How Complex Should an Explanation Be? A Distortion–Rate View of Time Series Attribution

Anonymous authors
Paper under double-blind review

Abstract

Feature attribution for time series is usually posed as a choice among solution concepts—Shapley, Owen, Banzhaf—estimated as accurately as possible. We argue, and measure, that this framing is mis-specified. First, the attributed object is not a property of the model: the coalition value function depends on a masking baseline and an output scale, and every summary of it on a coalition measure. Across seven trained models these choices move the additive share of the value function by up to 32 points, and on one real model the top-10% attributed cells agree only 32% of the time between a zero and a first-observation baseline; neither values nor rankings survive a change of convention. Second, at a fixed convention the answerable question is how much of the value function an explanation of a given size can carry. Treating an explanation as a lossy code, we measure its distortion against its *rate*, the numbers it emits per input. On five real models, additive attribution is inefficient even at its own rate, locality is the wrong structure—a lag-1 band buys 3–9 points where a rank-one factorisation buys 21–30 at 2–9× fewer numbers—and the optimal complexity is finite and set by the query budget. Perturbation-based faithfulness metrics do not resolve these distinctions on a game with exact ground truth (Barr et al., 2023; Zhang & Mcmillan, 2025; Oliveira e Silva et al., 2026). Finally, an amortized estimator that emits the coefficients of a structured surrogate ships each explanation with a measured fidelity: it reaches 99% of a 200,000-query per-input fit on one model and fails on two of six, which its audit, whose query cost we price, reports.

1 Introduction

An attribution method for a time series model answers: which of the $T \times D$ cells of this input mattered? The dominant formalism defines a cooperative game $v(S)$ over subsets S of cells and reports a solution concept of that game (Shapley, 1953; Lundberg & Lee, 2017), estimated cheaply by amortization (Jethani et al., 2022), by structured surrogates (Heidari & Rabusseau, 2026), or by architectures that expose the game by construction (Chen et al., 2023). Research then concentrates on the two questions the formalism makes salient—which solution concept, and how to estimate it cheaply. This paper reports a sustained attempt to answer them for time series, and the reasons we now think they are the wrong questions. Two findings drove the change.

The game is a modelling choice, and the choice dominates. $v(S) = f(x \odot \text{mask}(S))$ is fixed only after choosing what a masked cell becomes, what scalar of f's output to track, and what distribution over S to measure under. We show (§7) that these choices move the conclusions more than the method does: the additive share of the value function swings by 32.4 points on one model and 30.9 on another across $p \in \{0.25, 0.5, 0.75\}$, enough to reorder the models, and the ranking—which we expected to be the robust part, since perturbation metrics consume only the ranking—is not: switching the masking baseline from zeros to a first-observation baseline leaves a mean Kendall τ of 0.267 across five models.

At a fixed convention, size is the binding constraint. Once the convention is declared, an explanation is a lossy summary of v whose quality trades against how many numbers it contains. We

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.

make this literal (§4): under an orthonormal coalition basis the unexplained variance of a coefficient set is one minus its energy share, so *distortion* against *rate* is measurable and every structural assumption is a point on one plot. The additive class—the ceiling a Shapley or Owen value can reach—is one such point, and not a good one.

Contributions.

- A distortion–rate formulation of explanation, measured on five real time series models (§4).
- Evidence that on those models locality is the wrong structure and low rank the right one, with a separable (Kronecker) hypothesis refuted (§5).
- A finite optimal explanation complexity set by the query budget, reproduced by an independent experiment (§6).
- Measurements showing that neither attribution values nor rankings survive a change of convention (§7).
- A controlled demonstration that perturbation-based faithfulness metrics do not separate objectives that exact ground truth separates (§8).
- An amortized estimator with a priced per-input fidelity audit, and a diagnosis of where and why it fails (§9).

2 Related work

Amortized and structured estimation. FastSHAP (Jethani et al., 2022) trains a network to emit Shapley values in one forward pass, scored against a surrogate that supplies $v(S \mid x)$; HarsanyiNet (Chen et al., 2023) instead builds a network whose architecture exposes Harsanyi dividends exactly; TN-SHAP-G (Heidari & Rabusseau, 2026) fits a tensor-network surrogate to one input's coalition function, recovers values and interactions in closed form, and scores it by held-out R^2 on unseen coalitions of that input. Our estimator keeps FastSHAP's amortization and TN-SHAP-G's object: one network predicts a rank-r surrogate for every input (§9), so the coefficients cost no queries, while a held-out fidelity for a new input must be bought with fresh queries, which we price; its factored pairwise term is a factorization machine (Rendle, 2010) with input-conditioned, signed factors. Tree proxies fitted to masked outputs (Butler et al., 2025), with a residual correction for interaction indices (Thies et al., 2026), are the closest per-input alternative to our surrogate; we have not compared against them. Scoring fast Shapley approximations with conformal intervals (Alkhatib et al., 2024) addresses the same trust question as our audit, with a calibrated rather than an empirical guarantee.

Interaction indices and the additive ceiling. That an additive attribution is a lossy summary of an interacting game is well established: Shapley residuals (Kumar et al., 2021) quantify what the Shapley value cannot express; n-Shapley values (Bordt & von Luxburg, 2023) expose the full Möbius hierarchy; InstaSHAP (Enouen & Liu, 2025) is exact where the model is additive; SHAP-IQ and KernelSHAP-IQ (Fumagalli et al., 2023; 2024) estimate higher-order indices directly; and Owen values (Owen, 1977) exploit an a-priori grouping, which a time series supplies for free. Our additive explainability ratio is the degree-≤ 1 mass of that same decomposition, not a contribution. What we add is the second axis: these classes emit different numbers of coefficients, and plotting distortion against that count (§4) makes them commensurable. Rate–distortion itself is not new to explanation: RDE (Macdonald et al., 2019; Kolek et al., 2022) selects the smallest set of input features that preserves the output when the rest is randomised, so its rate is the size of a mask; ours is the number of coefficients of a decoder for the whole coalition function. The complexity–fidelity trade-off of surrogates is also known (Poyiadzi et al., 2021); our contribution is its measurement across structural classes.

Attribution for time series. Time series methods overwhelmingly take locality as the structuring assumption: contiguous windows with exact Shapley values per window (Nguyen & Ifrim, 2025), segment players (Kim & Park, 2026), shapelets (Huang et al., 2025) and transformer-specific Shapley pretraining (Cheng et al., 2025) all build units from temporally contiguous cells, FDTempExplainer decomposes a temporal model into timestep effects and their interactions (Yang et al., 2024), as graph-structured L- and C-Shapley do for general structured data (Chen et al., 2019); a parallel line

learns a mask rather than assigning coefficients (Liu et al., 2024a;b). Our band sweep (§5) tests the small-neighbourhood premise directly and it fails on every model we measured; it does not test every structured method. We place the learned-mask line by an argument about output format, not by running it; that comparison is open.

Conventions and evaluation. That removal-based explanations are defined by how features are removed, what is predicted and how it is summarised is the organising view of Covert et al. (2021); that the masking baseline changes conditional Shapley values is known (Aas et al., 2021; Olsen et al., 2024), as is the unreliability of perturbation-based faithfulness metrics (Barr et al., 2023; Zhang & Mcmillan, 2025; Oliveira e Silva et al., 2026). We make both operational for time series: we measure how far the coalition measure and the baseline move the conclusions, *rankings* included (§7), and give a controlled demonstration (§8) in which keep-loss does not separate an oracle attribution from a random one; Rao (2025) reaches a related ceiling from algorithmic information theory.

3 Setup

Let $x \in \mathbb{R}^{T\times D}$ and $P = TD$. A coalition $S \subseteq [P]$ induces $v(S) = g(f(x \odot m_S + b \odot (1 - m_S)))$ for a masking baseline b and an output scalarisation g; the conditional alternative to marginal masking is itself a live debate (Aas et al., 2021; Olsen et al., 2024). Coalitions are drawn i.i.d. Bernoulli(p) per player. Writing $z_i = (s_i - p)/\sqrt{p(1-p)}$, the products $\chi_A(z) = \prod_{i\in A} z_i$ are orthonormal, so

$$v = \sum_{A\subseteq[P]} \hat{v}(A)\,\chi_A, \qquad \mathrm{Var}(v) = \sum_{A\neq\emptyset} \hat{v}(A)^2. \tag{1}$$

We work with the centred game throughout: the constant $\hat{v}(\emptyset)$ is free to any explanation and is excluded from every ratio below. Recent time series attribution work varies the player set (Kim & Park, 2026; Huang et al., 2025), the training objective (Cheng et al., 2025), or abandons Shapley for learned masks (Liu et al., 2024a;b); all of them fix (b, g, p) without comment. Every quantity below is defined with respect to the triple (b, g, p); §7 is about what happens when it changes. Unless stated otherwise $b = 0$, g is the logit of the class predicted on the unmasked input, and $p = 1/2$.

Models and data. Nine trained black boxes: a GRU and a TCN on MIMIC-III mortality (Johnson et al., 2016) (T=48, D=16) and GRUs on seven UEA archives (Bagnall et al., 2018) spanning P from 16 to 600. Each experiment names its subset: the full pairwise basis is affordable only for $P \leq 300$ (five models), and the coalition-measure sweep uses seven. Every model is reported with its test accuracy against chance (Appendix C), since every measurement below is meaningless on a chance-level black box.

4 Explanations as lossy codes

An explanation that emits R numbers is a rate-R code for v. By (1) the distortion of a coefficient subset $\mathcal{C}$ is

$$D(\mathcal{C}) = 1 - \frac{\sum_{A\in\mathcal{C},\, A\neq\emptyset} \hat{v}(A)^2}{\mathrm{Var}(v)}, \tag{2}$$

the unexplained variance of the fit restricted to $\mathcal{C}$; the best *sparse* code at rate R keeps the R largest coefficients. Coefficients are least squares on n sampled coalitions (Appendix B), rescored on fresh ones in §6. The envelope is a reference, not a universal bound: a factored code spends Pr numbers on all P^2 interaction entries at once and can sit below it, which we observe.

Additive attribution is inefficient at its own rate. The additive class has rate P and its distortion is $1 - \mathrm{AER}$, where AER is the degree-≤ 1 energy share. Table 1 shows it sits 0.07–0.13 above the sparse envelope at the same rate on every dataset: on ERing the best 260 *mixed-order* coefficients reach distortion 0.087 where all 260 linear terms reach 0.156. AER bounds what the additive class can do; it does not make that class a good use of the budget.

