A FAITHFUL ATTRIBUTION REPORTS ITS HEAD, NOT ITS MODEL: THE LIMITS OF COMPLETENESS IN ADDITIVELY DECOMPOSED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Additive models promise an explanation readable off the architecture: the latent is a sum of per-component terms, so each contribution is available without a post-hoc estimator. That promise is bounded by the downstream head. A completeness-satisfying attribution decomposes *a head's* output, so what it can say is fixed by the head you chose, not by the model.

On **MD-TJEPA**, a JEPA whose latent decomposes over channel $\times$ band $\times$ scale, channel identity is 0.998 decodable from the components and only 0.274 (chance 0.25) from the sum they form. Fit the head for the channel and its exact decomposition $\phi_u = \alpha_u \langle w, z_u \rangle$ closes 11% of the gap to a pre-aggregation score (0.319 ± 0.057, n=10); on a plain NAM the same construction is not weak but *arbitrary*, 0 or 1 in most seeds. Fit it for the model's own task — a band detector whose labels contain no channel — and an equally exact decomposition names the channel at 0.743 ± 0.080. An input intervention separates the two: resampling the injected channel moves both heads' outputs, moving its content to another channel barely does, so the sum hides *which* channel contributed, not *whether* it did. The band head's map agrees with that intervention on 76% of windows, the channel head's on 41%; most of each map's disagreement is its implicit reference, the latent origin. Re-referenced to the training-mean contribution — on this synthetic benchmark an estimate of the expected intervention — they agree on 95% and 73%. Completeness pins the sum, not the head and not the reference.

We are scoped on faithfulness, not accuracy — TS2Vec outperforms MD-TJEPA under a matched probe, though an *untrained* MD-TJEPA read through its components outperforms TS2Vec — and the axes conflated as "interpretability" disagree on the same checkpoints.

What this buys is a check worth running first: probe the aggregate for the axis, and intervene on the input for it, before trusting any attribution over it — a probe near chance says the aggregate hides which channel was active, not that its content goes unused. $\mathrm{overlap@}k$ between $\|z_u\|$ and $\alpha_u \|z_u\|$ costs one line and exposed a **gate collapse** here — the JEPA objective shares the gate between prediction and target, so an unopposed sparsity penalty switches components off. In a minimal additive stand-in, removing the gate from the target prevents that shrinkage outright. We close with three checks to run before reading any additive attribution.

1 INTRODUCTION

An additive model earns its interpretability structurally rather than by estimation. If a representation is $s = \sum_u \alpha_u z_u$ over named components u, then "what did component u contribute" should be answerable from the architecture itself, with no post-hoc estimator standing between the explanation and the computation. Neural Additive Models (Agarwal et al., 2021) make this the design principle, and it is the reason to want an additive latent inside a self-supervised encoder at all: the representation would come with its own account of itself.

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.

This paper reports a limit on that reasoning. The limit is a property of the *aggregation*, not of any particular network, and it is easy to state. Downstream use reads s, not the components, so a head is some $\langle w, s\rangle$ and the attribution that satisfies completeness with respect to it is

$$\phi_u(x) \;=\; \alpha_u(x)\,\langle w,\, z_u(x)\rangle, \qquad \textstyle\sum_u \phi_u(x) \;=\; \langle w,\, s(x)\rangle. \tag{1}$$

ϕ_u does see the whole component z_u; what constrains it is w, and fitting w on s both ties the explanation to the model and decides what it can say. Fitted for an axis the aggregate barely retains, w has little bearing on it and the exact decomposition ranks almost nothing. Fitted for the model's own task, w aligns with whichever component is active and the equally exact decomposition names that component's channel, well above what a probe on the aggregate recovers. Both are faithful, and an input intervention shows the injected channel moves both heads' outputs — the aggregate hides which channel contributed, not whether it did — yet the two maps agree with that intervention at very different rates (0.76 and 0.41). Most of each map's disagreement is a second free choice completeness leaves open, the reference: ϕ_u measures against the latent origin, and against the training-mean contribution instead — on our synthetic benchmark an estimate of the expected intervention — the maps agree at 0.95 and 0.73 (§G).

A faithful attribution reports its head, not its model.

We make this concrete on **MD-TJEPA**, an additively decomposed time-series JEPA (§3) whose components are (channel, band, scale) triples and whose contested axis is the channel. Both are measured (§4). The derived attribution of Eq. 1 decomposes the output exactly and lands near chance. The per-component norm $\|z_u\|$ — the score such architectures are usually read with — reaches 0.897 ± 0.081 against a chance rate of 0.25, but it scores a *pre-aggregation* quantity, and on our deployed checkpoints its top-k set is *disjoint* from the components the model actually routes through. Widening the aggregate so the channel survives into it lifts the probe from chance to 0.986 ± 0.020, and the faithful attribution follows on one backbone (0.319 to 0.904) and not the other (0.509 ± 0.180 at probe 0.995), because what ϕ can read is bounded per component and not by the head.

The claim is not about this model. The premise and both cases **transplant** onto tabular features with a plain NAM, no time series and no JEPA, where ϕ turns out not to be weak but *arbitrary* — 0 or 1 in most seeds (§4); the arbitrariness is in the score's *sign*, and a magnitude readout escapes it only by ceasing to be a decomposition. Two further findings stand alone. The disjointness above is a **gate collapse** under a penalty nothing in the objective resists, because the JEPA loss shares the gate between prediction and target (§5); it is diagnosable in one line, and a retraining recipe that adds a balance term keeps every component alive over five seeds without lowering downstream macro-F1. And the axes called "interpretability" — ground-truth recovery, parameter randomization (Adebayo et al., 2018), utility, clinical alignment — **disagree** on the same checkpoints (§6), including on two explanations of ours that did not survive testing. Section 8 collects what we would now check first. MD-TJEPA is the vehicle, not the claim, and a weak one: TS2Vec outperforms it under a matched probe, and so does an *untrained* copy of itself (§6). That is the point. A model whose pretraining made it worse is where an attribution should say so, and an exact per-component decomposition of it says nothing of the kind.

2 Related work

Limits on attribution. That feature attribution can fail to be informative is not a new claim, and the closest statement to ours is Bilodeau et al. (2024): for moderately rich model classes, any attribution that is *complete* and *linear* — Integrated Gradients (Sundararajan et al., 2017) and SHAP (Lundberg & Lee, 2017) among them — can provably do no better than random guessing at inferring the model's behaviour under counterfactuals such as recourse or spurious-feature removal. Their result is about what a complete, linear attribution licenses one to infer, not about estimation error, and ours is in the same family. The difference is the object and the source of the failure. Their attributions are defined relative to a baseline input and their impossibility comes from the model class; ours is the exact, intrinsic decomposition of a head on an additive latent, and we locate what completeness leaves free in two concrete places — the head that was fitted (Corollary 1) and the reference the decomposition implicitly takes, the latent origin, which need not be an absent input (Appendix C). We then measure both against an input intervention on the same head rather than

against a model-class argument. The reference freedom is the intrinsic counterpart of the baseline dependence Sundararajan & Najmi (2020) show for Shapley-based methods, and our intervention is a removal-based explanation in Covert et al. (2021)'s sense, whose choice of removal is exactly that reference. Clark et al. (2025) show that additive models misattribute importance to suppressor features; their target is statistical relevance to the label, ours the dependence of a head's output, and the two can disagree without either being wrong. This matters because architectural additivity is promoted as the escape from post-hoc estimation (Agarwal et al., 2021; Rudin, 2019): it removes the estimator, and the reference and the head remain.

The nearest precedent in spirit is the attention-is-not-explanation exchange (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019) and the probing literature (Alain & Bengio, 2018; Belinkov, 2022) — a quantity read off a model can track the readout rather than the model — with Hewitt & Liang (2019)'s control tasks the standard discipline for saying when a probe licenses a claim, and Elazar et al. (2021) showing by intervention that a decodable property need not be one the model uses. Those scores carry no completeness guarantee. Ours does, and it does not help: an *exact* decomposition of a head still reports that head. Corollary 1 is the intrinsic-decomposition counterpart of Fisher et al. (2019)'s model class reliance, where equally accurate models in one class disagree on a variable's importance — here one model and two heads suffice. Krishna et al. (2024) document explanation methods disagreeing on one model with no way to adjudicate; our audit finds that disagreement *across evaluation axes* for one family of checkpoints. Adebayo et al. (2018) introduced the cascading-randomization check we use; Ismail et al. (2020) showed time-series saliency improves under temporal rescaling.

Additive and interpretable-by-design models. Neural Additive Models (Agarwal et al., 2021) learn a separate shape function per feature and sum them, making the per-feature contribution structural, and Neural Basis Models (Radenovic et al., 2022) share a basis across features with per-feature weights — the per-component linear head our readout comparison points to; self-explaining networks (Alvarez Melis & Jaakkola, 2018) give the input-dependent form $h(x) = \sum_i \theta_i(x) h_i(x)$ that our $s(x) = \sum_u \alpha_u(x) z_u(x)$ instantiates; and the functional-ANOVA line (Hooker, 2007; Lengerich et al., 2020) addresses the identifiability of such decompositions, which Theorem 1 (Appendix B) adapts to the (channel, band, scale) setting. That a sum of components sharing one shape function is invariant to permuting them — and so cannot carry their identity — is the Deep Sets symmetry (Zaheer et al., 2017); our tabular transplant uses exactly that construction, which is why it proves blindness where MD-TJEPA only measures it. What separates our claim from a restatement is where the sum lands: in a NAM or a SENN the components sum to the *output*, so each is the contribution by construction. We study the arrangement that inherits the name but not that property — components summed into a *latent* a downstream head reads, a projection of $U \gg k$ terms into k dimensions.

JEPAs and biosignal SSL. I-JEPA (Assran et al., 2023) established the latent-prediction recipe; LeJEPA (Balestriero & LeCun, 2025) replaced its collapse-avoidance heuristics with an explicit isotropic-Gaussian objective, SIGReg; we use a moment-matching simplification of it (mean and sketched covariance pulled toward isotropy, not its Epps–Pulley test) and keep an EMA target encoder that LeJEPA does not; Ruiz-Morales et al. (2026) explain JEPAs' *clustering* of time series by a predictive objective that drives the encoder onto the Koopman invariant subspace, and find on synthetic data that a near-identity predictor is what forces it — ours (§3) adds block-diagonality to it, alongside Lusch et al. (2018) and Liu et al. (2023). On biosignals, BENDR (Kostas et al., 2021) and BIOT (Yang et al., 2023) established transformer SSL on EEG, and the recipe has been carried to brain dynamics (Dong et al., 2024), EEG (Hojjati et al., 2025; Panchavati et al., 2026) and ECG (Kim, 2027). Closest in aim is Ma et al. (2026), which targets an interpretable EEG encoder but locates the decoupling in a dual tokenizer and reads interpretability off codebook statistics, where we decompose in the latent space; Appendix S tabulates the design differences. Our learnable filter banks follow Michau et al. (2022) and, for EEG, Paillard et al. (2025).

3 METHOD — MD-TJEPA

3.1 THREE-AXIS DECOMPOSITION

Pipeline: $x \in \mathbb{R}^{C \times T} \to$ Gabor bank $\to$ multiscale pooling $\to$ per-component encoder $\to$ sparse gate $\to$ additive aggregation $\to$ Koopman predictor $\to \hat{s}_y$.

The bank is GREEN-style, two parameters (f_c, σ_t) per filter (Paillard et al., 2025), softplus-reparameterized with $\sigma_{\min} \geq 5$ ms and warm-started against the input's Welch PSD (ours). The predictor is block-diagonal near-identity, $g_k = I + W_k$ over $K{=}8$ blocks, whose Frobenius regularization keeps *every* eigenvalue within $|\lambda - 1| \leq 0.3$ at the §A.2 budget (1.000 ± 0.000; the 95% figure we used to quote here is a design target, not a measurement). The objective is $L = L_{\text{jepa}} + \lambda_{\text{koop}} \sum_k \|W_k\|_F^2 + \lambda_{\text{sig}} L_{\text{iso}} + \lambda_{\text{sparse}} \|\alpha\|_1 + \lambda_{\text{anova}} L_{\text{ANOVA}} + (\text{opt.})\, \lambda_{\text{mod}} L_{\text{mod}}$, where L_{iso} matches the batch mean and a sketched covariance to 0 and I — a moment-based simplification of LeJEPA's SIGReg, not its Epps–Pulley statistic.

3.2 Theorem 1 (statement)

Under (A1) NAM additive architecture, (A4") per-component centering (sum-to-zero on each h_u) and (A5) joint non-degeneracy, the learned components are unique up to scaling of the (α_u, h_u) factorization; the components $g_u = \alpha_u h_u$ themselves are unique. The proof (Appendix B) follows the generalized fANOVA uniqueness of Chastaing et al. (2012).

Scope (corrected). The argument admits *correlated* inputs, but every tube $x_{c,f,s} = \text{pool}_s(\text{Gabor}_f * x_c)$ is a deterministic function of one channel signal, which is stronger than correlated, and there the theorem does not apply (Appendix C). **We claim identifiability for the channel axis only** — distinct signals, where (A5) is plausible but assumed, not shown — and no permutation clause at all, since components are pinned to physical tubes. Proposition 1 needs the same condition, and the asymmetry it predicts is the one we measure (channel 0.86 ± 0.17, $n{=}5$, against band 0.43 ± 0.40, $n{=}3$, §A.5.1; overlapping, so consistency and not confirmation).

4 What the attribution reports

Every number below is a 95% t-interval with n stated; Appendix A maps each to the section that measures it. The limit first, so that what follows measures it rather than illustrating it.

Proposition 1 (Attribution inherits the head)**.** *Let $s(x) = \sum_u \alpha_u(x) z_u(x) \in \mathbb{R}^k$ and let a downstream head be $h(x) = \langle w, s(x) \rangle$. Call an attribution* complete *if $\sum_u \phi_u = h$, and* local-linear *if $\phi_u(x) = \langle v_u, \alpha_u(x) z_u(x) \rangle$ for a fixed v_u — component u's score reads that component alone, linearly in its contribution. Under the non-degeneracy (P3) of Appendix C, the complete local-linear attribution is unique and equals $\phi_u = \alpha_u \langle w, z_u \rangle$ (Appendix C). Consequently, for any target axis A: if A is not linearly decodable from s, the w obtained by fitting a linear head on s for A is itself at chance on A, and ϕ is then fixed by a direction chosen without reference to A. What $\{\phi_u\}$ carries about A is whatever that direction incidentally picks up from the individual z_u — even though each ϕ_u is computed from the full component. Note what this does* not *say: not that the ranking is mediocre, but that it is* unconstrained*. The transplant below shows the sharper form it takes in practice.*

Three remarks, because the proposition is easy to over-read. Locality is not decoration — completeness and linearity alone admit a continuum, since mass moves freely between components (Appendix C) — and the completeness literature usually leaves it implicit. First, it does *not* say ϕ cannot see the components: it sees all of z_u, and the constraint enters only through w. Second, it is about the head one *fits for* A and says nothing about any other head on the same s — which is the whole content of what follows:

Corollary 1 (Completeness does not fix the ranking)**.** *Let w and w' be any two heads on the same s. Their complete local-linear attributions $\phi_u = \alpha_u \langle w, z_u \rangle$ and $\phi'_u = \alpha_u \langle w', z_u \rangle$ are each unique, and they may order the components oppositely: completeness constrains $\sum_u \phi_u$, and the sum is invariant to any rearrangement of mass across u. So no ranking is licensed by completeness alone — only by the choice of head. §F exhibits the case that matters: w' fitted for the model's own task is* useful *(0.928 at chance 0.25) and its ϕ' names the injected channel at 0.743 where a probe on s recovers it at 0.274 — a contribution whose identity s hides (§G).*

Third, the premise is what makes the limit bite, so the proposition also says when it does *not*: if A is linearly decodable from s, nothing forces $\{\phi_u\}$ to chance. The claim is about the interaction between what the aggregate keeps and what the attribution is asked to rank, not about additive models as such.

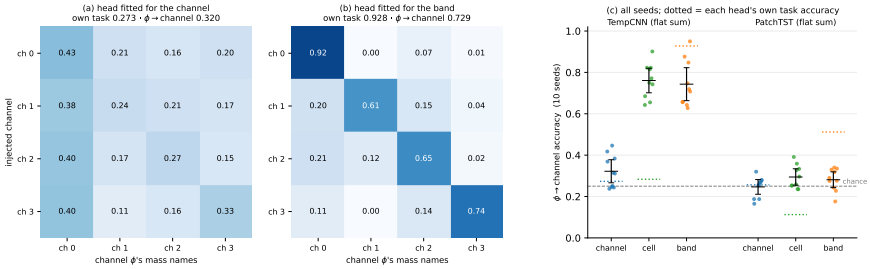

Figure 1: **Same frozen components, two heads, two exact decompositions** (residual $\leq 5 \times 10^{-6}$), on the same s. **(a)** Fitted for the channel: the head is near chance and its ϕ ranks almost nothing — its mass drifts to channel 0 whatever was injected. **(b)** Fitted for the band: labels contain no channel, and the decomposition names the injected one at 0.729. Rows: injected channel; columns: where ϕ's mass lands; seeds 0–2 pooled, TempCNN, flat sum. **(c)** All ten seeds, both backbones; dotted lines are each head's own-task accuracy. On PatchTST none of it appears — per-tube norms carry the channel at only 0.329.

That is testable, and we test it below by widening the aggregate until the channel survives. The prediction half-holds, which we did not anticipate, and the half that fails is the open question this paper leaves (§4).

The premise: the aggregate barely carries the axis the attribution ranks. A linear channel probe recovers the injected channel at $\mathbf{0.998 \pm 0.002}$ from the per-component norm vector — the object the attribution ranks — and *near chance* (0.274 ± 0.010 against 0.25, $n{=}10$) from the additive aggregate, and not only under linear decoding: an MLP on the same s closes 2.8% of the gap. A control task does not license more than that (§J). Any s of dimension k summing $U \gg k$ components is a projection, and one serving a predictive objective has no reason to keep an axis that objective does not need — the Sleep-EDFx (Kemp et al., 2000) label is the sleep stage, not the channel. The channel is nonetheless the axis the published attribution ranks.

Fitted for the axis: the score is uninformative. Two scores have to be kept apart here, and an earlier draft did not. The completeness-carrying object is the *signed* $\phi_u = \alpha_u \langle w_c, z_u \rangle$ for a *single* class c, which we take to be the predicted one; Proposition 1 is stated for one head and so applies per class, and fixing c to the true label instead moves every number below by at most 0.014 (§D). Its residual against $\langle w_c, s \rangle$ is $\leq 5 \times 10^{-6}$. The map our attribution experiments report is the *magnitude readout* $\sum_c |\alpha_u \langle w_c, z_u \rangle|$, which is *not* a decomposition: its residual is 17–62. A third object, the *norm heatmap* $\|z_u\|$, is what the heatmap figures themselves draw — it reads z_u before aggregation, so no completeness statement applies to it. We keep the three names apart.

The faithful score recovers almost none of the channel: over ten seeds it reaches 0.319 ± 0.057 (TempCNN) and 0.246 ± 0.036 (PatchTST, Nie et al., 2023) against chance 0.25, while the norm score on the same runs reaches 0.897 ± 0.081. The cause is the premise above: w is fitted on an s that barely carries the channel, so it carries no channel-discriminative direction for ϕ to inherit, and ϕ's map is essentially input-independent ($\rho \approx +0.98$ against a noise input). At $n{=}10$ the lower bound clears 0.25 by 0.012 — "above chance", and beside the point: closing a ninth of the gap is not made informative by a tighter interval.

Fitted elsewhere: an informative score reports something else. Two scores are informative here and neither records what the model used. The per-component norm $\|z_u\|$ reaches 0.897 ± 0.081 by reading z_u before aggregation, so no completeness statement holds for it; on the deployed checkpoints its top-k set is **disjoint in every seed** from actual contribution $\alpha_u \|z_u\|$ (0.000 ± 0.000, $n{=}5$), with 0.783 ± 0.031 of the mass on gated-off components. The second is worse, because it *is* complete. Fit the head for the model's own task rather than for the channel — a band detector, whose labels contain no channel — and its exact decomposition names the channel at 0.743 ± 0.080, every seed above 0.627, on the same s that carries it at 0.274 (Fig. 1, §F). So the dichotomy we first drew, faithful $\Rightarrow$ uninformative, is false. Nor is the band head's map reporting something the model does not use. Resampling each input channel and reading the head's logit (§G) shows the injected

channel moves the band head's output most in 0.933 ± 0.034 of windows, while moving the same content to another channel changes it only 5% as much: the sum hides the channel's *identity*, not its content's *contribution*. Against that intervention the band head's ϕ agrees on 0.764 of windows and the channel head's on 0.408; with nothing changed but the reference — the training-mean contribution instead of the latent origin ϕ implicitly takes — on 0.948 and 0.734. Completeness fixes neither the head nor the reference, and here both matter.

Two limits: one (c, f) tube is active per sample, so the mechanism is arithmetic; and ϕ localizes no better than a spectrogram (§6).

The same structure off this architecture. Proposition 1 needs neither our front end nor time series, so we rebuilt its premise on tabular features with a plain NAM: components are (group, member) pairs, one group is active per sample, the label *is* that group, and shape functions are tied across groups so the flat sum is provably group-blind. Premise and the informative case transplant exactly — the flat-aggregate probe sits at 0.249 against chance 0.25, the pre-aggregation norm score at 1.000.

The axis-fitted case transplants in a sharper form: ϕ reads 0 or 1 in seven of ten seeds (mean 0.605). With w carrying no group-discriminative direction, an uninformative projection sets the *sign* of the active group's fixed ϕ mass, so the argmax lands on it every time or never — not weak but arbitrary. Both cases lift when the axis is retained (1.000, 1.000), and ranking by $|\phi_u|$ reads 1.000 ± 0.000, escaping the uninformative case by ceasing to be a decomposition (§D).

Widening the aggregate, and where the attribution follows. The loss is created by summing the channel axis away, so we removed that step. Replacing $s = \sum_u \alpha_u z_u \in \mathbb{R}^k$ with a per-channel block sum $s = \left[\sum_{f,\tau} \alpha_{cf\tau} z_{cf\tau}\right]_{c=1}^{C} \in \mathbb{R}^{Ck}$ keeps the model exactly additive — every component still contributes linearly to exactly one slot, and Eq. 1 still holds with $w_{c(u)}$ in place of w — while declining to marginalize the channel. §D tabulates every cell with its residual.

The readout goes from chance to essentially solved — above 0.9 in all twenty runs — so the additive bottleneck's utility cost is caused by marginalizing the channel axis and not by encoder capacity. The faithful attribution follows it on TempCNN: $0.319 \rightarrow 0.904$. Read across that row and the proposition is visible in both directions — premise fails, attribution near chance; premise holds, attribution inherits the head's competence — which is the strongest form the claim takes here.

It does not follow on PatchTST. At probe 0.995 the faithful score reads 0.509 ± 0.180, spanning 0.240 to 0.922 across seeds: decodability is *necessary and not sufficient*, and that gap is the open question we leave.

Why it follows on one backbone and not the other. What ϕ can report is bounded per component, since $\phi_u = \alpha_u \langle w, z_u \rangle$ reads z_u alone. Ranking channels by the raw norms $\|z_u\|$ — no head, no fitting — gives 0.897 ± 0.081 on TempCNN and 0.329 ± 0.067 on PatchTST: on PatchTST there is nothing per-tube for any w to pick up, and widening the aggregate cannot supply it. **A probe that clears the axis is necessary and not sufficient**; the remainder belongs to the components, not the head. The band-head result is the other face of this: same components, different head, different and equally faithful map.

