., 2021), and the few that instrument more than one person at a time do so with far fewer channels on each (Yamasaki et al., 2026; Pyarelal et al., 2023).

STRUM (Small Team Reconnaissance Urban Missions) records 66 people as 33 pairs while they play a cooperative mission game, with both members instrumented at once across eleven stream families on one clock (Figure 1).

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.

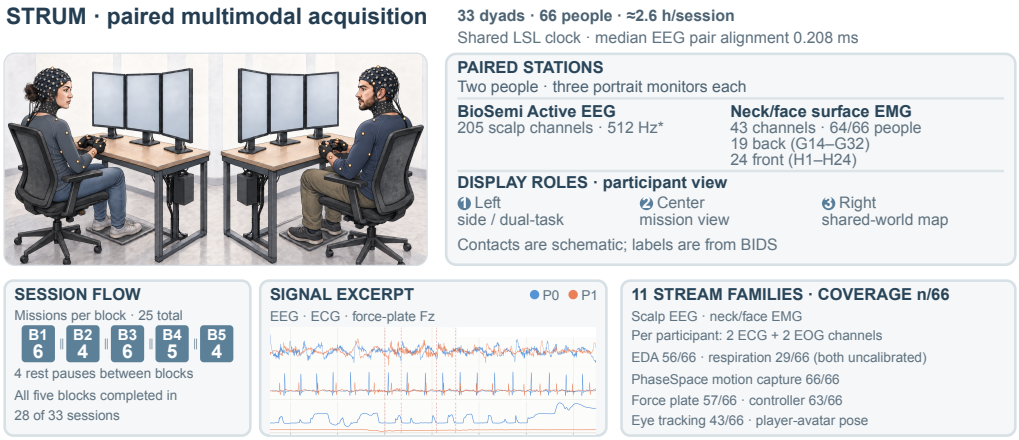


Figure 1: STRUM paired multimodal acquisition. **(A)** Two participants at paired workstations, each with three portrait displays. Sensor contacts are schematic; channel labels are from BIDS. BioSemi Active EEG has 205 scalp channels at 512 Hz*; the separate neck/face surface-EMG array has 43 channels (19 posterior and 24 anterior/face), present for 64/66 participants. The linked PhaseSpace BIDS example records 29 points per participant. **(B)** Five ordered blocks contain 25 missions (6/4/6/5/4) and four rest pauses; all five were completed in 28/33 sessions. **(C)** Example paired EEG, ECG, and force-plate signals. **(D)** Eleven stream families. Each participant has two ECG and two EOG channels; ECG is absent in two sessions. EDA and respiration are uncalibrated. *One amplifier ran at 2048 Hz; the small-cap EEG channel count remains unresolved.

Every stream sits on one synchronization framework (Kothe et al., 2025), and we record the rest periods between missions as well. We release it in BIDS (Gorgolewski et al., 2016) with fixed splits, inclusion lists and a loader.

Three research questions follow, which are answered in Section 5:

1. Which of the simultaneously recorded modalities carry information about a person's autonomic state, and which are still needed once the others are present?
2. What can a better model make up for, and what can it not?
3. What does recording two people at once make measurable that a single-person corpus cannot?

Across both targets, which modality we record matters more than which model we run on it, and combining modalities beats any single one even when every encoder is reused off the shelf. Pretraining pays within a modality: it gives a large gain over the same architecture trained from scratch. Instrumenting both partners at once supports two measurements that no single-person corpus admits. The two targets do not always agree and we report where they diverge.

Contributions.

- **The STRUM corpus**: both members of a pair instrumented across eleven stream families on one clock with the synchronization quantified, converted to BIDS with an annotated event timeline.
- **A benchmark and an evaluation protocol**: two continuous, event-free regression targets under leave-dyad-out with released splits and a loader, and a leave-one-modality-out necessity protocol reporting a minimum detectable effect on every cell.
- **Evidence that modality choice outweighs model choice, and that combining modalities beats any single one**, over various classical, from-scratch, and pretrained arms with matched controls.
- **Same-instant attribution, and a pair-level cardiac measurement**: how much of an apparent neural effect we can attribute to muscle and ocular channels from the same person at the same moment, and cardiac covariation between partners present in the signals and open to modeling.

Table 1: Positioning against peer resources. *Reported timing* is what each source states about inter-stream or inter-device alignment. *MIntRec2.0's parties are not simultaneously sensor-recorded.

Resource	At once	Neural	Other modalities	N / hours	Reported timing
STRUM *(ours)*	**2**	scalp, 205 ch	EMG, ECG, EOG, EDA, resp., eye, motion, avatar, force, controller	66 / 87 h; 2.6 h per dyad	11 families on one clock; inter-brain 0.208 ms
ToMCAT (Pyarelal et al., 2023)	**3**	scalp $\leq$32 ch + fNIRS	ECG, EDA, eye, face, audio, player position	40 teams / 3 h per team	LSL, unquantified
DUET (Yamasaki et al., 2026)	2	scalp, 64 ch (one pilot dyad at 32)	speech, facial EMG	18 dyads / 8 $\times$ 4 min per dyad	daisy-chained
EEGEyeNet (Kastrati et al., 2021)	1	scalp, 128 ch	eye tracking	356 / 47 h	$\leq$2 ms
EgoBrain (Lin et al., 2026b)	1	scalp, 32 ch	egocentric video	40 / 61 h	—
WildPPG (Meier et al., 2024)	1	none	PPG, ECG, accel., temp., altimeter	16 / 216 h	33 ms
StressID (Chaptoukaev et al., 2023)	1	none	ECG, EDA, resp., video, audio	65 / 39 h	qualitative only
Wearable corpus (LSM) (Narayanswamy et al., 2025)	1	none	HR, HRV, EDA, accel., temp., altimeter	>165,000 / $\sim$40M h	—
MIntRec2.0 (Zhang et al., 2024)	1*	none	text, video, audio	1,245 dialogues / 12.3 h	—

2 Related work

Multimodal physiological corpora generally fall into two groups. The first pairs a neural recording with one behavioral stream: scalp EEG with egocentric video, or scalp EEG with eye tracking (Lin et al., 2026b; Kastrati et al., 2021). Each is recorded one participant at a time and released as a supervised benchmark, with no autonomic channel on the person wearing the cap. The second group sets the neural signal aside and scales everything around it, from laboratory sessions to field recordings to population-scale wearable corpora (Chaptoukaev et al., 2023; Meier et al., 2024; Narayanswamy et al., 2025). Table 1 sets them side by side.

The literature on simultaneous multi-person recording is smaller. Hyperscanning corpora of natural conversation record both partners' brains to study how their activity couples. One such release covers 18 dyads at 64 scalp channels each, and records speech and facial EMG (Yamasaki et al., 2026). A triadic team-cognition corpus instruments all three subjects with scalp EEG and functional near-infrared spectroscopy, adding cardiac, electrodermal and eye-tracking channels, face video, audio and player position (Pyarelal et al., 2023). Neither corpus records optical motion capture of the participants' bodies, so movement cannot be weighed against physiology in either.