Table 1: Distortion at a given rate (numbers emitted per input), full cell-level pairwise bases, 16 inputs, $p = 1/2$. **Bold**: lowest of the four structured classes; all lie above the degree-2 floor. In-sample on these inputs, so the additive entry need not equal $1 - \mathrm{AER}$ of the 8-input atlas.

	rate / distortion				degree-2
dataset	additive	rank-1	banded($\ell \leq 1$)	Kronecker-1	floor
PenDigits (P=16)	16 / 0.347	32 / 0.267	52 / **0.205**	84 / 0.232	0.119
Libras (P=90)	90 / 0.445	180 / **0.318**	311 / 0.396	2119 / 0.303	0.182
RacketSports (P=180)	180 / 0.349	360 / **0.174**	1674 / 0.263	1116 / 0.286	0.100
ERing (P=260)	260 / 0.156	520 / **0.047**	1674 / 0.131	4501 / 0.074	0.022
JapaneseVowels (P=300)	300 / 0.522	600 / **0.290**	5406 / 0.466	1069 / 0.403	0.179

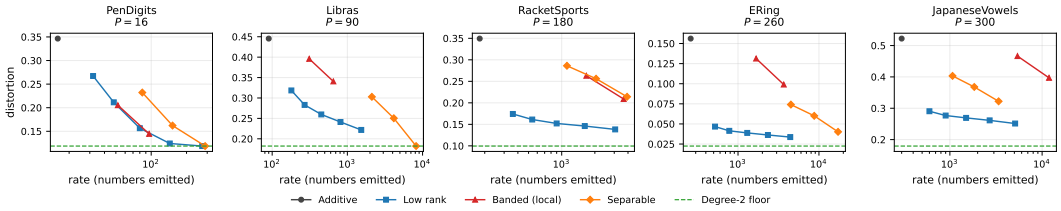


Figure 1: Distortion against rate for every structural class, five real models, full cell-level pairwise bases. Low rank (blue) dominates locality (red) and separability (orange) everywhere except PenDigits, whose $T = 8$ makes a lag-1 band nearly global. The additive point sits far above the degree-2 floor (dashed).

The rate accounting applies to learned saliency maps, in one direction only. Methods that abandon Shapley for learned masks (Liu et al., 2024a;b) still emit one number per cell, so read through the additive decoder $b + \sum_i a_i z_i$ they are rate-P, degree-≤ 1 codes, and Table 1 bounds them as it bounds a Shapley value: no such code accounts for more than AER of $\mathrm{Var}(v)$, however its numbers are obtained. The bound is about that decoder; it says nothing about how well these methods find one informative coalition, which is what they optimise, only that the difference is invisible to a top-k perturbation metric (§8). We did not run them; the claim concerns their output format.

5 Which structure? Locality is the wrong one

A degree-2 explanation over cells needs $O(P^2)$ numbers—295,000 per MIMIC patient—so structure is mandatory. We test three hypotheses.

Within-timestep structure is nearly empty. Restricting pairs to those sharing a timestep—the structure the Owen value (Owen, 1977) exists to exploit—buys $+0.2$ to $+3.4$ points of explained variance across five datasets. This is consistent with our Owen-based estimator never outperforming a flat sampler, though not a proof of it, since Owen values also depend on how higher-order interactions straddle the groups: the object it uniquely determines is a split of a quantity that is almost absent.

Locality fails, and it fails because interaction is long-range. A banded basis over lags ≤ 1 buys $+2.5$ to $+14.2$ points. Sweeping the band shows the mass is present but distant: on CharacterTrajectories the gain rises monotonically $+2.8, +9.6, +15.2, +18.3$ for lags $\leq 1, 4, 10, 18$ and reaches $+28.6$ at all lags, without saturating. A lag-1 band covers only 5–12% of pairs on the longer sequences. Aligning a surrogate to the input graph, as tensor-network surrogates do (Heidari & Rabusseau, 2026), has a temporal counterpart in aligning it to the time axis; on our models that is what measurement refutes. Per-sample interaction indices (Fumagalli et al., 2023; 2024) assume no structure but cost $O(P^2)$ output.

Low rank is the right structure; separability is not. Truncating the pairwise coefficient matrix to rank r retains exactly the top-r eigenvalue energy. That is a matrix statement and a conservative one for the estimator's function class: since z_i^2 is constant at $p = 1/2$, a quadratic head with an intercept

Table 2: Held-out distortion (rate / mean over 8 fresh inputs, 20,000 fresh coalitions) of each structure fitted as a decoder; bold is the lowest in the row. Rank-1 beats the lag-1 band on four of five models at fewer numbers, and full degree-2 is worse held-out than rank-4 on four of five.

dataset (P)	additive	banded($\ell \leq 1$)	Kronecker-1	rank-1	rank-1, free diag.	rank-4, free diag.	full degree-2
PenDigits (16)	16 / 0.303	52 / 0.178	84 / 0.214	32 / 0.225	32 / 0.212	80 / 0.130	136 / **0.114**
Libras (90)	90 / 0.384	311 / 0.355	2119 / 0.321	180 / 0.312	180 / 0.311	450 / **0.278**	4095 / 0.334
RacketSports (180)	180 / 0.350	1674 / 0.278	1116 / 0.291	360 / 0.138	360 / 0.135	900 / **0.130**	16290 / 0.166
ERing (260)	260 / 0.115	1674 / 0.093	4501 / 0.056	520 / 0.031	520 / 0.030	1300 / **0.019**	33930 / 0.024
JapaneseVowels (300)	300 / 0.491	5406 / 0.446	1069 / 0.340	600 / 0.250	600 / 0.249	1500 / **0.222**	45150 / 0.282

realises the off-diagonal of any rank-r matrix and can beat eigen-truncation (it fits the single-pair game $z_1 z_2 = (z_1 + z_2)^2/2 - 1$ exactly, where rank-1 truncation keeps half the energy), so the rank-r points below are upper bounds on the distortion that class attains. Rank-1 beats banded on four of five datasets using 1.7–9.0$\times$ fewer numbers (Table 1). The exception, PenDigits, confirms the mechanism: at $T = 8$ a lag-1 band already covers 43% of all pairs, so "local" is nearly "global". A separable $A \otimes B$ hypothesis—one factor over time, one over features, fitted by the rearrangement of Van Loan & Pitsianis (1993)—was our own prediction and is refuted: it loses to rank-1 on every dataset once rate is counted, on ERing by 8.7$\times$ in rate (0.074 at 4501 numbers against rank-1's 0.047 at 520). The ordering is not an in-sample artefact: rebuilt as fitted decoders and scored on fresh coalitions of eight inputs disjoint from Table 1 (Table 2), rank-1 still beats the lag-1 band on the same four models at the same rates (0.312 against 0.355 on Libras, 0.138 against 0.278 on RacketSports, 0.031 against 0.093 on ERing, 0.250 against 0.446 on JapaneseVowels), PenDigits is still the exception, and the full degree-2 fit is worse held-out than rank-4 on four of five models; the free-diagonal class the estimator's head realises is never worse than eigen-truncation (Table 2).

6 Optimal explanation complexity is finite

Each estimated coefficient removes bias worth its own energy share but adds variance of about $1/n$ for n coalitions, the hard-thresholding trade-off of Donoho & Johnstone (1994), so beyond some rate a longer explanation fits noise. Refitting on n coalitions and scoring on fresh ones confirms an interior optimum that grows with the budget: on RacketSports the best rate, chosen on a validation draw and scored on a separate test draw, moves $762 \to 1271 \to 1874$ as n goes $32.6\text{k} \to 65.2\text{k} \to 162.9\text{k}$, and at every budget the truncated explanation beats the full one, by 40% at the smallest (0.259 against 0.431); choosing on the test draw gives the same rates, so this is not a selection artifact.

The interior optimum appears once more in an experiment sharing no machinery with the above: a per-input rank-r fit on MIMIC degrades with rank at 4,000 coalitions $(0.788, 0.726, 0.663)$ and stops degrading at 30,000 $(0.942, 0.946, 0.945)$; a third experiment shows the budget side only (CharacterTrajectories, $0.352 \to 0.438 \to 0.446$ at $30\text{k} \to 100\text{k} \to 200\text{k}$). This is the paper's one directly actionable rule: *choose explanation complexity from the query budget*, and it explains without appeal to metric failure why short additive explanations are hard to beat in the small-budget regime the literature works in.

7 The game is a gauge

The coalition measure. With (b, g) fixed the game v is fixed, and so is its Shapley value; what p changes is the inner product under which v is projected onto the additive subspace, hence the order-1 coefficients and AER, not the game. The Fourier coefficients of (1) and the Möbius (Harsanyi) coefficients are one function in two bases; we use the orthonormal one. AER is defined at $p = 1/2$, the measure maximising each player's indicator variance. It does not follow that this is the measure most favourable to an additive fit; quantitative genetics, using the same decomposition over an allele frequency, finds the additive share largest at *extreme* frequencies (Hill et al., 2008). Our sweep agrees. In none of seven models is $p = 1/2$ the maximum (Table 3); it is a strict minimum in three and the response is monotone across the sweep in the other four. The swing exceeds 30 points on two models, enough that JapaneseVowels changes from the most interaction-dominated model in our atlas to comparable with MIMIC/TCN, while MIMIC/RNN moves 3.8; not every model is affected, and one cannot tell which without measuring.

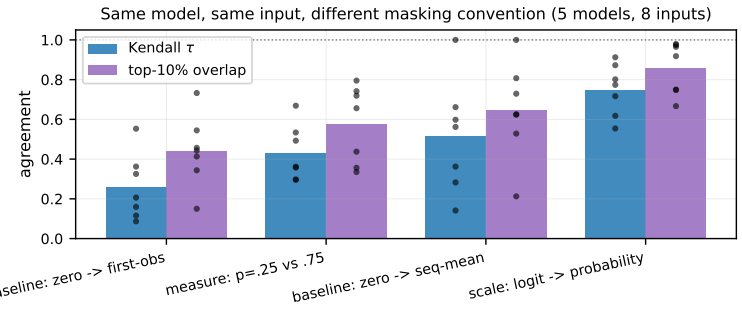


Figure 2: Agreement between order-1 attributions of the same model and input computed under different conventions. Bars are means over five models, dots are individual models. The masking baseline reorders the explanation; the output scale barely does.

Table 3: Additive share under a Bernoulli(p) coalition measure, cell level, 8 inputs. Swing is the signed range over the three p; the top two rows are not monotone in it.

model	p=0.25	p=0.5	p=0.75	swing
UEA-JapaneseVowels	0.410	0.370	0.694	+32.4
UEA-RacketSports	0.752	0.576	0.444	−30.9
UEA-PenDigits	0.691	0.734	0.793	+10.3
UEA-ERing	0.858	0.846	0.909	+6.2
MIMIC / TCN	0.789	0.799	0.849	+6.0
MIMIC / RNN	0.946	0.908	0.916	−3.8
UEA-Libras	0.556	0.541	0.523	−3.3

Rankings are not the invariant part. We expected values to be convention-dependent and rankings to survive, which would have justified the field's rank-based metrics. Computing order-1 attributions of the same model and input under six conventions and comparing pairwise, the mean Kendall τ is $0.598, 0.542, 0.393, 0.310$ and 0.219 on RacketSports, ERing, MIMIC/TCN, MIMIC/RNN and JapaneseVowels; on the last, the top-30 cells share fewer than a third of their members.