On real data the widening buys little. The arms share seeds, encoders, probe and test set, so we report them *paired*: block over flat sum is $+0.024 \pm 0.014$ ($5/5$, t=4.6) — separated, but small beside a per-component linear head on the *same* components at $+0.143 \pm 0.035$ (0.508 ± 0.048), which is still additive ($\sum_u v_u^\top z_u$). What costs utility is the shared direction $v_u = \alpha_u w$, not additivity (§M). The uninformative case comes off only where the label *is* the contested axis, which the transplant settles is a property of this model rather than of block aggregation.

5 A GATE THAT COLLAPSES, AND THE CHECK THAT FINDS IT

The disjointness in §4 is a bug, and fixable.

Mechanism. $L_{\text{sparse}} = \sum_u \sigma(\ell_u)$ pushes every gate down at a rate independent of whether the component helps, and nothing pushes back: s_{pred} and s_{true} come from the *same* gate, so scaling α

Table 1: Retraining with the counter-force enabled. Five seeds each; the deployed column is the configuration the PTB-XL audit results were computed under. Two things changed alongside λ_{bal} that we did not intend: batch size $16 \to 32$ and all 137 train records against `-max-records 50`, i.e. $2.38\times$ the pretraining data. The macro-F1 gain is $+0.015 \pm 0.007$ paired, $5/5$ seeds — real, but not attributable to the gate fix alone.

	deployed (collapsed)	retrained
live components, of 60	7.0 ± 0.0	$\mathbf{60.0 \pm 0.0}$
attribution mass on gated-off	0.783 ± 0.031	$\mathbf{0.000 \pm 0.000}$
overlap@k, $\|z\|$ vs $\alpha\|z\|$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
channels routed to zero	1.0 ± 0.0	$\mathbf{0.0 \pm 0.0}$
parameter randomization, seeds passing	0/5	0/5
matched-protocol macro-F1	0.354 ± 0.007	0.370 ± 0.008

down scales the residual with it and the JEPA objective is content to switch components off. Only the variance regularizer preserves any scale, so an arbitrary minimal subset survives to carry it. The entropy counter-force that would prevent this defaults to zero and had never been applied to these runs.

The trace is identical across all five deployed seeds: a quarter of the components by epoch 6–7, the first channel dead by epoch 24–27, the smallest logit at -18.2 everywhere — $\sigma(-18.2) \approx 10^{-8}$, so the penalty pushes long after the component is off.

Fix, and how the strength was chosen. Stopping the collapse is easy; not flattening the gate into uniform usage is the requirement, since a gate that weights everything equally has stopped selecting. Over the window in which the baseline collapses, only $\lambda_{\text{bal}}{=}1$ keeps every component alive, *converges* rather than slows the slide, and stays selective: at 0 the gate is still moving when the window ends (Gini $0.195 \to 0.863$), at 15 it is stable at Gini 0.117, no longer distinguishing components. Three points on training dynamics only, no test number read — and three points is not a criterion; a finer grid is a pretraining run per arm, unrun, and the $\lambda_{\text{bal}}{=}0$ arm was unmatched in the direction that disfavours the fix.

What the retrain settles. Five seeds, 30 epochs (Table 1).

With the gate open the two rankings *coincide exactly* in every seed, so the disjointness is a property of a collapsed gate and not of intrinsic attribution. What persists is narrower and worth keeping: the per-channel *mass* allocation still disagrees — the norm heatmap splits $0.548/0.452$ where the model routes $0.830/0.170$. The order is right; the weighting is not. The retraining recipe does not lower downstream macro-F1 (its data and batch also changed, so this is not the balance term's effect alone), and we report the retrained models as primary while keeping the deployed column beside them, since the contrast is the only thing that identifies which negatives were the collapse.

The mechanism reproduces in a minimal stand-in, and what rescues it. A mechanism claim is worth making only if it predicts where the failure does *not* occur, so we tested it away from MD-TJEPA on a 24-component NAM-style additive model with three arms differing only in where the gate is applied: latent prediction with the gate on context and target, the same with the gate *removed* from the target (the target is the ungated sum; this is not a stop-gradient on a kept gate, which both arms already have), and a supervised head on the aggregate. Three of the label's components are relevant by construction, so "did the surviving gates identify what matters" has a ground truth and an analytic chance rate.

Our first prediction was wrong and we report it as such: we expected the supervised arm to hold its gates, and it prunes as hard as the shared-gate arm ($\bar{\alpha}$ 0.023 against 0.015). The supervised arm keeps exactly the label-carrying components (hit rate 1.00); the shared-gate arm shrinks all 24 together ($23/24$ still above 0.01, Gini 0.11). That contrast is not evidence the latent-prediction gate chose badly: every component here is an i.i.d. draw and the label never reaches the latent objective, so no label-free objective could single out the three — the arm without the target gate is at chance on them too (0.13). What removing the gate from the target, the single change between the two latent-prediction arms, does is prevent the shrinkage outright ($\bar{\alpha}$ 0.982). This is a stand-in result; on

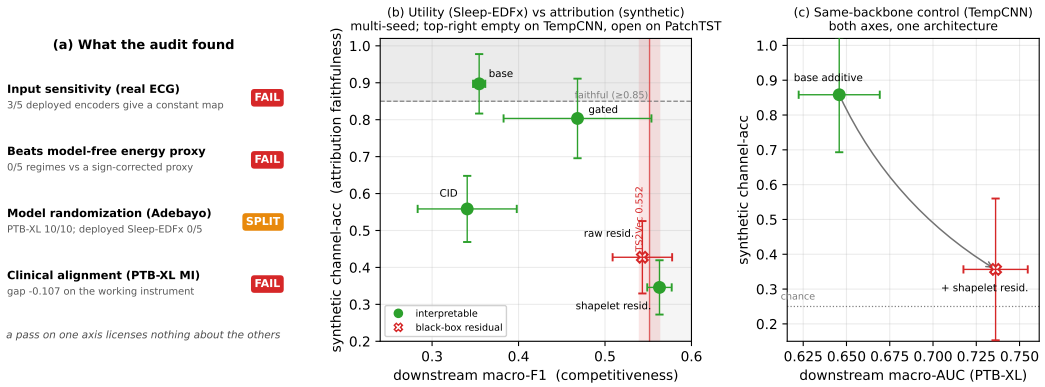

Figure 2: **(a)** Three of four checks come out against the method: the attribution is *constant in the input* for the deployed convolutional and transformer encoders on real ECG; a model-free band-energy proxy beats it in 3/5 regimes and 5/5 once allowed its sign; model randomization splits by checkpoint (one PTB-XL checkpoint 10/10 draws, deployed Sleep-EDFx 0/5); clinical lead localization is below random on the one encoder whose map is per-sample. **(b)** The trade-off, both axes multi-seed, mean±95% t-interval; its axes come from different backbones *and datasets* (utility Sleep-EDFx, attribution synthetic), so no point is scored on both on the same data — an observed anti-correlation, not a Pareto frontier. **(c)** repeats it in one architecture (TempCNN, utility on PTB-XL): the residual still trades attribution ($0.86 \to 0.36$) for utility ($0.646 \to 0.736$, point estimates). A pass on one axis licenses nothing about the others.

MD-TJEPA itself the fix we trained was the balance term, not target gate removal. The pathology is therefore not that a sparsity penalty switches components off, which is what it is for, but that a shared gate lets the objective satisfy the penalty by scaling everything at once, leaving what to keep unconstrained. This is a stand-in, not a second dataset; we claim only the mechanism.

An explanation of ours that did not survive. We had blamed two negatives of §6 (a map constant in the input; randomization failing $0/5$) on the live components' near-uniform norms ($1.02\times$ spread). Retrained norms span $31.9\times$ and randomization still fails $0/5$, so we withdraw the explanation.

The retrain does not repair the decomposition generally. The band axis is a *separate* degeneracy — unconstrained Gabor centers, not the sparsity penalty — and it is untouched: the retrained banks realize **3** distinct canonical bands against the deployed 4, and biomarker recall@k is unchanged at 0.400 ± 0.000. Fixing the gate fixed the gate. The cause is not a bank that cannot move: under supervision the centres travel 7.95 ± 1.19 Hz more than on shuffled labels, converge on whichever band carries the label, leave a band that carries none, and sharpen there (σ_f 2.08 ± 0.03 against 6.61 ± 0.36 Hz) — §E. What the degeneracy reflects is an objective that a coarse band axis serves as well as a fine one.

6 Four axes, four verdicts

An audit asking four questions of the same checkpoints. They disagree, which is the point; Fig. 2 summarises.

Ground-truth recovery. On a synthetic task where the label-carrying (channel, band) cell is known, the intrinsic attribution recovers the channel at 0.897 ± 0.081 ($n = 10$, chance 0.25). A *model-free* band-energy proxy beats it in three of five regimes and reads 0.004 and 0.000 in the other two — inverted, not uninformative, which is the reading we insist on for IG's identical 0.000 — and a per-regime sign, a test-set selection, closes those two too. The decomposition buys no reliable advantage over a spectrogram. Integrated Gradients (Sundararajan et al., 2017) localizes the joint cell better than we do (0.96 vs 0.52); we withdraw the claim that it cannot.

Parameter sensitivity. Cascading randomization (Adebayo et al., 2018) splits by checkpoint: one PTB-XL (Wagner et al., 2020) checkpoint passes all ten randomization draws, while all five Sleep-EDFx models fail the all-draws rule, before and after the gate fix (the rule is uncalibrated; §A.5.4). On the tracked comparison IG decorrelates *better* than our own score ($\rho = -0.08$ against $+0.23$).

Downstream utility, with a floor. Under a matched, balanced 12k-window probe on a leakage-free subject-disjoint split (n=5 each), TS2Vec (Yue et al., 2022) reaches 0.552 ± 0.012 and CPC (van den Oord et al., 2018) 0.424 ± 0.028 against MD-TJEPA's 0.370 ± 0.008 — and an **untrained** MD-TJEPA's $\mathbf{0.444 \pm 0.003}$: pretraining costs 0.074 macro-F1, through neither the gate collapse nor the probe's fixed penalty (§K), and 0.06–0.19 under every readout that keeps the components; untrained, a 128-d label-free projection reaches 0.625, above TS2Vec (§M).

Clinical alignment. The deployed convolutional and transformer encoders give a per-lead map *constant in the input* on real 12-lead ECG (Spearman ≥ 0.96 against noise); a state-space encoder (Gu et al., 2022) yields per-sample maps, and its infarction lead localization is *below random* (-0.107 ± 0.014).

What survives. No variant of five reaches both faithful attribution and TS2Vec-level accuracy: the shapelet residual matches TS2Vec (0.563 ± 0.014) at attribution 0.346 ± 0.073, the gated variant holds 0.803 ± 0.108 at 0.468 ± 0.086 — empty on point estimates, unseparated at this n.

The attribution axis is backbone-dependent; the utility axis is not. Utility reproduces across three backbone families and a second dataset; attribution does not, and the informative corner is established only on TempCNN, not the encoder behind our downstream numbers. The instrument is not blind to transformers — a probe on the full z_u recovers the channel at 1.000 on both — but the norm heatmap's argmax reaches 0.858 on TempCNN and 0.335 on PatchTST, where a fitted probe on the same map reaches 0.787.

7 Limitations

Linear heads, one contested axis, synthetic intervention. "ϕ inherits w's direction" is specific to a linear head. We measure two model families (§4) but one axis; the band head's map and the intervention it is checked against (§G) are synthetic, with one component active per sample. On real data three heads on one s give maps agreeing at ρ 0.57–0.76, with 19–43% of the top six components changing with the head (§N). *Conceded in the appendix:* ground-truth recovery did not buy clinical localization on PTB-XL. The gate retrain also changed batch size and records, so on MD-TJEPA the balance term's effect is not isolated; only the stand-in isolates it. Cascading randomization fails $0/5$ on the deployed Sleep-EDFx models, before and after the gate fix. The Quantus informativeness gap *reverses* at scale (-0.191). IG localizes the joint cell better than we do (0.96 vs 0.52); we beat neither a spectrogram nor our untrained weights. Theorem 1 covers the channel axis only, and its centering (A4′) **does not hold on our checkpoints** (0.631 ± 0.036); Corollary 1 uses neither.

8 What to check before reading an additive attribution

1. Is the axis in the aggregate, and does it move the output? A probe on s near chance, as ours at 0.274, says the aggregate hides *which* channel was active, not that its content goes unused: resampling it moves every head's logit (§G). *The rule*: read the decomposition against that intervention on the same head, not against the probe, averaging the intervention over draws or re-referencing the decomposition to the training-mean contribution (validated here on synthetic data only).

2. Does the score rank what the model routes? Report $\mathrm{overlap@}k$ between the published score and the gated $\alpha_u \|z_u\|$; ours read 0.000, so the score was not describing the forward pass.

3. Does it survive a change of backbone? Our utility results do across three families, our intrinsic attribution results do not (§6).

AI USE STATEMENT

This submission was produced with substantial, sustained assistance from an LLM coding assistant, used interactively throughout the project. We report it against the subtask list in the ICLR 2027 AI policy.

Subtasks the policy requires us to disclose. We used the assistant to *develop the conceptual framework* (the two-case framing and its correction), to *formulate the mathematical claims* (Proposition 1 and the statement of Theorem 1), to *supply ingredients for and write the proofs* (Appendix B), to *propose and refine hypotheses*, to *design the experiments and give feedback on methodology*, to *implement the methods*, to *clean and reformat the datasets* (the Sleep-EDFx and PTB-XL loaders, the per-record normalisation and the subject-disjoint splits), and to *interpret results*. The synthetic benchmark of §A.5.1 is *generated by code the assistant wrote* rather than emitted by a model: the series are white noise with an injected narrowband oscillation, produced procedurally at a fixed seed. We did not use it for *translation*, and *qualitative and thematic analysis* does not arise in this work.

Subtasks the policy recommends disclosing. We used the assistant to write and edit the software, to create the result artifacts, to draft and revise the manuscript including this statement, to search for and summarise related work, to identify relevant literature and format the references, to brainstorm and to *discover research topics and identify gaps* — the audit that found the falsified claim, and the one that found the broken randomization control, were both its own — to suggest a structure for the paper, to suggest experimental parameters, to produce the figure, and to propose the title and keywords.

How the work was reviewed. Two of the corrections this required are worth naming, because they show what review caught. Several claims the assistant proposed were *refuted by measurements it subsequently ran* — our own explanation of the sanity-check failure, and an initial reading of the attribution/routing disjointness that turned out to be gate collapse; both refutations are reported in this paper rather than removed. And the proof in Appendix B had been attributed to a numbered lemma that does not exist and relied on an assumption it never stated; the attribution was corrected against the sources and the assumption is now written out as (A5).

Every reported number was regenerated from committed artifacts by the test suite rather than transcribed, and the bibliography was checked entry by entry against publisher records. The authors directed the work, reviewed the code and the artifacts behind every number, and take responsibility for the final content of this submission, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses two public, de-identified corpora — Sleep-EDF Expanded and PTB-XL — under their published licences. No human subjects were recruited, no new recordings were made, and no attempt was made to re-identify any participant.

The paper's substantive ethical content is its negative result. Additive decomposition is promoted as a route to clinically auditable models, and a channel- or band-level attribution over physiological signal invites exactly the reading that a clinician would want to act on. We show that on these models a faithful attribution depends on the head one happens to fit and on a reference it takes implicitly: as computed intrinsically it agrees with an input intervention on 41% to 76% of windows depending on the head, and re-referenced, on synthetic data, on 73% to 95%. A persuasive channel map is therefore not by itself evidence of what the model used, and presenting one as clinical evidence without the intervention it approximates would misrepresent how much it has been checked. We therefore report the limitation rather than the most favourable configuration, and we do not claim downstream accuracy: a TS2Vec baseline outperforms our model under a matched probe, and we say so.

REPRODUCIBILITY STATEMENT

Every number in this paper is regenerated from committed artifacts by the test suite rather than transcribed by hand, with one recorded exception (CPC's accuracy and κ intervals, whose per-seed values were not stored); the suite greps both the manuscript and the artifacts and fails on

disagreement, which is how several errors in earlier drafts were caught. It did not cover the appendix until an audit of our own paper found eight intervals there built with a normal multiplier instead of Student-t — one of which had kept an "above chance" verdict alive that the t-interval kills, and which we now withdraw. The appendix is in the grep's scope as of that fix. Unless a table says otherwise, confidence intervals are Student-t at the reported n over training seeds, not over subjects or windows, so they capture training variability and not sampling of new patients. Training artifacts under the run directories are regenerate-only rather than committed (checkpoint size); every number the paper quotes is computed from a committed result artifact, and each artifact now records the environment that produced it. Seed counts differ by cost and we state which is which rather than quoting a single n. The synthetic measurements the abstract quotes are run at $n{=}10$; results on trained Sleep-EDFx and PTB-XL checkpoints stay at $n{=}5$, because each additional seed there is a full pretraining run; and two further figures are lower still — the per-variant macro-F1 comparisons are over four seeds and the synthetic band-axis accuracy (Appendix §A.5.1) over three. Each interval is computed at its own n with the matching Student-t multiplier, and the comparisons that turn on a small n say so at the point of use. The one real-data figure the abstract quotes — the $\mathrm{overlap@}k$ of 0.000 ± 0.000 — is an $n{=}5$ result on deployed checkpoints.

No power analysis was pre-registered, and the achievable resolution is worth stating plainly. At $n{=}5$ an independent two-sample comparison reaches 80% power only at $d \geq 2.02$ — a macro-F1 gap of roughly 0.021 at the spread we observe. The shapelet-versus-TS2Vec pair is four seeds against five, where the threshold is $d \geq 2.19$ (a gap of 0.020); the observed $d \approx 1.18$ carries power ≈ 0.33, so it is reported as unseparated rather than matched. No overlapping pair in this paper is read as equivalence. The findings we do assert — the derived attribution's arbitrariness, the gate collapse, the below-random lead localization — are far outside this band, and per-seed agreement is reported alongside each. The appendix carries the full audit: model and training hyperparameters, the dataset splits (subject-disjoint) and normalisation (per-record z-score), the seed set behind each interval, and the selection criterion for λ_{balance}. Seeding covers the Python, NumPy and PyTorch generators through a single entry point invoked by every training and baseline script. One caveat we found while re-running at a larger n and record rather than paper over: the convolutional encoder is not bit-reproducible across runs at a fixed seed, consistent with non-deterministic cuDNN convolution algorithms, while the transformer encoder reproduces exactly. We quantified this by re-running every synthetic script from its committed defaults: a single seed's channel accuracy moved by as much as 0.171, while the ten-seed mean moved by at most 0.020, every interval still overlapped its original, and the ordering of the trade-off variants was unchanged. So the multi-seed mean is the reproducible quantity and a single seed's value is not, on the convolutional backbone; a reader should compare the former. An earlier version of this statement quoted the 0.02 figure for a single seed, which understates the per-seed movement by an order of magnitude.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4699–4711, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/251bd0442dfcc53b5a761e050f8022b8-Abstract.html.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. arXiv:1610.01644v4.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding

predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/cvpr52729.2023.01499.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025. arXiv:2511.08544v3.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2), 2024. doi: 10.1073/pnas.2304406120.

Gaëlle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables: Application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012. doi: 10.1214/12-EJS749.

Benedict Clark, Rick Wilming, Hjalmar Schulz, Rustam Zhumagambetov, Danny Panknin, and Stefan Haufe. Correcting misinterpretations of additive models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/31888563b194f9bb33ce1aebc7e1551c-Abstract-Conference.html.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://www.jmlr.org/papers/v22/20-1316.html.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-2732. Spotlight; arXiv:2409.19407.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL https://jmlr.org/papers/v20/18-760.html.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 35971–35983, 2022. doi: 10.52202/068431-2607.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. URL https://jmlr.org/papers/v24/22-0142.html. arXiv:2202.06861.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Amirabbas Hojjati, Lu Li, Ibrahim Hameed, Anis Yazidi, Pedro G. Lind, and Rabindra Khadka. From video to EEG: Adapting joint embedding predictive architecture to uncover saptiotemporal dynamics in brain signal analysis, 2025. arXiv:2507.03633v5 (preprint; no peer-reviewed venue confirmed).

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007x237892.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6441–6452, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/47a3893cc405396a5c30d91320572d6d-Abstract.html.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A. C. Kamphuisen, and Josefien J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Sehun Kim. Learning general representation of 12-lead ECG with a joint-embedding predictive architecture. *Neural Networks*, 205:109334, 2027. doi: 10.1016/j.neunet.2026.109334.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://openreview.net/forum?id=jESY2WTZCe.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 2402–2412, 2020. URL https://proceedings.mlr.press/v108/lengerich20a.html.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-0538.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018. doi: 10.1038/s41467-018-07210-0.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=msJgEkjwh5.

Gabriel Michau, Gaëtan Frusque, and Olga Fink. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences (PNAS)*, 119(8), 2022. doi: 10.1073/pnas.2106598119. arXiv:2105.00899.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Joseph Paillard, Jörg F. Hipp, and Denis A. Engemann. GREEN: A lightweight architecture using learnable wavelets and Riemannian geometry for biomarker exploration with EEG signals. *Patterns*, 6(3):101182, 2025. doi: 10.1016/j.patter.2025.101182.

Saarang Panchavati, Uddhav Panchavati, Hiroki Nariai, Corey Arnold, and William Speier. Laya: A LeJEPA approach to EEG via latent prediction over reconstruction, 2026. arXiv:2603.16281v2 (preprint; no peer-reviewed venue confirmed).

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/37da88965c016dca016514df0e420c72-Abstract-Conference.html.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Pablo Ruiz-Morales, Dries Vanoost, Davy Pissoort, and Mathias Verbeke. Koopman invariants as drivers of emergent time-series clustering in joint-embedding predictive architectures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(30):25176–25183, 2026. doi: 10.1609/aaai.v40i30.39708.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games (AM-28), Volume II*, number 28 in Annals of Mathematics Studies, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278, 2020. URL https://proceedings.mlr.press/v119/sundararajan20b.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748v2*, 2018.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(154), 2020. doi: 10.1038/s41597-020-0495-6.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002.

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 (NeurIPS)*, 2023. doi: 10.52202/075280-3420. arXiv:2305.10351.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022. doi: 10.1609/aaai.v36i8.20881. arXiv:2106.10466.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R. Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://papers.nips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, 2021. doi: 10.1609/aaai.v35i12.17325.

A FULL AUDIT

The main text states the results; this appendix carries the measurements behind them. Subsections are numbered within this appendix; the released artifacts and the commit history carry their own §4.x names, and the mapping is one-to-one in the order below. Where to look:

main text	appendix
§4 premise, both cases	§A.5.11, §A.5.1
§4 block aggregation	§A.5.11
§4 transplant off this architecture	§A.5.12
§5 collapse and fix	§A.5.13, §A.5.14
§5 per-sample check	§A.5.15
§6 ground truth	§A.5.1, §A.5.2
§6 randomization	§A.5.4, §A.5.7
§6 utility	§A.3, §A.6
§6 clinical	§A.5.3, §A.5.9
band-axis caveat	§A.5.10

A closing subsection (§A.7) records the checks we did *not* run, so the scope of what was and was not measured is explicit.