Pretrained encoders carry the same limitations as the corpora behind them. Scalp-EEG encoders learn by masked prediction over aggregated clinical and research recordings of up to 25,000 participants, and are scored predominantly on classification (Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025). Away from the brain, cardiac encoders pretrain on large clinical databases (Li et al., 2025), and streams with no physiological encoder of their own fall to general time-series models (Goswami et al., 2024). Multimodal physiological encoders differ in which modalities they train together: SleepFM and OSF pretrain across the neural, cardiac and respiratory channels of polysomnography (Thapa et al., 2024; 2026; Shuai et al., 2026), while BIOT trains modality-siloed checkpoints (Yang et al., 2023) and PhysioWave trains separate single-modality models, with its muscle branch trained on forearm gesture corpora (Chen et al., 2025). That muscle branch is the closest pretraining to our own suite, yet it covers neither neck EMG recorded during whole-body movement, nor optical motion capture paired with physiology, nor two people at once. Several of the stream families we release therefore have no pretrained encoder to fine-tune. We ask whether pretrained encoders repay their cost (Yang et al., 2026) across modalities rather than within a single

modality. We benchmark the released encoders in both frozen and fine-tuned form, alongside from-scratch and classical baselines. LaBraM is absent from that roster because its pretrained checkpoint is not publicly available.

Most evidence about which signals predict autonomic state comes from studies that recorded the candidate signals apart. Earlier work decodes cardiac and electrodermal state continuously from neural and peripheral channels (Shehu et al., 2023); another study already regresses per-participant targets with whole dyads held out (Chowdhury, 2026). A parallel line of work removes person-specific structure by alignment or by adversarial invariance (Junqueira et al., 2024; Rayatdoost et al., 2021), and an audit of identity information in neural recordings reports that erasing it improves categorical decoding (Lin et al., 2026a). Recording them separately mixes the modality with its cohort and instrumentation, and separating them takes **one recording that carries every candidate signal at the same instants**; a target defined over a pair takes a second suite beside it, with the two suites' alignment measured. We know of no resource that carries both.

3 THE STRUM DATASET

The STRUM dataset consists of 33 sessions, one per dyad, each of which records two people working together for about 2.6 h. The streams run throughout for all 66 participants, so 87.4 h of unique recording is 174.8 subject-hours, and every stream from the pair sits on one **Lab Streaming Layer** clock (Kothe et al., 2025); Figure 1 shows the station, a session, and every modality from both partners at one moment. The corpus will be released at `https://nemar.org/dataset/nm000284`; code, splits and an interactive inventory of every session are at `https://strum-neuro.netlify.app`.

The two participants play a **cooperative urban-reconnaissance game** structured in five blocks separated by rest pauses. Five mission types recur across the blocks, two individual and three cooperative. In an example individual mission a player sweeps the camera and reports what they see while the other watches for a warning light; in one cooperative mission a player reads an aerial drone view and talks the other through the route while that partner drives blind on what they hear. The task software (SNAP) fixes that sequence, so it is the same in every dyad. All five blocks ran in 28 of the 33 sessions; the rest ended early. The recording is continuous, and we release it both whole and cut into one file per block and per pause, with a machine-generated timeline of 55 event classes annotated with Hierarchical Event Descriptors (Robbins et al., 2021) (Figure 1).

We recorded **eleven stream families**, with per-modality coverage in Table 2. One BioSemi amplifier per participant carries the scalp EEG, the neck EMG band, the cardiac and ocular externals, and electrodermal and respiratory auxiliaries. Room-scale PhaseSpace motion capture follows both bodies with active LED markers. An AMTI force plate sits under each participant, the task software writes a per-player in-game avatar pose stream, and the wired controller reports 36 axes, of which four to five carry signal. Participants wear head-mounted eye tracking over the EEG cap. The respiration belt is wired to one amplifier per session, so it reaches one member of each pair. Six montage variants and two versions of the task software appear across the corpus. The tree is valid BIDS under the EEG and motion extensions (Gorgolewski et al., 2016; Pernet et al., 2019; Jeung et al., 2024).

Across all 40 recordings, the two participants' EEG streams align to a median of 0.208 **ms**, and 35 of the 40 recordings fall below 0.5 ms (Figure 2). **The two brains fall inside a single sample in every recording**: the worst is 1.745 ms against a sample that spans 1.95 ms at 512 Hz, and the phenomena modeled here move four orders of magnitude slower.

We then followed blinks across the streams to show that they describe the same events: they appear in frontal EEG, in the ocular externals and as pupil dropout in the eye tracker. We showed that BLINKER (Kleifges et al., 2017) and the eye tracker agree on 65 % of one participant's 254 blinks within ± 300 ms (Figure 5 in the appendix).

4 BENCHMARK AND PROTOCOL

Two event-free regression targets. The benchmark's task is to predict phasic electrodermal arousal and mean heart rate, two measures of one person's autonomic state, from every other signal recorded

Table 2: STRUM at a glance. Coverage denominators are per participant out of 66.

Property	Value
Participants / dyads / sessions	66 / 33 / 33 (one session per dyad)
Demographics	mean age 20.3 (18-27); 54 male / 12 female; 58 right-handed / 8 left-handed
Session-runs / subject-runs	40 / 80 (seven sessions have a second run)
Duration	87.4 h unique; 174.8 subject-hours
Scalp EEG	205 channels at 512 Hz (160 channels for one dyad; 1 amplifier at 2048 Hz)
Neck EMG	43 channels, 64/66 participants
Amplifier externals	2 ECG + 2 EOG per amplifier; the ECG pair is absent for two sessions
Autonomic	electrodermal 56/66, respiration 29/66 — both uncalibrated
Motion / force / controller	optical motion capture 66/66, force plate 57/66, controller 63/66
Eye tracking	gaze, scene, and area-of-interest streams, 43/66
Task events	$\sim$960k rows on the shared timeline, 55 distinct event classes
Audio / video	present for most sessions; identifiable, restricted, not released as subject data
Format	BIDS: EEG and EMG as BDF; motion, physiology, and events as TSV with JSON sidecars

at the same moment. We score each as a Pearson correlation per participant, averaged over the held-out fold. Electrodermal arousal is analyzable for 54 participants and heart rate for 50. The electrodermal target is the phasic part of the uncalibrated bipolar channel at 32 Hz, separated from the tonic part by median smoothing (Makowski et al., 2021) and rebuilt independently with a convex-optimization decomposition (Greco et al., 2016). Predictors are 15 s windows, the target window sits $+2.5$ s later, and the grid overlaps, so 40–60 % of windows share samples with a neighbor. The heart-rate target comes from a bipolar ECG derivation, with the R-peak detection and interval criteria given in the appendix; both targets are z-scored per participant and winsorized. The **full suite** reads every stream group but eye: each pretrained encoder runs once per stream, the per-stream embeddings are concatenated before a trained head, and one alignment step uses the held-out participant's own unlabeled inputs (Junqueira et al., 2024).