Decomposing by axis, the masking baseline dominates and the output scale is benign—the reverse of what we assumed: zero→first-observation gives $\tau = 0.267$ and top-k overlap 0.473; p=0.25 vs p=0.75 gives 0.441; zero→sequence-mean 0.478; logit→probability 0.783 with overlap 0.895. *The axis nobody tunes is the one that reorders the explanation.*

Consequence. "Which attribution is correct" has no convention-free answer. What can be reported is a declared convention plus a measured fidelity inside it (§9); §4–6 are untouched, since distortion is defined relative to a stated game.

8 Perturbation metrics do not resolve the distinction they adjudicate

We train three explainers differing only in their coalition objective on a synthetic benchmark with exact Owen ground truth for the trained model's own game, and score them on every metric family. Over five seeds and two operating points, a paired test of Owen-consistent against flat training separates them on ground-truth Spearman at both ($p = 0.003, 0.007$); of the twelve perturbation-metric tests (keep and drop loss, insertion and deletion AUC, AUPD, AUPP) two reach $p < 0.05$ uncorrected ($0.017, 0.023$) and none survives a Bonferroni correction over the twelve (Appendix H). *Only the ground-truth-referenced metric reliably resolves the objectives.*

Since ground-truth Spearman is itself a rank correlation, the split is *ground-truth-referenced versus perturbation-proxy*, not values versus rankings.

Controls make this concrete rather than a null result. Submitting the exact Owen values as the attribution, and separately random attributions, AUPP and AUPD separate them at both operating points (paired $p \leq 0.008$), so the evaluation is not broken. But **keep loss does not separate exact ground truth from random attributions** in the interaction-dominated regime (oracle minus random -0.26, 95% CI $[-0.52, +0.01]$, $p = 0.058$), and insertion AUC fails the same test at the other operating point ($p = 0.12$). These are the two metrics this literature leads with. This is not evidence of equivalence—we fixed no equivalence margin, and five seeds are few—but a metric that cannot place an oracle outside random noise at a budget where AUPP does so at $p < 10^{-4}$ cannot be the one that adjudicates.

9 An amortized estimator with a measured fidelity

Given §7, an estimator should not assert that its values are correct. It should state its convention and measure its own fidelity. We therefore predict the coefficients of a structured surrogate rather than an attribution:

$$\hat{v}(S \mid x) = b(x) + \phi(x)^\top z + \tfrac{1}{2P} \sum_{k=1}^{r} s_k(x) \left(u_k(x)^\top z\right)^2, \tag{3}$$

with (ϕ, U, s, b) produced in one forward pass. Training uses gradient clipping at norm 1, without which a fifth of our configurations diverge. Because $\hat{v}$ is a degree-2 game, its Shapley, Owen and Banzhaf values are fixed linear maps of (ϕ, U, s) that coincide, and at $p = 1/2$ reduce to $2\phi_i$ (Appendix B.3); they are exact for the surrogate, not for v, and the interaction head changes fidelity and pairwise indices, not the order-1 ranking. And because $\hat{v}$ predicts the value function, it can be scored on held-out coalitions *of the input being explained*, giving each explanation a measured R^2 that needs no ground-truth attribution — but does need black-box evaluations of that input, which we price below.

Depth, not amortization, was the binding constraint—within one family. The amortization gap does not grow with P; within the dilated-convolution family it grows while the trunk's receptive field is shorter than T. At three layers (field 15) the surrogate on CharacterTrajectories (T=119) is worse than predicting the mean ($R^2 = -0.396$); at six layers (field 127) it reaches 0.272 and at seven, 0.314, closing the gap from $+0.771$ to $+0.061$; MIMIC (T=48) falls from $+0.263$ to $+0.076$, and PenDigits (T=8), already covered at three layers, is the control and barely moves (Table A.10).

Reach is necessary; capacity is conditional. Setting dilation independently of depth separates the two (Appendix I.2): at a field shorter than the series, adding depth makes things slightly worse (-0.063), while at an adequate field it helps ($+0.227$), and the reach main effect is $7\times$ the capacity effect. This also explains why a one-layer bidirectional GRU and transformer, which have reach but not depth, reach only -0.146 and 0.067.

Training must be stabilised before architectures can be compared. Without gradient clipping 3 of 18 runs diverged and single-seed comparisons of architectures were unreliable; clipping at norm 1 removes every divergence and lifts the attention trunk on MIMIC from 0.026 to 0.893 at rank 4 (Appendix J). All numbers below use it.

The dataset's additive ceiling dominates the architecture choice. With training stabilised (Table 4, three seeds each), the rank-4 head buys $+0.19$ to $+0.25$ on the low-AER model and $+0.006$ to $+0.020$ on the high-AER one: a between-dataset gap of 0.213, $3.6\times$ the largest between-architecture spread on either dataset (0.058), even though absolute additive fidelity varies by 0.18 across architectures on CharacterTrajectories. **The encoder strongly affects how well the surrogate fits and barely affects what the interaction head is worth.** This is not invariance: on MIMIC the GRU's $+0.020 \pm 0.006$ is 3.2 standard deviations from zero, so the head buys a little there.

The interaction head's payoff tracks the additive ceiling. Over three training seeds with the convolutional trunk, the rank-4 head's gain orders inversely with AER: $+0.227 \pm 0.007$ at AER 0.404, $+0.110 \pm 0.017$ at 0.734, and $+0.006 \pm 0.004$ at 0.907. The first two are 35 and 6.6 standard deviations from zero; the third is 1.3, that is, barely distinguishable from no effect — which is what the framework predicts for a model whose additive ceiling is already 0.907, and not the small negative

Table 4: Three encoder families on the low-AER model (CharacterTrajectories, T=119) and the high-AER control (MIMIC, T=48), three seeds each, gradient clipping at norm 1.

dataset	trunk	additive head	rank-4 head	head gain
CharTraj (AER 0.404)	dilated conv (7)	0.213 ± 0.018	$\mathbf{0.441 \pm 0.023}$	$+0.2272 \pm 0.0065$
	GRU (2)	0.036 ± 0.016	0.228 ± 0.022	$+0.1924 \pm 0.0349$
	attention (2)	0.123 ± 0.036	0.374 ± 0.011	$\mathbf{+0.2508 \pm 0.0476}$
MIMIC / GRU (AER 0.907)	dilated conv (6)	0.870 ± 0.003	0.876 ± 0.004	$+0.0055 \pm 0.0044$
	GRU (2)	0.853 ± 0.005	0.873 ± 0.007	$+0.0200 \pm 0.0061$
	attention (2)	$\mathbf{0.884 \pm 0.003}$	$\mathbf{0.892 \pm 0.002}$	$+0.0076 \pm 0.0035$

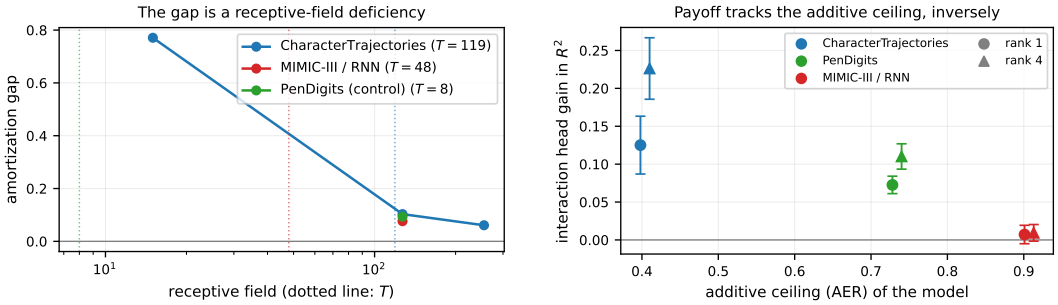


Figure 3: **Left:** within the convolutional family the amortization gap closes once the receptive field reaches T (dotted lines); PenDigits, already covered at three layers, is the control and barely moves. **Right:** with training stabilised, what the interaction head buys orders inversely with the model's additive ceiling; error bars are over three seeds and MIMIC's interval is close to zero.

gain an unclipped single seed had suggested. On CharacterTrajectories the amortized surrogate reaches $R^2 = 0.441$, above that model's additive ceiling of 0.404, so it captures structure no additive attribution could.

What amortization is worth. At rank 4 the amortized surrogate beat its per-input reference, but only because that reference was under-budgeted; raising its coalition budget restores its lead ($0.352 \to 0.438 \to 0.446$ at 30k $\to$ 100k $\to$ 200k). The defensible claim is conditional—amortization wins where the per-input fit is under-budgeted—and it yields an exchange rate: the stabilised rank-4 surrogate (0.441 ± 0.023, Table 4) emits its coefficients with no inference-time queries and matches a per-input fit only once that fit has more than 100,000, reaching 99% of its 200,000-query value. Training that surrogate costs 1.6×10^7 masked evaluations once per predictor (Appendix I.1).

The certificate flags the inputs the surrogate fails to capture. A measured fidelity number is only worth shipping if it tells the reader something, so we test the sharpest candidate available: does a low certificate flag the inputs whose explanation is convention-dependent? For each of 32 evaluation inputs we record two quantities independently. The *certificate* is the surrogate's R^2 on 8,000 held-out coalitions of that input: it needs no ground truth, but each coalition is one black-box evaluation. That audit is cheap to shrink: 500 coalitions reproduce the 8,000-coalition certificate at Spearman ≥ 0.90 on every model and keep the correlation below, but not every failure flag: one JapaneseVowels input passes at 500 and fails at 8,000 (Table A.15). The *gauge agreement* is the mean pairwise Kendall τ between that input's order-1 attributions under the six conventions of §7, which costs six full attribution runs. Across six models at three seeds each (Table A.16) the two rank-correlate wherever the estimator fits, and nowhere else. It is positive on all 12 seeds of the four models whose surrogate fits at least 26 of 32 inputs ($\rho = 0.41$–0.43 on MIMIC/RNN, 0.39–0.58 on CharacterTrajectories, both significant every seed; 0.32–0.36 on MIMIC/TCN, 0.19–0.49 on RacketSports, one seed of three each) and absent on the two it fits worst, Libras (-0.12 to -0.17) and JapaneseVowels (-0.07 to $+0.13$).

The ordering is exactly the fitted fraction: 32/32 on MIMIC/TCN, 31 on MIMIC/RNN, 29–30 on CharacterTrajectories, 26–27 on RacketSports, then 14–18 on Libras and 5–9 on JapaneseVowels

(median per-input R^2 -0.04 and -2.8). On those two the estimator fails and the certificate is uninformative because it reports exactly that—the point of shipping one. Neither scale nor data budget explains the failures: Libras has the *smallest* P (90) and two training inputs per player, and cutting CharacterTrajectories' pool $5\times$ leaves it unaffected (Table A.16). The failure is in amortization, not in the function class: on the same JapaneseVowels inputs a rank-4 surrogate fitted per input (about 1.8×10^5 queries each) reaches held-out distortion 0.18 with every input fitted, against about 4 amortized, and refitting just six amortized numbers—intercept, additive scale and four factor scales—on 750 queries per input brings it to 0.92 (Appendix I.6). The failure is mostly miscalibration of a few coefficients, not wrong directions. Whether amortization pays is model-dependent: at zero queries its coefficients beat a 750-query additive fit from scratch on MIMIC (0.13 against 0.50) and CharacterTrajectories (0.58 against 0.94), but not on JapaneseVowels (3.8 against 0.71).

