WHAT DO TIME-SERIES EXPLANATIONS ACTUALLY EXPLAIN?

Anonymous authors
Paper under double-blind review

ABSTRACT

An attribution over a time-series window may indicate that a cell matters, but not in what sense. Under autocorrelation, a cell can influence the prediction almost entirely through the later cells into which its effect propagates, yet a method that ranks by association credits it in exactly the same way as it credits a cell the model reads directly. Existing faithfulness scores cannot separate these cases, because they test an explanation against a single, undifferentiated notion of importance. We characterize these effects on the Model Dependency DAG (MDDAG): we define a direct effect as a directed input–output edge on the MDDAG, and a total effect as the corresponding input–output dependency obtained when later cells are allowed to respond. Against either reference, we define OMIC, a bounded concordance score between an explainer's claimed and denied edges (effects), together with a ranking curve and its normalized area. The mediation residual serves as a diagnostic: an explainer that emphasizes mediated relevance will score highly on total effect but poorly on direct effect. We validate the references and the score on a synthetic VAR with known coefficients and apply them to known time-series explainers.

1 INTRODUCTION

A recurring challenge in explainable AI (XAI) is defining what makes an explanation faithful (Jacovi & Goldberg, 2020; Dembinsky et al., 2026; Afroogh et al., 2026; Klingshirn et al., 2026). For time-series models, an attributed cell may matter because the model reads it directly or because it carries information about later cells that the model reads instead. These correspond to different notions of explanation—mechanistic dependence and predictive relevance—that standard evaluations do not separate. We therefore ask not only *whether* the model depends on claimed cells, but *how*.

Existing evaluations largely operationalize faithfulness through perturbation: features are removed, retained, or otherwise modified according to an explanation, and the resulting change in model output is measured. Comprehensiveness and sufficiency (DeYoung et al., 2020), ROAR (Hooker et al., 2019), and F-Fidelity (Zheng et al., 2025) address, in different ways, the out-of-distribution problem this creates, but they all use a baseline or masking strategy for feature occlusion, which creates two main problems: baselines are conceptually difficult to define and, in most cases, implicitly impose correlations, which can distort the assessment of feature importance; masking strategies also introduce a zero baseline that is rarely on-manifold for time series. Moreover, these methods are primarily benchmarked on image datasets, and extending them to time-series settings is non-trivial. The aforementioned problems are pronounced in time-series, where temporal and cross-variable dependence make isolated cell perturbations inconsistent with the observed data (Ismail et al., 2020; Loeffler et al., 2025), and this has motivated conditional-replacement schemes such as FIT (Tonekaboni et al., 2020), WinIT (Leung et al., 2023) and Dynamask (Crabbé & van der Schaar, 2021), as well as on-manifold variants of Shapley attribution (Frye et al., 2021; Janzing et al., 2020). But admissibility is not disambiguation: even an intervention that stays on the manifold leaves open whether the measured change travelled through the model's direct use of the source or through the later cells the source predicts. Structural criteria such as DAG-faithfulness (Tunyi, 2026a) make the distinction explicit, but are formulated for direct dependence only and as a binary criterion.

We therefore evaluate explanations against a model-side reference: the model's response to on-manifold interventions on the corresponding input cells. For a source cell s, we resample s with

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.

the remaining input held fixed to obtain a direct effect Δ_{dir}, and resample it with its later cells allowed to respond to obtain a total effect Δ_{tot}. Replacement values are drawn from the source's conditional distribution, avoiding arbitrary off-manifold baselines. The evaluation needs no ground-truth causal graph and is not an attribution method; the explainer supplies only claimed input-output relationships (edges in Definition 1) and their ranking.

Building on this reference, we introduce *On-Manifold Interventional Consistency* (OMIC), which compares claimed and denied edges by interventional dependence, and *Area Under the OMIC Curve* (AUOMIC), which evaluates the same consistency across the attribution ranking. Both apply to direct or total effects. We also report a per-edge residual to localize mediated rather than direct relevance.

Contributions.

- An edge-level model-dependency formulation for time-series XAI, extending the model-dependency graph to direct and total effects across forecasting, classification, and regression, with both effects expressed through one response functional (Definition 1, Section 2.3).
- OMIC and its ranking-based extension AUOMIC, providing graded, on-manifold faithfulness evaluation for arbitrary explainers, together with a standardised variant that controls for per-edge differences in conditional spread (Section 3).
- Direct and total-effect instantiations and the mediation diagnostic, validated against known direct coefficients and impulse responses on a synthetic VAR and applied to TIMING, TimeSHAP, and KARMA on Beijing PM2.5 (Sections 2.3–4).

2 PRELIMINARIES

2.1 SETUP

Throughout, let $\mathbf{X} = \{X_i^t : i \in [D], t \in [T]\}$ denote a multivariate time-series over D channels and T observed time steps, where $[N] = \{1, \ldots, N\}$ for any natural number N. Let $f : \mathbb{R}^{T \times D} \to \mathcal{Y}$ be any time-series predictor (regressor, classifier, forecaster) mapping an input window to an appropriate output space. We write $p(\mathbf{x})$ for the reference training distribution and $\mathrm{supp}(p)$ for its support.