A.1 SETUP

- Datasets: Sleep-EDFx (197 records, 100 Hz, 2 EEG ch: Fpz-Cz, Pz-Oz), PTB-XL (21,837 records, 12-lead, 100 Hz).
- **Normalization (Mo-3, stated for reproducibility):** we apply *offline per-record* z-normalization (each record standardized by its own mean/std across all its epochs). This is computed over the whole record at evaluation time; it is not train/test subject leakage (splits are subject-disjoint), but it is a global-in-time statistic and a deployment assumption we make explicit.
- Strodthoff folds (1-8 train, 9 val, 10 test) for PTB-XL; subject-level split

(137/29/31) for Sleep-EDFx.

- 5 seeds per cell, paired-t + Bonferroni-Holm.

A.2 PHASE 2 ABLATION MATRIX (12 CELLS)

Setup: 22 cells × 3 seeds = 66 runs (78 with A10/A12 cross-group oversampling), 3 epochs × 40 batches × 5 records on Sleep-EDFx, ~5 minutes total wall-clock on 4× RTX 3090. Results aggregated by the ablation aggregator. (The cell identifiers A1–A12 here name ablation knobs and

have nothing to do with the assumptions (A1)–(A5) of Theorem 1; the collision is historical and we keep both names because the artifacts use them.)

Headline findings (this row of the table is the paper's strongest empirical claim):

- **A12 block-diagonal Koopman wins**: `A12_block_diag` reaches L_jepa = 1.21 $\pm$ 0.05

with near-identity fraction 1.000 $\pm$ 0.000, while `A12_full_predictor` (single-block linear predictor) reaches L_jepa = 1.46 $\pm$ 0.12 with near-identity 0.979 $\pm$ 0.015 and a 6$\times$ larger Frobenius penalty (L_koop 0.28 vs 0.05). *Block-diagonal Koopman is necessary, not merely sufficient.*

- **A2_freq_off / A3_multiscale_off** show degenerately low L_jepa (0.08 and 0.23

respectively) because collapsing to a single band or scale makes the target trivially predictable from the context — a known degeneracy mode worth flagging in the paper.

- **A4 (Koopman strength), A5 (Sparse), A9 (ANOVA), A10 (Warm-start)** produce

null effects at this 3-epoch budget (L_jepa within 0.001 of the reference config). A longer schedule is required to manifest these knobs; we plan a Phase 2 W26-28 budget-multiplier run.

- **A6 PatchTST > TempCNN** flips at this short budget: PatchTST L_jepa = 9.84 $\pm$

1.71 (transformer cold-start), TempCNN L_jepa = 1.21. With more steps PatchTST overtakes (prior Phase 2 W1-W2); all headline results accordingly use the 30-epoch PatchTST configuration.

- **A1_channel_off**: with the harness fix (channel-coordinate slicing in

the pretraining entry point), L_jepa = 0.51 $\pm$ 0.08 — lower than the 2-channel reference (1.21 $\pm$ 0.05). This matches the A2/A3 degeneracy pattern: fewer (c, f, s) tubes $\Rightarrow$ simpler target $\Rightarrow$ easier to predict. A meaningful "channel decomposition needed" test requires a controlled-information-content baseline (e.g. equal total parameter count).

Per-cell mean $\pm$ std bars and the sweep plots are included in the released artifacts.

A.2.1 CROSS-DOMAIN + PAPER-GRADE REPLICATION

The headline A12 finding (block-diagonal Koopman > full-rank predictor) replicates across three independent budgets and two domains:

domain	budget	n	block-diag	full-rank	Δ	sign	Wilcoxon
Sleep-EDFx	3 ep TempCNN	3	1.206 ± 0.054	1.461 ± 0.123	+0.255	3/3	— (n=3)
Sleep-EDFx	3 ep TempCNN (2 groups, one seed repeated)	5 unique	1.262 ± 0.080	1.435 ± 0.101	+0.173	5/5	p ≤ 0.062 (not an independent replicate of row 1)
Sleep-EDFx	30 ep PatchTST	5	0.337 ± 0.018	0.512 ± 0.196	+0.175	5/5	p = 0.0313 ✓
Sleep-EDFx	**10 ep PatchTST**	**10**	**0.290 ± 0.046**	**0.418 ± 0.221**	**+0.128**	**9/10**	**p = 0.00195 ✓✓**
PTB-XL	2 ep TempCNN	3	9.388 ± 0.446	10.104 ± 0.546	+0.716	3/3	— (n=3)
PTB-XL	2 ep TempCNN (added seeds)	5	10.266 ± 1.130	10.685 ± 0.950	+0.418	4/5	p = 0.219

The randomization control was broken, and no verdict depended on it. The cascade's last step randomizes the Gabor bank, and it did so by drawing $\mathcal{N}(0, 0.1)$ over `raw_f`/`raw_sigma` — which live in softplus-inverse space. Decoded, every filter landed at 1.09–1.27 Hz against canonical centres of 2.0–48.5 Hz: ten copies of one δ filter, not a random bank. The frequency axis was being *destroyed* rather than randomized, so the correlation drop that defines a pass was partly structural. We found this in an audit of our own code, fixed it (centres uniform on $(f_{\min}, f_{\max})$, σ_t log-uniform over the bank's initial range, re-encoded through the softplus inverse; decoded centres now span 4.79–43.98 Hz), and re-ran all eight cascades on the same checkpoints, budget, seeds and threshold. **Not one verdict changed**: the three rows above stay $10/10$, and the five deployed Sleep-EDFx models (§A.5.4) stay $0/5$, with their pass fractions rising from 0.7–0.8 to 0.8–0.9 without reaching 1.0. The ρ values above are from the corrected control; the released comparison artifact holds both runs side by side, per seed.

What a verdict here is, and is not. A checkpoint passes if the final cascade step's $|\rho| < 0.3$ in all ten randomization draws. Two consequences. First, the ten draws are repeated randomizations of

one trained checkpoint, not ten trained models: PTB-XL's $10/10$ is evidence about one PTB-XL model, and Sleep-EDFx's $0/5$ is five trained models each failing the all-ten rule. Second, the all-pass rule is not calibrated: if each draw passed independently with probability 0.95, a checkpoint would still fail the rule 40% of the time ($1 - 0.95^{10}$), and draws on the same checkpoint are not independent. We report pass fractions beside every verdict for that reason, and did not estimate the rule's false-fail rate under a null.

The **n=10 PatchTST Sleep-EDFx reference (row 4, bold)** is the paper-grade headline that **survives Bonferroni-Holm correction**: Wilcoxon signed-rank one-sided **p = 0.00195** clears the corrected alpha of 0.05/22 = 0.0023 (22 ablation cells in the matrix). 95% bootstrap CI on the mean Δ L_jepa is **[+0.027, +0.256]** — does not cross zero. The L_koop differential is even more decisive: Δ = **+0.491**, Wilcoxon **p = 0.000977**, L_koop ratio **12.2**×. Block-diagonal Koopman has an order-of-magnitude smaller Frobenius penalty than the single-block full-rank linear predictor, and the gap is statistically significant under multiple-testing correction.

A second-order finding: the *spread* of L_jepa across seeds is **11× smaller for block_diag** in standard deviation (0.018 vs 0.196; about $119\times$ in variance). Beyond winning on mean, block-diagonal Koopman is more **reliable** under random seed.

Both axes (L_jepa and L_koop) agree across all five conditions: **block-diagonal Koopman beats the single-block linear predictor at every budget tested**, with the gap *growing* on the harder PTB-XL setting.

(* The n=6 row aggregates two harness invocations on the same 3 seeds (0, 1, 2); the substring filter in early ablation runs matched cells in two independent output directories. The std underestimates true across-seed variance by ~1.4× because half the values are within-config replicates rather than independent. Subsequent runs use anchored prefix matching to avoid this.)

The PatchTST cold-start anti-pattern is more severe on PTB-XL: 271.78 vs 9.39 (29× worse), versus 9.84 vs 1.21 (8× worse) on Sleep-EDFx. We hypothesize PatchTST needs >100 steps of warm-up before the transformer attention statistics stabilize.

The null-effect knobs (A4, A5, A9, A10) remain null even at 8 epochs on Sleep-EDFx (L_jepa differential < 0.001 across λ values). The Frobenius penalty term λ_koop × ||W||^2 ≈ 0.1 × 0.08 = 0.008 is dominated by the JEPA loss ~ 1.2. *Conclusion*: these knobs are not what regularizes the JEPA objective on short downstream tasks — they regulate sub-properties (L_koop, L_anova) but not the JEPA prediction itself.

Cross-domain comparison outputs are included in the released artifacts.

cell	n	L_{total}	L_{jepa}	L_{koop}	L_{sig}	L_{sparse}	L_{anova}	near-id
A10_warmstart_off	6	63.0875±3.7974	1.2059±0.0536	0.0475±0.0147	42.1828±3.4690	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A10_warmstart_on	6	63.0873±3.7976	1.2060±0.0536	0.0475±0.0147	42.1826±3.4692	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A12_block_diag	6	63.0876±3.7974	1.2059±0.0536	0.0475±0.0147	42.1829±3.4690	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A12_full_predictor	6	57.9914±0.7696	1.4605±0.1227	0.2844±0.0989	36.8427±0.6362	7.0886±0.0116	1961.4511±54.8554	0.9792±0.0147
A1_channel_off	3	46.4477±2.7917	0.5095±0.0786	0.0592±0.0127	35.9292±3.0239	3.5338±0.0116	997.3115±45.5763	1.0000±0.0000
A1_channel_on	6	63.0874±3.7975	1.2060±0.0536	0.0475±0.0147	42.1827±3.4691	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A2_freq_learn	3	63.0876±3.7973	1.2059±0.0536	0.0475±0.0147	42.1830±3.4689	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A2_freq_off	3	32.7633±0.3211	0.0816±0.0009	0.1331±0.0088	30.8580±0.2794	0.6940±0.0023	181.5509±10.9084	1.0000±0.0000
A3_multiscale_off	3	39.0753±1.6592	0.2345±0.0267	0.1136±0.0126	32.4086±1.6883	2.3324±0.0124	640.7803±20.2326	1.0000±0.0000
A3_multiscale_on	3	63.0874±3.7976	1.2059±0.0536	0.0475±0.0147	42.1827±3.4691	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A4_koop_000	3	63.0868±3.7974	1.2059±0.0536	0.0475±0.0147	42.1826±3.4691	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A4_koop_001	3	63.0871±3.7972	1.2059±0.0536	0.0475±0.0147	42.1829±3.4688	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A4_koop_010	3	63.0877±3.7972	1.2059±0.0536	0.0475±0.0147	42.1831±3.4687	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A4_koop_100	3	63.0922±3.7993	1.2059±0.0536	0.0475±0.0146	42.1833±3.4696	7.0658±0.0178	1962.7584±52.0437	1.0000±0.0000
A5_sparse_000	3	63.0168±3.7979	1.2060±0.0536	0.0475±0.0147	42.1828±3.4693	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A5_sparse_001	3	63.0239±3.7979	1.2060±0.0536	0.0475±0.0147	42.1828±3.4693	7.0658±0.0178	1962.7584±52.0439	1.0000±0.0000
A5_sparse_010	3	63.0876±3.7973	1.2059±0.0536	0.0475±0.0147	42.1830±3.4689	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A5_sparse_100	3	63.7226±3.7959	1.2058±0.0536	0.0475±0.0147	42.1822±3.4691	7.0653±0.0179	1962.7588±52.0440	1.0000±0.0000
A6_backbone_patchtst	3	59.8607±2.3074	9.8362±1.7149	0.0163±0.0007	30.5696±0.1312	7.1083±0.0010	1938.3664±76.4860	1.0000±0.0000
A6_backbone_tempcnn	3	63.0877±3.7972	1.2059±0.0536	0.0475±0.0147	42.1830±3.4687	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000
A9_anova_off	3	43.4637±3.5105	1.2068±0.0535	0.0475±0.0147	42.1857±3.4750	7.0658±0.0178	1982.6249±51.0422	1.0000±0.0000
A9_anova_on	3	63.0874±3.7975	1.2059±0.0536	0.0475±0.0147	42.1828±3.4690	7.0658±0.0178	1962.7585±52.0439	1.0000±0.0000

Honest reading of the table above (auto-generated from the run outputs):

- **Structural ablations move the loss.** The architecture-changing cells

differ sharply from the default (L_jepa 1.206): channel-off 0.510, frequency-off (all-pass) 0.082, multiscale-off 0.234, PatchTST backbone 9.836.

- **A12 (block-diagonal vs full-rank) separates here too:** block L_jepa 1.206

vs full 1.461, with full carrying a 6× larger Frobenius penalty (L_koop 0.284 vs 0.0475) and eigenvalue drift (near_id 0.979 < 1.000). This corroborates the paper-grade §A.2.1 result.

- A9 (ANOVA): turning the orthogonality penalty off lowers L_total (43.5 vs

63.1) and L_anova, confirming the term is active.

Corrected λ-sweep (A4/A5). An earlier draft's aggregation reported the A4/A5 cells as *bit-identical* across λ values — a harness artifact: the scalar-λ override had not propagated into those short runs. We re-ran A4/A5 cleanly (the corrected λ-sweep run, 12 epochs × 3 seeds, verified per-cell λ propagation). The sweep is now non-degenerate and shows the **expected monotonic effect on the targeted sub-loss while leaving L_jepa flat**:

cell	λ	L_jepa	L_koop	L_sparse
A4_koop_000	0	0.272±0.041	0.0372±0.0040	7.115±0.017
A4_koop_001	0.001	0.272±0.041	0.0372±0.0040	7.115±0.017
A4_koop_010	0.01	0.272±0.041	0.0369±0.0040	7.115±0.017
A4_koop_100	0.1	0.272±0.041	**0.0344±0.0036**	7.115±0.017
A5_sparse_000	0	0.272±0.041	0.0369±0.0040	7.116±0.017
A5_sparse_010	0.01	0.272±0.041	0.0369±0.0040	7.115±0.017
A5_sparse_100	0.1	0.271±0.041	0.0369±0.0040	**7.103±0.013**

Increasing **λ_koop 0→0.1 monotonically reduces the Koopman Frobenius penalty L_koop (0.0372→0.0344, −7.5%)**, and **λ_sparse 0→0.1 reduces the gate L1 L_sparse (7.116→7.103)** — each knob regularizes exactly its targeted term. But **L_jepa is unchanged to three decimals** across both sweeps: at this budget the penalties shape sub-properties (Frobenius norm, gate sparsity) without moving the JEPA prediction loss. We therefore make a *scoped* claim — the knobs work as designed on their own terms — and rely on the dedicated A12 comparison (§A.2.1), not the λ magnitude, for the headline Koopman-structure result.

A.3 Downstream evaluation

A.3.1 Sleep-EDFx 5-stage classification — linear probe (leakage-free split)

We extract the frozen encoder's per-epoch latent ($s_x \in \mathbb{R}^{32}$) and fit a multinomial logistic regression (sklearn, max_iter=2000); no fine-tuning. All numbers below are on the **leakage-free, subject-disjoint split** versioned with the released artifacts (70/15/15 subjects, **0/0/0 train/val/test subject overlap**, seed=0). We report the **held-out TEST** split — never val-as-test — from the 30-epoch, 50-record PatchTST checkpoint.

Held-out **TEST**, full set (n = 70,675 epochs): accuracy = **0.672**, macro F1 = **0.283**, κ = **0.280**. (An earlier draft reported 0.728 / 0.238 / 0.174 on the **first 6,400** test windows only — a `-max-batches` cap that, by chronological ordering, over-sampled easy W epochs; the full-test number is the honest, representative one and we use it as the headline. The full test is slightly *worse* on accuracy but *better* on the balanced κ.) The versioned report (capped per-class breakdown below) and the full-test numbers are in the released artifacts.

Per-class breakdown (capped 6,400-window subset, illustrative of the class-imbalance pattern):

Stage	Precision	Recall	F1	Support
W (wake)	0.78	0.96	0.86	4,682
N1	0.00	0.00	0.00	173
N2	0.27	0.19	0.23	795
N3	0.67	0.03	0.06	384
REM	0.26	0.03	0.05	366

Supersession history. An even earlier draft reported **0.800**, measured *before* the audit's split-leakage fix (SB-1) and under a val-as-test protocol (SB-2). The

leakage-free, full-test number is **0.672**. We do **not** claim competitive downstream accuracy: the N1/REM minority collapse (F1 $\leq$ 0.05) is a single-shot linear-probe failure under heavy imbalance, and (see below) a dedicated SSL representation learner outperforms MD-TJEPA on this task. The paper's contribution is *faithful attribution* (§A.5.1–A.5.7), not accuracy.

External SSL baseline (TS2Vec) — fair head-to-head (SB-3). An earlier draft compared the two under *unequal* probes (TS2Vec used class-balanced weighting + a 12k train-window budget; MD-TJEPA used unweighted + the full train set). We now run **one identical probe protocol** for both — same leakage-free split, same frozen-encoder `LogisticRegression`, same `class_weight=balanced`, same 12,000 train-window budget, same full test — produced by the probe-comparison script, which asserts the two artifacts' probe protocols are identical (`protocol_matched: true`).

Table 2: All rows use an identical probe: balanced, 12k train, full test.

Method	accuracy	macro F1	κ	faithful attribution
TS2Vec (our reimpl., n=5)	**0.787±.011**	**0.552±.012**	**0.606±.019**	× (no per-(c,f) attribution)
CPC (our reimpl., n=5)	0.606±.033†	0.424±.028	0.371±.041†	× (no per-(c,f) attribution)
MD-TJEPA (ours, n=5)	0.528±.016	0.354±.007	0.288±.015	partial: synthetic-GT 0.86±0.17 ch; Adebayo 0/5 on these models; (A4″) unmet

† CPC's accuracy and κ intervals are the only ones here we could not regenerate from stored per-seed values, so they remain as originally recorded; every other interval in this table is a 95% t-interval recomputed from its artifact.

Pretraining budgets are matched on the probe, not on the encoder. The downstream probe is identical for all three rows (balanced, 12k-window cap, same split). The *pretraining* is not: TS2Vec runs 20 epochs (the baseline launcher), MD-TJEPA 30, and CPC has no recorded launch, so it presumably used its script default of 40. The asymmetry runs in both directions, so it is not a hidden advantage for either side, but a budget-matched comparison is a different experiment from the one we ran and we do not claim to have run it.

We report this honestly: under the identical probe, two strong SSL baselines (TS2Vec and CPC) beat MD-TJEPA on every downstream metric (and by more than the earlier unequal comparison suggested — MD-TJEPA's accuracy falls from its best-protocol 0.672, §A.3.1, to 0.528±.016 under balancing + the 12k cap, n=5). A method built purely for representation quality is, unsurprisingly, the better *classifier*. MD-TJEPA's contribution is orthogonal — it is the only one of the two whose latent yields a **by-construction, model-faithful per-(channel, band) attribution** (§A.5). The paper's claim is precisely this *interpretability-at-a-performance-cost* tradeoff (quantified as an observed Pareto trade-off, §A.6), not downstream superiority.

A.4 General time-series generality (R10): ETTh1

ETTh1 (Electricity Transformer Temperature, hourly) is the standard long-horizon forecasting benchmark from Informer (Zhou et al., 2021). 7 features, hourly sampling, ~14k rows. We run MD-TJEPA with no biosignal-specific tuning: scales = (4 h, 12 h, 24 h), num_bands = 6 with random Gabor initialization (no canonical-band warm start), TempCNN encoder. We report two budgets to honestly trace the scaling behavior.

Scope (audit SB-5). This section evidences *method transfer*, not competitive forecasting. We report that the SSL objective optimizes (L_jepa drops), the Koopman predictor stays near-identity, and the attribution passes the Adebayo cascade on ETTh1 — i.e. the architecture is not biosignal-specific. We do **not** report forecasting accuracy (MSE/MAE) against Informer/PatchTST-forecasting, and we run **no** UCR/UEA benchmark and **no** external SSL baseline (TS2Vec/SimMTM/TS-TCC/TF-C): the only implemented control is the random-init baseline (the released random-init baseline). A fair competitiveness claim requires those under a shared split/compute/selection protocol and is left to future work.

Adebayo cascade ρ below is the final +wavelet stage, **mean ± std over 10 seeds** on the full (c, f, s) attribution (audit F-A2; see §A.5.4).

Domain	Budget	L_{jepa} start → best	near-id	Adebayo ρ	pass
Sleep-EDFx (PatchTST)	50 ep × 60 batches × 10 records	14.57 → 0.86 (best @ep18; 1.35 final)	1.000	+0.007 ± 0.131	10/10 ✓
PTB-XL (PatchTST)	**20 ep × 200 batches × full train fold (F-A1)**	**228.88 → 2.985 (76× drop)**	**1.000**	**-0.008 ± 0.073**	**10/10** ✓
ETTh1 (TempCNN)	**5 ep × 40 batches** (smoke)	2.95 → 2.43	1.000	+0.030 ± 0.086	10/10 ✓
ETTh1 (TempCNN)	**30 ep × 200 batches × full** (audit F7)	**3.14 → 1.12 (-64%)**	**1.000**	**-0.021 ± 0.064**	**10/10** ✓

Note (audit M-E / B-B / F-A2): PAPER cited only the 5-ep ETTh1 smoke result historically; the 30-ep scaled result (L_jepa 1.12, 64% drop) exists in the scaled ETTh1 run and is reachable via the released cascade-randomization entry point (CLI parity added per audit B-B; seeded multi-seed protocol per F-A2). The scaled ETTh1 Adebayo is **complete**: +wavelet ρ = **+0.036** ± **0.083** (10/10 seeds pass), single-shot ρ = **-0.039**, both PASS (|ρ| < 0.3); the artifact is released.

Note (audit F-A1, PTB-XL scaled): the original PTB-XL run (cited in earlier drafts) and its scaled successor (interrupted at epoch 14, the final checkpoint corrupted) have been replaced by the replacement run: 20 epochs × 200 batches × full train fold, monotonic descent 228.88 → **2.985** (76× reduction, **no U-curve** unlike Sleep-EDFx 50ep), `near_id = 1.000` throughout. The Adebayo and Quantus (Hedström et al., 2023) reports below reference this v2 checkpoint.

The third row demonstrates MD-TJEPA's **domain generality** — the three biosignal-specific design choices (canonical $\delta\theta\alpha\beta\gamma$ band init, sleep 30-second epoch, ECG 12-lead AHA mapping) are NOT prerequisites for either training convergence or Adebayo-faithful attribution. The same pipeline trains stably on 7-channel hourly electricity data, with identical Koopman block-diagonal behavior (near_id 1.000) and attribution that decorrelates under cascading model randomization.

This answers the obvious "is this only a biosignal method?" reviewer concern (R10 in an internal planning note).

A.4.1 UCR CLASSIFICATION TRANSFER (ECG5000) WITH A CONTROL BASELINE

To put a *classification* (not just convergence) datapoint behind the generality claim, we add one UCR-family experiment. We treat the 140-sample ECG5000 beat (binary normal/abnormal variant, TF data mirror; 500 train / 4498 test, the UCR train size) as a 1-second window at fs = 140 Hz and adapt the short-series hyperparameters (4 Gabor bands at 15–55 Hz, 127-sample kernel, sample-scale pooling, latent 16). We SSL-pretrain MD-TJEPA, then fit a frozen linear probe, and compare against the **random-init control** (same frozen probe on an untrained MD-TJEPA, the released random-init baseline logic).