The certificate is not per-input additivity restated. On MIMIC the two are collinear ($\rho = 0.74$) and partialling either on the other leaves neither significant; on CharacterTrajectories they are nearly independent ($\rho = 0.16$), yet each predicts gauge agreement at $\rho \approx 0.51$ and survives partialling the other ($p < 0.003$; Table A.17).

10 Negative results

Owen values. An amortized Owen-value estimator loses to a multi-scale sampler on MIMIC in $3/3$ seeds (0.319 against 0.294 keep loss); §5 gives the likely reason: the within-timestep interaction it uniquely resolves is nearly absent. **A direct keep-loss optimiser** failed to beat attribution methods across three relaxations, so our critique of that metric rests on the controls in §8, not on gameability. **AER as a predictor** of which method wins is unsupported; it is a ceiling only.

11 Limitations

Our estimator results use three seeds and three encoder families, with 16 evaluation inputs, and the certificate–gauge correlation is measured at one rank on 32 inputs of six models, and its mechanism separates from per-input additivity on one of the two where we could test both. The estimator fits four of those six; on Libras and JapaneseVowels it fails, and we have no rule that predicts in advance which models it will fit. Using the audit to choose which explanations to repair did not beat a random choice of the same cost (Appendix I.6). Absolute fidelity remains modest where interaction is strong (0.44 on CharacterTrajectories against a degree-2 reference of 0.45). The rate–distortion measurements need a full pairwise basis, so are confined to $P \leq 300$. Finally, the coalition-measure sweep rests on 8 inputs per model at a single seed, so the per-model swings in Table 3 carry a sampling error we do not quantify. The certificate inputs are the first 32 of each test split, not a random draw, and the correlations are exploratory: seeds re-fit the explainer on the same inputs and predictor, so they are not independent replications. Distortion does not measure whether a person can use the object, and a low certificate flags an estimator failure or a hard input under one convention, not an ill-posed explanation.

12 Conclusion

"Which attribution is correct" has no convention-free answer for time series models, and the perturbation metrics used to adjudicate it did not separate the candidates on a game with exact ground truth. What does have an answer is how much of a declared value function an explanation of a given size can carry. On the models we measured, additivity is an inefficient use of the budget, locality is the wrong structure, low rank is the right one, and the optimal complexity is finite and set by the query budget. Four changes to practice follow, none needing a new solution concept: declare the gauge (b, g, p), report the rate, set it from the query budget, and ship a priced fidelity number with each explanation.

AI USE STATEMENT

We used generative AI tools throughout this work: to search and summarise related literature, to propose research directions and the falsifiable hypotheses recorded in §K, to write and revise the measurement and estimator code, to run and analyse the experiments, and to draft and revise this paper, including this statement. We did not use generative AI to produce any reported number. Every figure and table here was written by a script in the supplementary material, run against the stated checkpoints and data, and the CSVs those runs emitted are shipped so that each number can be traced to the run that produced it; §A maps every claim to the script and artifact behind it. We have reviewed all AI-assisted work. Claims were checked against the artifacts that produced them rather than against the text asserting them, and §L lists the safeguards that checking rests on. Every reference was verified against the publisher record where one exists and the preprint abstract page otherwise. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

One of our datasets is a restricted-access clinical database of intensive-care admissions, used under its credentialed data-use agreement; no raw or preprocessed patient series are included in the supplementary material, and the released artifacts contain only aggregate measurements from which the reported numbers can be checked. The remaining datasets are public archives. We note that the results here are largely negative about the reliability of time series attribution, and that the setting in which such attributions are most often proposed is clinical: our finding that neither attributed values nor their rankings survive a change of masking convention is a reason for caution in exactly that setting, and we would not want a reader to take a single convention's ranking as a statement about a patient.

REPRODUCIBILITY STATEMENT

The supplementary material contains the library, the measurement scripts, the configurations, and the CSV outputs that every table and figure is built from. §A gives a claim-to-script map; §B states the estimator, the basis, the degree-of-freedom adjustment and the numerical choices the results depend on; §C reports each trained black box with its test accuracy against chance. Release mode, described in the supplementary README, makes a missing or shape-incompatible checkpoint raise rather than fall back to random weights, and the measurement scripts enable it by default; no number reported here was produced with it disabled. The clinical dataset requires its own credentialed access, so those runs are reproducible only by a reader who holds it — the released CSVs let the reported clinical numbers be checked without it.

REFERENCES

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021. doi: 10.1016/j.artint.2021.103502.

Amr Alkhatib, Henrik Boström, and Ulf Johansson. Estimating quality of approximated Shapley values using conformal prediction. In *Symposium on Conformal and Probabilistic Prediction with Applications (COPA)*, volume 230 of *PMLR*, pp. 158–174, 2024.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018. v1.

Brian Barr, Noah Fatsi, Leif Hancox-Li, Peter Richter, Daniel Proano, and Caleb Mok. The disagreement problem in faithfulness metrics. In *XAI in Action: Past, Present, and Future Applications (NeurIPS 2023 Workshop)*, 2023. arXiv:2311.07763.

Sebastian Bordt and Ulrike von Luxburg. From Shapley values to generalized additive models and back. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *PMLR*, pp. 709–745, 2023.

Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2426.

Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-Shapley and C-Shapley: Efficient model interpretation for structured data. In *International Conference on Learning Representations*, 2019. arXiv:1808.02610.

Lu Chen, Siyu Lou, Keyan Zhang, Jin Huang, and Quanshi Zhang. HarsanyiNet: Computing accurate Shapley values in a single forward propagation. In *International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 4804–4825, 2023. arXiv:2304.01811.

Qisen Cheng, Jinming Xing, Chang Xue, and Xiaoran Yang. Unifying prediction and explanation in time-series transformers via Shapley-based pretraining. In *21st IEEE International Colloquium on Signal Processing and its Applications (CSPA)*, 2025. arXiv:2501.15070.

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.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. doi: 10.1093/biomet/81.3.425.

James Enouen and Yan Liu. InstaSHAP: Interpretable additive models explain Shapley values instantly. In *International Conference on Learning Representations*, 2025. arXiv:2502.14177.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order Shapley interactions. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2303.01179.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. KernelSHAP-IQ: Weighted least square optimization for Shapley interactions. In *International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 14308–14342, 2024. arXiv:2405.10852.

Farzaneh Heidari and Guillaume Rabusseau. TN-SHAP-G: Graph-structured tensor network surrogates for Shapley values and interactions. In *International Conference on Machine Learning*, 2026. arXiv:2606.01540.

William G. Hill, Michael E. Goddard, and Peter M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4(2):e1000008, 2008. doi: 10.1371/journal.pgen.1000008.

Bosong Huang, Ming Jin, Yuxuan Liang, Johan Barthelemy, Debo Cheng, Qingsong Wen, Chenghao Liu, and Shirui Pan. ShapeX: Shapelet-driven post hoc explanations for time series classification models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2510.20084.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time Shapley value estimation. In *International Conference on Learning Representations*, 2022. arXiv:2107.07436.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Jinwoong Kim and Sangjin Park. GroupSegment-SHAP: Shapley value explanations with group-segment players for multivariate time series. *arXiv preprint arXiv:2601.06114*, 2026. v2.

Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. A rate-distortion framework for explaining black-box model decisions. In *xxAI – Beyond Explainable AI*, volume 13200 of *Lecture Notes in Computer Science*, pp. 91–115. Springer, 2022. doi: 10.1007/978-3-031-04083-2_6.

I. Elizabeth Kumar, Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle A. Friedler. Shapley residuals: Quantifying the limits of the Shapley value for explanations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Zichuan Liu, Tianchun Wang, Jimeng Shi, Xu Zheng, Zhuomin Chen, Lei Song, Wenqian Dong, Jayantha Obeysekera, Farhad Shirani, and Dongsheng Luo. TimeX++: Learning time-series explanations with information bottleneck. In *International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 32062–32082, 2024a. arXiv:2405.09308.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *International Conference on Learning Representations*, 2024b. arXiv:2401.08552.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jan Macdonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions. *arXiv preprint arXiv:1905.11092*, 2019. v1.

Thach Le Nguyen and Georgiana Ifrim. TSHAP: Fast and exact SHAP for explaining time series classification and regression. In *Machine Learning and Knowledge Discovery in Databases. Research Track (ECML PKDD)*, Lecture Notes in Computer Science, pp. 60–77. Springer, 2025. doi: 10.1007/978-3-032-06078-5_4.

Inês Oliveira e Silva, Sérgio Jesus, Iker Perez, Rita P. Ribeiro, Carlos Soares, Hugo Ferreira, and Pedro Bizarro. Rethinking XAI evaluation: A human-centered audit of Shapley benchmarks in high-stakes settings. *arXiv preprint arXiv:2604.22662*, 2026. v1.

Lars Henry Berge Olsen, Ingrid Kristine Glad, Martin Jullum, and Kjersti Aas. A comparative study of methods for estimating model-agnostic Shapley value explanations. *Data Mining and Knowledge Discovery*, 38:1782–1829, 2024. doi: 10.1007/s10618-024-01016-z. Preprint: arXiv:2305.09536, under the title "A Comparative Study of Methods for Estimating Conditional Shapley Values and When to Use Them".

Guillermo Owen. Values of games with a priori unions. In Rudolf Henn and Otto Moeschlin (eds.), *Mathematical Economics and Game Theory: Essays in Honor of Oskar Morgenstern*, volume 141 of *Lecture Notes in Economics and Mathematical Systems*, pp. 76–88. Springer, 1977.

Rafael Poyiadzi, Xavier Renard, Thibault Laugel, Raul Santos-Rodriguez, and Marcin Detyniecki. Understanding surrogate explanations: The interplay between complexity, fidelity and coverage. *arXiv preprint arXiv:2107.04309*, 2021. v1.

Shrisha Rao. The limits of AI explainability: An algorithmic information theory approach. *arXiv preprint arXiv:2504.20676*, 2025. v2.

Steffen Rendle. Factorization machines. In *IEEE International Conference on Data Mining*, pp. 995–1000, 2010. doi: 10.1109/ICDM.2010.127.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, number 28 in Annals of Mathematics Studies, pp. 307–317. Princeton University Press, 1953.

Santo M. A. R. Thies, Hubert Baniecki, R. Teal Witter, Eyke Hüllermeier, Maximilian Muschalik, and Fabian Fumagalli. Proxy-based approximation of Shapley and Banzhaf interactions. *arXiv preprint arXiv:2605.22738*, 2026. v2.

Charles F. Van Loan and Nikos Pitsianis. Approximation with Kronecker products. In Marc S. Moonen, Gene H. Golub, and Bart L. R. De Moor (eds.), *Linear Algebra for Large Scale and Real-Time Applications*, volume 232 of *NATO ASI Series E: Applied Sciences*, pp. 293–314. Springer, 1993.