2.2 EDGE-LEVEL INTERPRETATION OF TIME-SERIES ATTRIBUTIONS

Across mathematics, causal inference, and the physical sciences, dependencies among stochastic processes are routinely described by directed graphs with annotated edges that encode transition strengths. This representation transfers naturally to time-series prediction, with one shift in perspective: the dependencies of interest need not be those present in the data-generating process, *but rather those the model has internalized*. Tunyi (2026a) formalized this model-induced structure, but only for direct effects in forecasting; we recall and extend it.

Definition 1 (Model Dependency Directed Acyclic Graph (MDDAG)). *Let $f : \mathbb{R}^{T \times D} \to \mathcal{Y}$ be a predictor and $p(\mathbf{x})$ a reference distribution. The Model Dependency Directed Acyclic Graph $\mathcal{G}_f$ has node set*

- *for classification given $\mathcal{Y} \in \mathbb{R}^C$, $V = \{X_i^t\} \cup \{c_j : j \in [C]\}$;*
- *for regression given $\mathcal{Y} \in \mathbb{R}^R$, $V = \{X_i^t\} \cup \{r_j : j \in [R]\}$;*
- *for forecasting given $\mathcal{Y} \in \mathbb{R}^{W \times D}$, $V = \{X_i^{(t)} : i \in [D],\ t \in \{1, \ldots, T + W\}\}$, where $t \leq T$ are inputs and $t \in \{T + 1, \ldots, T + W\}$ are forecast targets,*

and, for each input cell X_i^t ($t \leq T$) and output node o_j (r_j for regression[1], c_j for classification, or the forecast cell X_j^{t+w} for forecasting), the edge set is defined under $p(\mathbf{x})$ by the model's response to on-manifold interventions on the source:

- *if resampling X_i^t changes o_j when the source's later cells are allowed to respond on-manifold, then there exists a* total edge *from X_i^t to o_j, written $\Delta_{\mathrm{tot}}(X_i^t \rightsquigarrow o_j) > 0$;*
- *the edge is* direct *if resampling X_i^t changes o_j even with the full complement fixed, written $\Delta_{\mathrm{dir}}(X_i^t \to o_j) > 0$.*

Every direct edge is also a total edge; the two coincide when there is no mediation. The interventional quantities $\Delta_{\mathrm{dir}}, \Delta_{\mathrm{tot}}$ that make this precise are defined in Section 2.3. Tunyi (2026a) define the direct case for forecasting; the total-effect case and the non-forecasting outputs are our extension.

We retain the name MDDAG for continuity with Tunyi (2026a), but note that acyclicity is not a substantive property here: $\mathcal{G}_f$ is bipartite from input cells to output nodes, with no edges among input cells. In particular, the graph itself carries no representation of the dependence *within* the window; that dependence lives in p and enters only through the interventional definitions below. Accordingly, a total edge is a shorthand for an input–output dependency defined through mediator-responsive interventions, not for a literal path in $\mathcal{G}_f$.

For any explainer ϕ we write the *claimed influence* of X_i^t on output o_j as

$$a(X_i^t \Rightarrow o_j) := \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\left|\phi(\mathbf{x})_{t,i}^{j}\right|\right], \tag{1}$$

and analogously $a(X_i^t \Rightarrow X_j^{t+w}) := \mathbb{E}_{\mathbf{x}\sim p}[|\phi(\mathbf{x})_{t,i}^{t+w,j}|]$ for forecasting, where $\phi(\mathbf{x})_{t,i}^{j}$ is the influence of X_i^t on o_j at the local instance $\mathbf{x}$. An explainer claims an effect of X_i^t on o_j if $a(\cdot) > \tau$ for a threshold $\tau \geq 0$ fixed in advance (Section 4.1). The same symbol a is used for the attribution that induces the ranking in Section 3.3; for a natively global explainer, equation 1 is the identity.

For forecasting, influence on a specific cell variable is not readily obtained. Methods specific to forecasting, such as TimeSHAP, ShapTime (Zhang et al., 2024), and TS-MuLe (Schlegel et al., 2021), treat the problem like regression, measuring sensitivity to an aggregated output. To date the only method producing direct cell-to-cell attributions is KARMA (Tunyi, 2026b), which fits a k-order Markov surrogate to the forecaster and reads edge strengths off the surrogate's transition structure.

The edge-level interpretation supplies a concrete target for model analysis and, as Tunyi (2026a) observe, makes evaluation tractable: for direct edges, faithfulness of an edge-level explanation is analogous to faithfulness in the causal-discovery sense. We push this further with two graded metrics: the first assesses how well an explainer recovers direct effects and their ordering, the second does the same for total effects. First, we make precise what direct and total effects mean on the MDDAG.

Remark 1. *For readability, we use the word "edge" for both direct and total input–output dependencies, and we use $\to$ for $\rightsquigarrow$ and $\Rightarrow$ when the intended meaning is clear from context.*

2.3 Direct and total effects on the MDDAG

Fix a source $s := X_i^t$ and a target o_j. Partition the complement by temporal precedence into the mediator set and the pre-context,

$$M(s) = \{X_k^{t'} : t < t' \leq T\}, \qquad \mathrm{pre}(s) = \{X_k^{t'} : t' \leq t\} \setminus \{s\}, \qquad M(s) \cup \mathrm{pre}(s) = \mathbf{X}_{-s},$$

with contemporaneous cells ($t' = t$) assigned to $\mathrm{pre}(s)$ unless a domain contemporaneous order is specified.

