TEMPORAL DISTRIBUTIONAL VALUES FOR PROBABILISTIC FORECASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic time-series forecasts describe a range of possible futures. When these forecasts inform health decisions, patients, clinicians and researchers need to understand how patient history shapes the predictions and their uncertainty. For example, in diabetes management individuals may wish to understand how behaviors (e.g., eating, exercise) shift predicted glucose trajectories and how they influence the range of future glycemic states. Specific patient behaviors may have similar effects on a predicted future glucose trajectory but induce substantially different levels of uncertainty in the range of glycemic outcomes predicted by the model. Scalar attribution methods explain a chosen summary of a forecast, such as its mean. However, a single aggregate statistic can conceal probabilistic information: an input may change the predicted variance without changing the mean, or shift the forecast from a uniform to a bimodal distribution. We introduce **Temporal Distributional Values (TDV)**, a model-agnostic method that extends Shapley attribution to probabilistic time-series forecasting. TDV computes distributions of input contributions directly from the model's native probabilistic output, including quantiles and sampled trajectories. We evaluate TDV on four state of the art forecasting models using long-term glucose, meal, and activity data from type 1 diabetes, type 2 diabetes, and health and wellness populations. Experiments demonstrate accurate forecast reconstruction and close agreement between TDV average contributions and exact scalar Shapley values; stability under trajectory and order sampling; and practical computational cost. Case studies reveal previously hidden distributional effects of meals, activity, and glucose history, demonstrating the value of TDV for uncovering a richer understanding of probabilistic forecasts. These explanations give practitioners a new way to examine how patient history shapes the range of outcomes a forecasting model predicts.

1 INTRODUCTION

Probabilistic time-series forecasts assign probabilities to future outcomes, representing uncertainty and distributional information in how a time series may evolve. In healthcare, these forecasts can help patients and clinicians anticipate changes in health related measurements. In diabetes management, for example, glucose forecasts describe possible future glucose levels given recent measurements, meals, and activity. Understanding how this history shapes the predictive distribution is important because similar mean forecasts can conceal different levels of variability or distinct likely outcomes.

SHAP assigns input contributions to a selected scalar output; TimeSHAP and WindowSHAP extend this approach to time-series inputs (Lundberg & Lee, 2017; Bento et al., 2021; Nayebi et al., 2023). For a probabilistic forecast, explaining its mean, variance, or one quantile answers a specific question, but none alone fully describes the underlying distribution. FIT compares predictive distributions through scalar relevance scores (Tonekaboni et al., 2020). These approaches summarize input effects, leaving a need for explanations that retain distributions of contributions and their dependence for forecast reconstruction.

We introduce Temporal Distributional Values (TDV), a model-agnostic, post-hoc method for explaining probabilistic time-series forecasts. It records changes as input groups are added to a reference history, pairing matching quantile levels or trajectory draws generated with shared randomness. Retaining these changes across input orders and forecast draws reveals contribution patterns that their

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.

averages can conceal. The same construction explains quantities spanning several future time points, such as the total time that predicted glucose levels spend within a specified range.

Our contributions are:

1. We formulate distributional attribution for temporal input groups and native forecast representations, distinguishing what marginal quantiles and joint trajectories support when explaining future outcomes.
2. We provide a sampling procedure that retains the baseline and contributions together to reconstruct the entire distributional forecast. Averaging contributions recovers scalar Shapley values for the corresponding expected output.
3. We evaluate TDV using PatchTST, TFT, Chronos-2, and TACTiS-2 on glucose data from diabetes and health-and-wellness cohorts. Results show reconstruction and scalar agreement to numerical precision and subsecond average attribution times using stored forecasts. Case studies illustrate distributional changes and input contributions to predicted time in range.

TDV gives researchers and practitioners a way to examine how recorded history shapes predicted outcomes across forecasting applications, with potential use in tools that help patients and clinicians interpret health forecasts.

2 Related Work

Temporal attribution methods connect predictions to observed history. TimeSHAP, WindowSHAP, and GroupSegment-SHAP assign scalar contributions to temporal groups; SHAPformer provides horizon-specific attributions, and TsSHAP explains a forecasting surrogate (Bento et al., 2021; Nayebi et al., 2023; Kim & Park, 2026; Hertel et al., 2025; Raykar et al., 2023). DynaMask and ContraLSP identify influential observations through perturbations, while PAX-TS measures local sensitivity (Crabbé & van der Schaar, 2021; Liu et al., 2024; Kreuzer et al., 2025). These methods summarize input effects for selected outputs; TDV retains contribution distributions, including opposing changes concealed by their average.

Explanations of predictive distributions differ in the information they retain. FIT yields scalar relevance scores; information-theoretic Shapley methods explain uncertainty or information measures, and SHAP has been applied to individual forecast quantiles (Tonekaboni et al., 2020; Watson et al., 2023; Chronopoulos et al., 2024). Tangent-SHAP produces function-valued explanations; its released quantile implementation averages over input orders to reconstruct quantile vectors relative to a background (Zhao, 2026; YuZhao20, 2026). GP-SHAP retains Gaussian-process posterior randomness in jointly distributed attributions while averaging over coalitions (Chau et al., 2023). Distributional Values retains distributions of paired payoff differences across input orders and shared randomness (Franceschi et al., 2024). TDV uses this construction for forecast quantiles, sampled trajectories, and quantities spanning forecast horizons, keeping each order's contributions paired with their baseline for reconstruction.

3 Background and Problem Setup

3.1 Probabilistic forecasts

For a fixed observed history x, we explain a trained model's forecast at H future horizons. A marginal forecast describes Y_t at one horizon; a joint forecast describes $Y_{1:H} = (Y_1, \ldots, Y_H)$ and its temporal dependence. Models may supply point forecasts, marginal quantiles at selected probability levels, or joint trajectory samples.

Finite quantiles leave values between supplied levels, unrepresented tails, and cross-horizon dependence unspecified. Appendix A gives the representation assumptions.

3.2 Input groups and reference values

We partition the input coordinates being explained into disjoint groups (players), indexed by $N = [n]$. A group can represent a time step, one variable across the history, or a time–variable pair. A coalition

$S \subseteq N$ retains its groups from x; omitted groups take values from a reference history $W \sim \rho_x$, which may be fixed.

Definition 1 (Forecasting game)**.** Let $r(x, S, W)$ retain groups in S, replace the others using W, and recompute derived covariates according to the specified preprocessing rules. Inputs outside the player partition stay fixed, and $r(x, N, W) = x$. The model output and selected representation define a distribution $\mu_{x,S}$ on $\mathcal{Y} \subseteq \mathbb{R}^d$ for each coalition. These choices define the forecasting game.

3.3 Scalar and stochastic payoffs

The coalition's model output is its payoff. For a scalar payoff $v(S)$, the Shapley value averages the change from adding group i over uniformly sampled input orders Π (Shapley, 1953; Lundberg & Lee, 2017). Writing $\mathrm{Pre}_i(\Pi)$ for the preceding groups gives

$$\phi_i(v) = \mathbb{E}_\Pi[v(\mathrm{Pre}_i(\Pi) \cup \{i\}) - v(\mathrm{Pre}_i(\Pi))] .$$

Distributional Values uses a random-variable payoff $V_S(Z)$ and retains the distribution of each contribution over input orders and shared randomness Z (Franceschi et al., 2024).

Definition 2 (Forecasts with shared randomness)**.** Let $Z = (W, \Omega) \sim \nu$, and let G produce the represented forecast from the model input and forecast randomness Ω.

$$V_S(Z) = G(r(x, S, W), \Omega) \in \mathcal{Y}, \qquad V_S(Z) \sim \mu_{x,S}.$$

Using the same Z for all coalition forecasts defines their joint distribution, or coupling. Write $v_x(S, Z) = V_S(Z)$ for this stochastic game.