Linxiao Yang, Yunze Tong, Xinyue Gu, and Liang Sun. Explain temporal black-box models via functional decomposition. In *International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 56448–56464, 2024.

Siyu Zhang and Kenneth Mcmillan. Beyond output faithfulness: Learning attributions that preserve computational pathways. *arXiv preprint arXiv:2509.04588*, 2025. v2.

A Artifact provenance

Every number in the main paper is produced by a committed script and written to a CSV, so the two cannot drift apart. Figures are generated from those same CSVs by `scripts/generate_v2_figures.py`.

Table A.1: Claim → script → artifact. Scripts are under `scripts/` and artifacts under `results/`; both prefixes are omitted below.

claim	script	artifact
Additive ceiling (AER) atlas	`measure_aer`	`aer/aer_*`
Distortion–rate curves (§4, Fig. 1)	`measure_rate_distortion`	`rate_distortion/rd_*`
Structured ceilings, band sweep, spectra (§5)	`measure_interaction_ceiling`	`aer/ceiling_*`, `aer/lowrank_*`
Held-out complexity curves (§6)	`measure_rate_distortion +rd.generalization=true`	`rate_distortion/gen_*`
Coalition-measure sweep (§7)	`measure_aer +aer.coalition_p`	`aer/aer_*_p0.25`, `aer/aer_*_p0.75`
Gauge agreement (§7, Fig. 2)	`measure_gauge_invariance`	`gauge/gauge_*`
Metric separation and controls (§8)	`run_value_rank_dissociation`	`dissociation/*`
Amortization gap (§9, Fig. 3)	`run_amortization_gap`	`amortization/gap_*`
Certificate utility (§9, Fig. A.1)	`run_certificate_utility`, `measure_per_input_aer`	`certificate/cert_*`, `aer_control_*`
Held-out structured decoders, fresh inputs (§5, Table 2)	`measure_heldout_structures`	`rate_distortion/heldout_*`
Amortization vs. per-input fit; repair policies (§9, Table A.18)	`run_repair_pilot`, `analyze_repair_pilot`	`repair/repair_*`
Certificate audit budget (§9, Table A.15)	`run_certificate_utility +cert.n_test_coalitions`	`certificate/cert_audit_budget`
Paired seed-level tests (§8)	`dissociation/*` (derived)	`dissociation/paired_contrasts`

B Estimators and numerical details

B.1 The p-biased orthonormal basis

For inclusion probability p we encode a coalition as $z_i = (s_i - p)/\sqrt{p(1-p)}$, which is mean zero and unit variance under Bernoulli(p), so products over distinct players are orthonormal for any p. At $p = 1/2$ this reduces to the Walsh ± 1 encoding. AER and the structured ceilings of §5 are degree-of-freedom adjusted R^2; the raw statistic is badly inflated once the column count approaches the number of coalitions (on MIMIC's 768 players, raw R^2 is 0.954 at $2{,}000$ coalitions against 0.901 at $8{,}000$). The distortion–rate curves of §4 and §6 are in-sample energy shares of the least-squares coefficients and carry estimation noise, which is exactly why §6 rescores them on fresh coalitions and chooses the rate on a third, separate draw.

B.2 FITTING

The design has $1 + P + |\mathcal{C}_2|$ columns, up to $45{,}151$ for JapaneseVowels. We accumulate the Gram matrix on the GPU in float32 — its entries are integers bounded by the sample count, so this is exact — and solve by Cholesky, falling back to a rank-revealing least squares when the factorisation fails, which happens exactly when a player is never toggled or columns are duplicated. Three details matter in practice and each was a bug before it was a detail:

- The Gram must be accumulated in place (`addmm_`); the naive form materialises a second $C \times C$ temporary and puts wide designs out of reach of a 12 GB card.
- The widening to float64 must happen after the transfer to host memory, not before.
- The residual sum of squares is a difference of large nearly-equal sums, so on a near-perfect fit cancellation can drive it negative and report a ratio above 1. We observed 1.0177 before clamping at zero.

B.3 FROM SURROGATE TO ATTRIBUTION

Write the estimator's pairwise coefficient matrix as $M = \frac{1}{P}\,\mathrm{offdiag}\left(\sum_k s_k u_k u_k^\top\right)$, so that in the z basis $\hat{v} = b' + \phi^\top z + \sum_{i<j} M_{ij} z_i z_j$ with the diagonal of each $u_k u_k^\top$ absorbed into b' (at $p = 1/2$, $z_i^2 = 1$) or into the linear term (otherwise). Substituting $z_i = (s_i - p)/\sigma$, $\sigma = \sqrt{p(1-p)}$, gives the Möbius coefficients of the same degree-2 game on $s \in \{0,1\}^P$:

$$a_{ij} = M_{ij}/\sigma^2, \qquad a_i = \phi_i/\sigma - \frac{p}{\sigma^2}\sum_{j\neq i} M_{ij} + \frac{1-2p}{2P\sigma^2}\sum_k s_k u_{ki}^2. \tag{4}$$

For a game of degree at most 2 the Shapley, Banzhaf and Owen values coincide, since every pairwise term is split equally between its two players under each of them, and equal $a_i + \frac{1}{2}\sum_{j\neq i} a_{ij}$; the pairwise Shapley interaction index is a_{ij} itself. At $p = 1/2$ the pairwise contributions cancel and the attribution is simply

$$\mathrm{Sh}_i = \mathrm{Bz}_i = \mathrm{Ow}_i = 2\phi_i, \tag{5}$$

so the solution-concept choice is immaterial for the order-1 attribution of this surrogate and the interaction head changes the fidelity and the pairwise indices, not the ranking. For $p \neq 1/2$ the map stays linear in (ϕ, U, s) but the head enters the order-1 value through the bracketed terms. All of these are exact for $\hat{v}$; their accuracy for v is the certificate's R^2, not a theorem. As a check, the single-pair game $z_1 z_2$ has $a_1 = a_2 = -2$, $a_{12} = 4$ and $\mathrm{Sh}_1 = 0 = 2\phi_1$.

B.4 VALIDATION

The estimator is checked against constructed games rather than assumed correct. An exactly additive game must return 1.0 at every order and under every p; a game whose interaction is confined to single timesteps must be saturated by the within-timestep basis but *not* by a lag-1 basis, and a lag-1 game the reverse; an order-3 game must be bounded away from 1 at order 2 and saturated at order 3; one isolated pair splits its energy across a $\pm$ eigenvalue pair, so a rank-1 truncation must retain exactly 0.5 and rank-2 exactly 1.0, and two disjoint pairs must need rank 4. Two regression tests tie the distortion curve's additive point to an independently computed $1 - \mathrm{AER}$ and its floor to $1 - R^2$ of the degree-2 fit; these exist because a missing n factor in the explained sum of squares once made every reported distortion ≈ 1.0 while leaving monotonicity intact, so the earlier tests passed.

C MODELS AND DATASETS

AER spans 0.24 to 0.9998 across real trained models, so it discriminates rather than saturating. Two consequences worth stating. It is a property of the *model*, not only the data: on identical MIMIC inputs a GRU reaches 0.907 where a TCN sits at 0.798. And it retrodicts the uninformative synthetic benchmarks common in this literature — the synthetic GRU's game is additive to four decimal places, so no attribution method could have differed on it.

Table A.2: Trained black boxes. Accuracy is on the held-out test split (1,951 admissions for MIMIC, whose 10% positive rate makes AUROC the informative number there); every measurement is meaningless on a chance-level model, so we report it as a validity precondition. AER is the cell-level additive ceiling at $p = 1/2$.

dataset / model	T	D	P	test acc	chance	AER
UEA-PenDigits / GRU	8	2	16	0.988	0.100	0.734
UEA-Libras / GRU	45	2	90	0.772	0.067	0.539
UEA-RacketSports / GRU	30	6	180	0.743	0.250	0.578
UEA-ERing / GRU	65	4	260	0.696	0.167	0.846
UEA-JapaneseVowels / GRU	25	12	300	0.978	0.111	0.370
UEA-CharacterTrajectories / GRU	119	3	357	0.963	0.050	0.405
UEA-BasicMotions / GRU	100	6	600	0.675	0.250	0.239
MIMIC-III / TCN	48	16	768	0.904 (AUROC 0.826)	0.900	0.798
MIMIC-III / GRU	48	16	768	0.909 (AUROC 0.827)	0.900	0.907
Synthetic / TCN	50	10	500	—	—	0.979
Synthetic / GRU	50	10	500	—	—	0.9998

D FULL DISTORTION–RATE RESULTS

Table A.3 gives every class measured, extending the four shown in the main paper.

Table A.3: Rate / distortion by class, 16 inputs, $p = 1/2$. "envelope" is the best sparse code at the same rate; a negative gap means the class exploits structure sparse coding cannot.

dataset	class	rate	distortion	envelope	gap
PenDigits	additive	16	0.347	0.218	+0.129
	within-timestep	24	0.316	0.176	+0.140
	banded($\leq$1)	52	0.205	0.131	+0.074
	rank-1	32	0.267	0.153	+0.114
	rank-4	80	0.157	0.120	+0.036
	Kronecker-1	84	0.232	0.120	+0.112
Libras	additive	90	0.445	0.356	+0.089
	within-timestep	135	0.442	0.342	+0.100
	banded($\leq$1)	311	0.396	0.285	+0.111
	rank-1	180	0.318	0.309	+0.009
	rank-4	450	0.260	0.256	+0.004
	Kronecker-1	2119	0.303	0.191	+0.112
RacketSports	additive	180	0.349	0.227	+0.122
	within-timestep	630	0.314	0.177	+0.137
	banded($\leq$1)	1674	0.263	0.143	+0.120
	rank-1	360	0.174	0.194	**−0.020**
	rank-4	900	0.152	0.161	**−0.009**
	Kronecker-1	1116	0.286	0.161	+0.125
ERing	additive	260	0.156	0.087	+0.069
	within-timestep	650	0.151	0.059	+0.092
	banded($\leq$1)	1674	0.131	0.042	+0.090
	rank-1	520	0.047	0.062	**−0.015**
	rank-4	1300	0.039	0.043	**−0.004**
	Kronecker-1	4501	0.074	0.032	+0.042
JapaneseVowels	additive	300	0.522	0.454	+0.069
	within-timestep	1950	0.507	0.331	+0.177
	banded($\leq$1)	5406	0.466	0.264	+0.202
	rank-1	600	0.290	0.404	**−0.114**
	rank-4	1500	0.269	0.340	**−0.071**
	Kronecker-1	1069	0.403	0.376	+0.028

E STRUCTURE: BAND SWEEP AND SPECTRA

E.1 THE BAND SWEEP

A lag-1 band buys little, but the mass is present and distant. Gains in points of explained variance over the additive class:

Table A.4: Gain over additive as the temporal band widens.

band	CharacterTrajectories	JapaneseVowels	BasicMotions
within-timestep only	+0.2	+1.2	+1.9
$\ell \leq 1$	+2.8	+4.5	+3.8
$\ell \leq 3$	—	+11.3	+5.8
$\ell \leq 4$	+9.6	—	—
$\ell \leq 5$	—	+16.4	+7.0
$\ell \leq 10$	+15.2	—	—
$\ell \leq 18$	+18.3	—	—
all lags (timestep game)	+28.6	+19.8	+20.3

E.2 HOW THE SEPARABLE (KRONECKER) FIT IS COMPUTED

The separable class asks for the nearest $A \otimes B$ to the pairwise coefficient matrix, with A over timesteps and B over features. Indexing the matrix by cells (t, d) and rearranging it so that the (t, t') block becomes a row — $(t, d, t', d') \to (t, t', d, d')$, reshaped to $T^2 \times D^2$ — turns the problem into a rank-k SVD, which is the construction of Van Loan & Pitsianis (1993). The reported share is the energy the top-k singular values retain, and the rate charged is $k(T^2 + D^2)$, the size of the factors actually emitted.