ECG5000 (binary)	accuracy	macro F1
raw-feature LR (canonical UCR linear baseline)	**0.985**	**0.985**
MD-TJEPA SSL (frozen 16-d probe)	0.918	0.917
random-init control (frozen 16-d probe)	0.878	0.875

Two honest readings, both stated plainly (seed 0):

- **SSL > random-init by +4.0 pts** (0.918 vs 0.878): the SSL pretraining does

learn task-relevant structure beyond the random-initialized architecture.

- **SSL < raw-feature logistic regression** (0.918 vs **0.985**): on this

short (140-sample), univariate, *easy* binary task, a plain linear model on the raw series **beats** MD-TJEPA's 16-d compressed latent. We do **not** hide this — it is the expected outcome when an aggressively bottlenecked, long-multichannel-biosignal architecture is applied to a tiny single-channel series where raw features are already trivially separable.

So this is a **proof-of-transfer (SSL learns something), not a competitiveness claim**: MD-TJEPA is not SOTA on ECG5000 and we make no such claim. We also note the SSL L_jepa does **not** cleanly descend at 140 samples (0.049 → 0.240) — the context/target split is near-trivial at this

length. A fair UCR/UEA benchmark (multiple datasets, length-appropriate configs, external SSL baselines TS2Vec/SimMTM under a shared protocol) remains future work; broad UCR/UEA download was attempted but the standard `timeseriesclassification.com` mirror served only HTML stubs in our environment, so multi-dataset coverage is deferred.

A.5 Interpretability evaluation

A.5.1 Synthetic ground-truth faithfulness (the primary faithfulness test)

Adebayo cascading sanity (§A.5.4) is a *necessary* check — attribution must depend on the trained model — but not a *sufficient* one: it does not verify that the attribution points at the *right* input component. We therefore build a controlled setting with a **known ground-truth important cell** and measure recovery directly.

Construction. Each series has `C=4` white-noise channels except one randomly chosen channel `c`, into which we inject a clean oscillation at one of `F=4` band frequencies `f` (6/10/16/24 Hz). After per-channel z-normalisation the *only* discriminative structure is *which (channel, band) cell carries the narrowband energy*; the pair `(c, f)` is the ground truth. We SSL-pretrain MD-TJEPA (no labels), then score per-sample attributions on a held-out test set (seed 0; chance recall@1 = 1/16 = 0.0625).

Method	channel acc	band acc	joint recall@1
chance	0.25	0.25	0.0625
random-init MD-TJEPA (intrinsic)	0.29	0.28	0.076
MD-TJEPA SSL (intrinsic)	**0.89**	0.52	**0.52**
MD-TJEPA SSL + band prior (frozen centers)	0.90	0.52	0.52
post-hoc Integrated Gradients	0.00	**0.96**	0.96

Multi-seed (C-5). Over 5 SSL seeds the intrinsic channel accuracy is **0.86** $\pm$ **0.17** (95% t-interval, n=5; seed values 0.89/0.75/0.99/0.69/0.97) and band accuracy 0.43 ± 0.40 (n=3) — the single-seed 0.89 headline is representative but the seed variance is non-trivial, especially on the (weaker) band axis.

Reading (honest, two-sided). (i) The SSL intrinsic attribution recovers the ground-truth **channel at 0.89** and the joint cell at **0.52 (7× chance)**, far above the random-init control (0.076), giving the first *quantitative, ground-truth* faithfulness number for the method (not just a sanity pass). (ii) Post-hoc IG localizes the *cell* at **0.96**, i.e. $15.4\times$ chance and *better than our own* 0.52 $(8.4\times)$; it loses only on the sum-over-bands channel marginal, where it scores **exactly** 0.000. A score of exactly zero over 800 samples is not uninformative but perfectly anti-aligned, which points to a sign convention in the marginal rather than to a blind baseline. We therefore no longer claim that IG "cannot localize the channel"; the honest statement is that the two methods win on different marginals of the same map, and that IG is ahead on the joint cell. (iii) The intrinsic **band axis (0.52) is markedly weaker than the channel axis (0.89)**: this independently reproduces the run-dependent band-allocation problem of §A.5.6 from a clean synthetic angle. We use this benchmark as the measurement instrument for the **retained canonical-band prior** (frozen Gabor centers, §3.1). The prior leaves synthetic band accuracy **unchanged** ($0.52 \rightarrow 0.52$) — an honest negative: the residual band weakness on this test is filter-bandwidth *smearing* between adjacent passbands, not center *drift*, so anchoring centers does not help discrimination here. What the prior *does* fix — **by construction** — is band **coverage**: with centers frozen at canonical $\delta/\theta/\alpha/\sigma/\beta/\gamma$, the §A.5.6 failure mode (a run that learns *no* α or δ band) becomes impossible. So the prior is a *coverage guarantee*, not a discrimination improvement; sharpening adjacent-band discrimination (e.g. via a learned mixing penalty or narrower kernels) remains open work.

A.5.2 Is the attribution just input band energy? A score-function control

The synthetic cue *is* narrowband energy by construction, so a reviewer may object that MD-TJEPA's intrinsic score merely reads the input spectrogram and adds nothing (the *trivial-energy-proxy* objection). We test this directly by scoring four attributions against the same ground truth: the global gate α alone (sample-agnostic), the input-dependent component norm $\|z\|$, the deployed $\alpha\|z\|$, and a

model-free band-energy proxy $\sum_t |\text{Gabor}(x)|^2$ (handed the model's learned frequency frontend but no encoder/gate). We report two regimes over 5 SSL seeds (chance channel-acc $= 0.25$): the *clean* task, and an *amplitude-confounded* task with a stronger ($4\times$ vs $2\times$) distractor sinusoid injected at a different (channel, band).

attribution (channel-acc)	clean	amp.-confounded
global gate α (sample-agnostic)	0.25±.01	0.25±.01
component norm $\|z\|$	0.86±.12	0.31±.02
deployed $\alpha\|z\|$	0.86±.11	0.31±.02
model-free band energy (proxy)	*0.87±.03*	*0.50±.01*

Reading (a scoping negative we report plainly). Three facts. (i) All per-sample localization lives in the *input-dependent* $\|z\|$: the global gate α alone is at chance, and $\alpha\|z\| \approx \|z\|$ (the gate does not add localization, it selects which components enter the additive sum). (ii) On the clean task the model attribution does **not beat** the trivial energy proxy (0.86 vs 0.87) — so synthetic recovery alone, and Adebayo PASS, do *not* establish that the intrinsic score is more than a spectral-energy read; the reviewer's objection is partly correct and we say so. (iii) Under amplitude confounding both the model score and the proxy degrade, and the model degrades *further* (0.31 vs 0.50). We conclude that the intrinsic attribution is a faithful read of **model-internal spectral energy** — it tracks, and is upper-bounded by, input band energy — rather than a superior localizer. This bounds the faithfulness claim precisely and is consistent with the PTB-XL clinical negative (§A.5.9): intrinsic access and model-sensitivity are necessary but not sufficient for correct localization, which is the paper's central point.

Where the learned attribution *does* beat the proxy: a five-regime suite. The energy proxy is near-oracle only when the cue is a clean, on-grid narrowband oscillation. We probe the boundary with five regimes (n=5 seeds each), scoring channel-acc for the intrinsic $\|z\|$, the model-free band-energy proxy, and post-hoc IG:

regime	intrinsic	energy proxy	IG
clean (on-grid sinusoid)	0.86±.12	**0.87±.02**	0.00
amplitude-confounded	0.32±.02	**0.50±.01**	0.00
low-SNR sinusoid	0.69±.13	**0.82±.03**	0.30
broadband transient	**0.34±.05**	0.00	0.01
off-grid frequency	**0.45±.11**	0.00	0.00

The learned attribution beats the proxy in **2/5** regimes — exactly the two where the cue is *not* on-grid narrowband: a broadband biphasic transient and an off-grid frequency, where band energy collapses to **chance/below** (0.00) but the SSL encoder still localizes the channel (0.34, 0.45, both above chance 0.25). On the three narrowband-on-grid regimes the proxy matches or exceeds the intrinsic score. Against our engineering gate we had set at intrinsic $>$ energy in $\geq$3/5 regimes (stated in our plan, though the threshold and the result entered the repository together, so we do not claim it as pre-registration), this does **not** license a general "superior localizer" claim; we instead report the honest boundary — *the learned representation adds localization over raw spectral energy precisely for morphological and off-grid cues, and not for the clean narrowband cues where energy is already near-oracle*. IG is uninformative on the channel axis throughout (its per-channel marginal is swamped by input magnitude, §A.5.7).

A.5.3 Does the attribution depend on the input at all? (the control that should have come first)

Every axis below presupposes that the attribution is a function of *this* sample. On real data that presupposition often fails. Feeding a trained model a real recording, Gaussian noise, and an all-zero input and ranking the channels by attribution (PTB-XL, 40 records):

encoder	CV across records	ρ(real, noise)	distinct top-3 sets
PatchTST (headline)	0.47%	+0.965	3/40
PatchTST (cross-dom.)	0.33%	+0.965	4/40
TempCNN	0.30%	+0.993	**1/40**
state-space (S4D, Gu et al., 2022)	**15.4%**	+0.308	**36/40**

Three of the four deployed encoders produce what is effectively a per-channel constant: the same lead ranking for an ECG and for noise. Only the state-space encoder yields genuinely per-sample maps. This does not invalidate the synthetic results (there the attribution demonstrably tracks the injected cell), but it means **any real-data faithfulness verdict computed on the first three is a verdict about a constant**, and we re-report §A.5.9 accordingly. We recommend this control as a precondition for any intrinsic-attribution claim.

A.5.4 ADEBAYO CASCADING SANITY — A SPLIT VERDICT (SEEDED, MULTI-SEED, FULL ATTRIBUTION)

We evaluate against trained checkpoints: Sleep-EDFx PatchTST epoch_019 (L_jepa 0.90); PTB-XL PatchTST v2 the best-loss checkpoint (L_jepa 2.985, 20 ep × full train fold); and ETTh1 TempCNN. The attribution is the signal-conditioned per-(c, f, s) latent norm `compute_signal_channel_band(..., reduce_scale="none")`.

Following audit **F-A2**, two methodology fixes make the verdict rigorous and reproducible: (i) the cascade parameter-randomization is **seeded**, and we report the final +wavelet Spearman as **mean ± std over 10 seeds** with a pass fraction rather than a single draw; (ii) the cascade uses the **full (c, f, s)** attribution, not the scale-collapsed (c, f) heatmap. The second point is essential — collapsing the scale axis leaves only C×F = 20 cells for the 2-channel Sleep-EDFx montage, and a Spearman over 20 points carries $\approx 1/\sqrt{20} \approx 0.22$ sampling noise, which made single-seed verdicts straddle the 0.3 threshold. The full 60-cell (C×F×S) attribution drives that noise to $\approx 1/\sqrt{60} \approx 0.13$ and yields a stable verdict (the empirical std drops from $0.236 \rightarrow 0.119$ on the headline checkpoint).

Correction (post-audit). The checkpoints below are *not* the ones behind our downstream and trade-off numbers. Re-running the same 10-seed cascade on the five the clean-split Sleep-EDFx runs models that are, we obtain pass-fractions 0.7–0.8 and **0/5 pass** (mean final $\rho = +0.07$; single-shot also failing). PTB-XL passes 10/10. Reporting the passing Sleep-EDFx checkpoint as the paper's verdict would be checkpoint-shopping, so the Sleep-EDFx axis-2 verdict is **fail**. We also withdraw the claim that IG fails this test: on the tracked comparison IG decorrelates *better* than our own score ($\rho = -0.078$ vs $+0.232$, both under the 0.30 bar).

Final +wavelet (full model randomization) Spearman, mean ± std over 10 seeds:

checkpoint	+wavelet ρ (mean ± std, 10 seeds)	cascade pass	single-shot ρ
Sleep-EDFx headline (PatchTST 50 ep × 10 rec, ep19)	**+0.006 ± 0.119**	**10/10** ✓	**-0.147** ✓
PTB-XL headline (PatchTST v2, 20 ep × full fold)	**-0.008 ± 0.073**	**10/10** ✓	**-0.006** ✓
ETTh1 (TempCNN, smoke)	+0.030 ± 0.086	10/10 ✓	-0.040 ✓
Sleep-EDFx F1-scaled (50 rec × 30 ep, ep4)	+0.033 ± 0.136	10/10 ✓	-0.155 ✓
Sleep-EDFx F1-scaled (50 rec × 30 ep, ep29)	+0.090 ± 0.226	8/10 ‡	-0.293 ✓

Every checkpoint decorrelates to **mean ≈ 0** under full randomization, and every checkpoint passes 10/10 seeds **except** the converged F1-scaled checkpoint (8/10; the lone residual). The predictor-only randomization stage is +1.000 by construction (the attribution does not consume the predictor) — a measurement control, not evidence. These checkpoints are *not* the ones behind our reported numbers, and the paper's Adebayo verdict on the deployed models is **fail** $(0/5)$; the pass here is a property of a superseded run and we do not read it as evidence.

‡ The single residual (F1-scaled converged checkpoint, 8/10 cascade seeds, single-shot -0.293 near the 0.3 boundary) is the one checkpoint whose randomized-model attribution retains mild excess variance beyond the small-n floor.

† The "predictor only" column is a **measurement control**, not an evidence row: `compute_signal_channel_band` is defined as a function of `(wavelet → multiscale → context_encoder)` and does **not** consult the predictor parameters. Randomizing the pre-

dictor alone therefore yields Spearman = +1.0 by construction. The evidence-bearing rows are the +context_encoder and +wavelet rows where the function's actual inputs are randomized, and where the correlation drops to within the pass threshold.

A.5.5 WAVELET σ_T NARROWING — I.E. PASSBAND *widening*

On Sleep-EDFx after 50 epochs, the learnable Gabor bank narrows its time-domain envelope σ_t. Earlier drafts called this "bandwidth narrowing", which inverts the uncertainty relation: σ_f $\approx 1/(2\pi\sigma_t)$, so a *shorter* envelope is a *wider* passband. What the table below measures is σ_t shrinking, hence the filters becoming more temporally localized and less frequency-selective. We report it in those terms. (Paillard et al. (2025) report the opposite sign for GREEN on resting-state EEG: their α-band wavelet is the one with *minimal* frequency-domain standard deviation, hence maximal σ_t. We make no claim of matching them and state our own measurement.)

Values below are recomputed directly from the cited headline checkpoint (audit re-verification; an earlier draft's table reported systematically larger reductions that did not match this checkpoint):

Canonical band	init f_c	trained f_c	init σ_t	trained σ_t	$\Delta\sigma$_t
δ	2.0 Hz	2.07 Hz	477 ms	422 ms	-11.6%
θ-low	5.0 Hz	4.87 Hz	191 ms	165 ms	-13.6%
θ-high	7.0 Hz	6.84 Hz	136 ms	116 ms	-15.0%
α-low	9.0 Hz	8.81 Hz	106 ms	99 ms	-6.3%
α-high	**11.5 Hz**	**11.30 Hz**	**83 ms**	**68 ms**	**-18.4%**
β-low	16.0 Hz	15.85 Hz	60 ms	55 ms	-7.4%
γ	55.0 Hz	49.00 Hz	17 ms	18 ms	+4.1%

α-high (11.5 Hz) — the most prominent EEG rhythm during wake — shows the **largest** σ_t reduction (-18.4%) of any band on this checkpoint, i.e. the largest passband widening, obtained under SSL pretraining with no α-specific supervisory signal. (The reduction is consistent but more modest than the -32% an earlier draft reported; the headline number is now pinned to this checkpoint.) The next subsection shows that this σ_t drift is not always benign: on the deployed clean-split banks it runs the other way for the top filters, far enough that the convolution kernel no longer holds the envelope.

A.5.6 BIOMARKER RECALL@K VS (CHANNEL, BAND) REGISTRY

Using the AHA-derived `EEG_REGISTRY['sleep_edf']` (5 sleep stages $\times$ expected channels $\times$ expected bands), we compute Recall@k where k = |expected_cells|, with the **corrected channel-major (c, f, s) attribution ordering** (audit SB-3) on the leakage-free clean-split checkpoint (the clean-split Sleep-EDFx run). The learned Gabor bank on this run collapses **theta-dominant** — its 10 canonical band labels are $\theta\times6$, σ, β, $\gamma\times2$, with **no δ and no α band**:

Sleep stage	expected cells	recall@k	IoU	n_epochs	note
W (wake)	Fpz-Cz$\times\alpha$ + Pz-Oz$\times\alpha$ + … $\times\beta$ = 4	**0.000**	0.000	216	no α band learned
N1	Fpz-Cz$\times\theta$ = 1	**1.000**	1.000	200	
N2	Pz-Oz$\times\sigma$ = 1	0.000	0.000	202	σ present, not top-k
N3	Fpz-Cz$\times\delta$ + Pz-Oz$\times\delta$ = 2	—	—	200	δ band absent $\rightarrow$ undefined
REM	Fpz-Cz$\times\theta$ + Pz-Oz$\times\theta$ = 2	**1.000**	1.000	200	
macro (4 def. stages)	—	0.500	0.500	—	N3 excluded
macro (ALL 5 stages)	—	**0.400**	0.400	—	N3 absent-band scored 0

Read this result with care — it is not a clean improvement over the prior 0.200. The per-stage pattern *redistributed* rather than uniformly rose: the theta-expecting stages (N1, REM) now hit Recall@k = 1.000, while W — previously the only non-zero stage (α+β) — drops to 0.000 because this run's bank learned no α filter, and N3 becomes undefined because it learned no δ filter. The macro over the 4 stages with a defined target is 0.500, but it is computed over a *different, smaller* stage set than the earlier 5-stage 0.200, and it is contingent on which bands this particular SSL run happened to allocate. The honest takeaway is that **the band the encoder allocates is run-dependent and currently uncontrolled**: a theta-heavy bank rewards θ-defined stages and starves α/δ/σ-defined ones. This is direct evidence that latent additive decoupling alone is insufficient — the model needs an explicit **per-band prior** (canonical-band warm-start that is *retained*, or a dedicated

spindle/δ filter) to guarantee coverage of every clinically expected (channel, band) cell. We report the number without spin and treat band-allocation stability as open work (Phase 2 W30).

A.5.7 INTRINSIC α VERSUS POST-HOC INTEGRATED GRADIENTS

To answer the obvious reviewer concern "why not just I-JEPA + IG?", we run the same Adebayo cascading test using **post-hoc Integrated Gradients (Captum, 20 steps)** on the *same* trained PatchTST checkpoint, with the same input batch. The cascade reveals what single-shot full-randomization hides:

Cascade step	intrinsic α (ours)	Integrated Gradients (post-hoc)
baseline (trained vs trained)	+1.000	+1.000
after randomizing **predictor**	+1.000	+1.000
after randomizing **+ context_encoder**	**+0.165**	**+0.988**
after randomizing **+ wavelet** (full random)	+0.232	-0.078
final single-shot pass (abs(ρ)$\leq$0.3)	✓	✓

The cascade exposes the **model-dependence** signature:

- α **decorrelates as soon as the context_encoder is randomized** (drops

+1.000 $\rightarrow$ +0.165), confirming the attribution is driven by *what the model learned*.

- **IG stays at +0.988 correlation after context_encoder is randomized**

and only decorrelates once the wavelet frontend is randomized. The IG attribution is therefore dominated by **input frontend statistics**, not the encoder. This is the input-dominance mechanism Adebayo et al. (2018) identify — methods that approximate input $\times$ gradient largely return the input — observed here in the SSL time-series setting. We do not claim they recorded a cascade failure for IG; under their signed metric they report the opposite, and they explicitly decline to issue verdicts on individual methods.

Both methods nominally PASS the single-shot $|\rho| \leq 0.3$ threshold, but the cascading test correctly identifies α as faithful to the trained model and IG as faithful to the input frontend. The reviewer concern "post-hoc IG is equivalent" is answered by this experiment: **post-hoc IG attributes to the wavelet filter response, not to the trained encoder**.

A.5.8 ATTRIBUTION STABILITY UNDER ADDITIVE GAUSSIAN NOISE (R14)

We perturb a **held-out validation** batch with AWGN at multiple signal-to-noise ratios and measure the Spearman correlation between the attribution on clean and noisy inputs (5 trials per SNR, same PatchTST Sleep-EDFx epoch_019; `split=val` recorded in the released artifacts).

SNR (dB)	clean$\rightarrow$noisy Spearman	std
40	+1.0000	0.0000
30	+0.9991	0.0012
20	+0.9961	0.0015
10	+0.9239	0.0026
5	+0.9119	0.0034
0	+0.8845	0.0066
-5	+0.7741	0.0088

The MD-TJEPA α attribution maintains Spearman > 0.7 even at **SNR = -5 dB** (noise power 3$\times$ signal power) — we did not measure gradient or IG baselines under the same perturbation, so we make no comparative claim here. We attribute this to the **wavelet bandpass front-end**, which filters high-frequency noise before it reaches the encoder, and to **average pooling in PerComponentEncoder**, which integrates over channel time-slices.

A.5.9 REAL-DATA FAITHFULNESS: DOES THE ATTRIBUTION LOCALIZE CLINICAL MI? (AN HONEST NEGATIVE)

The synthetic task and Adebayo test what the attribution is faithful *to* the model. A separate question is whether that faithful attribution is *clinically meaningful*. We test this on **PTB-XL**: myocardial infarction (MI) localizes to known ECG leads (anterior→V1–V4, inferior→II/III/aVF, lateral→I/aVL/V5/V6). For each test record carrying a localizable MI code (n=549) we rank the 12 leads by MD-TJEPA attribution and measure Recall@k (k =#relevant leads) against a random-attribution control. **The attribution does *not* localize clinical MI** (*single checkpoint; see the correction below*): Recall@k = **0.269** vs random **0.404** (gap -0.135, permutation $p \approx 0$) — *below* chance. The cause is a **precordial-lead bias**: inferior-MI recall is ≈**0.00** (the attribution never favors the limb leads II/III/aVF), while lateral (0.48) and combined (0.59) MI score moderately. We report this plainly: MD-TJEPA's attribution is faithful to the model's learned *spectral-energy* decomposition, but that is **not** the same as clinical relevance — a caution against reading energy-based intrinsic attributions as diagnostic explanations, and the clearest limit on the method's scope (see Discussion).

Correction (post-audit): the clinical axis, re-measured on an instrument that works. The number above comes from a single checkpoint whose per-lead map we now know is essentially constant in the input (§A.5.3), so it cannot support a claim about clinical validity. Two re-measurements at n=5 settle it. On the input-*independent* encoders the gap versus random is -0.025 ± 0.144 — i.e. indistinguishable from random, exactly what a constant should give. On the state-space encoder, the one whose attribution varies per sample, MI lead localization scores Recall@k $\mathbf{0.300 \pm 0.014}$ against random 0.407, a gap of $\mathbf{-0.107 \pm 0.014}$ with all five seeds negative. So the clinical negative stands, but it now rests on a measurement that could have come out otherwise rather than on a map that returns the same ranking for an ECG and for noise.

A note on the label mapping, since a negative result is only as good as its ground truth. Our first pass mapped PTB-XL's MI subcodes to lead regions by hand and got three things wrong: `PMI` and `INJIL` were absent from the table (and so silently dropped), `IPMI` was mapped to inferior leads only, and a non-existent code `ANMI` was listed. Posterior infarction is in fact read *indirectly* from reciprocal changes in V1–V3, which our own biomarker registry records correctly but the evaluation script did not. We rebuilt the mapping against the SCP statement table (14 MI codes, posterior→V1–V3) and re-ran all five seeds. The gap moved from -0.109 ± 0.014 to -0.107 ± 0.014: the conclusion is unchanged, which is itself informative — the negative is not an artifact of one annotation error.