Target reliability, measured by parallel forms for electrodermal arousal and a Spearman–Brown corrected split half for heart rate, caps what a perfect predictor could reach: the full suite reaches 59 % of that ceiling on electrodermal arousal and 67 % on heart rate.

Splits and leakage. We hold out both members of a dyad together, because the partners share a byte-identical task timeline, and any finer split trains on one and tests the other. We release five splits with the fold assignments, inclusion lists and a loader, all of them grouped by dyad or by block. Analysis populations differ by what each arm requires: the necessity comparison uses 26 dyads, the channel-density sweep on electrodermal arousal 27 because it needs EEG alone, any eye-inclusive analysis 17, and the cross-person comparison 50 participants on each target because it needs a per-subject model on both sides. Every number below uses **leave-dyad-out**; the other four splits cover a fixed partition, inner validation, within-participant blocks and calibration. A random split within a participant inflates the correlation by $+0.094$ to $+0.461$ across encoder–target pairings and window-overlap settings, so we report the inflation per pairing.

Three training populations. The **group model** trains on all training dyads and tests on the held-out dyad; every number in the paper comes from it unless labeled otherwise. The **own-person model** trains and tests inside one participant, split by block under the calibration family; the **single-donor model** instead trains on one other participant. The **cross-person gap** is the own-person model's correlation minus the group model's.

The circularity rule we enforce in code. Mean heart rate is computed from the participant's own ECG, so ECG may never predict it, and the scripts abort on a forbidden (predictor, target) pairing. The ban does not exclude every predictor that carries the target's own generating signal: the neck EMG passband retains cardiac content, and average-referenced EEG can carry it too. Every heart-rate claim below is therefore a paired contrast between arms scored on the same folds and windows.

Removing the schedule confound. All 33 dyads receive the same 25 missions in the same order, so the schedule confound survives leave-dyad-out. We remove the variance it explains by subtracting

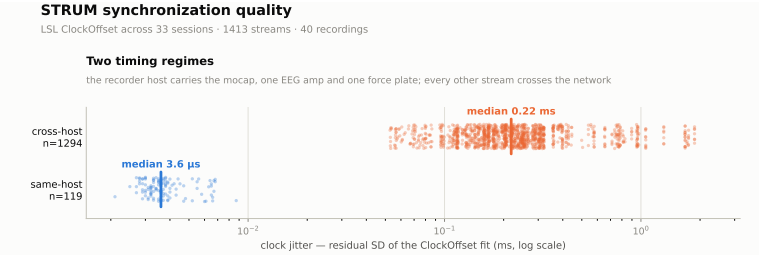


Figure 2: Two timing regimes over all 1,413 stream rows from the 40 recordings. Streams sharing the recorder's host align to a median of 3.60 µs, and streams crossing the network to a median of 0.218 ms. Horizontal position is the residual standard deviation of the per-stream ClockOffset fit; vertical position is cosmetic jitter.

each mission's own average, and then a smooth trend in elapsed time that we fit on the training dyads alone. Removing that variance costs almost nothing on electrodermal arousal ($0.3238 \rightarrow 0.3205$) and proportionally about ten times as much on heart rate ($0.4670 \rightarrow 0.4203$), both for a single pretrained encoder's embeddings scored pooled-window: **the schedule carries substantially more of the heart-rate signal than of the electrodermal one**. The strongest schedule-only predictor we can build predicts each held-out window from what the other dyads were doing at that moment, and stays an order of magnitude below the full suite on both targets.

The pair-level quantity. We correlate the two partners' target series within each game condition, and report that correlation minus its value under surrogate partner assignments, averaged over assignments, with the spread across them. We sweep the relative lag between the series and report the profile with every pair-level effect.

Statistical protocol. Every comparison is paired on per-fold differences, since the folds share dyads and evaluation windows. Wilcoxon signed-rank is primary, with a sign test as a robustness row. Intervals are percentile bootstraps over 10,000 resamples of the per-fold differences, and each interval states its n in folds, dyads or subjects, never windows. We control the false discovery rate with Benjamini–Hochberg (Benjamini & Hochberg, 1995) inside 71 families. Every null carries a minimum detectable effect at 80 % power, and a null counts as tight only when the interval bound in the direction of the hypothesized effect sits below that minimum.

5 Results

Table 3 gives each input under the model families we ran, and Table 4 shows what each modality is worth alone and what removing it costs the full suite. We comment on both throughout this section.

5.1 Muscle and movement carry autonomic state, and combining modalities beats any one of them. On both targets **the suite outperforms every modality inside it** (Tables 3 and 4).

The strongest single modality on mean heart rate is the neck EMG band: separated from the neural group, it reaches $\mathbf{0.4449}$ on the raw target, and 0.431 once each mission's own average is centered out. With the neck band separated out, EEG adds nothing detectable to mean heart rate, a tight null under the raw target and both schedule removals. That block reaches the model as eight scalp channels and the two ocular externals, through a sleep-pretrained encoder, so it bounds what those features carry and leaves open what the montage does. The force plate shows the same shape: it is third on heart rate by itself, while removing it from the full suite leaves an interval that spans zero, a tight null by its own minimum detectable effect.

Adding the behavioral modalities to a physiological baseline that already excludes ECG is worth $+\mathbf{0.0623}$ on mean heart rate over the strongest single group. Optical motion capture is the one behavioral modality that leave-one-modality-out finds necessary. The fixed mission schedule does not explain it: centering within mission raises its drop to $+0.0550$ ($q < 0.001$).

On phasic electrodermal arousal the ordering changes: optical motion capture alone leads every other modality at $\mathbf{0.376}$, and removing it costs $+\mathbf{0.0349}$ after the same centering ($q = 0.003$).

Table 3: Each input under three kinds of model, and the multi-input configurations. Pearson r under leave-dyad-out, pooled over folds. Cells come from separate runs over different window sets, so they are not paired and differences between them are not tested.