Both direct and total effects compare the model's output at the observed source value with its output under a resampled one. They differ in what is allowed to move alongside the source. We therefore

[1] In regression, when $R > 1$, we approach node construction via a binning strategy; i.e., separate $\mathcal{Y}$ into R quartiles and take $r_j \in [R]$.

define, for a candidate source value v and an observed window $\mathbf{x}$, two *response functionals*

$$g_j^{\mathrm{dir}}(v;\mathbf{x}) := f(\mathbf{x}^{\,s\leftarrow v})_j, \tag{2}$$

$$g_j^{\mathrm{tot}}(v;\mathbf{x}) := \mathbb{E}_{\mathbf{x}_M\sim P_M(v)}\big[\, f(v,\mathbf{x}_M,\mathrm{pre}(s))_j \,\big], \qquad P_M(v) = p\big(M(s) \mid s{=}v,\ \mathrm{pre}(s)\big). \tag{3}$$

g_j^{dir} moves the source with the full complement pinned at its observed value; g_j^{tot} moves the source and averages over the mediator configuration it induces, with only $\mathrm{pre}(s)$ pinned. Each is the model's output as a function of the source alone, under a different convention for what the rest of the window does in response.

The interventional effect is then the expected discrepancy between the response at the observed source value and at a resampled one:

$$\Delta_\bullet(s\to o_j) := \mathbb{E}_{\mathbf{x}\sim p}\,\mathbb{E}_{v\sim q_\bullet}\Big[\, d\big(g_j^\bullet(s;\mathbf{x}),\ g_j^\bullet(v;\mathbf{x})\big)\Big], \qquad \bullet\in\{\mathrm{dir},\mathrm{tot}\}, \tag{4}$$

with $q_{\mathrm{dir}} = p(s \mid \mathbf{X}_{-s})$ and $q_{\mathrm{tot}} = p(s \mid \mathrm{pre}(s))$, and $d(\cdot,\cdot)$ a discrepancy with $d(a,b) = 0 \iff a = b$ (absolute difference for scalar outputs, a distributional distance for probabilistic ones). Instantiating equation 4 at $\bullet = \mathrm{dir}$ and using $g_j^{\mathrm{dir}}(s;\mathbf{x}) = f(\mathbf{x})_j$ recovers the familiar form

$$\Delta_{\mathrm{dir}}(s\to o_j) = \mathbb{E}_{\mathbf{x}\sim p}\,\mathbb{E}_{v\sim p(s|\mathbf{X}_{-s})}\big[\, d\big(f(\mathbf{x})_j,\ g_j^{\mathrm{dir}}(v;\mathbf{x})\big)\big]. \tag{5}$$

and, instantiating at $\bullet = \mathrm{tot}$, gives

$$\Delta_{\mathrm{tot}}(s\to o_j) = \mathbb{E}_{\mathbf{x}\sim p}\,\mathbb{E}_{v\sim p(s|\mathrm{pre}(s))}\Big[\, d\big(\mathbb{E}_{\mathbf{x}_M\sim P_M(s)}[f(s,\mathbf{x}_M,\mathrm{pre}(s))_j],\ g_j^{\mathrm{tot}}(v;\mathbf{x})\big)\Big]. \tag{6}$$

The total-minus-direct residual. Define

$$\mathrm{Med}(s\to o_j) \;=\; \Delta_{\mathrm{tot}}(s\to o_j) - \Delta_{\mathrm{dir}}(s\to o_j).$$

The diagnostic regime is $\Delta_{\mathrm{dir}} \approx 0$ with Δ_{tot} large: the source barely moves the output when its later cells are held fixed, yet moves it substantially once they are allowed to respond. In this regime, the edge has predictive dependence under the total reference but not direct dependence under the direct reference; within our framework, the difference is attributed to mediation through later cells.

Remark 2. *The above definitions carry over to forecasting with o_j replaced by the forecast cell X_j^{T+w} and $f(\cdot)_j$ by $f(\cdot)_{j,w}$.*

Propositions 1 and 2 in Supplementary Material § A verify that $\Delta_{\mathrm{dir}} > 0$ characterizes conditional dependence of the output on the source given the full complement, that $\Delta_{\mathrm{tot}} > 0$ characterizes non-constancy of the mediator-averaged response g_j^{tot} in the source, and that the two conditions $\Delta_{\mathrm{dir}} = 0,\ \Delta_{\mathrm{tot}} > 0$ jointly characterize a purely mediated edge.

2.4 Relation to mediation analysis and to conditional-baseline attribution

Δ_{dir} is a magnitude-valued controlled direct effect, and Δ_{tot} is a magnitude-valued total effect, in the sense of Pearl (2001) and Robins & Greenland (1992), transported to the input cells of a fitted predictor and evaluated under conditional rather than fixed reference values. Med is deliberately *not* the Natural Indirect Effect (NIE): the NIE requires a nested counterfactual in which the source takes one value in f and another in P_M, and identifying it requires a cross-world independence assumption we do not want to impose on a model-side diagnostic. Causal mediation has been used to localize the mechanism inside neural models for gender bias in language models (Vig et al., 2020) and, in the interchange-intervention line, to align model internals with a causal abstraction (Geiger et al., 2021). Those analyses intervene on hidden units to explain the model; we intervene on input cells to *evaluate an explanation of* the model, and the mediators are other input cells rather than internal states.