We seek contribution distributions for the input groups together with the dependence needed to reconstruct the full-input forecast.

4 Temporal Distributional Values

4.1 Forecast outputs and input contributions

TDV compares forecasts as input groups are added to the reference history. For quantile forecasts at a fixed horizon, we compare values at matching quantile levels; Appendix A specifies interpolation and sampling for a finite grid. For joint forecasts, we pair complete trajectories $V_S(Z) = Y_{S,1:H}(Z)$ across coalitions using shared random numbers. Point forecasts have no forecast randomness.

Definition 3 (Temporal Distributional Values)**.** Draw a uniform permutation $\Pi = (\pi_1, \ldots, \pi_n)$ independently of Z, and set $S_0 = \varnothing$ and $S_j = \{\pi_1, \ldots, \pi_j\}$. Following the stochastic marginal-contribution construction of Franceschi et al. (2024), the contribution of group π_j is

$$D^{\Pi}_{\pi_j}(Z) = V_{S_j}(Z) - V_{S_{j-1}}(Z). \tag{1}$$

Group i's TDV is the random variable $\xi_i(v_x) := D^{\Pi}_i(Z)$, with probability distribution $q_i := \mathcal{L}(\xi_i(v_x))$. Here $\mathcal{L}$ denotes the probability distribution of a random variable.

For a scalar payoff, positive contributions raise the forecast value and negative contributions lower it. Variation across input orders, reference histories, and forecast draws can yield opposing contributions whose average is small.

4.2 Quantities across forecast horizons

Forecast quantities such as total time within a target interval summarize outcomes across several future horizons. To explain a trajectory summary h, we use the payoff $V^h_S(Z) = h(Y_{S,1:H}(Z))$. We apply h before taking differences: $D^{\Pi,h}_{\pi_j}(Z) = V^h_{S_j}(Z) - V^h_{S_{j-1}}(Z)$. For total duration within an interval I,

$$T_S(Z) = \sum_{t=1}^{H} \Delta t_t \, \mathbf{1}\{Y_{S,t}(Z) \in I\}, \tag{2}$$

where Δt_t is the duration represented by horizon t. For glucose forecasts, T_S is predicted time in range. Its expectation, $\mathbb{E}[T_S] = \sum_t \Delta t_t \Pr(Y_{S,t} \in I)$, follows from the horizon marginals. The distribution of T_S generally requires their dependence (Appendix B.1).

4.3 RECONSTRUCTING THE FORECAST

For the sampled contributions in Definition 3, adding the baseline and contributions from the same order and shared state reconstructs the forecast.

Definition 4 (Joint contribution record)**.** For $\mathbf{D}^{\Pi} = (D_1^{\Pi}, \ldots, D_n^{\Pi})$, the joint record has probability distribution

$$\Xi = \mathcal{L}\big(V_{\varnothing}(Z), \Pi, \mathbf{D}^{\Pi}(Z)\big)\,. \tag{3}$$

The joint distribution Ξ retains dependence among contributions and with the baseline. The successive changes telescope:

Proposition 1 (Additive reconstruction)**.** *Under Definitions 1–2, the following identity holds for each input order.*

$$V_N(Z) = V_{\varnothing}(Z) + \sum_{i=1}^{n} D_i^{\Pi}(Z). \tag{4}$$

Taking the distribution of the paired sum gives

$$\mu_{x,N} = \mathcal{L}\left(V_{\varnothing}(Z) + \sum_{i=1}^{n} D_i^{\Pi}(Z)\right). \tag{5}$$

For finite quantiles at one horizon, reconstruction covers the interpolated distribution on the supported rank interval; joint trajectory samples reconstruct the full forecast vector. Applied to V_S^h, the construction reconstructs the distribution of $h(Y_{N,1:H})$. Independent sampling from the separate q_i generally changes the distribution of the sum (Appendix B.1). Appendix B gives the proof and its relationship to the distributional efficiency property of Franceschi et al. (2024).

Different pairings can change contribution distributions while preserving every coalition's predictive distribution. The explanation therefore depends on the pairing rule, input groups, and reference values.

4.4 RELATION TO SCALAR SHAPLEY VALUES

For a real-valued payoff with finite absolute expectation for every coalition, the mean-recovery property of Distributional Values gives (Franceschi et al., 2024)

$$\mathbb{E}[\xi_i(v_x)] = \mathbb{E}_{\Pi,Z}[D_i^{\Pi}(Z)] = \mathrm{Sh}_i(S \mapsto \mathbb{E}_Z[V_S(Z)])\,. \tag{6}$$

This gives a scalar compatibility check: average TDV equals the Shapley value for the expected represented payoff, with the same input groups, reference distribution, replacement rule, and forecast quantity.

4.5 COMPUTING TDV

We implemented TDV in Python. Algorithm 1 computes paired contributions for the fixed forecasting game, forecast quantity, and pairing rule. With K input orders and R shared-state draws per order, let $\mathbf{D}_{kr}$ denote the contribution vector from order k and draw r.

Each order requires $n + 1$ coalition evaluations; batching draws and reusing identical evaluations can reduce repeated model calls. Orders may be sampled in reverse pairs, with equal weight for each realized order. Appendix C gives storage and validation details.

Estimating contribution distributions. The KR paired records in Algorithm 1 give the empirical estimate

$$\widehat{\Xi} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} \delta_{(V_{\varnothing}(Z_{kr}), \Pi_k, \mathbf{D}_{kr})}, \tag{7}$$

where δ_a is a point mass at a. Collecting one group's contributions estimates q_i and its mean. Adding the baseline and contributions in each record recovers the empirical distribution of the full-input payoff.

Algorithm 1 Computing paired TDV contributions

Require: Fixed forecasting game; K input orders; R shared-state draws per order

for $k = 1, \dots, K$ **do**
  Draw a uniform order $\Pi_k = (\pi_1, \dots, \pi_n)$ independently of shared states.
  Set $S_0 = \varnothing$ and $S_j = \{\pi_1, \dots, \pi_j\}$ for $j = 1, \dots, n$.
  for $r = 1, \dots, R$ **do**
    Draw $Z_{kr} = (W_{kr}, \Omega_{kr}) \sim \nu$; keep it fixed along the coalition chain.
    Evaluate the chosen payoffs $V_{S_j}(Z_{kr})$ for $j = 0, \dots, n$.
    Set $D_{\pi_j, kr} = V_{S_j}(Z_{kr}) - V_{S_{j-1}}(Z_{kr})$ for $j = 1, \dots, n$.
    Check $V_{\varnothing}(Z_{kr}) + \sum_i D_{i,kr} = V_N(Z_{kr})$ to numerical tolerance.
    Store $(V_{\varnothing}(Z_{kr}), \Pi_k, \mathbf{D}_{kr})$ together.
  end for
end for
return the paired records.

With a fixed reference, we linearly interpolate finite coalition quantiles on a common rank grid over $[a, b]$. For order k, let $\widetilde{T}_k(u) = (\widetilde{Q}_{\varnothing,t}(u), \Pi_k, \widetilde{\mathbf{D}}_t^{\Pi_k}(u))$ collect the baseline and contributions at rank u. Retaining these functions gives

$$\widehat{\Xi}^{\widetilde{Q}} = \frac{1}{K} \sum_{k=1}^{K} (\widetilde{T}_k)_{\#} \operatorname{Unif}[a, b], \tag{8}$$

where $(\widetilde{T}_k)_{\#} \operatorname{Unif}[a, b]$ is the distribution obtained by evaluating $\widetilde{T}_k$ at a uniform rank. This integrates over represented ranks and averages over orders.

Contribution spread reflects variation across input orders, reference histories, and forecast draws. Monte Carlo error instead measures uncertainty from estimating this distribution or its summaries with finitely many samples.