		Input	ridge	CNN	pretrained encoder
Phasic electrodermal arousal	unimodal	EEG	0.2179	0.1461	0.1981 (CBraMod)
		neck EMG	0.2434	0.1858	0.2030 (PhysioWave)
		ECG	0.2377	**0.3391**	0.2618 (ECGFounder)
		eye	0.1500	0.0837	0.1530 (MOMENT-1-large)
	multimodal	ECG + neck EMG	—	—	0.2219 (PhysioWave)
		every group but eye	—	—	**0.4004** (full suite)
Mean heart rate	unimodal	EEG	**0.3050**	0.1365	0.2377 (OSF)
		neck EMG	0.3024	0.3674	**0.4768** (SleepFM)
		eye	0.2199	0.1386	—
	multimodal	EEG, EOG + neck EMG	—	—	0.4954 (SleepFM)
		every group but eye	—	—	**0.5072** (full suite)

Physiology comes third alone and adds nothing detectable once the rest are present. We can ask whether a behavioral predictor reads an autonomic response or the movement itself. We swept the lag between predictor and target: the physiological banks peak 1.3 to 1.9 s ahead of the electrodermal target, where an autonomic response belongs, while motion capture peaks at 0.2 s there and at 2.5 s on mean heart rate. We therefore question whether the electrodermal behavioral result reflects autonomic coupling, while the heart-rate one carries the latency a physiological response requires. On electrodermal arousal one modality carries almost all of what is decodable, while on mean heart rate two groups genuinely combine.

5.2 Model quality and modality choice are separate levers. Within one modality, pretraining pays: across 25 leave-dyad-out folds on mean heart rate, with all arms on EEG alone, every pretrained encoder we tested beats a convolutional network trained from scratch on identical windows. SleepFM gains $+\mathbf{0.3667}$ (Thapa et al., 2024), CBraMod $+0.1384$ (Wang et al., 2025) and BIOT $+0.1086$ (Yang et al., 2023). The inputs we gave those encoders span more than an order of magnitude — CBraMod the full scalp montage, BIOT sixteen bipolar derivations, SleepFM eight — while the ordering of their gains does not follow that span.

The same from-scratch architecture reaches $\mathbf{0.3674}$ on the neck EMG band and $\mathbf{0.1365}$ on EEG, one stream at a time, over a different window set and aggregation from Table 4. On matched windows, no pretrained encoder clearly surpasses a small network trained directly on the neck band on either target; on each target at least one encoder is significantly worse. No encoder in the roster was pretrained on surface EMG recorded during movement; the closest is a sleep-polysomnography model carrying chin EMG (Thapa et al., 2024).

Holding one encoder fixed and cutting its scalp input to 19 channels changes the result by 0.004, so input width contributes little inside that encoder. A classical pipeline behaves differently. **Scalp channels do carry information that a wider model can use**: from eight channels to 192 (15 excluded for signal quality), a band-power pipeline with its ridge penalty tuned at each channel count gains $+\mathbf{0.0550}$ $[+0.0303, +0.0799]$ on mean heart rate ($p = 0.0002$, $22/26$ dyads) and $+\mathbf{0.0549}$ $[+0.0238, +0.0868]$ on electrodermal arousal ($p = 0.004$, $20/27$ dyads). The curve has not plateaued at the widest setting we ran. These are handcrafted features, so the result does not say what a pretrained encoder would do with the same width.

Twenty-two handcrafted kinematic features of head and chest, read by the same trained output layer as every other arm, reach $\mathbf{0.409}$ on electrodermal arousal and 0.403 on mean heart rate. On that first target they sit above every pretrained encoder they were measured against.

Against an identical-width shuffled-code control, a per-participant code gains $+\mathbf{0.1850}$ $[+0.1424, +0.2294]$ on mean heart rate, $26/26$ dyads, $q = 5.96\text{e}{-08}$ (Appendix Figure 7), and erasing that direction in a separate analysis costs 0.0891 ($q = 0.0012$). How the code enters matters more than the capacity it is given: an additive code gains exactly 0.0000, while a multiplicative gate captures about two and a half times what a hidden layer does, $+\mathbf{0.3589}$ on the raw target. An own-person model's advantage over the group model reverses between targets, so we scope personalization to cardiac decoding (Appendix Figure 6).

Table 4: Leave-one-modality-out necessity, over the 26 dyads with every modality. *Alone* is each modality's own correlation; *drop* is the paired loss from removing it, with a 95 % bootstrap interval, a Benjamini–Hochberg-corrected q, and the effect detectable at 80 % power (MDE80).

Group	dims	alone	drop	95 % CI	q	MDE80	k/n
Mean heart rate — full suite $r = 0.5072$, 26 dyads, neural block ECG-free							
neck EMG	128	0.4449	$+0.0973$	$[+0.0299, +0.1569]$	**0.020**	0.0962	19/26
motion capture	63	0.3692	$+0.0312$	$[+0.0069, +0.0536]$	**0.020**	0.0351	19/26
in-game avatar	13	0.1838	$+0.0049$	$[-0.0054, +0.0148]$	0.272	0.0153	18/26
force plate	19	0.3138	$+0.0041$	$[-0.0065, +0.0143]$	0.408	0.0160	16/26
controller	21	-0.0080	-0.0029	$[-0.0130, +0.0065]$	0.803	0.0146	12/26
neural	128	0.2478	-0.0111	$[-0.0242, +0.0009]$	0.233	0.0192	10/26
Phasic electrodermal arousal — full suite $r = 0.4004$, 26 dyads, neural block includes cardiac							
motion capture	63	0.3760	$+0.0403$	$[+0.0223, +0.0585]$	**0.002**	0.0275	21/26
force plate	19	0.3267	$+0.0085$	$[+0.0005, +0.0163]$	0.158	0.0118	18/26
controller	21	0.0186	$+0.0003$	$[-0.0070, +0.0075]$	0.727	0.0111	15/26
neck EMG	128	0.2858	-0.0004	$[-0.0082, +0.0080]$	0.580	0.0121	11/26
in-game avatar	13	0.1043	-0.0023	$[-0.0091, +0.0049]$	0.580	0.0106	12/26
neural	256	0.3200	-0.0049	$[-0.0176, +0.0088]$	0.580	0.0199	9/26

5.3 What recording two people at once makes measurable. Recording both partners makes a quantity defined over the pair measurable, and supplies controls that need the partner's signals from the same moment. Cardiac signals covary more between true partners than between surrogates during cooperative missions, by $+\mathbf{0.1114}$ with a standard deviation of 0.0286, positive under six of the seven surrogate assignments (Figure 3). It survives cross-session pseudo-dyads and a respiration control, and it is specific to the cardiac signal: the same analysis applied to motion finds no such coupling. No representation we tried recovers it, though each test's minimum detectable effect falls below $+0.1114$.

The second measurement masks a multimodal encoder's input slots one at a time. On mean heart rate the encoder starts at $\mathbf{0.6959}$ with all eight slots filled, two of which carry EEG. That starting point is a decomposition baseline, not a decoding result: masking the participant's own cardiac channels brings the level to 0.5484, and EEG alone to $\mathbf{0.2700}$, a retention of 38.8%. On electrodermal arousal the same three steps give 0.3503 with every slot filled, 0.3330 after the cardiac mask and 0.2373 with EEG alone, a retention of 67.7%. Masking the cardiac channels costs 21.2% on mean heart rate and 4.9% on electrodermal arousal, $\mathbf{4.3}\times$ more where the target derives from ECG, so the electrodermal decomposition is free of cardiac circularity. Muscle and ocular channels account for $\mathbf{50.8}\%$ of what survived the cardiac mask on mean heart rate.