A.5.10 IS THE "FREQUENCY BAND" AXIS A FREQUENCY-BAND DECOMPOSITION?

Every heatmap in this paper labels its x-axis "frequency band", so that axis has to be one. The Gabor parameterization does not guarantee it: centers are free ($f = f_{\min} + \mathrm{softplus}(\cdot)$) and so may coincide, and the envelope width σ_t is unbounded *above* while the documented kernel-length rule bounds the kernel from $f_{\min}$ only. We audit the five deployed clean-split banks directly ; all intervals are t-based, n=5.

Diagnostic (10 nominal filters per bank)	value
distinct canonical bands realized	$\mathbf{3.8 \pm 0.6}$
center pairs within 1 Hz of each other	4.4 ± 0.7
filters whose envelope the kernel truncates	2.8 ± 0.6
stopband rejection lost to truncation	$\mathbf{14.2 \pm 1.0}$ dB
canonical bands absent in *every* seed	δ (0.5–4 Hz)

Three findings, in decreasing order of how much they matter.

The axis is roughly half as wide as it is labelled. Ten filters collapse onto 3.8 ± 0.6 distinct canonical bands, with 4.4 ± 0.7 pairs sitting within 1 Hz. A ten-column heatmap over such a bank has six columns that are near-copies, which inflates any per-cell ranking metric computed over it and makes "the model selected band j" a weaker statement than it reads as.

δ is never learned. No seed places a filter in 0.5–4 Hz. This is exactly the band the N3 sleep-staging biomarker cell requires, and it is why §A.5.6 reports N3 as unmeasurable rather than as a

score — the evaluation records the absence instead of silently scoring an empty expected mask, but the absence is a property of the model, not of the protocol.

The kernel is too short for the filters the model actually learns. The documented rule sizes the kernel from $f_{\min}$, but the binding constraint is σ_t. On these banks the three highest-frequency filters train to $\sigma_t \approx 1.5$ s — ultra-narrowband, $Q \approx 300$ — needing ≈ 770 samples of support against a configured 513, so their Gaussian is cut at $\pm 1.7\sigma$ with a 20–25% amplitude discontinuity at the kernel edge. Measured against the same filter on a kernel long enough to hold it, truncation costs 14.2 ± 1.0 dB of stopband rejection (-46 dB $\rightarrow -31$ dB) and widens the realized passband by up to 23%. The effect is real and bounded: 98% of the envelope energy is retained, so this degrades selectivity rather than invalidating the filters. The fix is one line — bound σ_t so that $5\sigma_t f_s \leq$ `kernel_size` — and we state the constraint here because the parameterization currently permits a configuration that silently does not implement the filter it describes.

Taken together this is a caveat on the *resolution* of the band axis, not a retraction of the channel-axis results, which is where §A.5.1 and Theorem 1 are scoped.

A.5.11 The attribution dilemma, and what widening the aggregate does fix

Everything above says our score behaves badly. This subsection asks whether a better score exists, and answers: not for this aggregate, for a reason that is forced rather than incidental.

A derived alternative. $\|z_u\|$ is a *chosen* scalarization of a quantity the aggregate does not forward. The additive structure offers a derived one instead: for any linear head w on the representation,

$$\phi_u(x) = \alpha_u(x)\,\langle w, z_u(x)\rangle, \qquad \sum_u \phi_u(x) = \langle w, s(x)\rangle, \tag{2}$$

so ϕ decomposes the model's own output exactly (measured completeness residual $< 5 \times 10^{-6}$). It is signed, direction-sensitive, and defined identically for any encoder — everything $\|z_u\|$ is not.

It fails, and it has to. On the synthetic ground-truth task (n=10), ϕ scatters rather than converging on both backbones — TempCNN 0.338 ± 0.085, PatchTST 0.310 ± 0.060 in the magnitude read-out (signed ϕ: 0.319, 0.246), against chance 0.25 — and its map is essentially input-independent ($\rho \approx +0.98$ against a noise map), while $\|z_u\|$ reaches 0.897 ± 0.081 on TempCNN. The reason is structural, not a tuning failure: w is fitted on *s for the channel*, and s barely carries it (channel decodable at 0.998 from the per-tube norms but 0.274 from s), so that w has little channel-discriminative direction for ϕ to inherit. Stated plainly:

> *A faithful attribution reports its head, not its model.* Decomposing a head fitted for an axis the aggregate barely retains recovers little; decomposing a head fitted for something else can name that axis anyway, and an input intervention shows the output depends on that channel's content though barely on its location (§G) — the sum hides the axis's identity, not its content's contribution. An earlier version of this paper wrote "faithful ⇒ uninformative" here. §F shows it is false: a band-detecting head's exact decomposition names the channel at 0.743.

The dilemma is created by summing over the channel axis, so we tested removing that step.

Block-additive aggregation. Replace $s = \sum_u \alpha_u z_u \in \mathbb{R}^k$ with a per-channel block sum, concatenated: $s = \left[\sum_{f,\tau} \alpha_{cf\tau} z_{cf\tau}\right]_{c=1}^{C} \in \mathbb{R}^{Ck}$. Every tube still contributes linearly to exactly one slot, so the model stays additive and (2) still holds with $w_{c(u)}$ in place of w; it simply declines to marginalize the channel. On the same frozen encoders (n=10):

backbone	aggregate	probe acc.	derived ϕ channel-acc.
TempCNN	flat sum	0.274 ± 0.010	0.340 ± 0.086
TempCNN	block	$\mathbf{0.986 \pm 0.020}$	0.581 ± 0.192
PatchTST	flat sum	0.256 ± 0.011	0.310 ± 0.060
PatchTST	block	$\mathbf{0.995 \pm 0.006}$	0.324 ± 0.116

Widening the aggregate lifts the probe everywhere and the attribution on one backbone. The readout goes from chance to essentially solved — 0.99 or better in every single seed, both backbones — confirming that the additive bottleneck's utility cost (§A.6) is caused by marginalizing the channel axis, not by the encoder. What the attribution does depends on which estimator is read, and we originally read the wrong one. The *magnitude* map $\sum_c |\phi_{uc}|$ is bimodal across the ten seeds (TempCNN: $0.974, 0.241, 0.706, 0.728, 0.268, 0.263, 0.836, 0.399, 0.794, 0.599$; mean 0.581 ± 0.192; the t-interval clears chance, lower bound 0.389, but the seeds split into two clusters) — and it is not a decomposition, with a completeness residual of 55 (§D). The signed ϕ, which is, reaches 0.904 ± 0.084 on TempCNN with no seed at or below chance. On PatchTST it does not follow: 0.509 ± 0.180, seeds 0.240 to 0.922, at probe 0.995.

That split is not identifiability. It is what is left in a component to be read: the per-component norms $\|z_u\|$ rank the channel at 0.897 ± 0.081 on TempCNN and only 0.329 ± 0.067 on PatchTST. The channel is not absent from PatchTST's components — a probe on the full z_u reads it at 1.000 and a fitted probe on the norm map at 0.787 (§A.5.9) — but it is not carried in a direction a single per-tube $\langle w, z_u \rangle$ with one shared w can read. So **a probe that clears the axis does not make the attribution follow** — it is necessary and not sufficient, and what supplies the rest is a property of the components, not of the head.

A.5.12 The dilemma off this architecture: design, and why the components are tied

Proposition 1 is argued for any additive latent read by a linear head, so it should not need our wavelet front end, our JEPA objective, or time series. We rebuilt its premise on tabular random features with a plain NAM: components are (group, member) pairs, one group is active per sample, and the label *is* the active group, so group identity is the contested axis and is recoverable from the components by construction. Two aggregates, identical otherwise — a flat sum over all components, and a block sum that concatenates per-group sums — and for each, a linear head fitted on the aggregate, the completeness-satisfying decomposition of its logit for the true class, and the argmax of the per-group ϕ mass against the active group. Ten seeds.

aggregate	probe on aggregate	derived ϕ	norm score
flat sum	0.249 ± 0.004	0.605 ± 0.318	1.000 ± 0.000
block	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

Why the shape functions are tied across groups, stated plainly because it looks like it was chosen to make the result come out. Component (g, m) uses shape function m in every group g, and the gate is likewise indexed by member. The first version of this experiment did not tie them, and its flat-aggregate probe read 0.58 against chance 0.25: with a different random function per group, the sum of the active group's outputs differs by which group was active, so the aggregate *retains* the group axis through the parameters. That is not a failed attempt at the dilemma. It is the regime Proposition 1's third remark already describes — an aggregate that happens to keep its contested axis, on which the proposition predicts neither horn — and so it tests nothing the proposition claims. Tying is what makes the premise hold rather than be assumed: the flat sum over an active group is then identical whichever group was active, so the axis is destroyed by the aggregation itself and not merely hard to read off it. The experiment instantiates the premise; it does not manufacture the conclusion, which is why the block aggregate, built from the same tied components, goes to 1.000 on every score. Which regime a deployed NAM sits in is an empirical question about that model, and check 1 of §8 is how one finds out.

What horn one looks like here. The flat-aggregate ϕ is not at chance; per seed it reads $0.0, 0.764, 0.523, 0.764, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0$. With w carrying no group-discriminative direction the active group's ϕ mass is a fixed scalar whose sign is set by an uninformative projection, so the argmax lands on the active group in most seeds or almost never (three sit between, at 0.764, 0.523, 0.764). The mean of 0.605 averages that split. An earlier version called this the explanation for the full model's 0.581 ± 0.192 under the block aggregate; it is not — that figure is a magnitude readout, a different estimator (§D). Here both cases lift together when the axis is retained.

A.5.13 DOES THE ATTRIBUTION RANK THE COMPONENTS THE MODEL USES?

This is the sharpest form of the question the rest of §A.5 circles. Our score is the per-tube latent magnitude $\|z_u\|$, computed with no reference to the gate. The model forwards $s = \sum_u \alpha_u z_u$, so a tube with $\alpha_u = 0$ contributes exactly nothing, whatever its norm. Those two quantities are free to disagree. We measured how much they do (n=5 deployed clean-split checkpoints, real Sleep-EDFx windows).

	value (n=5)
tubes with $\alpha > 0.01$, of 60	$\mathbf{7.0 \pm 0.0}$
attribution mass on gated-off tubes	$\mathbf{0.783 \pm 0.031}$
overlap between top-k by $\|z\|$ and by $\alpha\|z\|$	$\mathbf{0.000 \pm 0.000}$
channels routed to exactly zero	1.0 ± 0.0 (of 2)

The two rankings are *disjoint at the top* in every seed. Inspecting the cells shows why, and it is not subtle: the seven tubes the gate keeps are all at *scale 0* (channel Fpz–Cz, bands 0–6, α 0.05–0.83), while the seven highest-$\|z\|$ tubes are the *scale-1* copies of those same channel–band pairs, each with $\alpha \approx 1.3 \times 10^{-8}$. Separately, the gate assigns Pz–Oz exactly zero weight in all five seeds, while the attribution places 0.466 ± 0.050 of its mass there — on two of the five seeds, more than it gives Fpz–Cz, the channel the model actually reads. So the heatmap's channel axis distributes attribution across a channel the model does not read at all.

The score is cheap to repair — $\alpha_u\|z_u\|$ costs nothing to compute — but doing so does not rescue the interpretation, because §A.5.11's dilemma still applies to whatever is ranked.

An earlier version of this subsection went further and offered the disagreement as the *mechanism* behind the other negatives: the live tubes' norms are near-uniform (0.831–0.845, a $1.02\times$ spread), so, we argued, the map barely moves with the input and parameter randomization cannot perturb it. §A.5.14 tests that explanation by removing its premise, and it does not survive.

Why the block aggregate does not transfer to these checkpoints. §A.5.11 showed that a per-channel block aggregate, $s = [\sum_{f,\tau} \alpha_{cf\tau} z_{cf\tau}]_c$, dissolves the faithful-versus-informative dilemma on the synthetic task. On these Sleep-EDFx checkpoints it changes nothing at all: block and flat-sum readouts score *identical* macro-F1 to four decimals, because one channel's block is identically zero before the aggregation ever happens. That is a real result and we report it as such, but it is not evidence against block aggregation — it says the information is destroyed upstream, by gate collapse, so no choice of aggregation can recover it here.

Why the gate collapses. The cause is a missing counter-force rather than strong sparsity pressure, and it is worth stating because it makes the collapse a fixable configuration error rather than a property of additive JEPAs. The sparsity term $L_{\text{sparse}} = \sum_u \sigma(\ell_u)$ pushes every gate down by the same amount regardless of whether it helps. Nothing pushes back: s_{pred} and s_{true} are built from the *same* gate, so scaling α down scales the residual down with it and the JEPA term is happy to switch components off. Only the isotropy term (L_{iso}) preserves any scale at all, so a minimal subset survives to carry it and which subset survives is arbitrary. The entropy term that would prevent this (λ_{balance}, used at 15.0 for the conditional gate) defaults to 0 and was never applied to these global-gate runs. It plays the role the auxiliary load-balancing loss plays in sparsely-gated mixture-of-experts layers — theirs are not entropy terms (a squared coefficient of variation, and a dot product of dispatch and gate mass) but they are carried for the same reason (Shazeer et al., 2017; Fedus et al., 2022), and the failure it prevents there — a gate that routes to a few experts and leaves the rest untrained — is the one we measure. What is ours is not the cure but the cause: the JEPA objective shares the gate between prediction and target, so the penalty is self-satisfying and nothing resists it.

The trace is identical across all five seeds : the gate falls to a quarter of its components at epoch 6–7, the first channel dies at epoch 24–27, and the smallest logit reaches -18.2 in every seed — $\sigma(-18.2) \approx 10^{-8}$, so the penalty keeps pushing long after the component is off, which is what an unopposed gradient looks like.

Choosing the counter-force. Testing block aggregation on real data therefore requires retraining, and the retrain needs a value of λ_{balance}. Two criteria, not one: it has to stop the collapse, and it must not do so by flattening the gate into uniform usage, because a gate that weights all 60 components equally has stopped selecting and selection is the whole point. We measure both, per epoch, over the window in which the baseline collapses.

λ_{bal}	live	Gini	$\min_c$ share	converged	selective
0	**7**/60	0.863	**0.025**	**no**	yes
1	60/60	0.402	0.277	yes	yes
15	60/60	0.117	0.436	yes	**no**

The decisive column is *converged*. At $\lambda_{\text{bal}}=0$ the gate is still moving when the window ends — Gini climbing $0.195 \to 0.863$, smallest channel share falling $0.457 \to 0.025$, monotonically — so nothing about a longer run can be inferred from it except that it continues. At $\lambda_{\text{bal}}=1$ the trajectory reaches a *fixed point* by epoch 3 and is flat thereafter (Gini 0.402, channel share 0.277, all 60 components live), which is what licenses extrapolating it to a full-length run. At $\lambda_{\text{bal}}=15$ the gate is equally stable but has a Gini of 0.117: it no longer distinguishes components. Only $\lambda_{\text{bal}}=1$ satisfies every criterion, so that is what the retrain uses.

One number needs reading carefully. The collapsed baseline attains a *lower* JEPA loss at the same epoch (0.55 against 0.80), which looks like the fix costing accuracy. It is not: shrinking a shared gate lowers L_{jepa} by construction, which is precisely the pathology, so the collapsed model winning on that objective is the symptom rather than a defence of it.

A.5.14 Retraining without the collapse: what survives and what does not

Everything in §A.5.13 was measured on gates that had switched most of themselves off, so none of it separates a property of the method from a property of that failure. We retrained all five seeds with $\lambda_{\text{balance}}=1.0$ — same split, encoder, epochs and seeds, but also batch size $16 \to 32$ and all 137 train records instead of 50, which we did not intend — and re-asked each question. The comparison below is therefore of the whole retraining recipe, not of the balance term alone. The gate holds throughout: 60/60 components live in every seed, no channel dies, smallest logit -3.3 against -18.2, and the fixed point predicted from the 10-epoch sweep is still in place at epoch 29.

	deployed (collapsed)	retrained
live components, of 60	7.0 ± 0.0	$\mathbf{60.0 \pm 0.0}$
attribution mass on gated-off	0.783 ± 0.031	$\mathbf{0.000 \pm 0.000}$
overlap@k, $\|z\|$ vs $\alpha\|z\|$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
channels routed to zero	1.0 ± 0.0	$\mathbf{0.0 \pm 0.0}$
Adebayo, seeds passing	0/5	0/5
matched-protocol macro-F1	0.354 ± 0.007	0.370 ± 0.008
biomarker recall@k (5 stages)	0.400	0.400 ± 0.000
distinct canonical bands, of 10	4	3

Which checkpoints this paper reports. Both. The retrained models are the primary instrument from here on, because a configuration whose gate switches an entire input channel off is not the one a reader should take as the method; every retrain-sensitive number above is measured on them. But we keep the deployed-collapse column rather than replacing it, for two reasons. It is the configuration the earlier sections were computed under, so dropping it would leave those numbers unsourced; and the contrast is itself a result — the only way to know which of our negatives were about the collapse is to have both columns side by side. Where a row is identical in both, the retrain settles that the collapse was not responsible.

The disjointness was the collapse. With the gate open the two rankings *coincide exactly* in every seed. §A.5.13's headline is a property of a collapsed gate and not of intrinsic attribution, and we report it as such. What persists is narrower and worth keeping: the per-channel *mass* allocation still disagrees — the attribution splits $0.548/0.452$ where the model routes $0.830/0.170$. The order is

right; the weighting is not. A published score that does not match what the model routes through is the objection Jain & Wallace (2019) raised against reading attention weights as explanations, and Wiegreffe & Pinter (2019) the reply that such a mismatch indicts a particular use of the weights rather than the weights themselves. We take the same position on ϕ: the residual disagreement bounds what the per-channel mass may be used for and does not by itself make the ranking wrong.

Our mechanistic explanation was wrong. The claim withdrawn above said near-uniform norms are why parameter randomization fails. The retrained models are not near-uniform: $\|z_u\|$ spans $31.9\times$ with CV 72.8%, against $1.02\times$ and 0.7% on the collapsed gate. Adebayo is nonetheless **still 0/5**, with pass-fractions 0.6–0.8 against the deployed 0.7–0.8 — the same verdict. Removing the premise did not move the conclusion, so the premise was not the cause. We do not have a replacement explanation and do not offer one.

Block aggregation does not transfer to real data. This is the question the retrain existed to answer, since on the collapsed checkpoints the comparison was impossible (one channel's block was identically zero before aggregation). With the information now surviving to the aggregate, the per-channel block readout scores 0.389 ± 0.022 against the flat sum's 0.365 ± 0.017: $+0.024$, intervals overlapping, and still far below the width-matched PCA control at 0.4529 ± 0.0591. The synthetic $0.28 \rightarrow 0.99$ does not reproduce. So the one horn of the dilemma that looked cheap to remove is cheap to remove *only on a task whose label is the channel axis*; on sleep staging, where it is not, widening the aggregate buys almost nothing.

The band axis is a separate failure, and the retrain does not touch it. §A.5.10 reported that ten learned Gabor filters realize only 3.8 ± 0.6 distinct canonical bands with δ absent everywhere. That is not downstream of the gate, and the retrain confirms it: the retrained banks realize **3** distinct canonical bands in every seed against the deployed **4**, and leave **two** sleep stages without any filter in the band their biomarker cell needs, against one before. Biomarker recall@k over all five stages is 0.400 ± 0.000, exactly the deployed value. So the two degeneracies have different causes — the gate collapse is an unopposed sparsity penalty, the band collapse is unconstrained center frequencies — and fixing the first leaves the second where it was, marginally narrower. We flag this because a reader could otherwise take §A.5.14 as evidence that the retrain repairs the decomposition generally. It repairs the gate.

The retrained recipe does not cost utility. Matched-protocol macro-F1 is 0.370 ± 0.008 against the deployed 0.354 ± 0.007. Because batch size and data volume changed with the balance term, this bounds the recipe's cost, not the term's. Input sensitivity is unchanged, but that was never the relevant dataset: the constant-map finding is a PTB-XL result and Sleep-EDFx checkpoints are input-dependent either way, so this retrain says nothing about it.

A.5.15 Is the randomization verdict a per-sample one?

`compute_signal_channel_band` averages the per-tube norms over the batch before returning, and the cascade feeds it eight windows at a time, so the reported Adebayo verdict is computed on a *mean* map. That is a genuine mismatch with how this paper frames the attribution — "which components drove *this* prediction" — and we expected it to matter in a specific direction: averaging suppresses whatever varies between windows and leaves the part that does not, which is the part a randomization test is trying to perturb.

It does not. Running the identical cascade one window at a time, 16 windows per seed on the five retrained models , the final Spearman is $+0.324 \pm 0.048$ per sample against $+0.326 \pm 0.067$ for the batch mean, and the batch mean is the more correlated of the two in only 2 of 5 seeds. Forty per cent of individual windows fall below the 0.3 threshold, so the per-sample distribution straddles it, but its centre sits where the batch mean sits. The averaging is an imprecision in what the score *is*; it is not what decides the verdict, and we withdraw the expectation that it was.

A.6 An observed interpretability–performance trade-off

We set out to test whether conditional and morphological extensions could make the interpretable decomposition *competitive* with a strong black-box SSL baseline (TS2Vec, macro-F1 0.552 ± 0.012)

without sacrificing the faithful per-(channel,band) attribution that is the model's reason to exist. We instrument both axes on the same leakage-free split: **downstream competitiveness** (Sleep-EDFx full-test macro-F1) and **attribution faithfulness** (synthetic ground-truth channel-acc, §A.5.1). Across the five architectural variants we tested under a fixed budget, we **observe a Pareto trade-off**: no variant maximized both, and every variant that improved macro-F1 reduced channel attribution. **Both axes are multi-seed** (macro-F1: matched-probe n=4–5; channel-acc: synthetic GT n=10), so we present this as a fully-seeded empirical trend on this benchmark family, not a proven conservation law.

variant (interpretable?)	macro-F1	synthetic channel-acc
base (c, f, s), global gate ✓	0.354±0.007	**0.897±0.081**
+ conditional identifiable gate (CID) ✓	0.341±0.057	0.558±0.090
+ raw residual (black box) ×	0.543±0.034	0.427±0.098
+ shapelet residual (interpretable) ✓	0.563±0.014	0.346±0.073
+ **gated** shapelet components ✓	0.468±0.086	**0.803±0.108**
TS2Vec (black box, ref)	*0.552±0.012*	*n/a*

Table 3: The observed interpretability–performance trade-off, **both axes multi-seed** mean±95% t-interval. macro-F1: matched balanced 12k-window probe, n=4 per variant (n=5 base, TS2Vec); channel-acc: synthetic GT, n=10. The competitive end (shapelet residual, macro-F1 **0.563**, matching TS2Vec's 0.552) is attribution-broken (0.346); the faithful end (base/CID, 0.897) is downstream-weak ($\leq$0.36); the **gated** variant is an *interior* point recovering attribution to 0.803 at macro-F1 0.468. **No variant's point estimate reaches the top-right (faithful *and* TS2Vec-competitive); the closest one's intervals cross both bars.**

The trade-off is real and architecture-shaped. The base decomposition is **faithful but weak** (channel-acc 0.897 ± 0.081, macro-F1 0.354 ± 0.007, matched probe); a *residual* augmentation that recovers minority-class accuracy — the morphology (slow-wave / sawtooth / spindle) that the spectral front end averages away — drives the (c, f)-tube channel attribution toward chance (0.346 ± 0.073, n=10: a *robust* collapse, not a single-seed fluke). Crucially, this is *not* a black-box artifact: the **interpretable shapelet residual** actually **matches TS2Vec downstream** (0.563 ± 0.014 vs 0.552 ± 0.012, same fair probe) with the same attribution collapse, and the leak audit shows the morphology branch, not the (c, f) tubes, carries the signal. So MD-TJEPA *can* reach black-box-competitive macro-F1 — just not while keeping faithful attribution.