Conditional replacement is well established as a remedy for off-manifold occlusion (Frye et al., 2021; Janzing et al., 2020); FIT (Tonekaboni et al., 2020), WinIT (Leung et al., 2023), and Dynamask (Crabbé & van der Schaar, 2021) use related conditional schemes for time series. What remains missing is the separation between direct and propagated relevance: WinIT, for example, is motivated by information propagation and therefore reports a quantity closer to Δ_{tot} than Δ_{dir} without separating the two.

3 THE TWO FAITHFULNESS EVALUATION METRICS

We now build a graded consistency score for either reference effect. The premise is that an explanation can be evaluated with regard to direct dependence (whether it identifies edges with large direct effects) or total dependence (whether it identifies edges with large total effects), and that these are different claims that deserve separate scores.

3.1 THE REFERENCE EFFECT AS A PER-EDGE QUANTITY

For a candidate edge $e = (X_i^{(t_1)} \to X_j^{(T+w)})$ in forecasting, or $e = (X_i^{(t_1)} \to o_j)$ in classification or regression, write $\Delta_\bullet(e)$ for the quantity defined in equation 4 with source $s = X_i^{(t_1)}$ and the corresponding target. Everything in this section is stated for a generic $\Delta_\bullet$ and applies verbatim to Δ_{dir} and Δ_{tot}.

The conditional sampling replaces the source with plausible values. The direct reference conditions on the full complement; the total reference conditions on the pre-context and resamples the mediator configuration. Thus $\Delta_\bullet(e)$ measures model response to an on-manifold change, not arbitrary occlusion. It is a reference measure of model dependence, not ground-truth causality. Because it scales with the spread of $q_\bullet$, we also report the standardized effect

$$\widetilde{\Delta}_\bullet(e) \;=\; \frac{\Delta_\bullet(e)}{\mathrm{sd}\big(q_\bullet(s)\big)}, \tag{7}$$

which converts the reference into output movement per unit of admissible source displacement, and treat agreement between the two rankings as a robustness check rather than assuming it.

3.2 ON-MANIFOLD INTERVENTIONAL MODEL-DEPENDENCY CONSISTENCY

Let E denote the candidate edge set and let an XAI method assign a non-negative attribution $a(e) \geq 0$ to each candidate edge. After thresholding, let E^+ denote the set of claimed-present edges and $E^- = E \setminus E^+$ the claimed-absent ones.

We first ask a support-level question: *does the explainer claim as present the edges to which the model is more interventionally dependent than the edges it denies?* This motivates the on-manifold interventional consistency (OMIC) score,

$$\mathrm{OMIC}_\bullet = \frac{1}{|E^+|\,|E^-|} \sum_{e^+ \in E^+} \sum_{e^- \in E^-} \psi\big(\Delta_\bullet(e^+), \Delta_\bullet(e^-)\big), \qquad \psi(x,y) = \begin{cases} 1, & x > y, \\ \frac{1}{2}, & x = y, \\ 0, & x < y, \end{cases} \tag{8}$$

the tie-aware convention assigning half credit when two edges have equal interventional effects. Equation 8 is the Mann–Whitney U statistic of $\Delta_\bullet$ between the claimed and denied sets, equivalently the AUROC of $\Delta_\bullet$ as a score for membership in E^+; this fixes $\mathrm{OMIC}_\bullet \in [0,1]$ with 0.5 the chance level, so the usual rank-based variance estimates are available. A value close to one indicates that the claimed support is consistently more model-dependent than the denied support; a value close to zero indicates systematic disagreement.

OMIC answers whether an XAI method identifies edges on which the model depends, but it does not evaluate the relative ordering within E^+: it is unchanged by any permutation of the attributions inside the claimed set. An explainer should identify the right edges *and* order them correctly, which motivates a ranking-based extension.

3.3 AREA UNDER THE OMIC CURVE

Rank the selected edges by attribution magnitude,

$$|a(e_{(1)})| \geq |a(e_{(2)})| \geq \cdots \geq |a(e_{(|E^+|)})|, \tag{9}$$

rather than evaluating over the complete candidate space, we evaluate consistency over the most highly attributed portion of the explanation. Let $\pi_{\max} \in (0,1]$ be the maximum retained fraction; we use $\pi_{\max} = 0.75$. For each level $\pi_k = k/|E^+|$, $k = 1, \ldots, \lfloor \pi_{\max}|E^+| \rfloor$, set

$$E_k^+ = \{e_{(1)}, \ldots, e_{(k)}\}, \qquad E_k^- = E \setminus E_k^+, \tag{10}$$