6 Discussion and limitations

Five readings carry over from the results. The **full suite outperforms every single modality** on both targets, and leave-one-modality-out shows which of them cannot be replaced; because the suite reaches much of what target reliability permits (Section 4), correlations that look modest sit closer to the attainable ceiling than they appear. A **better model and a better sensor are separate levers**: pretraining lifted EEG at every step we measured, while no pretrained encoder clearly cleared a small network trained directly on the neck band, so the larger gain came from the sensor. The encoders differ not by channel count but by how much of it the model can use, since a regularized classical pipeline is still improving at 192 channels. And a person code has to enter a model in the right stage: gating the features yields substantial gains, whereas appending the same code to the input yields none. And recording both partners splits an apparent neural result into the muscle and ocular signal inside it, half of what survives the cardiac mask on heart rate, and exposes a pair-level coupling no representation we tried recovers.

Some questions lie beyond what this corpus can settle. Each participant contributed one session, so every personalization result here is within-session, and these data cannot test cross-session transfer. Every encoder reads one modality at a time, so the joint configuration these models were pretrained in stays untested. The channel-density result rests on one encoder and one substitution, so it bounds that encoder's input width. Delaying either partner's signal weakens the pair-level coupling equally, so neither leads the other, and we describe it as co-variation under a shared task state.

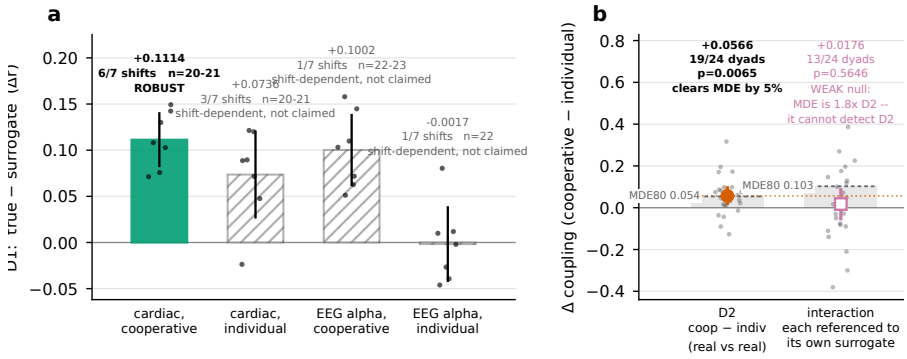


Figure 3: Partner cardiac coupling exceeds its surrogate, and only during cooperation. (a) True-partner minus surrogate-partner coupling, for cardiac and EEG-alpha signals in cooperative and individual missions; bars are the mean over seven surrogate assignments with ± 1 SD. (b) The cooperative-versus-individual contrast over $n = 24$ dyads, with a $95\,\%$ bootstrap interval and the dashed line at the effect detectable at $80\,\%$ power. The two panels are different contrasts and are not read off one another, and a surrogate control bounds the shared-timeline confound from below.

Coverage is uneven across modalities; the electrodermal and respiratory channels are uncalibrated, so every quantity taken from them is relative. On electrodermal arousal the behavioral result peaks at zero lag where physiology peaks later, so we do not read it as autonomic coupling. The eye-restricted analyses fail their own positive control, so whether eye tracking is necessary remains open. The mission sequence is identical across dyads, so a schedule confound survives leave-dyad-out, and we report the necessity results both raw and with each mission's own average centered.

Every pretrained arm here is a transfer result from encoders trained on other corpora, so the clearest next step is to train on these signals instead of around them. Surface EMG is the sharpest case: the strongest single modality on mean heart rate, with no pretrained encoder to fine-tune. Fusion is the second: these encoders were built for cross-modal attention over simultaneous streams, which this corpus provides. The conditioning ablations point to the third, a shared encoder in which a person or context code modulates the features multiplicatively. No model in the benchmark reads the corpus's annotated event timeline, and integrating it is the natural next step for tasks defined over events. The single clock makes inter-brain modeling and joint-state decoding addressable, and both come next.

7 Conclusion

STRUM records eleven stream families from both members of a pair on one clock, so any modality can be withheld with the rest held fixed, and four findings follow. The first is that the full suite outperforms every modality inside it on both targets: pretraining lifts EEG at every step we measured, yet no pretrained encoder clearly beats a small network trained on the neck EMG band. An apparent neural result can itself be decomposed, since masking a participant's own cardiac input leaves muscle and ocular channels carrying half of what survives, so it can be largely peripheral. Recording both partners also exposes a pair-level coupling no representation we tried recovers. Finally, a person code gains accuracy only where it gates the features, and nothing when appended to the input. Taken together, these make sensor and model separate levers, and the sensor showed the greater impact on accuracy. What a model trained directly on these signals would do therefore remains open, since every encoder we ran came from elsewhere. The logical next step is to train on STRUM itself, beginning with surface EMG, then to fuse streams that arrive together, read the annotated event timeline, and decode across both members.

Ethics Statement

STRUM comprises human neural and biometric recordings collected under a protocol approved by the authors' institutional review board (protocol 130372), with written informed consent from every

participant.[1] The protocol assigned each participant a study identifier at consent and required that every recording carry that identifier rather than a name; it permits release once no identifying link remains between a participant and the released data. The release goes a step further: participant and session labels are renumbered sequentially, so they carry neither a name nor the laboratory identifier under which the data were collected, and the mapping to the original identifiers will not be released. Demographics are limited to age, sex, and handedness; date of birth and every other survey field were dropped at the source.

Consent for audio and video was collected in two tiers, and some participants allowed those streams to be used only by the research team. We therefore hold audio, video, and microphone streams in a restricted directory outside the participant data; they stay restricted when the corpus is published, and no analysis in this paper uses them.

This corpus supports decoding autonomic arousal from neural and behavioral signals, a capability that could be used to infer someone's internal state without their consent. Restricting the identifiable streams and releasing relative rather than absolute physiological quantities are necessary but not sufficient mitigations; the electrodermal and respiratory channels are uncalibrated, which further limits what may be inferred about any individual.