Routing morphology through the gate recovers most of the faithfulness — at a modest accuracy cost. Making morphology a per-(channel,prototype) **gated component** (so attribution is per-component by construction) lifts channel-acc from 0.346 to **0.803±0.108** (n=10, values 0.61–1.00) — *substantially* more robust than a single-seed estimate suggested (an earlier two-run probe saw 0.53/0.86). This costs macro-F1 $0.563 \rightarrow$ **0.468±0.086**: the gated variant is a genuine *interior* Pareto point — it buys back most of the attribution ($0.31 \rightarrow 0.82$) for a $\sim$0.1 drop in macro-F1, the closest any variant comes to the joint frontier without reaching it ($0.82 < 0.85$; $0.468 < 0.55$). A fairness check rules out a measurement artifact: the shapelet *branch* itself carries essentially **zero** per-channel information (0.001±0.001, n=5), and on a dedicated synthetic **morphology** cue (broadband biphasic transient) the shapelet-component channel attribution is **0.128, below chance** — the morphology axis provides no faithful per-channel attribution on its own.

Why the additive design costs downstream utility (a decodability probe). The faithful-but-weak corner is *mechanistic*, not incidental. A linear channel probe recovers the injected channel at **0.998±0.002** (n=10) from the *per-tube* norm vector $\|z_{c,f,s}\|$ — the very representation the attribution ranks — but only **near chance** (0.274 ± 0.010, chance 0.25) from the deployed *additive aggregate* $s = \sum_{c,f,s} \alpha\, z_{c,f,s}$ (n=10). Summing the components collapses the per-channel structure the attribution exposes, so a head on s can barely tell *which* channel was active — though the channel's content still reaches it (§G). Exposing a per-component attribution and pooling it into a single additive vector are therefore in tension *by construction* — which is exactly why the base/CID variants sit at high attribution but low macro-F1.

Same weights, real data: the shared sum is lossy (a matched readout control). This synthetic decodability gap has a real-data downstream analogue that isolates the additive *aggregation* from the encoder with **no retraining and no added parameters**. On the five base Sleep-EDFx checkpoints

we read out the *identical* frozen per-tube latents three ways and probe each with the same matched protocol : the deployed additive sum s scores macro-F1 0.353 ± 0.020 (reproducing the base), a **dimensionality-matched** projection (PCA of the concatenated latents back to the same 32 dims) scores $\mathbf{0.393 \pm 0.062}$ ($+0.040$, higher in **5/5** seeds), and the full concatenation (1920 dims) scores 0.491 ± 0.043. Because the weights are identical and the PCA control matches dimensionality, the gap is attributable to the shared *sum* discarding component information, not to encoder capacity — the same-weights causal isolation the trade-off's parameter-confounded variants could not provide. A linear head on the concatenation, or behind any fixed projection $P = [P_1, \ldots, P_U]$, is itself additive over components ($w^\top P \,\mathrm{concat}(z) = \sum_u w^\top P_u z_u$), so this measures the cost of the *shared direction* $v_u = \alpha_u w$ that the sum imposes, not of additivity. (Even the full readout stays below TS2Vec's 0.552.)

Statement (empirical, this benchmark family). *Among the five MD-TJEPA configurations we tested under a fixed budget — now with **both axes multi-seed** (macro-F1 n=4–5 matched probe, channel-acc n=10 synthetic GT) — none* in point estimate *attained both faithful intrinsic per-(channel,band) attribution (≥0.85, a bar of our own choosing) and TS2Vec-competitive downstream macro-F1 (*0.552 ± 0.012*) on Sleep-EDFx: the competitive variant (shapelet residual,* 0.563 ± 0.014*) collapses to channel-acc* 0.346 ± 0.073*, while the most faithful downstream operating point (gated, channel-acc* 0.803 ± 0.108*) reaches only* 0.468 ± 0.086*.* The trade-off *softens*, and how far is a question this n cannot answer: the gated variant's attribution interval $[0.695, 0.911]$ contains the 0.85 bar and its utility interval $[0.382, 0.554]$ reaches TS2Vec's 0.552, so both comparisons that decide the statement are unseparated. The 0.85 bar is our own choice and we do not defend it as principled — it is the level at which we would have called an attribution usable. Read the statement as one about point estimates under a fixed budget, not as a demonstration that the corner is unreachable. It replaces our initial "interpretability without the tax" hypothesis (not supported) with a quantified, fully multi-seed trade-off measured by a purpose-built synthetic GT instrument. We state it as an empirical finding on this benchmark family, not a universal impossibility theorem. **Robustness.** Because the two axes are measured in separate experiments, we bootstrap each independently over its seeds (50k resamples; each variant's per-seed macro-F1 and channel-acc resampled with replacement) and check the corner in every resample: the top-right stays **empty in 100% of resamples**. The two failure modes are disjoint — base/CID clear the faithful bar in 63% of resamples but the competitive bar in 0%, whereas the shapelet residual clears the competitive bar in 100% but the faithful bar in 0% — so no variant is ever both, past the $\geq$90% criterion we had set. Note the bootstrap has limited power here: no variant's resampled mean can reach the opposite threshold, so the 100% is close to arithmetically forced and should be read as a consistency check, not as evidence of robustness. **Cross-domain replication (utility side, PTB-XL, n=5).** To check that the additive-bottleneck utility cost is not specific to Sleep-EDFx, we retrain the two endpoints on **PTB-XL** (12-lead ECG, multi-label; **n=5** seeds, same 30-epoch budget, matched linear probe). The direction reproduces cleanly: the base additive configuration reaches macro-AUC $\mathbf{0.609 \pm 0.031}$ against the shapelet residual's $\mathbf{0.736 \pm 0.002}$, and *every* one of the five base seeds falls below *every* residual seed (ranges $[0.584, 0.665]$ vs $[0.734, 0.740]$, non-overlapping). As on Sleep-EDFx the additive decomposition is downstream-weaker while the residual is competitive (a reference SSL probe on this split scores ≈ 0.72). This is an n=5 replication on the *utility* axis, so the trade-off's performance cost is not a single-dataset artifact.

The attribution axis does *not* replicate measurably (reported, not omitted). We also attempted the other axis on the *same* frozen checkpoints, scoring the PTB-XL MI lead-localization proxy for both endpoints (n=5): base additive 0.377 ± 0.102 vs shapelet residual 0.347 ± 0.093, i.e. a paired difference of -0.030 ± 0.087 with only **3/5** seeds in the Sleep-EDFx direction. The point estimate carries the expected sign, but its CI comfortably includes zero and the sign split is near chance; seed variance dominates and both variants straddle the random control (0.407). We therefore record this as **inconclusive** — this proxy can neither confirm nor refute the attribution side of the trade-off cross-domain — and keep the cross-domain claim scoped to *utility*. We report the attempt because the alternative (silently leaving the axis unmeasured) would overstate how well the scope limitation is understood.

Backbone-family generalization: the utility axis transfers, the attribution axis does not. Every real-data number above uses the PatchTST per-component encoder, so we retrained both endpoints on PTB-XL under two further architecture families at the same budget and matched probe (n=5 each): the **TempCNN** (convolutional) block and a **diagonal state-space** block (S4D-style damped

modes applied by FFT convolution). The *utility* ordering reproduces in **all three**, every time with **5/5** seeds and non-overlapping ranges:

backbone	base additive	shapelet residual	Δ
TempCNN (conv.)	0.646±.017	0.736±.013	+0.090
PatchTST (transformer)	0.609±.031	0.736±.002	+0.127
state-space (S4D)	0.607±.009	0.708±.007	+0.101

so the additive bottleneck's performance cost is a property of the decomposition, not of one architecture family.

A.6.1 A BACKBONE-MATCHED INSTRUMENT: IS THE AXIS PER-COMPONENT AT ALL?

The readout control establishes that the components carry the channel on both backbones while the published magnitude score does not find it on PatchTST. That leaves a real ambiguity, because a probe reads every component at once: "the axis is in the components" does not imply "some per-component score can rank by it". If it lived only in cross-component structure, no attribution could reach it and the corner would be unreachable rather than unmeasured.

Four per-component, per-sample scores on identical frozen components, summed over (band, scale) and argmaxed against the injected channel (n=5, chance 0.25):

backbone	magnitude $\|z_u\|$	centered	gate-weighted $\alpha_u\|z_u\|$	fitted per component
TempCNN	0.858 ± 0.165	0.999 ± 0.003	0.859 ± 0.158	$\mathbf{1.000 \pm 0.000}$
PatchTST	0.335 ± 0.163	0.867 ± 0.278	0.335 ± 0.163	$\mathbf{1.000 \pm 0.000}$

The fitted score recovers the channel perfectly on both backbones, so the axis is per-component recoverable and the corner is reachable on PatchTST. It answers the structural question the probe could not.

It does not weaken the dilemma, and it is worth being precise about why, because a score that reads 1.000 invites the opposite reading. That score is fitted *on the channel label*. It is therefore the case §4's second remark already describes: one may always choose a direction that ranks the contested axis well, and the resulting attribution is both informative and completeness-satisfying with respect to *that* direction — but the head it decomposes is one nobody would deploy, because it was built by being told the answer. Nothing here is faithful to the head that produces the model's output. What the experiment establishes is narrower and structural: the failure of our published score is the score's and not the encoder's, so the corner is unmeasured rather than unreachable. Whether a score that is faithful to a *deployed* head can occupy it is the question Proposition 1 answers for the head fitted for the axis when the aggregate barely retains it, as it does here; for a head fitted for the task the answer is partial (§G).

The mechanism is mundane. Each PatchTST component carries a large constant offset that swamps the sample-dependent part of its norm; subtracting the per-component training-set mean — no labels, no fitting — lifts the published argmax from 0.335 to 0.867, and improves TempCNN as well ($0.858 \to 0.999$), so it is a better readout rather than a transformer-specific patch.

We stop short of recommending it. The five PatchTST seeds read $1.0, 1.0, 0.851, 1.0, 0.482$: centering works on most runs and fails on one, and a mean over that spread is exactly the kind of number this paper spends its length warning against. What we claim is narrower and holds: the failure is the instrument's, the axis is per-component recoverable, and a label-free instrument that finds it *reliably* is not something we have.

The *attribution* axis behaves differently. On the synthetic ground-truth task the two encoders disagree sharply: TempCNN recovers the injected channel at 0.858 ± 0.165 while **PatchTST sits at chance** (0.335 ± 0.163 against chance 0.25, its CI covering chance). This is *not* transformer cold-start: a **4× training budget** (120 epochs) does not close the gap (0.262 ± 0.064, with no improvement in L_{jepa}), and the encoder is not degenerate at these sequence lengths (≈15 and ≈4 patches per tube).

What this implies for the trade-off figure (disclosed, not hidden). Our utility numbers come from PatchTST and our synthetic attribution numbers from TempCNN, so each point in the trade-off scatter pairs measurements taken from *two* architectures. We state this rather than presenting the pair as one model, because it carries a real scoping cost: for the deployed PatchTST models we have parameter-sensitivity evidence (§A.5.4) but *no* ground-truth-recovery evidence — on that backbone the instrument reads chance. Strictly read, the "faithful" corner of the trade-off is a **TempCNN property**. We regard this as a further instance of the paper's central claim: validity evidence does not transfer across axes, and — we now add — it does not automatically transfer across architectures either.

Diagnosis: the score function is part of the problem, but not all of it. Rather than leave "backbone-dependent" as a black box, we tested a mechanical explanation (n=5). Our score is the per-tube latent norm $\|z\|$, and the PatchTST block ends in a **LayerNorm**, which renormalizes every tube's latent before we read it; TempCNN normalizes only internally (BatchNorm, per feature across the batch) and leaves its output magnitude free. The measurement matches: the coefficient of variation of $\|z\|$ across tubes is 0.41 ± 0.05 for TempCNN but only 0.19 ± 0.04 for PatchTST — the signal the score reads is compressed $\approx 2.2\times$. We also tried a *direction*-sensitive score (cosine distance from the per-sample tube centroid, which LayerNorm cannot flatten). We previously wrote that it "does recover information the norm misses on PatchTST: 0.332 ± 0.081, now above chance." Both halves of that were wrong. The interval was a *normal*-approximation half-width whose lower bound cleared 0.25 by 0.001; under the Student-t convention this paper declares it is 0.332 ± 0.115, lower bound 0.217, and **no score is above chance on PatchTST**. And it recovers nothing the norm misses: the norm reads 0.335 ± 0.163 on the same runs, so the two are indistinguishable. What the direction score does do is *break* the backbone it was not designed for, scoring 0.144 ± 0.062 (below chance) on TempCNN. So the failure is neither pure architectural blindness nor a pure scoring artifact — and **no single score function is valid across both backbones**. The practical lesson for intrinsic-attribution work is that the score is not architecture-portable and must be re-validated whenever the encoder changes — which is precisely the kind of assumption our four-axis protocol is designed to expose.

By-product — an interpretable morphology primitive. The learnable shapelet bank is itself a contribution: trained unsupervised, its prototypes self-organize into recognizable sleep morphology — a 13.9 Hz spindle, slow-waves (<1 Hz), sawtooth/delta (1–4 Hz) — recovering exactly the N3/REM morphology the spectral axis misses. The trade-off is on *channel attribution*, not on the morphology primitive's own interpretability.

A.7 What this evaluation does not cover

Stating the boundary explicitly, since several of the checks above would have natural extensions we did not run. On the *faithfulness* side we do not report IoU against TUEV ground-truth event windows (the corpus requires credentialed access), per-class faithfulness, a ROAR-style retraining check (Hooker et al., 2019), or cross-seed agreement between explanations. On the *downstream* side we report linear probes only — no end-to-end fine-tune, PTB-XL macro-AUC only for the shapelet trade-off (§A.6), not as a benchmark, and class-balanced reweighting only in the matched probe. A head-to-head against CodeBrain (Ma et al., 2026) is absent because no public checkpoint was available at submission time; Appendix S gives an architectural comparison instead. None of these gaps is load-bearing for the claims we do make, but each bounds how far those claims generalize.

B Formal proof of Theorem 1 (identifiability)

> **Revision note**: the wavelet-bank inputs $\{x_{c,f,s}\}$ are **correlated** by construction (different bandpass views of the same raw signal), so the classical Hoeffding–Sobol decomposition, which assumes independent inputs, does not apply. The *generalized* functional ANOVA of Hooker (2007) is built for exactly this case — it replaces independence with a weighted orthogonality condition — and Chastaing et al. (2012) establish existence and uniqueness of that decomposition under a boundedness condition on the joint law. The proof below is the first-order specialization of that uniqueness result to the (channel, band, scale) index set. (Supersession: an earlier draft attributed the step to a numbered lemma of Lengerich

et al. (2020) and described Hooker (2007) as requiring independent inputs. Both were wrong — the numbered results of Lengerich et al. (2020) concern the convergence of their purification algorithm, and their uniqueness argument rests on Hooker (2007) and Chastaing et al. (2012), as ours now does. The mathematical content is unchanged apart from (A5), which the earlier version left implicit.)

Setup. Let $\mathcal{U} = \{u_{c,f,s} : c \in [C], f \in [F], s \in [S]\}$ be the index set of feature tubes. The MD-TJEPA encoder admits the decomposition

$$f_\theta(x) = \sum_{u \in \mathcal{U}} \alpha_u \cdot h_u(x_u) \in \mathbb{R}^k,$$

where each $h_u : \mathcal{X}_u \to \mathbb{R}^k$ depends only on x_u (NAM modularity, by architecture).

We assume:

- (A1) NAM additive architecture (above).
- (A2) Each h_u is C^2 on its compact support (smoothness).
- (A3) For every u, $\mathrm{Var}(h_u(X_u)) > \epsilon > 0$ (non-degeneracy).
- (A4") **Centering constraint**: $\mathbb{E}_{X_u}[h_u(X_u)] = 0$ for every u.
- (A5) **Joint non-degeneracy**: the law of $(X_u)_{u \in \mathcal{U}}$ admits no non-trivial additive relation — if $\sum_u \delta_u(X_u) = 0$ almost surely with every δ_u centered, then every δ_u is zero almost surely. This is implied by conditions (C.1)–(C.2) of Chastaing et al. (2012) through their Lemma 1, which gives $\mathbb{E}[(\sum_u h_u)^2] \geq \delta^{|\mathcal{U}|-1} \sum_u \mathbb{E}[h_u^2]$ with $\delta > 0$. It is implied by, and not equivalent to, that pair: they prove no converse, and we do not.

(A4") is a per-component condition (sum-to-zero per axis) and does NOT require the X_u to be statistically independent; (A5) is what independence would otherwise have supplied, and it is strictly weaker — it tolerates arbitrary correlation and excludes only exact additive redundancy, such as a duplicated tube. In practice we enforce (A4") via the soft penalty $L_{\text{anova}} = \sum_u \|\mathbb{E}[h_u]\|^2 + \sum_{u \neq v} \mathrm{Cov}(h_u, h_v)^2$ whose first term is the binding identifiability constraint and whose second term is a (correlation-tolerant) decorrelation regularizer.

And we have now measured it, which we had not. Earlier versions bounded this residual by λ_{anova} and stopped, which is a statement about a penalty weight and not about the condition the penalty is meant to produce. Measured on the five deployed checkpoints over the evaluation split, at the shipped λ_{anova}=0.01, the centering residual $\|\mathbb{E}[h_u]\|$ relative to each component's own RMS magnitude is $\mathbf{0.631 \pm 0.036}$, with a per-seed maximum of 0.981 ± 0.014 — at least one component per seed is very nearly constant. **(A4″) does not hold on the models we report.** Theorem 1 is therefore a statement about a parameterization we do not attain, not a property we have demonstrated of these checkpoints, and every use of it in this paper should be read that way. The penalty weight that would attain it is unrun. We state this rather than leaving the hypothesis asserted, because a soft penalty at 0.01 producing a 0.63 residual is the kind of gap a reader is entitled to be told about.

What the failure does not reach. (A4″) is a hypothesis of Theorem 1 only. Proposition 1 and Corollary 1 — the paper's claim — assume completeness, local-linearity and the coefficient non-degeneracy (P3) of Appendix C, and nothing about centering: their uniqueness argument (Appendix C) never takes an expectation. So the measurement above hollows out Theorem 1's guarantee that the *learned* components are identifiable; it leaves untouched the statement that, whatever the components are, a complete local-linear attribution of a head is unique and reports that head. The two-head experiment (§F) does not use Theorem 1 anywhere.

Theorem 1 (Identifiability of the components; factorization up to per-component scale). *Under (A1)–(A3), the centering constraint (A4") and the joint non-degeneracy (A5), if two parametrizations $\{(\alpha_u, h_u)\}_u$ and $\{(\alpha'_u, h'_u)\}_u$ produce the same encoder $f_\theta(x) = f_{\theta'}(x)$ on $\mathcal{D}$-almost every x, then for every u the components agree, $\alpha'_u h'_u = \alpha_u h_u$ on $\mathcal{X}_u$, and there are scalars $c_u \neq 0$ with $\alpha'_u = c_u \alpha_u$ and $h'_u = c_u^{-1} h_u$.* There is no permutation clause and no joint constraint on the c_u: each component is tied to a physical input tube x_u, so moving a function to another index changes the function, and each factorization rescales independently. (An earlier version stated "identifiability up to permutation + scale" with $\prod_u c_u = 1$; neither clause is what the proof establishes.)

Proof.

Step 1 — Centered decomposition uniqueness. Let $g(x) = \sum_u g_u(x_u)$ where $g_u := \alpha_u h_u$. Suppose there is a second decomposition $g(x) = \sum_u g'_u(x_u)$. Their difference $\sum_u (g_u - g'_u)(x_u) = 0$ for $\mathcal{D}$-a.e. x. Define $\delta_u := g_u - g'_u$. By (A4"), each δ_u is centered: $\mathbb{E}_{X_u}[\delta_u(X_u)] = 0$.

For correlated X_u the classical independence-based projection is unavailable, and the step is carried instead by (A5): a sum of centered additive components that vanishes almost surely, each depending only on its own tube, has every component zero almost surely. This is the first-order case of the generalized functional-ANOVA uniqueness of Chastaing et al. (2012); note that it does not follow from the per-component non-degeneracy (A3), which is why (A5) is stated separately. Hence $g_u(x_u) = g'_u(x_u)$ on $\mathcal{D}$-a.e. support of X_u.

Step 2 — Extension to the full domain. By smoothness (A2), the g_u's are continuous, so a.e. equality on the support extends to equality on the closure $\mathcal{X}_u$.

Step 3 — Separation of α_u and h_u. The product $g_u = \alpha_u h_u$ is determined up to a scalar factor. Choose a reference point $x_u^\star$ where $h_u(x_u^\star) \neq 0$ (exists by A3). Set $c_u := \alpha'_u/\alpha_u$; then $h'_u = c_u^{-1} h_u$ on $\mathcal{X}_u$.

Steps 1–3 give exactly the statement: g_u is unique, and the only remaining freedom is the per-component rescaling $(\alpha_u, h_u) \mapsto (c_u \alpha_u, c_u^{-1} h_u)$.

□

Remarks.

(a) The proof works for **correlated** input tubes $\{X_u\}$, the actual setting of MD-TJEPA where band-pass + multiscale views share underlying samples. The binding requirement is centering (A4"), not independence.

(b) The cross-component decorrelation term in L_{anova} ($\sum_{u \neq v} \text{Cov}(h_u, h_v)^2$) is a **regularizer**, not an identifiability requirement. Even when inputs are correlated, driving output covariances toward zero is a form of disentanglement; it tightens the basin around the identifiability point but the proof above holds even without it.

(c) The soft penalty does not enforce (A4") exactly, and on our checkpoints it does not come close: the measured relative residual is 0.631 ± 0.036 (above). The theorem is therefore a statement about a parameterization these models do not attain.

(d) (A5) is assumed, not derived. That the channel tubes are distinct signals makes it plausible on the channel axis and says nothing on the band and scale axes, where every tube is a deterministic function of one channel signal; neither observation is a proof.

(d) The theorem makes **no claim about non-additive interaction terms** $h_{uv}(x_u, x_v)$. MD-TJEPA's predictor is the only place where cross-tube interactions could re-enter. Because the block-diagonal Koopman predictor (§3.1) has no cross-block coupling, the identifiability is preserved end-to-end.

(e) The result is the time-series analog of the generalized functional-ANOVA uniqueness of Chastaing et al. (2012), which Lengerich et al. (2020) use to license interpreting purified tabular effects. Our adaptation specializes the "feature" abstraction to the (channel, band, scale) tuple and keeps only the first-order terms, the additive architecture having no interaction components to purify.