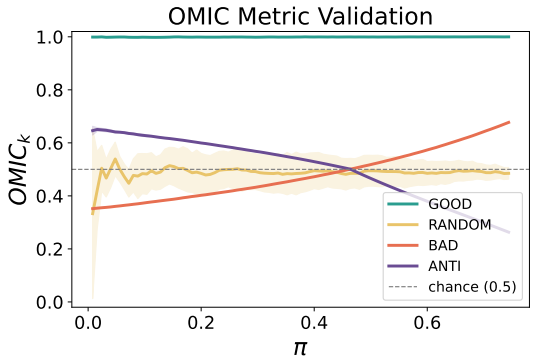


Figure 1: Metric behaviour on a synthetic VAR with known ground truth and an exact Gaussian conditional. π here denotes the fraction of retained edges. Pointwise $\mathrm{OMIC}_{\mathrm{dir},k}$ for four attributions scored against the same reference: GOOD (true support, ranked by true coefficient magnitude), RANDOM (a random claim of the same size), BAD (the correct support, rank-inverted), and ANTI (the lowest-coefficient edges, ranked normally). Bands are ± 1 std over seeds. The four separate by the *shape* of the curve rather than by level: GOOD sits at the ceiling, RANDOM at chance, ANTI decays from above chance as its wrongly-claimed edges are exhausted, and BAD *rises*, because it claims the right set and so wins every comparison against the non-support majority at each level. That BAD claims the correct support is why the ordering test in Section 3.3 is reported separately from the support test.

and compute the pairwise interventional consistency between the retained edges and everything outside them,

$$\mathrm{OMIC}_{\bullet,k} = \frac{1}{|E_k^+|\,|E_k^-|} \sum_{e^+ \in E_k^+} \sum_{e^- \in E_k^-} \psi\big(\Delta_\bullet(e^+), \Delta_\bullet(e^-)\big)\,. \tag{11}$$

The sequence $\{\mathrm{OMIC}_{\bullet,k}\}_{k=1}^{\lfloor \pi_{\max}|E^+| \rfloor}$ is the *OMIC ranking curve*. If the method identifies the model-dependent edges and ranks them highly, $\mathrm{OMIC}_{\bullet,k}$ stays high as more edges are retained; a ranking unrelated to model dependence approaches 0.5, and a systematically reversed ranking falls below it.

We summarize the ranking consistency by the normalized area under the OMIC curve, computed by the trapezoidal rule:

$$\mathrm{AUOMIC}_{\bullet,\pi_{\max}} = \frac{\sum_{k=1}^{K-1} \frac{1}{2}\left(\mathrm{OMIC}_{\bullet,k} + \mathrm{OMIC}_{\bullet,k+1}\right)(\pi_{k+1} - \pi_k)}{\pi_K - \pi_1}, \tag{12}$$

$$K = \left\lfloor \pi_{\max}|E^+| \right\rfloor, \quad \pi_1 = \frac{1}{|E^+|}.$$

Because the levels π_k are equally spaced, equation 12 is the trapezoidal mean of the curve values and inherits their $[0, 1]$ range and 0.5 chance level.

Direct recovery. Instantiating at Δ_{dir} gives $\mathrm{OMIC}_{\mathrm{dir}}$ and $\mathrm{AUOMIC}_{\mathrm{dir}}$: these measure whether the claimed edges tend to have larger direct effects than denied edges and whether higher-ranked claimed edges tend to have larger direct effects. Within our framework, this is the reference for an explainer intended to report direct model dependence.

As Figure 1 shows, a high AUOMIC indicates that highly attributed edges consistently exhibit larger on-manifold interventional effects than lower-ranked edges. A value near 0.5 indicates that the ranking carries little information about relative model dependence, and values below 0.5 indicate systematic disagreement.

Total recovery. Instantiating at Δ_{tot} gives $\mathrm{OMIC}_{\mathrm{tot}}$ and $\mathrm{AUOMIC}_{\mathrm{tot}}$: these measure whether the claimed edges tend to have larger total effects than denied edges and whether higher-ranked claimed edges tend to have larger total effects. Within our framework, this is the reference for predictive dependence that includes temporal mediation.

The gap is the diagnostic, not either score alone. An explainer whose claimed edges are mediated can score higher against the total reference than against the direct reference. Neither number

alone identifies the source of the difference; the pair does, and the per-edge residual $\mathrm{Med}(e)$ localizes it. We therefore additionally report a $\mathrm{Med}(e)$ map as a spatial account of the total-minus-direct residual and the mediation (mf) ratio $\frac{1}{|E^+|}\sum_{e\in E^+}\frac{\mathrm{Med}(e)}{\Delta_{\mathrm{tot}}(e)}$. A claimed edge with $\mathrm{Med}(e)\approx 0$ can correspond either to strong direct dependence or to weak dependence under both references, and only $\Delta_{\mathrm{dir}}(e)$ separates those.

3.4 Properties of the OMIC metrics

The OMIC scores have the following properties.

- **Bounded, with a chance level.** $\mathrm{OMIC}_\bullet \in [0,1]$, and $\mathrm{AUOMIC}_\bullet$ inherits the range as a trapezoidal mean. A claim unrelated to $\Delta_\bullet$ scores 0.5 in expectation, so 0.5 is the reference point, not 0.
- **Invariant to monotone rescaling, not to per-edge rescaling.** $\mathrm{OMIC}_\bullet$ sees $\Delta_\bullet$ only through the pairwise ordering, so d matters only up to monotonicity; a per-edge rescaling does change the ordering, which is why the standardized reference equation 7 is a substantive check rather than a reparametrization.
- **Insensitive to false negatives.** $\mathrm{OMIC}_\bullet$ compares only claimed against denied edges, so any E^+ contained in the set of edges with $\Delta_\bullet > 0$ scores 1, regardless of size. Thus τ and $|E^+|$ should be reported, and comparisons across explainers are most direct at matched support size.