E.3 SPECTRA

Share of pairwise energy retained by a rank-r truncation, timestep level, 8 inputs:

Table A.5: Rank-r energy share of the pairwise coefficient matrix (timestep game).

model	rank 1	rank 2	rank 4	rank 8	rank 16
UEA-JapaneseVowels	0.767	0.872	0.915	0.961	0.995
UEA-CharacterTrajectories	0.617	0.763	0.817	0.852	0.893
UEA-BasicMotions	0.613	0.677	0.725	0.769	0.833
MIMIC-III / GRU	0.531	0.841	0.947	0.977	0.991
MIMIC-III / TCN	0.245	0.417	0.634	0.818	0.945
Synthetic / GRU	0.621	0.817	0.918	0.969	0.991

The shape does not transfer to the cell game. Measured directly on JapaneseVowels (300 players, a 45,151-column design, 180,604 coalitions), the cell-level shares are $0.702, 0.726, 0.748, 0.773, 0.804, 0.844$ at ranks $1, 2, 4, 8, 16, 32$: one component already carries 70% and the spectrum is then nearly flat, where the timestep game reaches 0.961 by rank 8. Conclusions about rank must be drawn at the granularity the explanation is reported at.

E.4 HELD-OUT STRUCTURED DECODERS ON FRESH INPUTS

Table 1 scores each class by the in-sample energy share of its coefficients, and scores low rank by eigen-truncation. Neither is the error of a decoder one could ship. Here every class is rebuilt as a function, fitted on one draw of coalitions and scored on 20,000 fresh ones, on eight inputs (test-split indices 16–23) disjoint from those behind Table 1. "Free diagonal" is the class the estimator's quadratic head realises, since $z_1 z_2 = \frac{1}{2}\big((z_1 + z_2)^2 - 2\big)$ at rank 1 once the diagonal is unconstrained.

The full comparison is in Table 2 in the main text.

F HELD-OUT COMPLEXITY CURVES

Table A.6: Best rate against query budget. The rate is chosen on a validation draw of $20{,}000$ coalitions and D is scored on a separate test draw of $20{,}000$; choosing on the test draw gives the same rates to the nearest grid point. "predicted" is the count of coefficients above the noise floor $\mathrm{Var}(v)/n^2$; it captures the direction but overestimates the magnitude by 1.3–$3.5\times$ and does not track the budget, so we report it as a diagnostic rather than a calibrated rule.

dataset	n	predicted	best rate	D(best)	D(full)
RacketSports	32,582	2650	762	0.2587	0.4307
	65,164	2266	1271	0.2329	0.2836
	162,910	2502	1874	0.2216	0.2359
Libras	20,000	893	748	0.2801	0.2974
	40,960	1082	1124	0.2579	0.2618
PenDigits (control)	20,000	97	130	0.1125	0.1125

PenDigits is the control: with 136 coefficients and $20{,}000$ coalitions the design is $147\times$ overdetermined, nothing is worth truncating, and $D(\text{best}) = D(\text{full})$.

Counting coefficients above the noise floor at a fixed budget of $4\times$ the column count gives $97/136$ (71%) on PenDigits, $893/4095$ (22%) on Libras, $2266/16290$ (14%) on RacketSports, $1087/33930$ (3.2%) on ERing and $11442/45150$ (25%) on JapaneseVowels.

G GAUGE: FULL TABLES

Table A.7: Pairwise agreement between order-1 attributions under six conventions, averaged over the 15 pairs, 8 inputs, top-10% overlap.

model	Kendall τ	top-k overlap	relative L_2
UEA-RacketSports	0.598	0.760	0.357
UEA-ERing	0.542	0.707	0.529
MIMIC-III / TCN	0.393	0.694	0.478
MIMIC-III / GRU	0.310	0.632	0.520
UEA-JapaneseVowels	0.219	0.316	0.836
by axis (mean over 5 models)			
baseline: zero → first-observation	0.267	0.473	0.705
measure: p=0.25 vs 0.75	0.441	0.640	0.575
baseline: zero → sequence-mean	0.478	0.633	0.512
scale: logit → probability	0.783	0.895	0.157

The first-observation baseline replaces every masked cell by that input's value at its first timestep, i.e. the constant series $x_{1,:}$; it is the degenerate case of last-observation-carried-forward in which nothing earlier survives, and we name it for what it computes (`measure_gauge_invariance.py` still calls it `locf`). The six conventions are the Cartesian selection {zero, sequence-mean, first-observation} $\times$ {logit, probability} $\times\ p \in \{0.25, 0.5, 0.75\}$ varied one axis at a time from the default (zero, logit, p=0.5).

H EVALUATION: SEPARATION AND CONTROLS

Three explainers differing only in coalition objective (Owen-consistent, multi-scale, flat) are trained on a synthetic interaction benchmark (T=6, D=3, black-box validation accuracy 0.82–0.93; a smoke-sized configuration produced 0.542 and was refused by the accuracy floor). Ground truth is the exact Owen value of the trained model's own game under a logit value function.

Table A.8: Separation of the three objectives' means, in pooled seed standard deviations, 5 seeds. We use ≥ 2 as a descriptive threshold, not as a significance test; the paired seed-level tests and confidence intervals that the main text reports are in `dissociation/paired_contrasts.csv`.

metric	w=0.0	w=0.3	metric	w=0.0	w=0.3
gt Spearman	**2.08**	**3.32**	deletion AUC	0.26	0.15
gt Pearson	1.73	0.71	insertion AUC	1.16	0.46
gt MSE	1.44	0.61	AUPD	0.87	0.41
keep loss	0.18	0.38	AUPP	1.14	0.48
drop loss	0.32	0.17			

Table A.9: Control: separation of the exact Owen ground truth, submitted as the attribution, from random attributions. A metric that cannot reject random noise cannot adjudicate anything finer.

metric	w=0.0	w=0.3
keep loss	**0.74**	2.56
drop loss	0.55	1.09
insertion AUC	2.25	**0.62**
deletion AUC	1.15	1.86
AUPD	4.36	3.31
AUPP	**11.12**	**6.18**

The controls cut both ways, which is why they were added. AUPP and AUPD do separate oracle from random, so the evaluation harness is not broken; keep loss does not, at the operating point where interaction dominates.

I AMORTIZED ESTIMATOR

I.1 ARCHITECTURE AND TRAINING

A stack of dilated 1-D convolutions (kernel 3, dilation 2^ℓ, GELU, width 128) over the (T, D) input, with a per-cell head for ϕ, a per-cell head for the r factors, and pooled linear heads for the scales s and intercept b. Trained with Adam (10^{-3}, cosine schedule) for $8{,}000$ steps on batches of 32 inputs $\times$ 64 fresh coalitions, on the *training* split; evaluation uses 16 held-out test inputs and $8{,}000$ held-out coalitions each. The training cost is therefore $8{,}000 \times 32 \times 64 \approx 1.6 \times 10^7$ masked black-box evaluations per explainer, about one to two minutes on a single Titan Xp for the models here, paid once per predictor; a per-input fit at the 200,000-query budget of Table A.14 recovers that outlay after roughly 80 explanations. Emitting coefficients for a new input is one forward pass with no black-box calls; auditing them is not, and is priced in §I.5.

Two parameterisation choices matter. The projection is divided by $\sqrt{P}$, because $u^\top z$ has standard deviation $\|u\|$, which grows like $\sqrt{P}$, so the squared term grows like P and swamps the additive part on wide games. And the scale head is zero-initialised, so the network begins as exactly the additive surrogate and grows the interaction as a correction. Without these, MIMIC's $P = 768$ gave held-out R^2 of -0.70 where the additive head alone reached 0.64.

I.2 DEPTH, REACH AND CAPACITY

Within the dilated-convolution family the gap closes as the receptive field reaches T:

This sweep alone cannot say whether reach or capacity is responsible, because depth moves both. Dilation separates them — for kernel width 3 the field is $1 + 2\sum_\ell d_\ell$ — at identical parameter count *within* a depth (the two depths differ in parameters, and the shallow far field $(1, 32, 64)$ covers its span sparsely). Crossing them on CharacterTrajectories (T=119, additive head, three seeds) gives Table A.11.

The reach main effect ($+0.576$) is $7\times$ the capacity effect ($+0.082$) and the two interact ($+0.145$): **at a field shorter than the series, adding depth makes things slightly worse** (-0.063), while at

Table A.10: Amortization gap against trunk depth, convolutional trunk. PenDigits is the control: its series is already covered at three layers.

dataset	T	layers	receptive field	amortized rank 1	gap
CharacterTrajectories	119	3	15	-0.396	$+0.771$
		6	127	0.272	$+0.103$
		7	255	$\mathbf{0.314}$	$\mathbf{+0.061}$
MIMIC-III / GRU	48	3	15	0.679	$+0.263$
		6	127	$\mathbf{0.866}$	$\mathbf{+0.076}$
PenDigits (control)	8	3	15	0.630	$+0.112$
		6	127	0.658	$+0.093$

Table A.11: Reach crossed with capacity at matched parameter count, CharacterTrajectories additive head. Dilations are $(1, 2, \ldots, 64)$, $(1, \ldots, 1)$, $(1, 32, 64)$ and $(1, 2, 4)$ respectively.

	far (field 195–255)	near (field 15)	reach effect
deep (7 layers)	$\mathbf{+0.213 \pm 0.018}$	-0.507 ± 0.130	$+0.721$
shallow (3 layers)	-0.013 ± 0.031	-0.444 ± 0.039	$+0.431$
capacity effect	$+0.227$	$\mathbf{-0.063}$	

an adequate field it helps ($+0.227$). Reach is necessary and capacity is only useful once it is met, which also explains why encoders with unbounded field did not simply win: a bidirectional GRU and a transformer reach -0.146 and 0.067 at one layer, having reach but not depth. The ordering it gives is to cover the series first and add depth second.