REPRODUCIBILITY STATEMENT

Every statistic in this paper is computed from persisted per-fold prediction vectors rather than from a rerun of the models: the paired tests, the bootstrap intervals and the multiplicity corrections all read those arrays from disk, and every figure is produced by a script that reads them. The fold assignments, the fixed partition, the inclusion lists and a loader accompany the corpus on release. The corpus will be released at `https://nemar.org/dataset/nm000284`, a public neuroimaging archive that mints a persistent identifier and serves the tree over HTTP. The conversion pipeline that produced it is versioned and re-runnable from the read-only sources. The inner-validation rule is deterministic, so the partition reconstructs exactly; moving its offset changed no headline contrast, at a minimum detectable effect of 0.0134. The analysis container runs Python 3.11.10 with numpy 2.1.2 and scipy 1.17.1 (Virtanen et al., 2020), which computes the exact signed-rank test at our fold counts; one helper in our own code takes the normal approximation above $n = 25$, so every panel states which variant produced it. We run three seeds per arm and report the spread beside the mean. The modeling and analysis used 540.8 GPU-hours and $6{,}146$ CPU-hours, every GPU job on a V100. A datasheet and Croissant metadata accompany the corpus on release. The code, the loader and the split files are available to reviewers at `https://strum-neuro.netlify.app`. That site provides three interactive views of the corpus: a per-session and per-participant inventory covering modality coverage, EEG montage and channel counts; the canonical five-block task schedule with mission onsets and durations measured across every complete instance; and the mapping of all 55 event types onto validated HED-3G tags. Every figure shown there is measured from the recordings. It also hosts a metadata-only BIDS sample, one complete run (`sub-013/ses-08/task-blockB`), covering EEG, EMG, motion capture from two tracking systems and six behavioral physiological streams, together with `participants.tsv`, the event schema and its HED sidecar. The sample carries no signal data and is a structure preview rather than runnable access; institutional metadata fields are removed from its sidecars for review and restored on release. A starter package of worked examples accompanies it, covering dyad-aware splitting, segment labeling and fixed-length windowing. The tree is BIDS-valid with zero errors under `@bids/validator` 3.0.1, and the corpus, the conversion pipeline and the experiment software are released publicly at publication.

A roster of the controls accompanies the released code, including the controls we specified and did not run; each entry names what that control isolates and what it cannot detect. Two rules came out of building it: ask what a control destroys before treating survival as validation, since a shuffle that breaks alignment within a recording block also breaks what a clock-based predictor uses; and treat an effect that goes null under a control as ruled out only when the new interval excludes the original estimate. A register of the software defects found and corrected during the analysis accompanies the released code; each entry names the defect, how it surfaced and the guard that prevents it. Result

[1] The approving institution is withheld for double-blind review and will be named at camera-ready.

paths encode the embedding, the target and the condition, and no run overwrites an existing results file.

USE OF LARGE LANGUAGE MODELS

We used large language models in four roles: retrieval and discovery, including finding and checking related work; research ideation and execution, including code for analysis and figures; drafting sections of the manuscript; and polishing the writing throughout. They can fabricate a reference or an unsupported number, and we guard against both. We verified every reference here against a fetched source before citing it, and took none from a model's recollection. No number reaches the paper without a file on disk it can be traced to, results are read from the persisted per-fold arrays, and a conformance checker enforces that over the analysis code. A preflight refuses to run the pipeline when it finds a significance verdict derived from a confidence-interval bound, which is invalid for paired folds. The authors take full responsibility for the contents of this paper, including every passage that began as model output.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca D'Alpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma Ismailova, Nicholas Evans, François Bremond, Massimiliano Todisco, Maria A. Zuluaga, and Laura M. Ferrari. StressID: a multimodal dataset for stress identification. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29798–29811. Curran Associates, Inc., 2023. doi: 10.52202/075280-1297.

Yanlong Chen, Mattia Orlandi, Pierangelo Maria Rapa, Simone Benatti, Luca Benini, and Yawei Li. PhysioWave: A multi-scale wavelet-transformer for physiological signal representation. In *Advances in Neural Information Processing Systems*, 2025.

Tahiya Chowdhury. Predicting cognitive load from speech and interaction dynamics in dyadic conversations. In *Proceedings of Interspeech 2026*. ISCA, 2026. arXiv:2606.12971.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean-Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3(1):160044, 2016. doi: 10.1038/sdata.2016.44.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 16115–16152, 2024.

Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2016. doi: 10.1109/TBME.2015.2474131.

Sein Jeung, Helena Cockx, Stefan Appelhoff, Timotheus Berg, Klaus Gramann, Sören Grothkopp, Elke Warmerdam, Clint Hansen, Robert Oostenveld, Julius Welzel, and BIDS Maintainers. Motion-BIDS: An extension to the brain imaging data structure to organize motion data for reproducible research. *Scientific Data*, 11(1):716, 2024. doi: 10.1038/s41597-024-03559-8.

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, 2024.

Bruna Junqueira, Bruno Aristimunha, Sylvain Chevallier, and Raphael Y. de Camargo. A systematic evaluation of Euclidean alignment with deep learning for EEG decoding. *Journal of Neural Engineering*, 21(3):036038, 2024. doi: 10.1088/1741-2552/ad4f18.

Ard Kastrati, Martyna Beata Płomecka, Damián Pascual, Lukas Wolf, Victor Gillioz, Roger Wattenhofer, and Nicolas Langer. EEGEyeNet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Ian R. Kleckner, Rebecca M. Jones, Oliver Wilder-Smith, Jolie B. Wormwood, Murat Akcakaya, Karen S. Quigley, Catherine Lord, and Matthew S. Goodwin. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. *IEEE Transactions on Biomedical Engineering*, 65(7):1460–1467, 2018. doi: 10.1109/TBME.2017.2758643.

Kelly Kleifges, Nima Bigdely-Shamlo, Scott E. Kerick, and Kay A. Robbins. BLINKER: Automated extraction of ocular indices from EEG enabling large-scale analysis. *Frontiers in Neuroscience*, 11:12, 2017. doi: 10.3389/fnins.2017.00012.

Christian Kothe, Seyed Yahya Shirazi, Tristan Stenner, David Medine, Chadwick Boulay, Matthew I. Grivich, Fiorenzo Artoni, Tim Mullen, Arnaud Delorme, and Scott Makeig. The lab streaming layer for synchronized multimodal recording. *Imaging Neuroscience*, 3, 2025. doi: 10.1162/IMAG.a.136.

Jun Li, Aaron D. Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), 2025. doi: 10.1056/AIoa2401033.

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

Nie Lin, Yansen Wang, Dongqi Han, Weibang Jiang, Jingyuan Li, Ryosuke Furuta, Yoichi Sato, and Dongsheng Li. EgoBrain: Synergizing minds and eyes for human action understanding. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2506.01353.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2021. doi: 10.3758/s13428-020-01516-y.

Manuel Meier, Berken Utku Demirel, and Christian Holz. WildPPG: A real-world PPG dataset of long continuous recordings. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2246–2266. Curran Associates, Inc., 2024. doi: 10.52202/079017-0073.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 59382–59413, 2025.

Cyril R. Pernet, Stefan Appelhoff, Krzysztof J. Gorgolewski, Guillaume Flandin, Christophe Phillips, Arnaud Delorme, and Robert Oostenveld. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, 6(1):103, 2019. doi: 10.1038/s41597-019-0104-8.