The above properties motivate the following interpretation of the scores. Because OMIC compares claimed against denied edges, it is a *support/ranking consistency* measure rather than a completeness measure: a small but well-chosen claimed set can score highly. We therefore avoid using OMIC alone to define a complete notion of faithfulness. Instead, we refer to an explainer as *reference-consistent* when its claimed edges systematically exhibit larger reference effects than its denied edges, and as *partially reference-consistent* when this holds only up to a chosen tolerance α. These labels are reference-specific and do not imply that the explainer has recovered all model-dependent edges. Further evaluation can use other XAI properties, such as completeness, complexity, and sensitivity.

4 Experiments

4.1 Protocol

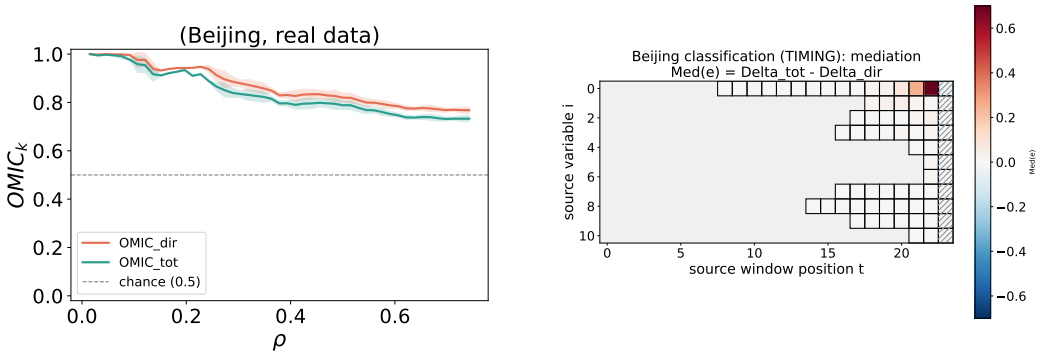
We evaluate three scenarios on the Beijing PM2.5 dataset with a GRU-based architecture: classification with TIMING (Jang et al., 2025), regression with TimeSHAP (Bento et al., 2021), and forecasting with KARMA (Tunyi, 2026b). An edge source is an input cell variable in the input window. An attribution assigned to a source–target relationship therefore determines the claimed strength of the corresponding edge; throughout, we reserve *cell* for a source variable and *edge* for the source–target relationship. Results are reported as mean $\pm$ standard deviation over five cross-validation folds. Code and additional results are available at https://github.com/anonymous-researcher914/WhatDoTimeseriesExplanationExplain.

4.2 Results

Task (explainer)	$\mathrm{OMIC}_{\mathrm{dir}}$	$\mathrm{OMIC}_{\mathrm{tot}}$	$\mathrm{AUOMIC}_{\mathrm{dir}}$	$\mathrm{AUOMIC}_{\mathrm{tot}}$	mf ratio	τ
Forecasting (KARMA)	0.987 ± 0.000	0.964 ± 0.001	0.987 ± 0.000	0.964 ± 0.001	0.167 ± 0.000	0.000
Regression (TimeSHAP)	0.611 ± 0.020	0.622 ± 0.022	0.809 ± 0.011	0.803 ± 0.011	0.473 ± 0.020	$1e-10$
Classification (TIMING)	0.738 ± 0.023	0.736 ± 0.021	0.853 ± 0.006	0.853 ± 0.006	0.379 ± 0.021	$1e-10$

Table 1: OMIC (support) and AUOMIC (ranking) scores, mean $\pm$ standard deviation over five cross-validation folds. Each row is a separate audit of one explainer on the task it targets; rows are not comparable to one another (Section 4.1). Chance is 0.5.

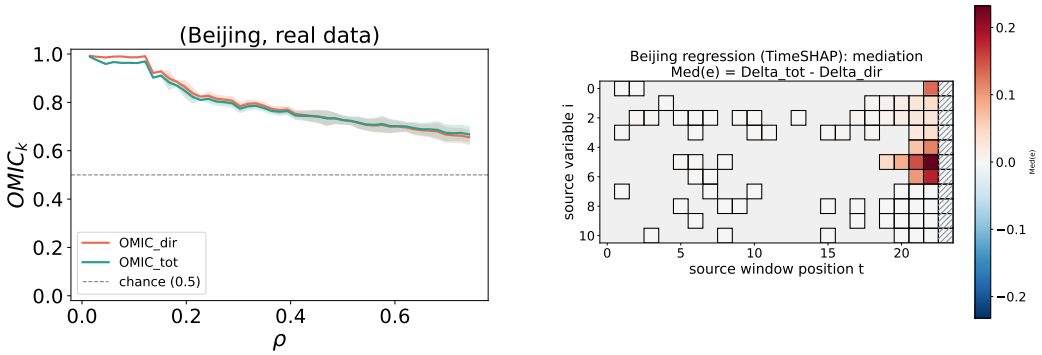
A. Classification (TIMING). Both curves in Figure 2(a) are above chance, while their separation is within the across-fold spread. TIMING is therefore partially reference-consistent at $\alpha = 0.3$. The residual map (b) is small in absolute terms for most claimed edges, but Table 1 gives a ratio of approximately 0.379, indicating a non-negligible average total-minus-direct residual. Since $\mathrm{Med}(e) \approx 0$ is compatible with either strong direct dependence or weak dependence under both references, $\Delta_{\mathrm{dir}}(e)$ remains necessary.