5 Evaluation

Using glucose forecasts, we evaluate TDV's numerical reconstruction and sampling stability (Section 5.2), computational cost (Section 5.3), scalar agreement (Section 5.4), distributional explanations (Sections 5.5–5.6), and additional diagnostics (Section 5.7).

5.1 Datasets, forecasting models, and experimental settings

We evaluate TDV on two private cohorts, Type 2 diabetes and health and wellness (H&W), and the OhioT1DM Type 1 diabetes dataset (Table 1). Private glucose data were collected with Dexcom G7 and Stelo sensors; meal and activity logs provide context for selected explanations. H&W comprises people without Type 1 or Type 2 diabetes. OhioT1DM provides five-minute glucose, insulin, and life-event records, available for research under a data-use agreement (Marling & Bunescu, 2020). Private analyses use 30 days (8,640 time steps) per person. All data were de-identified and obtained with appropriate consent and institutional approvals.

Table 1: **Source cohort characteristics.** Counts and recording spans describe the source cohorts, including the published OhioT1DM release; forecast-case counts are given in Table 2.

Cohort	Access	Participants	Span per person
Type 2	Private	10,000	6 months
H&W	Private	10,000	6 months
OhioT1DM (Type 1)	Research agreement	12	8 weeks

We evaluate TDV using PatchTST (Nie et al., 2023), TFT (Lim et al., 2021), Chronos-2 (Ansari et al., 2025), and TACTiS-2 (Ashok et al., 2024), which cover three forecast interfaces (Table 2), testing deterministic compatibility, marginal-distribution explanations, and quantities spanning forecast horizons.

Table 2: **Forecast interfaces and scalar targets.** Each model–dataset pair uses 12 forecast cases at each input-group count. Quantile means use the interpolated distribution over the supported ranks (Appendix A).

Model	Native output used	Scalar target	Dataset
PatchTST	Point forecast	Selected forecast coordinate	Type 2
TFT	Marginal quantiles	Represented quantile mean	Type 2
Chronos-2	Marginal quantiles	Represented quantile mean	H&W
TACTiS-2	Joint trajectories	60-minute glucose, mean over 256 paths	H&W; OhioT1DM

Experimental settings. All experiments ran on a virtual machine with 32 vCPUs, 208 GB RAM, and four NVIDIA T4 GPUs. For the TACTiS-2 numerical comparisons, $n \in \{2, 4, 6, 8\}$ consecutive windows partition the most recent 120 minutes of glucose history. In these comparisons, the primary reference is the training-data median at each history position; a mean-based reference tests sensitivity. Omitted groups take reference values, derived covariates are recomputed, and other inputs remain fixed, with shared forecast randomness across coalitions (Section 4).

5.2 Numerical reconstruction and sampling stability

We check TDV reconstruction with TACTiS-2, four 30-minute input windows, 12 forecast cases per cohort (OhioT1DM and H&W), and 256 trajectories per case. Maximum residuals of 4.3×10^{-14} and 6.2×10^{-15} mg/dL, respectively, show that paired contributions and baseline reconstruct the forecast payoff to numerical precision.

Omitting each of four 64-trajectory groups tests sensitivity within the fixed trajectory bank. Wasserstein-1 distances between estimated contribution distributions have medians of 0.0029 and 0.013 mg/dL for OhioT1DM and H&W, but maxima of 0.85 and 0.84 mg/dL, respectively: sensitivity varies across comparisons.

With increasing input-order budgets $K \in \{32, 128, 512\}$, median root-mean-square error of TDV means against exact scalar Shapley values decreases in both cohorts, indicating improved typical mean-attribution accuracy (Figure 5).

5.3 Computational cost and scalability

To assess scalability, we measure wall time separately for forecast generation and attribution over all 2^n coalitions for $n \in \{2, 4, 6, 8\}$ input groups. At eight groups, generating and storing TACTiS-2 forecasts took approximately 59 minutes for OhioT1DM and 41 minutes for H&W, each over 12 cases with 256 trajectories per case, excluding model loading and warm-up. Subsequent TDV analysis averaged 0.099 and 0.072 seconds per case, respectively.

Forecast generation dominates these exhaustive runs; attribution from stored forecasts is inexpensive. Table 3 reports per-case attribution times averaged over six- and eight-group settings for all four models. Figure 4 shows TACTiS-2 scaling; Appendix D.3 details timer scopes, including initial TimeSHAP package loading.

5.4 Agreement with scalar Shapley values

We test whether averaging TDV contributions recovers the corresponding scalar Shapley explanation (Equation (6)). For each model–dataset pair in Table 2, we evaluate all 2^n coalitions for $n \in \{2, 4, 6, 8\}$ input groups. Exact scalar Shapley values, TimeSHAP, and the TDV mean use the same forecasting game, scalar target, and stored coalition forecasts. Errors are the largest absolute deviations from exact scalar Shapley values across evaluated cases and groups. For six and eight groups, TDV mean errors are zero for PatchTST and reach a maximum of 5.68×10^{-14} mg/dL across the other models; the maximum TimeSHAP error is 1.71×10^{-13} mg/dL. Summed mean contributions recover the forecast change with a maximum residual of 9.59×10^{-14} mg/dL. These residuals are consistent with floating-point rounding.

Table 3: **Scalar agreement and attribution computation time.** Errors are maxima across the six- and eight-group comparisons, using the scalar targets in Table 2. *Sum* is the residual between the summed mean contributions and the scalar forecast's change from its reference. Times are average seconds per forecast case and input-group count, calculated from stored coalition forecasts and excluding forecast generation. TACTiS-2 errors take the maximum over both datasets; its times average over both.

	Maximum absolute error (mg/dL)			Attribution time (s)	
Model	TimeSHAP	TDV mean	Sum	TDV	TimeSHAP
PatchTST	0 (exact)	0 (exact)	0 (exact)	2.7×10^{-2}	3.5×10^{-3}
TFT	2.84×10^{-14}	3.20×10^{-14}	9.59×10^{-14}	4.2×10^{-1}	1.1×10^{-2}
Chronos-2	4.26×10^{-14}	3.20×10^{-14}	2.84×10^{-14}	1.5×10^{-1}	1.0×10^{-2}
TACTiS-2	1.71×10^{-13}	5.68×10^{-14}	8.53×10^{-14}	4.7×10^{-2}	5.2×10^{-3}