I.3 What is invariant across encoder families

The quantity the paper is about — what the interaction head buys over the additive head — is stable where absolute fidelity is not. On CharacterTrajectories at rank 4 the three families differ by 0.21 in absolute R^2 and by 0.06 in head gain.

Table A.12: Three encoder families, three seeds each, gradient clipping at norm 1. Uncertainties are standard deviations over seeds.

dataset	trunk	additive head	rank-4 head	head gain
Character-Trajectories (AER 0.404)	dilated conv (7)	0.213 ± 0.018	$\mathbf{0.441 \pm 0.023}$	$+0.227 \pm 0.007$
	GRU (2)	0.036 ± 0.016	0.228 ± 0.022	$+0.192 \pm 0.035$
	attention (2)	0.123 ± 0.036	0.374 ± 0.011	$\mathbf{+0.251 \pm 0.048}$
MIMIC-III / GRU (AER 0.907)	dilated conv (6)	0.870 ± 0.003	0.876 ± 0.004	$+0.006 \pm 0.004$
	GRU (2)	0.853 ± 0.005	0.873 ± 0.007	$+0.020 \pm 0.006$
	attention (2)	$\mathbf{0.884 \pm 0.003}$	$\mathbf{0.892 \pm 0.002}$	$+0.008 \pm 0.004$

Absolute additive fidelity spans 0.036–0.213 on CharacterTrajectories, a range of 0.18, while the head gain spans 0.192–0.251, a range of 0.058. The between-dataset gap in head gain is 0.213, $3.6\times$ the larger of the two within-dataset architecture spreads. The encoder therefore determines how well the surrogate fits and barely determines what the interaction head is worth.

I.4 Rank sweep and the crossover

At seven layers, over three training seeds:

The two low-AER models separate from zero at 3.3–6.6 standard deviations; MIMIC does not, at 0.6–0.9. A single seed had shown a small *negative* gain on MIMIC at rank 4 (-0.039), which three seeds do not support: the head does nothing there rather than hurting, which is what the framework predicts at AER 0.907. The amortization gap is stable across ranks within a dataset (± 0.02 or better) and, on MIMIC, across seeds to ± 0.004.

Table A.13: Interaction head gain over the additive head, and the amortization gap, three seeds.

dataset	AER	rank	head gain	gap	gain / sd
CharacterTrajectories	0.404	1	$+0.125 \pm 0.038$	$+0.057 \pm 0.016$	3.3
		4	$\mathbf{+0.226 \pm 0.041}$	-0.070 ± 0.012	**5.6**
PenDigits	0.734	1	$+0.073 \pm 0.012$	$+0.074 \pm 0.006$	6.3
		4	$\mathbf{+0.110 \pm 0.017}$	$+0.137 \pm 0.014$	**6.6**
MIMIC-III / GRU	0.907	1	$+0.007 \pm 0.012$	$+0.078 \pm 0.004$	0.6
		4	$+0.009 \pm 0.011$	$+0.079 \pm 0.002$	0.9

At rank 4 on CharacterTrajectories the amortized surrogate (0.411) exceeded its per-input reference (0.351). Raising the reference's coalition budget shows this was a budget artifact:

Table A.14: Per-input rank-4 reference against its query budget, CharacterTrajectories.

per-input queries	30,000	100,000	200,000
held-out R^2	0.352	0.438	0.446
vs amortized (0.411, 0 queries)	amortized	per-input	per-input

That single-seed 0.411 was superseded by the clipped three-seed result 0.441 ± 0.023 (Table A.12), which a per-input fit matches only past 100,000 queries and which is 99% of the 200,000-query value. The 200,000-query point is the best we measured, not a verified ceiling.

I.5 Does the certificate predict convention sensitivity?

Protocol: rank 4, seven-layer convolutional trunk on CharacterTrajectories and six on MIMIC, 32 evaluation inputs drawn from the test split, three seeds, six models.

What the audit costs, and how far it can be shrunk. The certificate is not free: each held-out coalition is one masked black-box evaluation of the input being explained, so the 8,000 used above are 8,000 queries per audited input, on top of the coefficients, which cost none. The surrogate retrained from the same seed (42) and audited at smaller budgets shows how much of that is needed (Table A.15): 500 coalitions reproduce the full certificate at Spearman 0.90–0.99 on every model, flag the same inputs as failures ($R^2 < 0$) on MIMIC and CharacterTrajectories and 31 of 32 on JapaneseVowels, and keep the certificate–gauge correlation (0.575 against 0.581 on CharacterTrajectories, 0.419 against 0.403 on MIMIC); 100 coalitions still agree on 91–100% of failure flags but rank inputs less reliably on the hardest model (0.69). The one disagreement matters more than its rate suggests: JapaneseVowels input 18 scores $R^2 = 0.084$ at 500 coalitions and -0.041 at 8,000, so it is one of the six inputs that pass at 500 — a false pass among passes of $1/6$ on this model. Because only a few inputs pass there, 32 inputs cannot bound the false-pass rate tightly, and a near-zero threshold is a weak trust criterion. A practitioner who only needs a coarse trust signal can audit at a few hundred queries; one who needs the ranking of inputs by fidelity should spend about 500. (The CharacterTrajectories 1,000-coalition run was lost to a GPU out-of-memory error and is omitted rather than re-run on a different device.)

For each input the certificate is the surrogate's R^2 on 8,000 held-out coalitions of that input, and the gauge agreement is the mean of the 15 pairwise Kendall τ values between its order-1 attributions under the six conventions of Table A.7 (default, sequence-mean baseline, first-observation baseline, probability scale, $p = 0.25$, $p = 0.75$), each from 20,000 coalitions. The control is the same input's degree-≤ 1 mass under the default gauge from 20,000 coalitions — degree-1 only, so no pairwise basis is needed and $P = 768$ is affordable.

Three things are worth separating. First, the primary correlation replicates on both datasets and in both cases survives a Bonferroni correction over the two tests. Second, the Pearson coefficients are much weaker ($+0.232$ and $+0.174$, neither significant), so the relationship is monotone but not linear: on MIMIC a single input with certificate -0.73 and otherwise ordinary agreement is enough to flatten a linear fit, and dropping it leaves the rank correlation at $+0.360$. We therefore report rank

Table A.15: Certificate against audit budget. One trained surrogate per model (seed 42), audited on n held-out coalitions of each of 32 inputs. "vs 8,000" is the Spearman correlation of the n-coalition certificate with the 8,000-coalition one across inputs; "flags" is the fraction of inputs on which the two agree about $R^2 < 0$.

model	n	median R^2	vs 8,000	flags	ρ(cert, τ)
MIMIC-III / GRU	100	0.857	0.955	1.00	+0.349
	250	0.859	0.979	1.00	+0.403
	500	0.873	0.990	1.00	+0.419
	1000	0.872	0.993	1.00	+0.390
	2000	0.871	0.997	1.00	+0.403
	8000	0.867	1.000	1.00	+0.403
CharacterTrajectories	100	0.431	0.693	0.94	+0.290
	250	0.465	0.720	1.00	+0.341
	500	0.403	0.901	1.00	+0.575
	2000	0.432	0.995	1.00	+0.566
	8000	0.440	1.000	1.00	+0.581
JapaneseVowels	100	−1.697	0.927	0.91	+0.066
	250	−2.251	0.953	0.94	+0.052
	500	−1.983	0.982	0.97	−0.019
	1000	−2.030	0.989	0.97	−0.015
	2000	−2.030	0.996	1.00	−0.032
	8000	−2.026	1.000	1.00	−0.034

statistics throughout. Third, the mechanism is dataset-dependent. On MIMIC the certificate and per-input AER are collinear and neither survives partialling on the other, so they are two readings of one quantity; on CharacterTrajectories they are nearly independent and both survive, so the certificate carries information about convention sensitivity that additivity does not. Two datasets cannot say which case is typical. What holds in both is the cost asymmetry: the control needs about 20,000 model evaluations per input and the certificate, audited at 500 coalitions, needs 500.

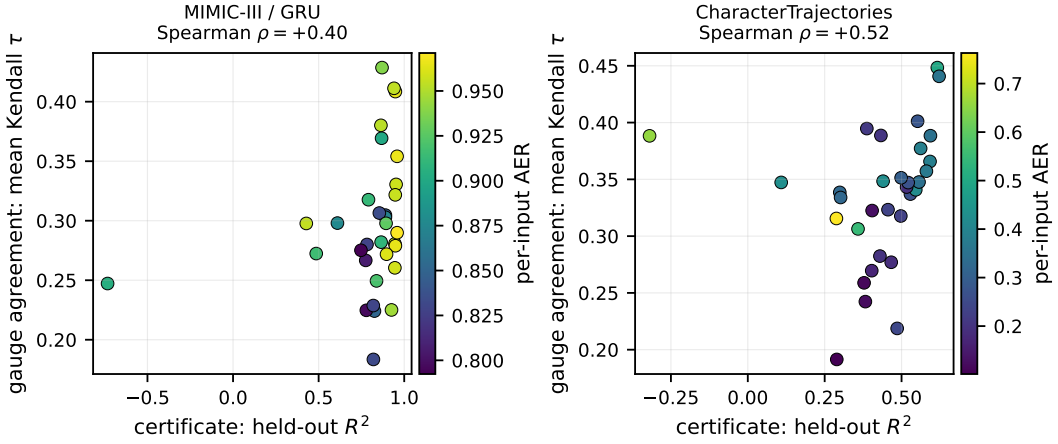


Figure A.1: Certificate against gauge agreement, 32 inputs each, coloured by per-input AER. The colour makes the dissociation visible: on MIMIC it tracks the horizontal axis (certificate and additivity are the same reading), on CharacterTrajectories it does not. In both panels one input has a negative certificate and otherwise unremarkable agreement, which is why we report rank statistics.

Two caveats on this experiment specifically. The 32 inputs are the first 32 of the test split rather than a random sample, and the three seeds re-train the explainer on the same inputs and the same predictor, so they measure explainer variability, not input or predictor variability. The correlations are exploratory; we fixed no primary outcome in advance beyond the original two-dataset test.