(a) OMIC ranking curves. **(b)** $\mathrm{Med}(e)$ over the window.

Figure 2: Classification with TIMING on Beijing PM2.5, mean ± 1 standard deviation over five cross-validation folds. **(a)** Pointwise OMIC ranking curves for the direct (Δ_{dir}) and total (Δ_{tot}) references against the retained fraction ρ; dashed line is the pairwise chance level of 0.5. **(b)** $\mathrm{Med}(e) = \Delta_{\mathrm{tot}}(e) - \Delta_{\mathrm{dir}}(e)$ over the input window; black outlines mark claimed edges, and hatched columns indicate positions where $M(s)$ is empty by construction.

B. Regression (TimeSHAP). Both curves in Figure 3(a) remain above chance, with separation within the across-fold spread, placing TimeSHAP in the partially reference-consistent regime at $\alpha = 0.35$. The residual map (b) is concentrated near the end of the window; the reported ratio is approximately 0.5, so the average residual is about half of the corresponding total effect among claimed edges. The scores alone do not distinguish direct dependence from temporal mediation.

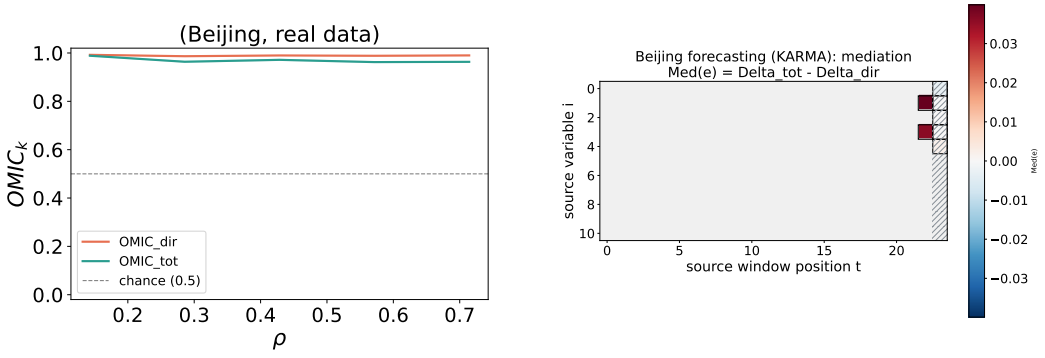


(a) OMIC ranking curves. **(b)** $\mathrm{Med}(e)$ over the window.

Figure 3: Regression with TimeSHAP on Beijing PM2.5, mean ± 1 standard deviation over five cross-validation folds. Panels follow Figure 2. Note that the colour scale of the $\mathrm{Med}(e)$ map differs from the classification figure.

C. Forecasting (KARMA). KARMA's claimed edges are closely aligned with both references, with curves near the ceiling across the retained range (Figure 4(a)). KARMA is also the only method here that produces cell-to-cell attributions natively; moreover, this task is not directly comparable to the classification and regression rows. The residual is negligible for almost all claimed cells and

confined to two cells near the boundary. Because KARMA is sparse, the reported ratio is computed over few claimed edges; subject to this limitation, its residual is small.



(a) OMIC ranking curves. **(b)** $\mathrm{Med}(e)$ over the window.

Figure 4: Forecasting with KARMA on Beijing PM2.5, mean ± 1 standard deviation over five cross-validation folds. Panels follow Figure 2.

5 Conclusion

We argued that a single faithfulness score cannot specify what a time-series explanation explains. On the model-induced dependency graph we defined two on-manifold interventional references through a common response functional: a direct effect, with the source's complement held fixed, and a total effect, with its later cells allowed to respond, and built a bounded consistency score, OMIC, and its ranking extension, AUOMIC, against either. Instantiated at both references, the pair separates direct model dependence from total predictive dependence, and the per-edge residual $\mathrm{Med}(e)$ localizes where the two diverge. Neither component suffices alone: $\mathrm{Med}(e) \approx 0$ does not distinguish a strong direct edge from a cell the model ignores, and a high consistency score does not reveal which kind of dependence was recovered.

Four limitations qualify these conclusions. The total effect requires estimating the joint mediator response P_M, so unlike the direct effect it inherits the error of a learned conditional generator, and Med inherits this error from both effects. The direct/mediated split is defined relative to a mediator set, which we fix by temporal precedence; a different admissible ordering moves the split. Identification becomes weaker near the end of the window, where $M(s)$ is nearly empty and $\Delta_{\mathrm{tot}} \to \Delta_{\mathrm{dir}}$ by construction. And both references are magnitudes, so the evaluation is blind to the sign an explanation assigns.