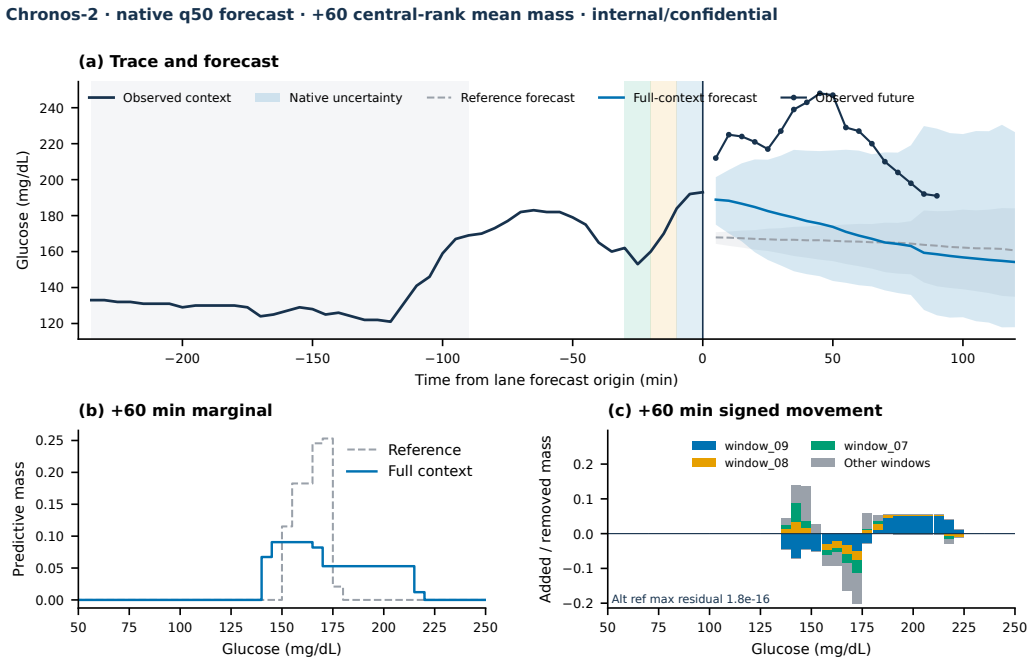


Figure 1: **Distributional changes beyond the forecast mean.** (a) Observed history and future, with reference and full-input median forecasts and predictive bands; vertical shading marks the three newest 10-minute windows. (b) Forecast probability masses represented by the available quantiles at 60 minutes. (c) Signed probability changes allocated to these windows and the other six combined; positive and negative bars add and remove probability in 5 mg/dL glucose bins.

Table 3 reports model-specific errors and cached attribution times. Appendix D.3 gives the recovery plot and full results for two, four, six, and eight groups.

5.5 How recent history changes a forecast distribution

TDV retains information beyond each input group's mean contribution. We illustrate why this matters with a selected Chronos-2 forecast from H&W (Figure 1). Nine 10-minute windows partition the most recent 90 minutes of input history; the explained distribution represents glucose 60 minutes ahead using the available quantiles.

The represented mean rises from 164.58 to 173.80 mg/dL (+9.22 mg/dL): the newest window contributes +17.81 mg/dL, partly offset by −8.59 mg/dL from the earlier eight. The full-input

distribution places probability both below and above the concentrated reference peak, showing a shape change beyond the mean increase.

The signed probability changes locate this redistribution: the newest window removes probability around 140–180 mg/dL and adds it mainly around 185–220 mg/dL, partly offset by earlier windows. These bars allocate changes in bin probabilities (Appendix D.4); TDV's q_i instead describes contribution sizes in mg/dL. This illustrates an intended use of TDV: tracing which parts of the history drive or offset a shift toward higher predicted glucose that the mean increase alone does not describe.

Appendix D.4 gives nine-window examples for PatchTST, TFT, and TACTiS-2.

5.6 Time-in-range case studies

Time in range measures how long glucose stays within a target interval, commonly $[70, 180]$ mg/dL in diabetes care (Battelino et al., 2019). We explain its predicted counterpart in three selected TACTiS-2 cases, counting in-range minutes along each forecast trajectory (Equation (2); Table 4). Figure 2 shows the Type 2 case, selected after examining attributions (Appendix D). Separate explanations group its preceding 120 minutes into twelve 10-minute windows or glucose, meal and nutrition, and activity inputs. Both use 256 trajectories, with 64 temporal orders and all six variable-group orders.

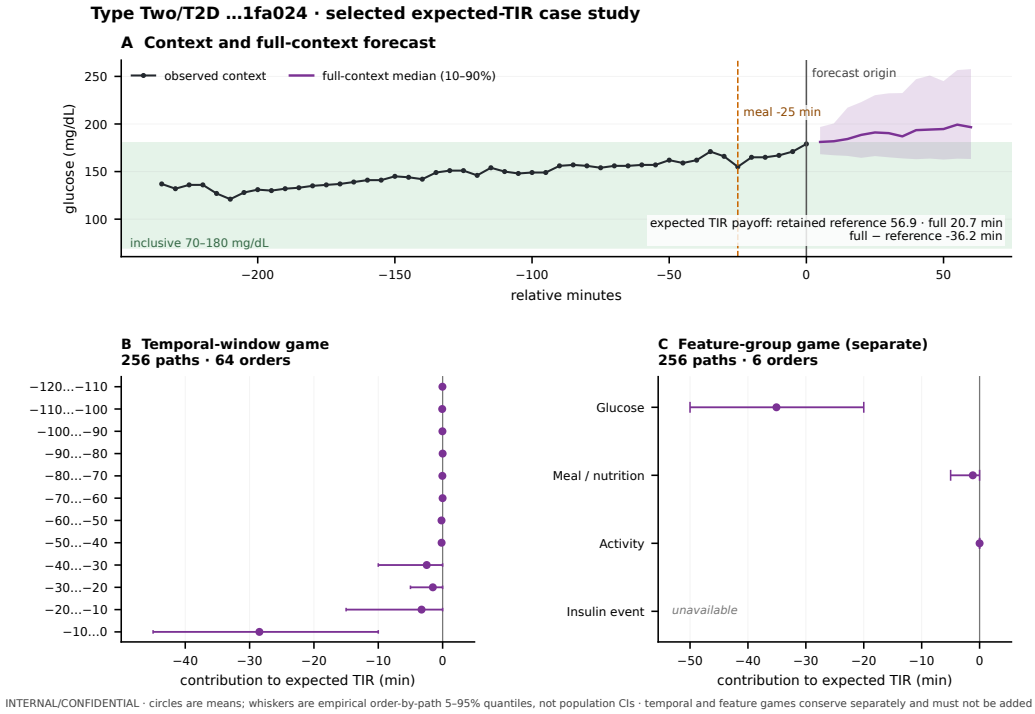


Figure 2: **Selected Type 2 time-in-range example.** (A) Observed glucose, a meal recorded 25 minutes before forecasting, and the full-input median forecast with its 10–90% predictive interval; green shading marks $[70, 180]$ mg/dL. The reference forecast's expected time in range is reported numerically; its trajectory is not shown. (B) Contributions from twelve 10-minute history windows. (C) A separate explanation using glucose, meal and nutrition, and activity; insulin is unavailable. Circles and whiskers show means and empirical 5th–95th percentiles across forecast trajectories and input orders, in minutes.

In the Type 2 case, expected time in range falls by 36.25 minutes relative to the reference. The newest window has the largest negative mean contribution; among variables, glucose history accounts for most of the reduction, the meal contribution is smaller, and activity is near zero.

The broad negative contribution intervals for the newest window and glucose history show variation across trajectories and input orders, rather than uncertainty in the estimated mean or differences between participants.

Table 4: **Selected time-in-range cases.** All values are minutes; reference and full-input columns report expected time in $[70, 180]$ mg/dL for one selected case per cohort.

Case	Horizon	Reference	Full input
Type 2	60	56.91	20.66
OhioT1DM	60	60.00	30.06
H&W	90	90.00	45.92

The OhioT1DM and H&W cases also have small mean meal contributions relative to glucose history: -0.07 versus -29.87 minutes and $+0.68$ versus -44.77 minutes, respectively (Appendix D.5). These examples motivate examining how the model uses meal records and whether glucose history already carries related information.

5.7 Additional diagnostics

For the OhioT1DM and H&W cases, Appendix D.6 examines terminal in-range event switches, sensitivity to trajectory count, and dependence on contribution alignment (Figures 11–13). Shuffling alignment changes reconstructed endpoint distributions; three independent 256-trajectory banks preserve the signs of glucose and meal means and the larger magnitude of glucose contributions.

6 Discussion and Conclusion

Limitations TDV explanations can change with the input groups, reference values, and forecast pairing; broader sensitivity studies can assess which conclusions persist. Marginal quantile grids do not generally identify omitted tails or dependence across horizons, limiting explanations to their represented distributions.

Selected glucose cases do not establish the prevalence of observed patterns or benefits for model diagnosis. Randomly sampled forecast origins, controlled problems with known input effects, and studies with model developers would address these questions.