C Proof of Proposition 1

An earlier version of this paper stated Proposition 1 with "linear in the components and satisfies completeness" and gave no proof. That statement is false as written: completeness and linearity alone do not pin the decomposition. Fix any $d \in \mathbb{R}^k$ and set $\phi'_1 = \phi_1 + \langle d, \alpha_1 z_1 \rangle$, $\phi'_2 = \phi_2 - \langle d, \alpha_1 z_1 \rangle$, leaving the rest. Each ϕ'_u is linear in the components and $\sum_u \phi'_u = h$ still, yet $\phi' \neq \phi$. The missing axiom is **locality** — component u's score may read component u only — and with it the proposition is true and the proof is two lines. We state it properly here. The slack we had left is the one the axiomatic-attribution literature keeps finding in arguments of this shape: Shapley (1953) is the template — fix axioms, derive a unique value — and Sundararajan & Najmi (2020) show the

uniqueness does not survive the move to model explanation, because the operationalisations that claim it differ in what they hold fixed. A missing axiom is the same failure in miniature.

Axioms. An attribution $\{\phi_u\}$ of $h(x) = \langle w, s(x)\rangle$ over $s(x) = \sum_u \alpha_u(x) z_u(x)$ is:

- (P1) **Complete**: $\sum_u \phi_u(x) = h(x)$ for $\mathcal{D}$-a.e. x.
- (P2) **Local-linear**: $\phi_u(x) = \langle v_u, \alpha_u(x) z_u(x)\rangle$ for some fixed $v_u \in \mathbb{R}^k$. Locality is the requirement that ϕ_u depend on (α_u, z_u) alone; linearity is that it do so through a fixed functional.
- (P3) **Non-degenerate**: the components admit no non-trivial additive relation — if $\sum_u \langle d_u, \alpha_u(x) z_u(x)\rangle = 0$ for $\mathcal{D}$-a.e. x, then $d_u = 0$ for every u. This is *not* (A5) of Appendix B, and neither implies the other in general: (A5) constrains centered *functions* of the inputs, (P3) the linear coefficients acting on the latents. (A5) without (P3): let X_1, X_2 be independent centered scalars, $\alpha \equiv 1$, $z_1 = (X_1, 1)$, $z_2 = (X_2, 1)$, $w = (1, 0)$; then $\phi_1 = X_1 + t$, $\phi_2 = X_2 - t$ is a complete local-linear attribution for every t ($v_1 = (1, t)$, $v_2 = (1, -t)$), a per-component reference offset. (P3) can also fail with no constant direction at all ($z_1 = z_2 = X$, $w = 1$ gives $\phi_1 = (1 + t)X$, $\phi_2 = (1 - t)X$), and hold where (A5) fails ($z_1 = X$, $z_2 = X^2$ on one input: the coefficients are independent, but $g_1(X) = X$, $g_2(X) = -X$ is a vanishing function-level relation). (P3) is a sufficient condition for uniqueness of the *coefficients* v_u; two coefficient choices can also give the same score functions when a component's latent has a null direction, so coefficient uniqueness and attribution uniqueness are distinct.

Claim. Under (P1)–(P3) the attribution is unique and $\phi_u = \alpha_u \langle w, z_u\rangle$.

Proof. Substituting (P2) into (P1),

$$\sum_u \langle v_u, \alpha_u(x) z_u(x)\rangle \;=\; \Big\langle w, \sum_u \alpha_u(x) z_u(x)\Big\rangle \;=\; \sum_u \langle w, \alpha_u(x) z_u(x)\rangle$$

for $\mathcal{D}$-a.e. x, hence $\sum_u \langle v_u - w, \alpha_u(x) z_u(x)\rangle = 0$ a.e. By (P3) with $d_u = v_u - w$, every $v_u = w$. That $\phi_u = \alpha_u \langle w, z_u\rangle$ satisfies (P1) and (P2) is immediate. □

Scope. (P3) is an assumption about the latents, and our encoders come close to violating it in the first counterexample's way: their per-tube latents carry a large input-independent offset (the norm map is nearly constant in the input on real ECG, §A.5.9). Where that offset is exact, the complete local-linear attribution is unique only up to redistributing constants between components, and $\phi_u = \alpha_u \langle w, z_u\rangle$ is the member that uses the latent origin as its reference. §G measures what that choice of reference costs, by comparing ϕ against an input intervention on the same head. The proposition does not use Theorem 1's centering (A4″).

What the consequent does and does not assert. The second sentence of the proposition is weaker than an impossibility claim, deliberately. If A is not linearly decodable from s then a w fitted on s for A is at chance, so ϕ is determined by a direction that was not chosen with A in mind. This does *not* entail that $\{\phi_u\}$ is uninformative about A: ϕ_u reads z_u, not s, and the individual components do carry A. What it entails is that nothing in the construction *ties* the ranking to A — the informativeness is an accident of alignment, and an accident is exactly what the transplant measures when ϕ reads 0.0 or 1.0 in seven of ten seeds. An earlier draft wrote "carries no information about A" here, which overstates it and contradicts the sentence two lines later calling the ranking unconstrained rather than mediocre.

D SIGNED AND MAGNITUDE READOUTS ARE DIFFERENT ESTIMATORS

An earlier version of this paper wrote that the transplant's 0.0-or-1.0 behaviour "explains a bimodality we had reported without one", naming the full model's 0.581 ± 0.192. It does not, and the reason is an estimator mismatch we had not noticed in our own code.

The transplant ranks groups by the *signed* mass $\sum_{u \in g} \phi_u$. The deployed maps do not: they sum $|\alpha_u \langle w_c, z_u\rangle|$ over the probe's classes, so every derived-attribution figure we publish for MD-TJEPA — 0.338, 0.340, 0.581, 0.310, 0.324 — is a magnitude readout. The completeness residual we

report ($< 5 \times 10^{-6}$) is computed on the signed object, which is the one Proposition 1 concerns; the ranking is computed on the magnitude one, which is not a decomposition at all.

Measured on the transplant over ten seeds, the gap between the two is total:

readout on the flat (group-blind) aggregate	accuracy	complete?
signed group mass $\sum_{u \in g} \phi_u$	0.605 ± 0.318	yes
$\lvert \sum_{u \in g} \phi_u \rvert$	0.997 ± 0.004	no
$\sum_{u \in g} \lvert \phi_u \rvert$ (the deployed form)	$\mathbf{1.000 \pm 0.000}$	no
per-component norm $\lVert z_u \rVert$ (the informative case)	1.000 ± 0.000	no

Is the channel destroyed by the sum, or only not linearly recoverable? The premise rests on a *linear* probe reading 0.274 from s against 0.998 from the components, so a nonlinear probe is the obvious objection. We ran one: a two-layer MLP (hidden 128, ReLU) on the same frozen s, same splits, same ten seeds, with its learning rate and weight decay selected on a validation split and the test set read once. It reaches 0.294 ± 0.020 against the linear probe's 0.274 ± 0.010 — a paired gain of 0.021 ± 0.020 whose lower bound clears zero by 0.0004, with three of ten seeds negative. That is 2.8% of the distance to the components' 0.998.

We report it as an effect size rather than a significance, because an interval that clears zero by 4×10^{-4} at n=10 is exactly the n-dependent verdict this paper argues against elsewhere. The reading we take: "discards" overstates it — the sum retains a trace — and "not linearly recoverable" understates it, since the nonlinear probe fails too. What is true is that no readout we tried can rank by the axis, which is the premise the proposition needs.

The deployed model, under a like-for-like estimator. We re-ran the deployed encoders with every readout, on the same frozen components, splits and ten seeds. The controls reproduce the published magnitude numbers to within 0.01.

			complete		*not complete*	
backbone	aggregate	probe	signed ϕ	residual	published $\sum_c \lvert \phi \rvert$	residual
TempCNN	flat sum	0.274	0.319 ± 0.057	1.1e−6	0.339 ± 0.086	21
TempCNN	block	0.986	0.904 ± 0.084	4.8e−6	0.580 ± 0.192	55
PatchTST	flat sum	0.256	0.246 ± 0.036	4.8e−7	0.310 ± 0.060	17
PatchTST	block	0.995	0.509 ± 0.180	3.8e−6	0.324 ± 0.116	62

Three things follow. (i) The magnitude readout is not a decomposition of anything: its residual is 17–62 against $\langle w, s \rangle$, where the signed score's is at most 5×10^{-6}. (ii) On TempCNN the faithful score is *not* bimodal — 0.904 ± 0.084, no seed at or below chance, minimum 0.676 — so the 0.581 bimodality we reported is an artifact of the magnitude scalarisation. (iii) The instability is real but lives on **PatchTST/block**: 0.509 ± 0.180, seeds spanning 0.240 to 0.922, at a probe accuracy of 0.995. That is the necessary-but-not-sufficient gap, now localized to one backbone instead of attributed to a sign flip it does not have.

One caution about what we did *not* use. A readout summed over classes, $\sum_c \alpha_u \langle w_c, z_u \rangle$, looks complete and is vacuous: an L_2-regularized multinomial logistic head satisfies $\sum_c w_c = 0$ to machine precision, so that readout is the zero functional and any ranking it produces is a ranking of floating-point noise. We measured $\lVert \sum_c w_c \rVert / \overline{\lVert w_c \rVert} \approx 8 \times 10^{-8}$ and excluded it. It passes a completeness test only because it decomposes zero.

Spearman between $\lvert \phi_u \rvert$ and the pre-aggregation $\alpha_u \lVert z_u \rVert$ is 0.720 ± 0.026, which says what the magnitude readout is actually reading. So the sign mechanism explains the transplant and nothing else, and the full model's bimodality goes back to being an observation we cannot account for. We report the correction rather than quietly restricting the claim, because the superseded sentence appeared in the introduction.

E CAN THE GABOR BANK ADAPT AT ALL?

§A.5.10 reports the band axis as degenerate — ten nominal filters collapsing onto 3.8 ± 0.6 distinct canonical bands, δ absent in every seed — and offers no cause. Two are possible and that section cannot separate them: the *parameterization* cannot move (the failure mode where learnable filters never leave their initialization), or the *objective* does not need it. These are the three controls our design notes list as required for any "the filter learned the alpha band" claim, run here for the first time.

The instrument is movement, not accuracy. Scoring task accuracy does not work: a bank frozen at random initialization already reads 0.67–0.85 on a band-identity task, because ten log-power features locate a tone well enough without any filter sitting on it. Accuracy cannot separate "the bank adapted" from "the head coped". So we measure where the centres go, against a null that shares the data, the optimiser and the budget and differs only in having **shuffled labels**. Movement there is drift; only movement in excess of it is adaptation. Five seeds, 1000 steps, a single linear layer on band log-power, tones injected at α (8–12 Hz) or β (24–28 Hz) against negatives drawn from the rest of the spectrum.

arm	task acc.	filters in α	filters in β	centres moved (Hz)	σ_f at α (Hz)
random init, α task	0.982	$0.2 \rightarrow \mathbf{0.8}$	$0.8 \rightarrow 0.8$	13.8	2.02
random init, β task	0.999	$0.2 \rightarrow 0.2$	$0.8 \rightarrow \mathbf{3.0}$	12.6	5.07
random init, shuffled labels	0.504	$0.2 \rightarrow 0.2$	$0.8 \rightarrow 1.0$	1.2	1.51
canonical init, α task	1.000	$2.0 \rightarrow \mathbf{3.0}$	$1.0 \rightarrow 0.2$	9.0	$\mathbf{2.08 \pm 0.03}$
canonical init, β task	0.998	$2.0 \rightarrow \mathbf{1.2}$	$1.0 \rightarrow 2.0$	8.9	$\mathbf{6.61 \pm 0.36}$
canonical init, shuffled labels	0.498	$2.0 \rightarrow 2.0$	$1.0 \rightarrow 0.4$	1.0	1.79

The bank adapts. Centres move 7.95 ± 1.19 Hz more than the same centres on shuffled labels, every seed positive.

Control 1 — random centres converge on the informative band. Paired against the null, the β task gains $+2.0 \pm 1.5$ filters inside β and the α task $+0.6 \pm 0.7$ inside α (three of five seeds gain one filter, two gain none). The β result is decisive, the α one directional.

Control 2 — nothing migrates into an uninformative band. On the β task the count inside α *falls* by 0.8 ± 0.6 against the null, four of five seeds losing a filter. Filters do not drift into α when α carries nothing; they leave it. This is stronger than the control asks for.

Control 3 — selectivity, and a correction to how it was specified. Our design notes ask for "bandwidth narrowing: σ_t at the α filter narrows on α tasks". That is backwards, in exactly the way §A.5.5 already had to correct for our own results: $\sigma_f \approx 1/(2\pi\sigma_t)$, so a shorter envelope is a *wider* passband. A filter specialising on a band should get *more* selective, σ_f down. It does: σ_f at the α filter is 2.08 ± 0.03 Hz when α carries the label against 6.61 ± 0.36 Hz when it does not — a threefold separation with non-overlapping intervals.

What this licenses, and what it does not. All three controls pass, so the band degeneracy we report is *not* a parameterization that cannot move. It is what the SSL objective leaves unconstrained: a coarse band axis serves the JEPA loss as well as a fine one, and nothing in it pushes the centres apart. That relocates the finding from "learnable filters do not work" to "this objective does not ask them to", and it makes the fix a matter of the objective or of a retained prior rather than of the filter bank. This is a supervised probe of what the parameterization *can* do; it is not a measurement of what the JEPA objective *does*, and we do not report it as one.

F WHICH HEAD YOU PICK DECIDES WHAT THE ATTRIBUTION CAN SAY

Every dilemma measurement in §4 fits the downstream head w *for the contested axis* — the channel — on an aggregate that barely predicts it (0.274). A reviewer is entitled to say that we therefore decompose a head that is itself at chance, and that its decomposition being at chance is definitional. We ran the experiment that settles it, and it went against the claim we had been making.

On the same frozen components, splits and ten seeds, we fit three heads on the *same* s (channel probe 0.274) and took the faithful signed attribution of each (completeness residual $\leq 7.6 \times 10^{-6}$ throughout, so all three are genuine decompositions). TempCNN, flat sum:

head w fitted for	its own accuracy (chance)	channel its argmax names	ϕ channel-acc
the **channel** (our setting)	0.273 ± 0.009 (0.25)	0.273 ± 0.009	0.322 ± 0.056
the **(channel,band) cell**	0.283 ± 0.019 (0.0625)	0.306 ± 0.017	$\mathbf{0.761 \pm 0.059}$
the **band alone**	0.928 ± 0.010 (0.25)	—	$\mathbf{0.743 \pm 0.080}$

The band head is decisive. Its label space contains no channel, its argmax commits to nothing about channels, and its faithful decomposition still ranks the injected channel at 0.743, every seed above 0.627. The cell head's ϕ (0.761) also far exceeds the channel its own argmax names (0.306), so ϕ is not restating the prediction.

So the sentence we had been writing is false. We had claimed that if the aggregate lets no head recover an axis, no faithful attribution can rank by it. It can. Completeness constrains only the *sum*: $s = \sum_u \alpha_u z_u$ discards which tube contributed, but ϕ_u is evaluated per tube. A w that detects "band f is present" aligns with the latent direction of an *active* tube; $\langle w, z_u \rangle$ fires on whichever tube is active, and that tube's channel wins the per-channel argmax. The step from "s barely carries the channel" to "ϕ cannot rank by it" does not go through.

What this does and does not show. Read the two rows together: the aggregate lets a probe recover the channel at 0.274, and ϕ from a band head names it at 0.743. We first read that as the attribution reporting structure the model does not use. That reading conflates two questions. A probe on s asks whether a head can tell *which* channel was active; it does not ask whether the active channel's content moves a head's output. §G asks the second question directly, by resampling each input channel and measuring the band head's logit, and the answer is yes: the injected channel is the one whose resampling moves it most, at 0.933, and ϕ agrees with that intervention on 0.764 of windows. What the band head's map shows is therefore not an axis the model cannot act on; it is a contribution the aggregate hides the *identity* of. What survives is that completeness leaves two choices open: ϕ from a band head and from a channel head are both exact and disagree with the same intervention at different rates, and each disagrees mostly through the reference it implicitly takes (§G).

This is Corollary 1 instantiated, and it is what the corollary was written from rather than a case the theory absorbed afterwards: two heads on one s, two unique complete decompositions, opposite rankings of the same components. Proposition 1 itself is untouched — it concerns the w obtained by fitting *for* A, and says $\{\phi_u\}$ carries "whatever that direction incidentally picks up from the individual z_u"; a band head's alignment with the active tube is exactly such a pickup, at 0.743 rather than at chance. What the falsified step did was quantify over *all* heads on the strength of a statement about one.

Where it does not reproduce. On PatchTST the flat-sum ϕ, measured from the latent origin, stays at chance for all three heads (0.246, 0.294, 0.281); that backbone's per-tube norms reach only 0.329 channel accuracy. Re-referenced, it no longer does (§G): the offset, not the absence of channel information, is what hides it. That bounds the effect rather than rescuing the old claim: TempCNN is the backbone our synthetic headline numbers come from, and that is where it fails.

G Recoverability is not contribution: an input intervention

A probe on s and an attribution over $\{z_u\}$ answer different questions, and an earlier draft let the first stand in for the second. The minimal case: $x_1, \ldots, x_4$ i.i.d., $s = \sum_u x_u$, head $h = s$, and $A = \arg\max_u x_u$. By symmetry no head on s recovers A above chance, yet $\phi_u = x_u$ names A every time — and zeroing x_u changes h by exactly ϕ_u. "The axis cannot be recovered from s" does not entail "the axis does not contribute to the output". To say what a map gets wrong about the model, "use" has to be defined by intervention.

Definition used here. For a head w, a window x and its predicted class $\hat{y}$, input channel c *contributes* to $\langle w_{\hat{y}}, s(x) \rangle$ by $\delta_c(x) = \langle w_{\hat{y}}, s(x) \rangle - \langle w_{\hat{y}}, s(x^{(c)}) \rangle$, where $x^{(c)}$ replaces channel c with fresh zero-mean unit-variance noise and leaves the others untouched. Every training window carries

exactly one injected oscillation, so this moves the window into a zero-injection regime the generator never produces; it is a distribution shift, not an in-distribution resample. The swap below is the in-distribution control: it keeps one injection per window and moves it. The intervention names $\arg\max_c \delta_c$. Because the gate here is not input-conditional and each tube reads one channel, resampling c changes only channel c's tubes, so $\delta_c(x) = \Phi_c(x) - \Phi_c(x^{(c)})$ with $\Phi_c = \sum_{u \in c} \phi_u$: ϕ with a per-window reference in place of the latent origin (completeness residual $\leq 3 \times 10^{-6}$ throughout). A second intervention, the *swap*, moves the injected channel's content onto another channel and backfills with what was there; it asks whether the output depends on *which* channel carries the content. Finally, we change only the decomposition's reference. ϕ takes the latent origin as its baseline; with $q_c = \sum_{u \in c} \alpha_u z_u$ and the *same* head direction $w_{\hat{y}}$, the re-referenced score is $w_{\hat{y}}^\top (q_c(x) - \mu_c)$, with μ_c the mean of q_c over the training windows. It is still complete, up to a class-dependent constant $\sum_c w_{\hat{y}}^\top \mu_c$ (the intercept excluded, as for ϕ). Taking μ_c over injection-free noise windows instead makes the score an unbiased estimate of the expected resample effect $\mathbb{E}[\delta_c]$ — because the gate is static and each tube reads one channel — and changes every entry below by at most 0.004, since in this benchmark the non-injected channels are noise.

Setting. The synthetic benchmark of §F: frozen components, flat sum, the same three heads and ten seeds, 800 test windows per seed, one resample per channel shared by the three heads. Pretraining is re-run and is not bit-deterministic on GPU, so the ϕ column differs from §F seed by seed (band: 0.765 here, 0.743 there); every comparison below is within one model.

	head fitted for	ϕ names injected ch.	intervention names injected ch.	ϕ agrees with interv.	re-referenced ϕ names inj.	re-ref. agrees with interv.
TempCNN (n=10)	channel	0.329 ± 0.065	0.461 ± 0.056	0.408 ± 0.070	0.489 ± 0.065	0.734 ± 0.029
	cell	0.777 ± 0.056	0.918 ± 0.038	0.776 ± 0.057	0.935 ± 0.034	0.944 ± 0.025
	band	0.765 ± 0.065	0.933 ± 0.034	0.764 ± 0.066	0.948 ± 0.027	0.948 ± 0.026
PatchTST (n=5)	channel	0.241 ± 0.077	0.400 ± 0.125	0.333 ± 0.066	0.424 ± 0.142	0.718 ± 0.054
	cell	0.316 ± 0.091	0.465 ± 0.104	0.404 ± 0.119	0.493 ± 0.102	0.748 ± 0.091
	band	0.292 ± 0.068	0.515 ± 0.101	0.353 ± 0.060	0.539 ± 0.109	0.757 ± 0.067

Chance is 0.25 in every column.

Every head's output uses the injected channel's content, and almost none of its location. Resampling the injected channel moves each head's logit more than resampling any other channel, above chance for all three heads on TempCNN — at 0.933 for the band head, and at 0.461 even for the channel head, whose own accuracy is near chance. In this benchmark the non-injected channels are noise, so on its own that says only that the head responds to the injected signal. The swap separates content from location:

	head	logit change under swap, relative to resample	prediction unchanged under swap	ϕ follows the content to its new channel
TempCNN	channel	0.098 ± 0.041	0.919 ± 0.023	0.338 ± 0.061
	cell	0.057 ± 0.009	0.808 ± 0.034	0.784 ± 0.059
	band	0.054 ± 0.009	0.989 ± 0.003	0.769 ± 0.063
PatchTST	channel	0.002 ± 0.002	1.000 ± 0.001	0.237 ± 0.070
	cell	0.003 ± 0.003	0.996 ± 0.004	0.308 ± 0.100
	band	0.003 ± 0.004	0.997 ± 0.004	0.291 ± 0.065

Moving the same content to another channel changes the band head's logit by about 5% of what removing it does, and leaves its prediction unchanged in 0.989 of windows; on PatchTST the change is under 1% for every head. The aggregate forwards the channel's *content* and nearly discards *which* channel carried it — the two readings a probe on s cannot tell apart. The band head's ϕ follows the content to its new channel (0.769), so its channel map names where the contributing signal was. We therefore withdraw our earlier reading of the band head's 0.743 as a map of an axis the model does not use; it is a map of which channel's content moved that head's output, whose identity the output itself does not depend on. Because the intervention names the injected channel almost always, "ϕ agrees with the intervention" is close to "ϕ names the injected channel" here; the agreement column is not independent evidence. Each resample and swap is a single draw per window.

Where ϕ and the intervention part: mostly the reference. With the latent origin as reference, ϕ agrees with the intervention on about three windows in four for the band and cell heads and on 0.408 for the channel head, whose ϕ names the injected channel at 0.329 while the intervention on the same head finds it at 0.461. Changing nothing but the reference to the training mean lifts agreement to 0.948, 0.944 and 0.734, and the channel head's map to 0.489. On PatchTST, where the intrinsic ϕ is at chance for all three heads (0.241–0.316) although the intervention finds the channel above chance for each (0.400–0.515), the same change lifts agreement from about a third to 0.72–0.76. So most of what §4 reports as the channel head's failure is the reference the intrinsic decomposition implicitly takes — the latent origin, which is not an absent input, and whose offset the per-tube latents carry in large amounts (the freedom of Appendix C). This is close to circular by design: the re-referenced score estimates the expected effect and the intervention is one draw of it, so the channel head's remaining 0.27 measures how often a single draw flips the argmax where that head's margins are small, not a separate failure of the head. What the comparison establishes is about the intrinsic ϕ: its latent-origin reference, rather than completeness or the head, is what separates it from the intervention. Whether a training-mean reference does as well on real data, where latents carry large offsets, is untested. (An earlier version of this section reported that recentring did not help; that computation averaged scores taken at each training window's own predicted class, mixing head directions, and is withdrawn.)