Adarsh Pyarelal, Eric Duong, Caleb Jones Shibu, Paulo Soares, Savannah Boyd, Payal Khosla, Valeria A. Pfeifer, Diheng Zhang, Eric Andrews, Rick Champlin, Vincent Raymond, Meghavarshini Krishnaswamy, Clayton Morrison, Emily Butler, and Kobus Barnard. The ToMCAT dataset. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40872–40923. Curran Associates, Inc., 2023. doi: 10.52202/075280-1781.

Soheil Rayatdoost, Yufeng Yin, David Rudrauf, and Mohammad Soleymani. Subject-invariant EEG representation learning for emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3955–3959. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9414496.

Kay Robbins, Dung Truong, Stefan Appelhoff, Arnaud Delorme, and Scott Makeig. Capturing the nature of events and event context using hierarchical event descriptors (HED). *NeuroImage*, 245: 118766, 2021. doi: 10.1016/j.neuroimage.2021.118766.

Harisu Abdullahi Shehu, Matt Oxner, Will N. Browne, and Hedwig Eisenbarth. Prediction of moment-by-moment heart rate and skin conductance changes in the context of varying emotional arousal. *Psychophysiology*, 60(9):e14303, 2023. doi: 10.1111/psyp.14303.

Zitao Shuai, Zongzhe Xu, David Yang, Wei Wang, and Yuzhe Yang. OSF: On pre-training and scaling of sleep foundation models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

SNAP. SNAP: Simulation and neuroscience application platform, 2014. Task-presentation software, `lse` branch.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Zou. SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48019–48037. PMLR, 2024.

Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore IV, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, Emmanuel Mignot, and James Zou. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32(2):752–762, 2026. doi: 10.1038/s41591-025-04133-4.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Christopher Wang, Vighnesh Subramaniam, Adam Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations*, 2023. arXiv:2302.14367.

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.

Hiroyoshi Yamasaki, Philippe Blache, and Daniele Schön. A dual EEG hyperscanning dataset of natural french face-to-face conversation, 2026. bioRxiv 2026.05.13.724780.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 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 *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=5Xwm8e6vbh`.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *International Conference on Learning Representations*, 2024. arXiv:2403.10943.

A SUPPORTING DETAIL AND CONTROLS

A.1 ACQUISITION AND CROSS-MODAL ALIGNMENT

Every device writes into one Lab Streaming Layer session, and we divide that recording per participant and per modality only afterward (Figure 4). We never align one modality to another after the fact, so withholding a modality in Table 4 leaves the rest covering the same instants.

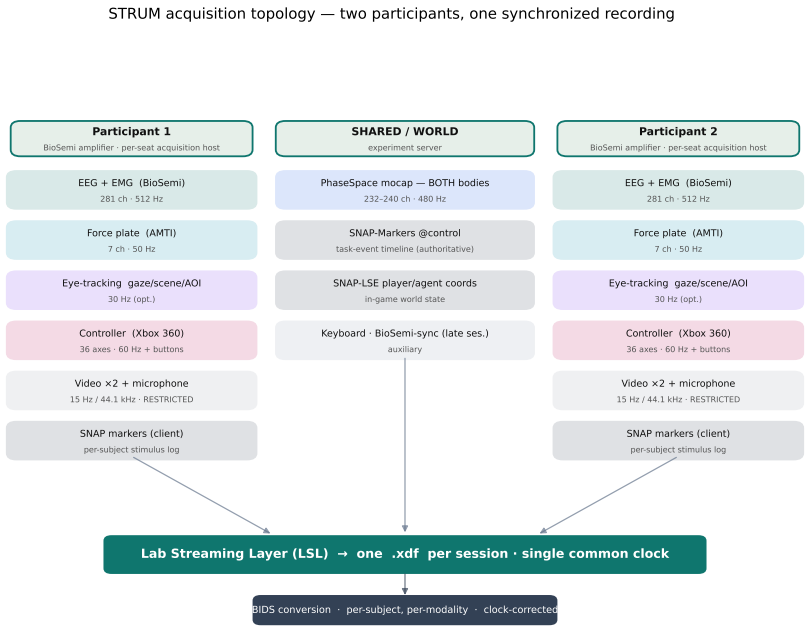


Figure 4: Acquisition topology. Two instrumented participants feed eleven stream families into a single Lab Streaming Layer session. Each session yields one recording on a common clock, which is then converted per participant and per modality.

Figure 5 carries the evidence behind the blink agreement rate reported in Section 3, from two detectors that share no signal. The check covers one participant, so we read it as an existence proof rather than a characterization of the corpus.

A.2 TARGET CONSTRUCTION

We find R-peaks in the bipolar ECG derivation with `scipy.signal.find_peaks` (Virtanen et al., 2020) under refractory and amplitude criteria, drop implausible intervals, and take the rate as 60 divided by the window's mean interval. We apply no response-rate quality control to the electrodermal target: the absolute-microsiemens thresholds used for calibrated recordings (Kleckner et al., 2018) do not apply to an uncalibrated channel.

A.3 CONTROLS

Where the leakage sits. Our windows overlap, so the inflation that Section 4 reports for a random split within a participant could come from temporal proximity alone. Windows merely adjacent to a training window span -0.0056 to $+0.0240$ and reach significance in no pairing after correction, which locates the inflation in the duplicated windows.

How coarsely we group the held-out unit. A subject-wise split groups by person but not by pair, so a held-out subject's partner can appear in training; it raises performance by $+0.0221$ in an unpaired comparison, which bounds what the looser grouping would gain.