TDV explains probabilistic time-series forecasts through paired input contributions that retain variation beyond their means. Numerical checks show reconstruction of represented forecasts and scalar Shapley agreement to numerical precision; glucose examples illustrate changes in forecast shape and contributions to predicted time in range. TDV enables researchers to examine how recorded history shapes possible forecast outcomes.

Ethics statement

This work adheres to ethical standards in data collection, model evaluation, and experimental design. All data used were de-identified and obtained with appropriate consent and institutional approvals, ensuring participant privacy and compliance with relevant regulations (e.g., HIPAA). The explanations characterize specific forecasting games and should not be interpreted as causal, physiological, diagnostic, or treatment advice.

Reproducibility statement

Appendices A–D provide the definitions, proofs, computation, and evaluation protocols. Code is not released with this submission; sensitive source data and trained checkpoints are not redistributed. Appendix E describes the information needed to reproduce the numerical explanations.

Statement on the use of generative AI

Generative AI tools assisted with software implementation and tests, experimental-methodology feedback, evidence-ledger reconciliation, literature discovery and synthesis, result-interpretation

checks, mathematical exposition and proof editing, manuscript drafting and editing, and adversarial review. AI tools did not serve as authors or independently authorize experiments or scientific claims. AI-assisted outputs were checked against primary literature, executable validators, privacy reviews, and immutable aggregate artifacts. The authors take responsibility for the final paper.

REFERENCES

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Nicolas Chapados, and Alexandre Drouin. TACTiS-2: Better, faster, simpler attentional copulas for multivariate time series. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xtOydkE1Ku.

Tadej Battelino, Thomas Danne, Richard M. Bergenstal, Stephanie A. Amiel, Roy Beck, Torben Biester, Emanuele Bosi, Bruce A. Buckingham, William T. Cefalu, Kelly L. Close, Claudio Cobelli, Eyal Dassau, J. Hans DeVries, Kim C. Donaghue, Klemen Dovc, Francis J. Doyle III, Satish Garg, George Grunberger, Simon Heller, Lutz Heinemann, Irl B. Hirsch, Roman Hovorka, Weiping Jia, Olga Kordonouri, Boris Kovatchev, Aaron Kowalski, Lori Laffel, Brian Levine, Alexander Mayorov, Chantal Mathieu, Helen R. Murphy, Revital Nimri, Kirsten Nørgaard, Christopher G. Parkin, Eric Renard, David Rodbard, Banshi Saboo, Desmond Schatz, Keaton Stoner, Tatsuiko Urakami, Stuart A. Weinzimer, and Moshe Phillip. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, 42(8):1593–1603, 2019. doi: 10.2337/dci19-0028. URL https://doi.org/10.2337/dci19-0028.

João Bento, Pedro Saleiro, André F. Cruz, Mário A. T. Figueiredo, and Pedro Bizarro. TimeSHAP: Explaining recurrent models through sequence perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021. doi: 10.1145/3447548.3467166. URL https://doi.org/10.1145/3447548.3467166.

Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2209. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9f0b1220028dfa2ee82ca0a0e0fc52d1-Abstract-Conference.html.

Ilias Chronopoulos, Aristeidis Raftapostolos, and George Kapetanios. Forecasting value-at-risk using deep neural network quantile regression. *Journal of Financial Econometrics*, 22(3):636–669, 2024. doi: 10.1093/jjfinec/nbad014. URL https://academic.oup.com/jfec/article/22/3/636/7163191.

Jonathan Crabbé and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2166–2177. PMLR, 2021. URL https://proceedings.mlr.press/v139/crabbe21a.html.

Luca Franceschi, Michele Donini, Cédric Archambeau, and Matthias Seeger. Explaining probabilistic models with distributional values. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 13840–13863. PMLR, 2024. URL https://proceedings.mlr.press/v235/franceschi24a.html.

Matthias Hertel, Sebastian Pütz, Ralf Mikut, Veit Hagenmeyer, and Benjamin Schäfer. Explainable time-series forecasting with sampling-free SHAP for transformers. *arXiv preprint arXiv:2512.20514*, 2025. URL https://arxiv.org/abs/2512.20514v1.

Jinwoong Kim and Sangjin Park. GroupSegment-SHAP: Shapley value explanations with group-segment players for multivariate time series. *arXiv preprint arXiv:2601.06114*, 2026. URL https://arxiv.org/abs/2601.06114v1.

T. Kreuzer, J. Zdravkovic, and P. Papapetrou. PAX-TS: Model-agnostic multi-granular explanations for time series forecasting via localized perturbations. OpenReview preprint, 2025. URL https://openreview.net/forum?id=HIM0n8rsBZ.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37 (4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012. URL https://doi.org/10.1016/j.ijforecast.2021.03.012.

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*, 2024. URL https://arxiv.org/abs/2401.08552v2.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Cindy Marling and Razvan Bunescu. The OhioT1DM dataset for blood glucose level prediction: Update 2020. In *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data*, volume 2675 of *CEUR Workshop Proceedings*, pp. 71–74, 2020. URL https://ceur-ws.org/Vol-2675/paper11.pdf.

Amin Nayebi, Sindhu Tipirneni, Chandan K. Reddy, Brandon Foreman, and Vignesh Subbian. WindowSHAP: An efficient framework for explaining time-series classifiers based on shapley values. *Journal of Biomedical Informatics*, 144:104438, 2023. doi: 10.1016/j.jbi.2023.104438. URL https://doi.org/10.1016/j.jbi.2023.104438.

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

V. C. Raykar, A. Jati, S. Mukherjee, N. Aggarwal, K. Sarpatwar, G. Ganapavarapu, and R. Vaculín. TsSHAP: Robust model agnostic feature-based explainability for time series forecasting. *arXiv preprint arXiv:2303.12316*, 2023. URL https://arxiv.org/abs/2303.12316v1.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953. doi: 10.1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K. 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. URL https://proceedings.neurips.cc/paper/2020/hash/08fa43588c2571ade19bc0fa5936e028-Abstract.html.

David S. Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining predictive uncertainty with information theoretic shapley values. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0320. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/16e4be78e61a3897665fa01504e9f452-Abstract-Conference.html.

YuZhao20. wasserstein-sensitivity: Code and data for GDI and Tangent-SHAP. Zenodo software archive, 2026. URL https://zenodo.org/records/20761059. Version 1.0.0.

Yu Zhao. Unified sensitivity analysis for distributional predictions: A wasserstein geometric framework. *Machine Learning with Applications*, 25:100936, 2026. doi: 10.1016/j.mlwa.2026.100936. URL https://doi.org/10.1016/j.mlwa.2026.100936.

A PREDICTIVE REPRESENTATIONS, NOTATION, AND ASSUMPTIONS

This appendix specifies which forecast distribution each representation supports and the conditions used in the proofs. Fix the observed history x, input groups $N = [n]$, and shared state $Z = (W, \Omega)$ from Definition 2. The uniform input order Π is independent of Z; for an order π, write $S_j^\pi = \{\pi_1, \dots, \pi_j\}$ and $S_0^\pi = \varnothing$. Let $Y_S(Z) \in \mathcal{Y} \subseteq \mathbb{R}^d$ denote the represented coalition forecast.

Definition 5 (Transformed-payoff reconstruction)**.** For a measurable $h : \mathcal{Y} \to \mathbb{R}^q$, set $V_S^h = h(Y_S)$ and

$$D_{\pi_j}^{\pi,h} = h(Y_{S_j^\pi}) - h(Y_{S_{j-1}^\pi}).$$

Let $\Xi_h = \mathcal{L}(V_\varnothing^h, \Pi, D_1^{\Pi,h}, \dots, D_n^{\Pi,h})$. The summation map

$$\mathsf{R}_h(v_0, \pi, d_1, \dots, d_n) = v_0 + \sum_{i=1}^{n} d_i \tag{A.1}$$

satisfies $(\mathsf{R}_h)_\# \Xi_h = \mathcal{L}(h(Y_N))$ by the same telescoping identity as Proposition 1.