The natural next step is an attribution method designed to target the direct reference rather than merely be evaluated against it; such a method would require a conditional rather than a fixed baseline. We leave that to future work, noting that a method designed against these metrics would need evaluating by machinery it does not share.

AI use statement

We used generative AI tools for grammar refinement, coding assistance for experimental artifacts, and related-work search. We did not use them for full-text or idea generation or mathematical formulae. We reviewed all AI-assisted work; an LLM also assisted with some supplementary proofs. All generated code, proofs, and modified text were inspected by the corresponding author, who accepts responsibility for the final content.

References

Saleh Afroogh, Syed Ishtiaque Ahmed, Petra Ahrweiler, David Alvarez-Melis, Mansur Maturidi Arief, Emilia Barakova, Falco J. Bargagli-Stoffi, Erdem Biyik, Hanjie Chen, Xiang 'Anthony'

Chen, Robert Alan Clements, Keeley Crockett, Amit Dhurandhar, Fethiye Irmak Dogan, Mollie Dollinger, Motahhare Eslami, Aldo A Faisal, Arya Farahi, Melanie F. Pradier, Saadia Gabriel, Diego Garcia-Olano, Marzyeh Ghassemi, Shaona Ghosh, Hatice Gunes, Ehsan Hajiramezanali, Stefan Haufe, Biwei Huang, Angel Hwang, Md Tauhidul Islam, Junfeng Jiao, Amir-Hossein Karimi, Saber Kazeminasab, Anastasia Kuzminykh, William La Cava, Brian Y. Lim, Xiaofeng Liu, Mohammad R. K. Mofrad, Alicia Parrish, Maria Perez-Ortiz, Shriti Raj, Swabha Swayamdipta, Salmonn Talebi, Kush R. Varshney, Mihaela Vorvoreanu, Lily Weng, Alice Xiang, Yiming Xu, Ding Zhao, and Jieyu Zhao. Beyond explainable ai (xai): An overdue paradigm shift and post-xai research directions, 2026. URL https://arxiv.org/abs/2602.24176.

J. Bento, P. Saleiro, A. F. Cruz, M. A. T. Figueiredo, and P. Bizarro. TimeSHAP: Explaining recurrent models through sequence perturbations. In *Proceedings of KDD*, 2021.

Jonathan Crabbé and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning (ICML)*, 2021.

David Dembinsky, Adriano Lucieri, Stanislav Frolov, Hiba Najjar, Ko Watanabe, and Andreas Dengel. Unifying vxai: A systematic review and framework for the evaluation of explainable ai, 2026. URL https://arxiv.org/abs/2506.15408.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations (ICLR)*, 2021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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*, volume 32, 2019.

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*, volume 33, pp. 6441–6452, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/47a3893cc405396a5c30d91320572d6d-Abstract.html.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020. URL https://aclanthology.org/2020.acl-main.386/.

Hyeongwon Jang, Changhun Kim, and Eunho Yang. Timing: Temporality-aware integrated gradients for time series explanation, 2025. URL https://arxiv.org/abs/2506.05035.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Kai Jonas Klingshirn, Christoph Garth, and Achim Ebert. Xai beyond reality: Identifying key research gaps and future directions. In Camille Fayollas, Pieter Van Gorp, Mahmoud Baghdadi, Achim Ebert, Jun Hu, Shah Rukh Humayoun, Sapna Jaidka, Kris Luyten, Tilo Mentler, Philippe Palanque, Parvaneh Parvin, Lucio Davide Spano, Simone Stumpf, Gerrit van der Veer, Luciana Zaina, and Jürgen Ziegler (eds.), *Engineering Interactive Computer Systems. EICS 2025 International Workshops*, pp. 167–184, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-26051-2.

Kin Kwan Leung, Clayton Rooke, Jonathan Smith, Saba Zuberi, and Maksims Volkovs. Temporal dependencies in feature importance for time series prediction. In *International Conference on Learning Representations (ICLR)*, 2023.

Christoffer Loeffler, Wei-Cheng Lai, Bjoern Eskofier, Dario Zanca, Lukas Schmidt, and Christopher Mutschler. Don't get me wrong: How to apply deep visual interpretations to time series, 2025. URL https://arxiv.org/abs/2203.07861.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 411–420, 2001.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

Udo Schlegel, Duy Vo Lam, Daniel A. Keim, and Daniel Seebacher. Ts-mule: Local interpretable model-agnostic explanations for time series forecast models, 2021. URL https://arxiv.org/abs/2109.08438.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David Duvenaud, and Anna Goldenberg. What went wrong and when? Instance-wise feature importance for time-series black-box models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Amadeo Tunyi. The failures of marginal influence-based attribution methods for global time series explanations, 2026a. URL https://arxiv.org/abs/2607.16236.

Amadeo Tunyi. Global explanations for multivariate time series forecasting models via k-order markov approximations, 2026b. URL https://arxiv.org/abs/2606.27599.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yuyi Zhang, Qiushi Sun, Dongfang Qi, Jing Liu, Ruimin Ma, and Ovanes Petrosian. Shaptime: A general xai approach for explainable time series forecasting. In Kohei Arai (ed.), *Intelligent Systems and Applications*, pp. 659–673, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-47721-8.

Xu Zheng et al. F-Fidelity: A robust framework for faithfulness evaluation of explainable AI. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.02970.