Table A.16: Certificate utility across six models, three seeds each, 32 inputs per run. "fitted" counts inputs whose certificate R^2 exceeds 0. The correlation tracks that count and the rows are ordered by it: positive on all 12 seeds of the four models that fit at least 26 of 32 inputs, significant on every seed of the best two, and absent on Libras and JapaneseVowels, where the estimator fails and the certificate reports the failure through its own negative median. The last block cuts CharacterTrajectories' training pool to the size JapaneseVowels has, showing the failure is not a training-set-size artifact.

model	pool	median R^2	fitted	Spearman ρ	p
MIMIC-III / GRU (P=768)	full	0.868	31	+0.411	0.019
	full	0.871	31	+0.411	0.019
	full	0.875	31	+0.425	0.015
MIMIC-III / TCN (P=768)	full	0.709	32	+0.327	0.075
	full	0.711	32	+0.319	0.068
	full	0.707	32	+0.363	0.041
CharacterTrajectories (P=357)	full	0.440	30	+0.581	0.0005
	full	0.418	29	+0.505	0.003
	full	0.426	29	+0.388	0.028
RacketSports (P=180)	full	0.595	27	+0.260	0.151
	full	0.603	27	+0.186	0.308
	full	0.484	26	+0.495	0.004
Libras (P=90)	full	−0.010	18	−0.163	0.374
	full	−0.041	16	−0.171	0.350
	full	−0.063	14	−0.121	0.510
JapaneseVowels (P=300)	full	−2.027	5	−0.034	0.855
	full	−3.939	5	+0.125	0.494
	full	−2.402	9	−0.068	0.711
CharacterTrajectories (training pool cut)	600	0.479	32	+0.479	0.006
	600	0.467	30	+0.497	0.004
	600	0.461	31	+0.510	0.003
	270	0.517	32	+0.506	0.003
	270	0.515	32	+0.554	0.001
	270	0.521	32	+0.463	0.008

Table A.17: Certificate against gauge agreement, 32 inputs per dataset. Partial correlations are Spearman on rank residuals. AER here is the per-input degree-≤ 1 mass, not the dataset mean.

	MIMIC-III / GRU	CharacterTrajectories
certificate R^2	0.792 ± 0.306	0.431 ± 0.180
gauge τ	0.296 ± 0.058	0.335 ± 0.058
per-input AER	0.905 ± 0.058	0.313 ± 0.153
certificate $\sim \tau$	$+0.397$ ($p = 0.025$)	$+0.522$ ($p = 0.002$)
AER $\sim \tau$	$+0.378$ ($p = 0.033$)	$+0.514$ ($p = 0.003$)
certificate $\sim$ AER	$+0.743$ ($p < 10^{-4}$)	$+0.161$ ($p = 0.380$)
certificate $\sim \tau \mid$ AER	$+0.188$ ($p = 0.304$)	$+0.519$ ($p = 0.002$)
AER $\sim \tau \mid$ certificate	$+0.135$ ($p = 0.461$)	$+0.510$ ($p = 0.003$)
bottom-quartile certificate: τ	0.270 ± 0.025	0.310 ± 0.061
top-quartile certificate: τ	0.316 ± 0.049	0.391 ± 0.037
separation (pooled sd)	$+1.17$	$+1.61$

I.6 Where the amortized estimator fails, and whether an audit can direct a repair

Representable, not amortizable. On the first 24 JapaneseVowels test inputs — the same inputs the certificate uses — a rank-4 surrogate fitted per input on 180,604 coalitions and scored on 20,000 fresh ones reaches median held-out distortion 0.180 (0.193 at rank 1), and every input beats predicting

the mean; the amortized rank-4 surrogate reaches a median of 4.09 on the same inputs, with 5 of 24 beating the mean. Inputs 24–31 were not run: the per-input fit needs about 10 GB of GPU memory, which was not available. The low-rank class therefore represents this model's games; the estimator fails to predict them.

A six-number repair. To see how much of that failure is in a few coefficients, we keep the amortized factor directions fixed and refit only the intercept, a multiplier on the additive term and the four factor scales, by ridge regression on the amortized surrogate's residual. Every input gets four disjoint coalition draws from independent generators: 500 to audit, 500 to fit the repair, 250 to choose the ridge strength and whether to repair at all (the identity is always a candidate), and 8,000 that only score. The surrogate is trained once per seed and its checkpoint reused by every policy. Table A.18 reports each policy's cost in black-box queries per input and its median reference distortion and fraction of inputs with $R^2 > 0$, averaged over three seeds.

Table A.18: Repair policies at matched cost (queries per explained input), three seeds, 32 inputs. Entries are median reference distortion / fraction of inputs with $R^2 > 0$. Selecting inputs by the audit never beats repairing a random set of the same total cost: the 500-query audit costs two thirds of a repair, so spending it on more repairs is at least as good. The value of the amortized coefficients themselves is model-dependent: at zero queries they beat a 750-query scratch fit on MIMIC and CharacterTrajectories and lose to it on JapaneseVowels.

policy	queries	MIMIC-III / GRU	CharacterTrajectories	JapaneseVowels
amortized, no queries	0	0.132 / 0.97	0.584 / 0.94	3.806 / 0.21
repair every input	750	0.121 / 1.00	0.525 / 1.00	0.919 / 0.71
additive ridge from scratch	750	0.502 / 1.00	0.937 / 1.00	0.710 / 0.98
audit all, repair lowest 25%	688	0.132 / 1.00	0.554 / 1.00	1.561 / 0.27
repair at random, same cost	680	0.122 / 1.00	0.528 / 0.99	0.930 / 0.66

We report the selection result as negative. An audit is useful for deciding whether to trust an explanation, but at these budgets it is not a cost-effective way to decide which explanation to repair.

J THE ATTENTION DIVERGENCE

Without gradient clipping, 3 of 18 estimator runs diverged — a GRU seed on CharacterTrajectories (-0.294 head gain) and two attention seeds on MIMIC (-0.609, -2.286) — and the surviving single-seed numbers were not representative of their own architecture. Clipping at norm 1 removes every divergence and raises absolute fidelity substantially where the instability had been worst: the attention trunk on MIMIC goes from 0.026 to 0.893 at rank 4, and its additive head from 0.635 to 0.884. Every estimator number in the paper uses clipping; it is part of the method, not a tuning detail. The rest of this appendix traces the divergence itself.

Before gradient clipping was added, 3 of 18 runs diverged: a GRU seed on CharacterTrajectories (head gain -0.294, per-input standard deviation 1.95) and two attention seeds on MIMIC (-0.609 and -2.286, standard deviations 0.894 and 3.65). The divergences were seed-dependent rather than configuration-wide, which is why single-seed runs of the same settings had looked healthy.

Clipping the gradient norm at 1 removes all of them, and also lifts fidelity where the instability had been worst:

Table A.19: The MIMIC attention trunk at rank 4, with and without gradient clipping.

seed	clipping	additive head	rank-4 head
42	none	0.635	0.026 (sd 0.894)
42	norm 1	**0.882**	**0.893**
123	none	0.715	-1.571 (sd 3.65)

Two consequences we did not anticipate. Every number measured before clipping is a lower bound on what the architecture can do — the additive head alone moves from 0.635 to 0.882 here — so the whole estimator section was re-measured with clipping on rather than patched. And without it, single-seed comparisons between architectures are unreliable, because divergence is seed-dependent.

K NEGATIVE RESULTS AND REFUTED HYPOTHESES

Owen values. An amortized Owen-value estimator was this project's original contribution. It loses to a multi-scale sampler on MIMIC in $3/3$ seeds (keep loss 0.319 against 0.294; a per-sample grouped baseline reaches 0.272). The mechanism is in Table A.4: the within-timestep interaction the Owen value uniquely resolves is nearly absent on real models.

Separable (Kronecker) structure. Predicted to beat low rank because it is the structure a temporal model most naturally induces; refuted on all five datasets once rate is counted, by up to $8.7\times$ in rate on ERing. A Kronecker term costs $T^2 + D^2$ numbers and forces every feature pair to share one time-coupling pattern.

Gauge invariance of rankings. Predicted; refuted (§G).

AER as a predictor of which method wins. On synthetic games the Owen-versus-flat gap appeared to grow with interaction mass, predicting a larger Owen advantage on MIMIC-TCN (AER 0.798) than MIMIC-GRU (0.907). Measured at 3 seeds per arm the deficit *doubles* on the lower-AER model ($+0.0513 \pm 0.0064$ against $+0.0250 \pm 0.0173$). AER is a ceiling and screening statistic; it does not order methods.

A player-subsampling proxy. To measure wide games cheaply we tried making a subset of cells the players and holding the rest present. Validated against the one dataset where both measurements exist, it failed: JapaneseVowels' additive ceiling reads 0.903 in the 120-player sub-game against 0.348 in the real 300-player one, and BasicMotions moves from 0.116 to 0.987. Conditioning on most cells makes the residual game far more additive. The proxy was abandoned; the finding that AER is a property of the *(model, player set)* pair was kept.

A direct keep-loss optimiser. Three relaxations (straight-through top-k; soft mask with a budget penalty; the same with temperature annealing) all failed to beat attribution methods ($2.079, 1.954, 1.687$ against 0.423). Each better relaxation helped monotonically, which suggests an optimisation difficulty rather than a property of the metric, so our critique of keep loss rests on the controls in §H and not on any claim of gameability.

L VERIFICATION

Every reported number was checked against the artifact that produced it rather than against the text asserting it. The safeguards behind the results are these.

- **Fail-closed checkpoints.** Every measurement script that loads a trained black box sets a release-mode flag by default (see the supplementary README), under which a missing or shape-incompatible checkpoint raises instead of falling back to random weights. Appendix C reports each black box's test accuracy against chance, and AUROC for MIMIC-III.
- **Scoped artifacts.** Checkpoints and result files are named by dataset, and multi-seed results additionally by seed and training-pool size, so runs on different datasets or seeds cannot overwrite one another; the repair pilot also refuses to overwrite an existing result.
- **Numerical regression tests.** The additive point of a distortion curve must equal $1 - \mathrm{AER}$, its degree-2 point the measured order-2 share, the sparse envelope must be monotone and bound every structural class, and an additive game must have near-zero held-out distortion (`tests/test_utils/test_interaction_aer.py`). All 205 tests pass on the released code.
- **Split hygiene.** Amortized surrogates are trained on the training split and scored on test inputs; a rate is chosen on one draw of coalitions and scored on another; audit, repair, validation and reference coalitions come from independent generators.
- **Provenance.** Appendix A maps every claim to the script and CSV behind it, and the figures are regenerated from those CSVs.

- **References.** Every entry was checked against the publisher or proceedings record where one exists and the arXiv abstract page otherwise, including author order and the latest version.

M REPRODUCTION

```
python scripts/measure_aer.py experiment=uea_racketsports \
    +aer.coalition_p=0.25
python scripts/measure_interaction_ceiling.py experiment=uea_japvowels \
    +ceiling.max_columns=46000 '+ceiling.bands=[1,3,5]' \
    +ceiling.spectrum=true
python scripts/measure_rate_distortion.py experiment=uea_racketsports \
    +rd.max_columns=16400 +rd.generalization=true
python scripts/measure_gauge_invariance.py experiment=mimic_mortality
python scripts/run_value_rank_dissociation.py \
    '+diss.seeds=[42,123,456,789,1024]'
python scripts/run_amortization_gap.py experiment=uea_chartrajectories \
    +gap.rank=4 +gap.num_layers=7 +gap.n_train_coalitions=30000
python scripts/run_certificate_utility.py experiment=mimic_mortality \
    +cert.rank=4 +cert.num_layers=6 +cert.n_eval=32
python scripts/measure_per_input_aer.py experiment=mimic_mortality \
    +cc.n_eval=32
python scripts/generate_v2_figures.py
```

Set the release-mode environment variable documented in the supplementary README for any run whose output is reported: it makes a missing or shape-incompatible checkpoint raise instead of silently falling back to random weights. The measurement scripts force it on by default.