Thus contributions to time in range reconstruct time in range; contributions to the identity payoff reconstruct the forecast itself. The transformation precedes subtraction: $h(Y_{S\cup\{i\}}) - h(Y_S)$ generally differs from $h(Y_{S\cup\{i\}} - Y_S)$.

Definition 6 (Predictive representation scope)**.** A joint-trajectory representation supplies $Y_S = (Y_{S,1}, \dots, Y_{S,H})$ and supports reconstruction of its joint distribution. A fixed-horizon representation supplies only $Y_{S,t}$ and supports reconstruction of that marginal distribution. A deterministic model with fixed reference w_0 gives the point mass $\delta_{f(r(x,S,w_0))}$. With a random reference W, its conditional distribution given W is a point mass, while its unconditional distribution can vary across reference draws.

Definition 7 (Complete and finite quantile representations)**.** At a fixed horizon and reference, let $F_{S,t}$ be the coalition forecast CDF, with generalized inverse

$$Q_{S,t}(u) = \inf\{y : F_{S,t}(y) \geq u\}, \qquad u \in (0, 1).$$

A complete quantile representation supplies $Q_{S,t}$ throughout $(0, 1)$. Evaluating every coalition at the same $U \sim \mathrm{Unif}(0, 1)$ defines their common-rank coupling. For a finite grid $a = u_1 < \cdots < u_m = b$, specified monotone postprocessing and interpolation define $\widetilde{Q}_{S,t}$ on $[a, b]$. Its *central-rank distribution* is $(\widetilde{Q}_{S,t})_\# \,\mathrm{Unif}[a, b]$, with rank density $1/(b - a)$.

For each order and rank,

$$Q_{N,t}(u) = Q_{\varnothing,t}(u) + \sum_i D_{i,t}^\pi(u), \qquad D_{\pi_j,t}^\pi(u) = Q_{S_j^\pi,t}(u) - Q_{S_{j-1}^\pi,t}(u). \tag{A.2}$$

Replacing Q by $\widetilde{Q}$ gives exact reconstruction of the interpolated central-rank distribution. This does not generally identify the model's original between-knot values or omitted tails. For random references, the construction instead uses conditional quantiles $Q_{S,t}(u \mid W)$ and retains the same W across coalitions.

Definition 8 (Player and horizon marginals)**.** A player marginal q_i retains one group's contribution distribution and discards its dependence with other groups and the baseline. A horizon marginal retains one coordinate of a predictive trajectory and discards dependence across future times. These are different projections of the joint record.

The results use the following conditions.

1. The input groups partition the coordinates being explained; all other inputs remain fixed. The reference distribution, replacement rule, payoff, coupling, and order distribution stay fixed within each game.
2. The replacement map, forecast realization map, and payoff are measurable, and $r(x, N, W) = x$ almost surely. Each Y_S has the specified represented forecast distribution.
3. All coalitions along an order use the same Z. Contributions use ordinary subtraction and addition in the payoff's ambient Euclidean space.

4. Statements involving means require finite absolute expectations for every coalition payoff.

All distributions are conditional on the fixed observed history x; pooling cases defines a different target. Variation across input orders, references, and forecast draws contributes to the spread of q_i. Only forecast-draw variation describes predictive uncertainty at a fixed reconstructed input; finite-sample estimation error is separate.

B PROOFS AND IDENTIFICATION BOUNDARIES

Proof of Proposition 1. For every order π and shared state z, adjacent differences telescope:

$$V_{\varnothing}(z) + \sum_{j=1}^{n} \big(V_{S_j^\pi}(z) - V_{S_{j-1}^\pi}(z)\big) = V_N(z).$$

This proves Equation (4). Taking the distribution of both sides gives Equation (5), because Ξ retains the baseline and contributions at that same order and state. □

Proposition 2 (Distributional efficiency after averaging over order)**.** *For integrable contributions, define* $\Phi_i(Z) = \mathbb{E}[D_i^\Pi(Z) \mid Z]$. *Then*

$$V_N(Z) - V_{\varnothing}(Z) = \sum_i \Phi_i(Z) \quad \textit{almost surely}. \tag{B.1}$$

Proof. Condition Equation (4) on Z and average over the independent uniform order. The endpoints are independent of the order, so linearity of conditional expectation gives the identity. □

This is the Shapley case of distributional efficiency in Franceschi et al. (2024, Proposition 3.9(iv)). The random variable $\Phi_i(Z)$ retains forecast and reference variation but averages over orders; its distribution can therefore differ from $q_i = \mathcal{L}(D_i^\Pi(Z))$. In Counterexample 1, $\Phi_1 = \Phi_2 = 0$, whereas both q_i assign equal mass to -1 and $+1$.

Proposition 3 (Relation to Distributional Values)**.** *For the shared realization in Definition 2,* $D_i^\Pi(Z)$ *is the distributional Shapley random variable of Franceschi et al. (2024). The joint record* Ξ *also retains the baseline, common order, and dependence among contributions.*

Proof. For a uniform order and $S \subseteq N \setminus \{i\}$,

$$\Pr\{\mathrm{Pre}_i(\Pi) = S\} = \frac{|S|!(n - |S| - 1)!}{n!}.$$

Conditional on this predecessor coalition, $D_i^\Pi(Z) = V_{S\cup\{i\}}(Z) - V_S(Z)$. Mixing these paired differences over S and Z gives the construction in Definitions 3.2–3.3 of Franceschi et al. (2024). □

Corollary 1 (Mean recovery)**.** *If every real-valued coalition payoff is integrable, Equation* (6) *holds.*

Proof. Expanding over predecessor coalitions gives

$$\mathbb{E}[D_i^\Pi] = \sum_{S \subseteq N\setminus\{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(\mathbb{E}[V_{S\cup\{i\}}] - \mathbb{E}[V_S]\right),$$

the scalar Shapley formula for the expected payoff. □

Proposition 4 (Null player)**.** *If* $V_{S\cup\{i\}}(Z) = V_S(Z)$ *almost surely for every* $S \subseteq N \setminus \{i\}$, *then* $D_i^\Pi = 0$ *almost surely and* $q_i = \delta_0$. *A zero mean alone is insufficient.*

Proof. Every possible paired marginal contribution is zero. Their mixture over orders is therefore concentrated at zero. Counterexample 5 shows why cancellation in expectation does not imply this premise. □

Proposition 5 (Distributional symmetry). *Let τ exchange players i and j. If $V_{\tau S}(Z) = V_S(Z)$ almost surely for every $S \subseteq N$, then, under a uniform permutation, D_i^Π and D_j^Π have the same probability distribution; equivalently, $q_i = q_j$.*

Proof. The bijection $\pi \mapsto \tau\pi$ preserves uniform measure and maps the predecessor contribution of i to the corresponding contribution of j. The assumed game invariance makes those paired quantities equal almost surely. □

Proposition 6 (Deterministic forecast reduction). *With no forecast randomness, Ξ varies only over references and input orders. With a fixed reference as well, it is a finite mixture over the $n!$ orders.*

Proof. Set $\Omega = \omega_0$ in Definition 2; fixing W then leaves only Π random. □

Proposition 7 (Coupling identification). *Separate coalition payoff distributions do not generally determine Ξ or the player marginals q_i.*

Proof. Counterexample 2 gives identical coalition distributions with different paired contribution distributions. □

Proposition 8 (Joint-process invariance). *Two realization maps with the same joint distribution of $(V_S)_{S\subseteq N}$ induce the same Ξ and every q_i. When the payoffs are integrable, they also induce the same joint distribution of $(\Phi_1, \ldots, \Phi_n)$.*