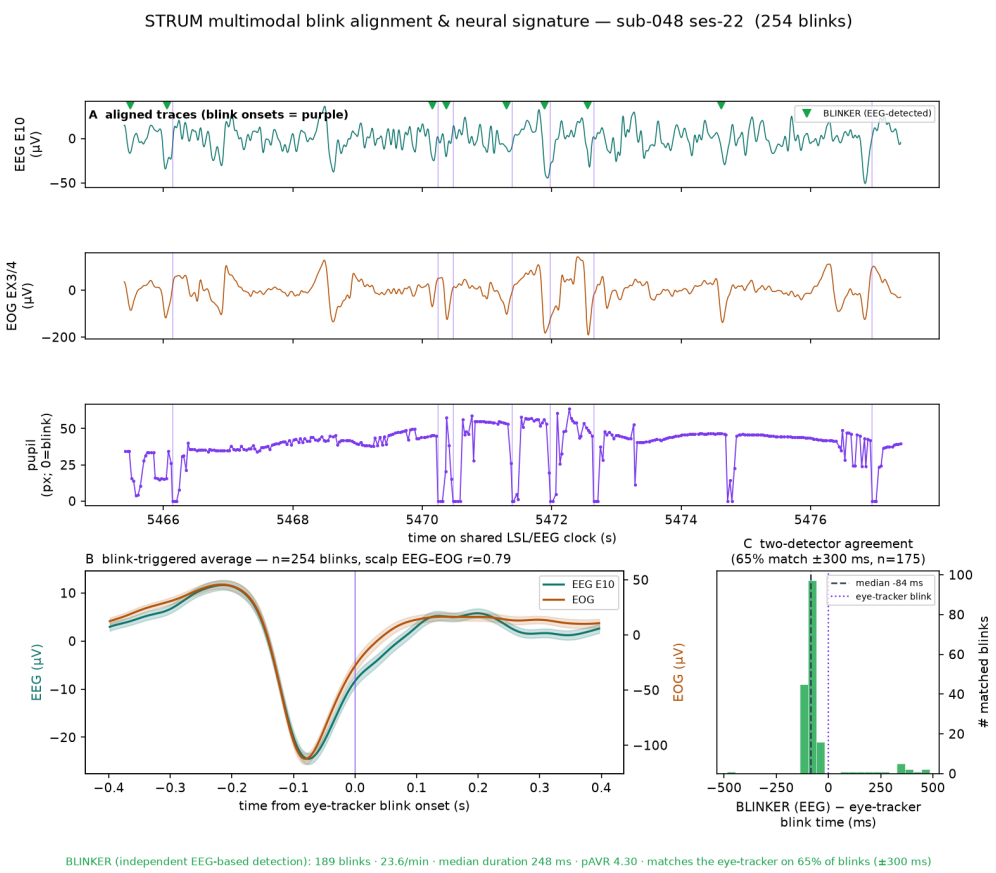


Figure 5: Cross-modal coherence in one participant. (A) Frontal EEG, the EOG externals and eye-tracker pupil size on the shared clock, with blinks detected independently from EEG marked above. (B) Blink-triggered averages of scalp EEG and EOG correlate at $r = 0.79$. (C) BLINKER (Kleifges et al., 2017) and the eye tracker agree on $65\,\%$ of that participant's 254 blinks within ± 300 ms, at a median offset of -84 ms — the lag between eyelid movement and pupil occlusion.

A free choice inside the protocol. Moving the inner-validation offset changes the result by less than 0.0134, the minimum detectable effect at $80\,\%$ power.

Whether the neck band stands in for respiration. The respiration belt reaches one member of each pair, so it is absent from Table 4 and we compared the two directly instead. A respiration arm adds nothing over the neck EMG features on mean heart rate (-0.0008, minimum detectable effect 0.0032), while those features add $+0.4805$ over respiration.

What the target's own past provides. Persistence and autoregressive predictors read the target's own recent history, so we report them as context for a forecasting question the benchmark does not ask; adding the learned representation to a first-order autoregressive predictor gains $+0.1903$ $[+0.1384, +0.2405]$ on electrodermal arousal and $+0.2699$ on heart rate, and against a fourth-order causal predictor the full suite's heart-rate margin is $+0.0885$.

A.4 Personalization and the person code

Figure 6 is the measurement behind scoping personalization to cardiac decoding, with the window-duplication confound of Appendix A.3 separated out.

Figure 7 separates the code from the parameters that carry it: its arms share an input and a target, differ in where the code reaches the features, and sit beside a capacity arm with no code.

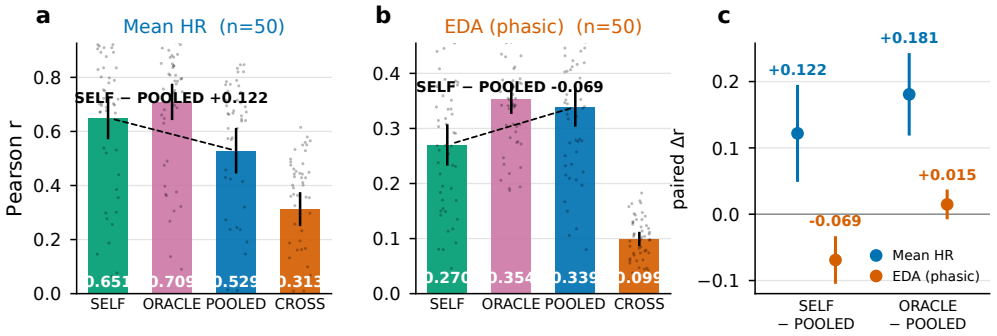


Figure 6: The cross-person gap reverses between targets. A participant's own model beats the group model by $+0.1219$ on mean heart rate but loses to it by -0.0689 on phasic electrodermal arousal, so we scope personalization to cardiac decoding. Heart rate is computed from an ECG-free embedding, and separating the window-duplication confound moves the gap only from $+0.1354$ to $+0.1219$. SELF is a participant's own model on their own held-out windows, POOLED one model trained on all other participants, CROSS the mean over other participants' models, and ORACLE the best single donor, picked with test labels and therefore a ceiling rather than a result. Bars are means over held-out subjects with ± 1.96 SEM.

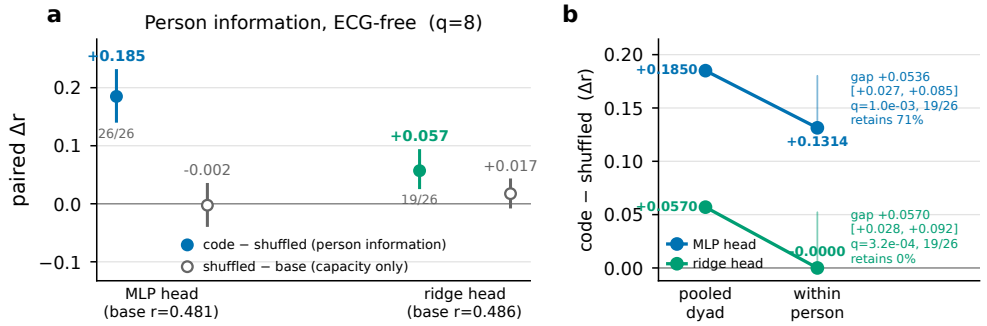


Figure 7: A person code modulates the feature-to-target mapping, and a linear head cannot use it. Mean heart rate from an ECG-free embedding. Concatenating the code to the input gains nothing; gating the features with it gains several times what extra capacity does. Bars are means over 26 folds with 95 % percentile bootstrap intervals, and the fractions on each arm are the folds with a positive effect.

We reran that contrast under both schedule removals of Section 4. The gain from conditioning a model on a per-participant code survives both removals on heart rate ($p = 3.8\text{e}{-05}$ within mission, $p = 1.3\text{e}{-06}$ against elapsed time) and survives the time removal on electrodermal arousal while falling short within mission ($p = 0.105$, uncorrected); mission centering also takes out genuine between-mission arousal, $19.1\,\%$ of the target's variance, whose square root bounds any correlation with it at 0.437, so a reduced effect is a lower bound.