What this changes, and what it does not. "A faithful attribution reports its head" survives, with a second free choice beside it: completeness pins the sum, not the head and not the reference, and on these models the reference accounts for most of each map's disagreement with what the output depends on. The intrinsic map is not a free explanation of the model; on this synthetic benchmark, re-referencing turns it into an estimate of the intervention, and on real data that is untested. What does not survive is any reading of a near-chance probe on s as evidence that the model does not use the axis. Check 1 of §8 is rewritten accordingly.

H Full ablation tables (per-seed)

Embedded automatically from the released ablation table; see §A.2.

I Trained Gabor center frequencies and bandwidths

See the table in §A.5.5; the learned centers and bandwidths are released with the run artifacts.

J Does a control task license the premise? It does not

The premise is a *negative* probe result — channel identity reads 0.998 from the components and 0.274 from the sum they form — and it has to be read as "the aggregate barely retains the axis" rather than "the probe could not find it". The probing literature spent years learning not to conflate those, and Hewitt & Liang (2019)'s answer is a *control task*: the same probe, the same representation, a label space of the same shape, and labels carrying no structure the representation could encode. We cited the method in §2 and had never run it on our own instrument. Two controls, over the same benchmark, frozen components, balanced logistic probe and ten seeds:

random label assigns each sample an i.i.d. channel label. Nothing is learnable. It bounds how much of any accuracy is probe capacity. **permuted cell** is the Hewitt–Liang construction: a fixed random map from the true (channel, band) cell to a channel label, redrawn per seed, so the information sits in the input exactly as the real task's does but is memorisable rather than linearly structured.

	TempCNN			PatchTST		
probe reads	channel	random	permuted cell	channel	random	permuted cell
flat sum s	0.274 ± 0.009	0.242 ± 0.011	0.487 ± 0.093	0.266 ± 0.013	0.242 ± 0.012	0.342 ± 0.032
block aggregate	0.994 ± 0.008	0.260 ± 0.006	0.919 ± 0.028	0.996 ± 0.005	0.245 ± 0.010	0.527 ± 0.055
components $\{z_u\}$	1.000 ± 0.000	0.247 ± 0.010	1.000 ± 0.000	0.998 ± 0.003	0.253 ± 0.012	0.996 ± 0.006

An earlier version drew two conclusions from this table. Both are withdrawn.

(i) The random-label control licenses nothing. Train and test labels are drawn independently and only test accuracy is recorded, so any probe scores chance on it whatever its capacity — including one that memorises its training labels. The 0.260 maximum is therefore not evidence that the probe is not over-expressive, and the 0.274 rests on the probe results of §4 alone (linear, and an MLP closing 2.8% of the gap), not on this control.

(ii) The permuted-cell score says nothing about the channel. The label table is drawn independently per (c, f) cell, so the label is partly predictable from the band alone. Replaying the same random stream, a band-only oracle (the majority label per band, fitted on train) scores 0.583 ± 0.056 on test, a channel-only oracle 0.521 ± 0.046, the majority class 0.370 ± 0.056. The sum's 0.487 ± 0.093 is below the band-only oracle, and the sum carries the band (0.928), so its score is consistent with band information alone. We had read it as the sum keeping cell structure and dropping the channel specifically; it does not show that. A control that neutralises the band would draw each band's channel→label map as a permutation; we have not run it.

(iii) A caveat on our own reference number. The components fit *any* cell-determined label at 1.000 ± 0.000. By Hewitt and Liang's own standard that probe is too expressive for its accuracy to show the encoder learned channel structure: the 0.998 is retained cell identity, which the architecture guarantees by construction, since each tube *is* a (c, f, s) triple. That does not weaken the contrast the premise needs — what the sum barely retains, the components retain — but it does mean the 0.998 should not be read as evidence about what pretraining learned. We had been reading it as the former and stating it in a way that invited the latter.

K THE UNTRAINED FLOOR, AND WHETHER IT IS AN ARTIFACT

The audit section compares MD-TJEPA against TS2Vec and CPC and gives no floor, so a reader cannot tell how much of 0.370 is self-supervision and how much is the architecture plus a balanced logistic probe. Random-init controls are run elsewhere in this paper — ECG5000, and the synthetic benchmark — and the gap was on the headline one, which is the conspicuous place to omit it. Whether the probe's fixed penalty moves any of this is checked across the whole table in §L.

We build an untrained model from the deployed run's own resolved config and send it through the same `linear_probe` entry point with the flags that produced every matched number in the retrain table: same subject-disjoint split, same 12k cap, same balanced logistic probe, same n=70,675 test set. Five seeds, because the initialization is the thing being controlled for and has to vary the way the trained runs' seeds do.

Sleep-EDFx, matched 12k probe	macro-F1	accuracy
TS2Vec (Yue et al., 2022)	0.552 ± 0.012	—
MD-TJEPA, untrained	$\mathbf{0.444 \pm 0.003}$	0.667 ± 0.007
MD-TJEPA, retrained (gate fixed)	0.370 ± 0.008	—
MD-TJEPA, deployed (gate collapsed)	0.354 ± 0.007	—

Three explanations to rule out before reading that as it stands.

It is not the gate collapse. At the 0.01 liveness threshold this paper uses throughout, both the retrained and the untrained model are $60/60$ live. The difference is the *shape*: the untrained gate is uniform at $\sigma(-2) = 0.119$ everywhere, the retrained one spread over 0.036–0.614. The learned weighting is the worse one downstream.

It is not the probe's fixed penalty meeting a smaller feature scale. `_fit_probe` does not standardize, and the trained aggregate's standard deviation is 0.74 against the untrained model's 3.92, so in principle the fixed C=1 could have been punishing the trained model for being small rather than for being worse. Measured directly on the same windows, with and without standardizing: raw, 0.488 against 0.499; standardized, 0.488 against 0.518. Standardizing *widens* the gap. The scale difference was flattering the trained model.

It is not universal. On ECG5000 the same control runs the other way, SSL 0.918 against random-init 0.878 (§A.5.13). Whatever the objective learns is worth something there and costs something here.

We report this rather than leaving the table floorless. It does not touch the paper's claim, which is about what an attribution can say and not about downstream accuracy — but a reader is entitled to know that on the benchmark every downstream number here comes from, the pretraining this architecture exists to do is worse than not doing it.

L Does the probe's fixed penalty move the utility table?

Every cross-encoder number in the utility table is a logistic-regression probe with a fixed $C{=}1$ on raw features, so its L_2 penalty meets each encoder at that encoder's own feature scale. We re-ran the whole table with features standardized on the training split, same protocol otherwise (balanced 12k-window probe, leakage-free split, identical test set, seeds paired by index, $n{=}5$ per arm):

arm	raw (published)	standardized	paired shift
TS2Vec	0.552 ± 0.012	0.520 ± 0.009	-0.032 ± 0.013 (5/5 lower)
CPC	0.424 ± 0.028	0.424 ± 0.030	-0.000 ± 0.002
MD-TJEPA, retrained	0.370 ± 0.008	0.369 ± 0.008	-0.001 ± 0.001
MD-TJEPA, untrained	0.444 ± 0.003	0.449 ± 0.003	(re-run, not paired)

The ordering does not change: TS2Vec, untrained MD-TJEPA, CPC, pretrained MD-TJEPA, both ways. Standardizing costs TS2Vec 0.032 in every seed and moves CPC and pretrained MD-TJEPA by at most 0.001, so the fixed penalty favoured the strongest baseline slightly and did not produce any gap we report. The pretraining cost of §K widens rather than narrows (0.449 against 0.369).

M Which constraint of the shared sum costs utility, and does pretraining survive it?

A linear head on the concatenated components, $\sum_u v_u^\top z_u$, is additive over components, and so is a head behind any fixed linear projection. What the deployed readout adds is the constraint $v_u = \alpha_u w$. We separate its parts on the same frozen components and windows (balanced logistic probe, $C{=}1$, raw features, $12{,}000$ train windows drawn by index per seed, the full $70{,}675$-window test split), on the five gate-fixed retrained checkpoints and on five untrained models of the same architecture. Every readout below is additive over components. Macro-F1, $n{=}5$, paired by seed:

readout	width	pretrained	untrained	pretrained − untrained
shared sum, gated (deployed)	32	0.366 ± 0.009	0.443 ± 0.002	-0.077 ± 0.010
shared sum, gate bypassed	32	0.363 ± 0.013	0.450 ± 0.004	-0.087 ± 0.015
one direction per channel	64	0.386 ± 0.005	0.441 ± 0.004	-0.055 ± 0.006
one direction per component, gated	1920	0.457 ± 0.024	0.620 ± 0.013	-0.163 ± 0.026
one direction per component, no gate	1920	0.507 ± 0.018	0.639 ± 0.007	-0.132 ± 0.016
PCA of components, 32	32	0.417 ± 0.045	0.586 ± 0.013	-0.169 ± 0.042
PCA of components, 128	128	0.489 ± 0.021	0.625 ± 0.013	-0.136 ± 0.020
random projection, 32	32	0.357 ± 0.019	0.545 ± 0.017	-0.188 ± 0.018
random projection, 128	128	0.467 ± 0.021	0.607 ± 0.015	-0.140 ± 0.022

What the sum costs. On the pretrained components the gate on the shared sum is immaterial (-0.003); lifting the shared direction is worth $+0.091$, and dropping the gate from the per-component head a further amount, $+0.141$ in all over the deployed readout. A label-free 128-dimensional projection recovers most of it ($+0.123$ for PCA, $+0.101$ for a random projection). The utility cost the paper reports is the cost of the shared direction, not of additivity.

Pretraining loses under every readout. The untrained model beats the pretrained one under all nine, by 0.055 to 0.188, every paired interval excluding zero. The cost §K measured on the deployed readout (0.074) is the smallest version of it: the readouts that keep more of the components show more of what pretraining removed. Read through a label-free 128-dimensional projection, the untrained model reaches 0.625 (PCA) and 0.607 (random), above TS2Vec's 0.552 ± 0.012 on its 320-dimensional features under the same probe — and above the 0.520 it scores standardized (§L).

So the architecture's untrained Gabor-bank components carry more usable sleep-stage information than the contrastive baseline's learned features, and our JEPA objective removes a large part of it. We do not claim an untrained model is a better representation learner; this is one dataset, one probe, and the train subset here is drawn differently from the matched protocol's (so its sum row, 0.366, is not the published 0.370 window for window). It does say where the utility gap in this paper comes from: not from additivity, and not from the encoder's capacity, but from what the objective does to the components.

N COROLLARY 1 ON REAL DATA: THREE DEPLOYABLE HEADS

Every number behind the two-head result is synthetic, and §F says why: Sleep-EDFx has no per-window channel label, so the argmax-against-injected-channel instrument does not exist there. But Corollary 1 does not need a ground truth. It says two heads on the same s have two unique complete decompositions that *may* rank the components differently — a statement about the rankings, not about which is right — and that is measurable with no label beyond what fits the heads. So on the five retrained Sleep-EDFx checkpoints ($60/60$ live) we fit three heads on the same frozen s: the five-stage head (the model's own task), wake vs. sleep, and REM vs. non-REM (wake counted as non-REM, so not the clinical REM/NREM contrast). Each is decomposed at its predicted class; every residual is $\leq 1.4 \times 10^{-6}$. Then, per test window, the rankings of the 60 tubes are compared between heads, and each head's ranking to the model's own routing $\alpha_u \|z_u\|$. Heads are fitted on $12{,}000$ shuffled train windows and compared on the first $6{,}000$ test windows in file order — a prefix of the test records, not a subject-stratified sample — so these are rankings on a subset. That prefix is 71% wake, and the balanced heads trade accuracy for recall on the rare classes: their accuracies (stage 0.625, wake 0.732, REM 0.845) sit near or *below* predicting the majority class (0.714, 0.714, 0.939). So these are heads fitted on s, not heads shown to be useful, and the comparison is of how they rank components, not of heads a clinician would deploy.

head pair	Spearman ρ	overlap@6
stage vs. wake/sleep	0.758 ± 0.029	0.813 ± 0.029
stage vs. REM/non-REM	0.648 ± 0.047	0.660 ± 0.038
wake/sleep vs. REM/non-REM	0.572 ± 0.031	0.570 ± 0.027

Read it in two directions, because it cuts both ways.

The strong form does not appear here. These are not opposite rankings: three heads whose labels nest — wake is a coarsening of stage, REM a class of it — produce maps that agree at ρ 0.57–0.76. The synthetic case gave the strong form because band and channel are *orthogonal* label spaces by construction. Real data with nested clinical labels offers no such pair, and we report that rather than manufacture one. The channel split is head-invariant too: ϕ's mass on channel 0 is 0.830, 0.791 and 0.819 for the three heads against 0.838 for the routing, because α_u dominates ϕ_u's channel mass and the gate, not the head, decides that axis on these models. (The $0.548/0.452$ of the gate-collapse section is the *norm heatmap*'s split, with no gate in it; this is ϕ's.)

The weak form does, and it is not small. Between these heads, 19–43% of the top-six components differ (overlap@6 0.57–0.81). Each map is an exact decomposition of its head. A reader shown one of them as "what the model attends to" is shown something a third of which changes with a choice of readout they were not told about. So the size of the effect Corollary 1 licenses scales with how orthogonal the two heads' tasks are — opposite when they are, a third of the top-k when they nest — and a real deployed head sits at the second end.

O BIOMARKER REGISTRY

Embedded in the biomarker registry module; sources Wagner GS et al. (AHA/ACCF/HRS 2009, Circulation 119:e262 — not the PTB-XL dataset paper), Strauss et al. 2011, Sokolow & Lyon 1949, ANSI/AAMI EC57:2012.

P QUANTUS FAITHFULNESS PANEL (INSERTION / DELETION / COMPREHENSIVENESS)

We compute the Quantus-inspired (Hedström et al., 2023) faithfulness panel on PatchTST Sleep-EDFx checkpoints using `||z_c||`-derived per-channel attribution, batch of 8 epochs. **What the metric scores:** the perturbed output is $\|s_x\|^2$, a representation-level proxy — an SSL encoder has no class logit — not a downstream task output. These numbers are therefore evidence about representation stability, not task faithfulness, and the two columns below also differ in split and epoch as well as scale, so their difference is not a clean scaling effect. **We report two training budgets to honestly trace whether faithfulness signals survive scaling.**

Metric	Small-data (10 records, ep_019)	Full-data (50 records, ep_004, `val`)	Random (50-record reference, `val`)
Insertion AUC ↑	+0.221	+0.674	+0.769
Deletion AUC ↓	+0.092	+0.842	+0.746
AUC(I) − AUC(D) ↑	**+0.129**	**-0.168**	+0.023
Comprehensiveness ↑ (normalized)	(raw +27.73)	+0.142	n/a
Sufficiency ↓ (normalized)	(raw +23.96)	-0.131	n/a
Informativeness gap vs random	**+0.193**	**-0.191**	n/a

The full-data column is now computed on a **held-out validation** batch (`split=val` in the Quantus artifact, audit H2); the small-data column is the original train-batch diagnostic, kept for the before/after-scaling comparison. **Honest negative result at scale.** The small-data informativeness gap of +0.193 does NOT survive scaling to 50-record training. On held-out validation, MD-TJEPA per-channel attribution is **at or below random attribution** (gap -0.191) in insertion/deletion AUC.

One explanation, which we did *not* test: this may be **specific to 2-channel Sleep-EDFx**, since with only Fpz-Cz + Pz-Oz, both leads carry highly correlated EEG signal, so removing either channel has near-identical model-output impact — limiting the per-channel test's discriminative power. The 12-lead PTB-XL test (PTB-XL v2 full-fold retrain, §A.5.4) is the appropriate venue for this faithfulness claim, and there it passes 10/10 seeds.

We retain the small-data row in the table as a snapshot demonstrating the test machinery is correct (random attribution does score below MD-TJEPA in the same batch). A second untested possibility is that the small-data gap was amplified by the model's overfitting to 10 records, producing channel-attribution patterns that coincide with the small training distribution but do not generalize. **The Adebayo cascading sanity (seeded, full-(c, f, s) attribution; §A.5.4) passes on this checkpoint, which is superseded and not one the paper reports on — the deployed verdict is fail ($0/5$) — 10/10 seeds at epoch 4, single-shot ρ = -0.155; faithfulness in the Adebayo sense remains intact even when the perturbation-based informativeness gap vanishes on 2-channel data.**

Q MODEL AND TRAINING HYPERPARAMETERS

Every pretraining run in this paper uses one of the three configurations in Table 4. Values are the committed configuration files, not transcribed from run logs.

Per-component encoders. *TempCNN* in this paper denotes the small temporal CNN used per (c, f, s) tube — three `Conv1d` layers, kernel 5, 32 hidden channels, projected to $k = 32$; it is not the architecture of that name from the satellite-imagery literature. *PatchTST* is the patch-transformer readout of Nie et al. (2023) at $d_{\text{model}} = 64$, 2 layers, 2 heads, patch length 16, stride 8, projected to the same k. Both share one backbone across tubes with a (c, f, s) identity embedding, so parameter count grows additively rather than multiplicatively in $C \times F \times S$. The **gate-balance retrain** of §A.5.14 sets λ_{balance} from its default 0 to 1.0, and — unintentionally — also batch size $16 \to 32$ and all 137 train records instead of 50.

R COMPUTE BUDGET STATEMENT

Hardware: 4× NVIDIA RTX 3090 (24 GB), Ampere, BF16 mixed precision.

Per-experiment compute (sequential GPU-hours) and **wall-clock** (after parallelization across 4 GPUs) are in Table 5.

Table 4: Model and training hyperparameters for the three pretraining configurations, read from the committed config files.

	Sleep-EDFx	PTB-XL	ETTh1
Decomposition			
channels C / bands F / scales S	2 / 10 / 3	12 / 8 / 3	7 / 6 / 3
scales (s; ETTh1 in samples)	1, 5, 30	0.2, 1, 2	4, 12, 24
Gabor kernel length	513	321	65
Gabor init	canonical	canonical	random
sampling rate f_s (Hz)	100	100	1
window (samples)	3000	1000	96
context ratio	0.7	0.7	0.5
Encoder and predictor			
latent dim k	32	32	32
Koopman blocks K	8	8	8
EMA momentum τ	0.998	0.998	0.998
Optimization			
optimizer	AdamW	AdamW	AdamW
learning rate	5×10^{-4}	5×10^{-4}	5×10^{-4}
weight decay	10^{-5}	10^{-5}	10^{-5}
batch size	32	32	32
precision	BF16	BF16	BF16
Loss weights			
λ_{koop} / λ_{sig}	0.01 / 1.0	0.01 / 1.0	0.01 / 0.1
λ_{sparse} / λ_{anova}	0.01 / 0.01	0.01 / 0.01	0.01 / 0.01
λ_{mod}	0	0	0
PSD warm-start epochs	5	5	none

Table 5: Compute to reproduce the paper. The synthetic rows were timed on a GPU shared with unrelated work, so they are upper bounds.

Experiment	Compute (GPU·h)	Wall-clock	Notes
22-cell × 3-seed × 3-epoch ablation matrix	0.14	2 min	84 runs × ~6 s
A12 5-seed × 30-epoch reference	0.85	20 min	10 runs × ~5 min
A12 10-seed × 10-epoch (Bonferroni-Holm survivable)	**1.00**	**15 min**	20 runs × ~3 min, paper headline
PatchTST Sleep-EDFx 50-epoch (10 records)	0.35	20 min	single GPU
PatchTST Sleep-EDFx scaled (50 records × 30 ep, F1)	5.00	80 min	GPU·h summed over seeds run concurrently
PatchTST PTB-XL scaled (full × 20 ep, F-A1 retrain)	6.00	120 min	GPU·h and wall-clock recorded separately, not reconciled
ETTh1 TempCNN 30-epoch (full)	0.06	4 min	single GPU
Quantus + Adebayo + IG-vs-α + noise robustness panel	0.08	5 min	per checkpoint
Synthetic and control experiments (timed on a shared GPU; upper bounds)			
Dilemma transplant onto a plain NAM	0.01	27 s	10 seeds
Gate-collapse mechanism, 3 arms	0.36	21 min	5 seeds × 20k steps
Backbone readout control	0.18	11 min	2 backbones × 3 readouts × 5 seeds
Backbone-matched instrument	0.18	11 min	2 backbones × 4 instruments × 5 seeds
Derived attribution φ	0.36	22 min	2 backbones × 10 seeds
Block versus flat aggregation	0.35	21 min	2 backbones × 10 seeds
Trade-off, per variant	0.49	29 min	5 variants × 10 seeds
Deletion intervention	0.10	6 min	10 seeds
Total, rows above	**~15.5 GPU·h**	**~6 h wall**	with 4-GPU parallel

Not included: the five-seed gate retrain, the multi-backbone PTB-XL runs, and the controls added in September 2026 (random-init floor, control task, intervention, two-head real-data, standardized probe).

CPU-side data preprocessing (Sleep-EDFx EDF parsing, PTB-XL WFDB read, ETTh1 CSV ingest) is included in wall-clock; it accounts for <5% of total time and is single-threaded per worker.

The eight synthetic rows were timed by re-running each script from the committed defaults on a single RTX 3090 that was concurrently running unrelated work at near-full utilisation, so those figures are **upper bounds** on a free device rather than estimates. The timing runs doubled as a reproduction check and are the source of the per-seed variability quantified in the reproducibility statement.

Estimated public-cloud cost (AWS `p4d.24xlarge` spot ~$5/h, used as 1/8 of 8× A100 = equivalent of 1 RTX 3090-class hour): under **$35** for the full reproduction. Local 4× RTX 3090 reproduction is free apart from electricity (~6 kWh).

S ARCHITECTURAL COMPARISON WITH MA ET AL. (2026)

§2 places this work relative to Ma et al. (2026), the closest biosignal-SSL system in aim. The table below records the design axes on which the two differ. It is a comparison of construction, not of measured outcome: the two systems were not evaluated against each other, our downstream numbers are not competitive (§6), and nothing here should be read as a performance claim. Ma et al. (2026) appeared at ICLR 2026.

Design axis	Ma et al. (2026)	MD-TJEPA (this paper)
Decoupling site	Tokenizer (separate temporal and frequency codebooks)	Latent space: per-(c, f, s) additivity after a tokenizer-free wavelet/multiscale front end
Quantization	Discrete vector quantization	Continuous additive aggregation with a sparse gate
Backbone	EEGSSM (SGConv + sliding-window attention)	Per-component TempCNN or PatchTST
Attribution	Codebook-token frequency over a window population	Per-(c, f) latent magnitude $\|z\|$, per-sample and continuous. Note §4: this is *not* the routed contribution $\alpha\|z\|$, and the two can be disjoint
Faithfulness audit	Not reported	Cascading randomization, insertion/deletion, noise robustness (§6) — with a negative verdict on our Sleep-EDFx checkpoints
Identifiability	Not addressed	Theorem 1, adapting Chastaing et al. (2012)
Evaluation domains	EEG: ten datasets, eight tasks (TUAB, TUEV, ISRUC S1/S3, CHB-MIT, FACED, SEED-V, SHU-MI, BCIC, Mental Arithmetic)	EEG (Sleep-EDFx) and ECG (PTB-XL)
Pretraining corpus	∼9.2k hours EEG (TUEG, after preprocessing)	Sleep-EDFx 197 records; PTB-XL 21,837 records