Proof. The baseline and contribution record are measurable functions of the coalition-payoff vector and an independent uniform order. Each Φ_i is a finite weighted sum of that payoff vector. Equality of the joint input distributions therefore gives equality of the resulting distributions. □

Proposition 9 (Horizon-marginal scope). *Separate predictive distributions at each horizon do not generally identify a joint trajectory distribution or the distribution of a path-dependent payoff.*

Proof. Counterexample 3 gives two trajectory distributions with the same horizon marginals and different probabilities for the same path event. □

Proposition 10 (Bernoulli marginal contribution). *For fixed $p^-, p^+ \in [0,1]$, let $V^- = \mathbf{1}\{U \le p^-\}$ and $V^+ = \mathbf{1}\{U \le p^+\}$ share $U \sim \mathrm{Unif}(0,1)$. Then $D = V^+ - V^-$ has masses*

$$\Pr(D=1) = (p^+ - p^-)_+, \quad \Pr(D=-1) = (p^- - p^+)_+, \quad \Pr(D=0) = 1 - |p^+ - p^-|, \tag{B.2}$$

where $(z)_+ = \max(z, 0)$ (Franceschi et al., 2024, Section 3.1).

Proof. If $p^+ > p^-$, $D = 1$ on $(p^-, p^+]$ and is zero elsewhere. If $p^+ < p^-$, $D = -1$ on $(p^+, p^-]$ and is zero elsewhere. The interval lengths give the stated probabilities. □

B.1 Counterexamples

Counterexample 1 (Independent convolution loses reconstruction). Let B be equally likely -1 or $+1$ and define a two-player game by

$$V_\varnothing = V_{\{1,2\}} = 0, \qquad V_{\{1\}} = V_{\{2\}} = B,$$

using the same B throughout a chain. In player-ID order, the contributions are $(B, -B)$ for permutation $(1,2)$ and $(-B, B)$ for permutation $(2,1)$. Every joint row sums to zero, so the reconstructed grand-coalition payoff distribution is δ_0. Each player marginal is nevertheless $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$. Convolving independent draws from those marginals assigns probabilities $1/4, 1/2, 1/4$ to $-2, 0, +2$ and therefore does not reconstruct the grand-coalition payoff distribution.

Counterexample 2 (Equal payoff marginals, different couplings). Let $U \sim \mathrm{Unif}(0,1)$ and give the present and absent payoffs the same $\mathrm{Ber}(1/2)$ distribution. The comonotone realization

$$V^+ = V^- = \mathbf{1}\{U \le 1/2\}$$

gives $D = 0$ almost surely. The antimonotone realization

$$V^+ = \mathbf{1}\{U \le 1/2\}, \qquad V^- = \mathbf{1}\{U > 1/2\}$$

gives equal probabilities to $D = -1$ and $D = +1$. The coalition payoff distributions and base-noise distribution agree, but the paired contribution distributions do not.

Counterexample 3 (Equal horizon marginals, different path events)**.** For $B \sim \mathrm{Ber}(1/2)$, consider

$$Y^{\text{persistent}} = (B, B), \qquad Y^{\text{alternating}} = (B, 1 - B).$$

Both constructions have $\mathrm{Ber}(1/2)$ marginals at each horizon, but

$$\Pr(Y_1 > 1/2,\, Y_2 > 1/2) = 1/2 \quad \text{and} \quad 0,$$

respectively. For the duration $T = \mathbf{1}\{Y_1 > 1/2\} + \mathbf{1}\{Y_2 > 1/2\}$, the persistent construction gives $T \in \{0, 2\}$ with equal probabilities, whereas the alternating construction gives $T = 1$ surely. Both have $\mathbb{E}[T] = 1$. Thus horizon marginals identify expected duration but need not identify its distribution.

Counterexample 4 (Finite quantiles leave interpolation and tails unspecified)**.** Suppose only $Q(0.1) = 0$ and $Q(0.9) = 1$ are supplied. One compatible quantile function rises linearly between them; another stays at zero through rank 0.5 and equals one above that rank. An interpolation rule selects a representation without identifying which original curve generated the supplied values.

For a separate tail comparison, let two quantile functions agree with the linear interpolant on $[0.1, 0.9]$. Complete the first with zero below 0.1 and one above 0.9. Let the second equal -3 on $(0, 0.05]$ and 3 on $(0.95, 1)$, and otherwise match the first. These completions have the same supplied quantiles but give $\Pr(Y \geq 2) = 0$ and 0.05, respectively.

Counterexample 5 (Zero mean is not a null contribution)**.** With one player, let $V_{\varnothing} = 0$ and let $V_{\{1\}} = 2B - 1$ for $B \sim \mathrm{Ber}(1/2)$. Then $D_1 \in \{-1, +1\}$ with equal probabilities and $\mathbb{E}[D_1] = 0$, although adding the player changes every realized payoff.

C Implementation and predictive interfaces

Run specification and retained records

Each run specifies the model, forecast case, input groups, reference, replacement rule, payoff, forecast representation, coupling, and sampling budget. The input groups are disjoint, and the replacement rule identifies derived covariates that must be recomputed. Study-specific references are described in Appendix D; the case studies and numerical comparisons use different references. Algorithm 1 computes the paired contributions.

For case c, input order Π_k, and shared draw Z_{ckr}, the reconstruction record contains the baseline, order, and contribution vector from Definition 4. Validation also retains the adjacent coalition payoffs and the full-input payoff. Case, order, shared-state, payoff, and coordinate identifiers link these quantities when stored in separate rows. With a random reference, the shared-state identifier must also identify the reference draw.

Validation and storage

Before inference, the implementation checks which input coordinates are replaced, fixed, or recomputed. Before returning contributions, it checks coalition-chain connectivity, equality of shared adjacent outputs, and reconstruction for each case, order, shared state, payoff, and coordinate. Some specialized runners repeat these checks after reloading stored records; the generic interface validates records in memory.

For finite quantiles, equality at every knot implies equality between knots when all terms share a grid and piecewise-linear interpolation. A stored representation must therefore retain the common grid, baseline and contribution functions, supported rank interval, and treatment of crossings and tails. The reported in-memory checks do not establish that every interface preserves complete records after reloading.

Sampling weights. The budget K counts all realized orders. Reverse pairing draws $K/2$ independent uniform orders and appends each reversal, requiring even K. Each realized order receives weight $1/K$, and each of its R sampled states receives weight $1/R$, giving the atom weights in Equation (7).

For finite quantiles, Equation (8) integrates over ranks within each order. Reverse-paired orders are dependent and must be treated as pairs when estimating sampling error; reuse of the same forecast bank introduces additional dependence.

PREDICTIVE INTERFACES

TACTiS-2 supplies joint trajectory draws. Using shared random numbers across coalitions pairs their paths; retaining all horizons under each draw identifier preserves the trajectory needed for duration payoffs. This is an algorithmic pairing, whose interpretation depends on the chosen sampler.

TFT and Chronos-2 supply finite marginal quantiles. We use a common knot grid and linear interpolation over the supported ranks, as defined in Appendix A. The baseline and every contribution are evaluated at the same rank. Their joint representation supports reconstruction; separate player summaries discard the required dependence. These quantile interfaces do not supply the complete quantile function on $(0, 1)$.

PatchTST supplies deterministic forecasts. With a fixed reference, each coalition has one point or trajectory vector, so reconstruction requires no forecast-draw budget. The same order-based calculation applies.

C.1 EVENT AND DURATION PAYOFFS

A binary payoff asks whether a forecast satisfies an event, such as glucose being in range at one future time.

Definition 9 (Aligned binary-event contribution)**.** For a measurable event A, set $B_S(Z) = \mathbf{1}\{Y_S(Z) \in A\}$ and $p_S = \Pr(B_S = 1)$. The contribution

$$E^{\pi}_{\pi_j}(Z) = B_{S^{\pi}_j}(Z) - B_{S^{\pi}_{j-1}}(Z) \in \{-1, 0, +1\}$$

records whether adding a group switches the event off, leaves it unchanged, or switches it on. The paired record reconstructs the full-input indicator:

$$B_N(Z) = B_{\varnothing}(Z) + \sum_i E^{\pi}_i(Z).$$

Taking expectations allocates the event-probability change: $p_N - p_{\varnothing} = \sum_i \mathbb{E}[E^{\Pi}_i]$.

Definition 10 (Shared-uniform Bernoulli game)**.** Given the coalition event probabilities, a different coupling is

$$B^{\mathrm{unif}}_S(\Lambda) = \mathbf{1}\{\Lambda \le p_S\}, \qquad \Lambda \sim \mathrm{Unif}(0, 1).$$

This common selector gives the contribution probabilities in Proposition 10. It preserves each coalition's event probability and expected contribution, but can change the contribution distribution relative to the original forecast-induced coupling. The TACTiS-2 event study uses indicators of paired forecast trajectories, as in Definition 9.

Definition 11 (Expected time in range)**.** For equally spaced forecasts with interval Δ and target range $I = [\ell, u]$, write $p_{S,t} = \Pr(Y_{S,t} \in I)$. The expected duration is

$$\theta(S) = \Delta \sum_{t=1}^{H} p_{S,t}. \tag{C.1}$$

Its scalar Shapley allocation is

$$\tau_i = \Delta \sum_{t=1}^{H} \mathbb{E}_{\Pi}\left[p_{\mathrm{Pre}_i(\Pi)\cup\{i\},t} - p_{\mathrm{Pre}_i(\Pi),t}\right]. \tag{C.2}$$

By linearity, $\sum_i \tau_i = \theta(N) - \theta(\varnothing)$. Both forecast-induced indicators and shared-uniform indicators give these expected allocations when they use the same probabilities $p_{S,t}$. For unequal forecast spacing, replace Δ by Δt_t inside the sums, as in Equation (2).

A finite quantile grid instead identifies an event probability for its interpolated central-rank distribution:

$$\widetilde{p}_{S,t}^{\text{central}} = \frac{1}{b-a}\int_a^b \mathbf{1}\{\widetilde{Q}_{S,t}(u) \in I\}\, du.$$

If the representation is extended outside $[a, b]$ with all omitted ranks known to fall outside I, its full-rank event probability is $(b-a)\widetilde{p}_{S,t}^{\text{central}}$. This remains the probability under the specified interpolant and tail assumptions; it equals the original model probability only if those assumptions preserve the original in-range rank mass.

Horizon probabilities suffice for expected duration without an independence assumption. The distribution of realized duration $T_S = \Delta \sum_t \mathbf{1}\{Y_{S,t} \in I\}$ generally also depends on the joint horizon indicators (Counterexample 3). Its paired contribution is $T_{S\cup\{i\}} - T_S$, which can take more than the three values of a single event contribution. The case studies use this trajectory-duration payoff; the terminal-event diagnostics use a single horizon (Appendix D.2).

C.2 Dependence across groups and horizons

Definition 12 (Cross-player and cross-horizon dependence)**.** Cross-player dependence is the dependence among contributions $(D_1^\Pi, \ldots, D_n^\Pi)$ under a common order and shared state. Cross-horizon dependence is the dependence among future coordinates $(Y_{S,1}, \ldots, Y_{S,H})$ of one coalition forecast.

The first dependence matters when adding contributions to reconstruct a forecast; the second matters when explaining quantities across future times. A complete trajectory record retains both, together with dependence on the baseline. Player marginals alone cannot reconstruct the paired sum, and horizon marginals alone cannot generally reconstruct a duration distribution.

Using one common rank across horizons imposes a comonotone trajectory, whereas independent ranks impose horizon independence for a fixed input. Neither choice is determined by marginal quantiles. The TFT and Chronos-2 analyses therefore concern fixed-horizon distributions. The TACTiS-2 duration analysis instead uses the model's joint trajectories, paired across coalitions through shared randomness. Applying a scalar summary to those trajectories reconstructs that summary's distribution, as specified in Definition 5.

D Evaluation and results

The experiments address three questions: whether TDV reconstructs the represented forecast and recovers scalar Shapley values, how its estimates and cost change with sampling choices, and what its distributions reveal in selected glucose forecasts. Each comparison fixes the input groups, reference, payoff, forecast representation, and pairing rule. The protocols below distinguish case selection before attribution inspection from the outcome-aware selection used for the Type 2 example.

D.1 Numerical evaluation

Reconstruction and sampling sensitivity. For each sampled record, we check the residual

$$e_{kr} = V_N(Z_{kr}) - V_\varnothing(Z_{kr}) - \sum_i D_i^{\Pi_k}(Z_{kr}).$$

The TACTiS-2 reconstruction study uses 12 cases per cohort, four 30-minute glucose-history groups, and 256 trajectories per case. Cases were selected without inspecting attribution results. All 16 coalitions and 24 orders are evaluated for the exact comparison, with a 10^{-9} mg/dL tolerance for scalar-mean and reconstruction residuals.

We assess sensitivity to forecast draws by omitting each of four 64-trajectory groups and measuring Wasserstein-1 distances between the resulting contribution estimates. Separately, we hold the trajectory bank fixed and estimate mean contributions using $K \in \{32, 128, 512\}$ reverse-paired orders. Root-mean-square error against exact scalar Shapley values measures sensitivity to the order budget (Figure 5). These comparisons vary the estimator within fixed forecast cases.

Computation time. The TACTiS-2 sweep uses $n \in \{2, 4, 6, 8\}$ consecutive groups over 120 minutes of glucose history, 12 cases per cohort, and 256 trajectories per coalition. Its scalar target is glucose at 60 minutes, averaged over those trajectories. Forecast generation and storage are timed separately from attribution using the stored forecasts. The first two-group TimeSHAP call includes package loading. Table 5 states the timer scope for each model; Figure 4 shows the TACTiS-2 scaling.

Agreement with scalar Shapley values. For the models and targets in Table 2, we compare the TDV mean and TimeSHAP with exact Shapley values computed from the same coalition forecasts. Each model–dataset pair uses 12 cases at each $n \in \{2, 4, 6, 8\}$, with all 2^n coalitions evaluated. PatchTST supplies a point forecast, TFT and Chronos-2 supply represented quantile means, and TACTiS-2 supplies a mean over sampled trajectories. The expected payoffs differ across interfaces; all comparisons use the same scalar Shapley identity.

For the exhaustive TACTiS-2 comparison, the tolerances are 10^{-8} mg/dL for TimeSHAP and summed-attribution residuals, and 10^{-9} mg/dL for TDV means. At eight groups, each cohort contributes 96 case–group comparisons. The reported residuals test agreement conditional on stored forecasts.

D.2 Time-in-range case-study protocol

We explain predicted minutes in the inclusive interval $[70, 180]$ mg/dL, using the duration payoff in Equation (2). The interval follows the usual diabetes time-in-range convention (Battelino et al., 2019); here it defines a forecast quantity, including for the H&W example. These selected cases illustrate explanations for particular forecasts.

Input groups and reference. Each case uses 120 minutes of input history and two separate partitions. The temporal partition comprises twelve consecutive 10-minute windows; replacing a window replaces its mutable input variables. The variable partition groups inputs across the full 120 minutes. The reported feature calculations use four groups for OhioT1DM and three for H&W and Type 2. Each partition independently allocates the same full-input minus reference change.

For OhioT1DM and H&W, the glucose reference is the median of observed values in the stable first 60 minutes of the explanation window, with derived glucose inputs recomputed. Other mutable variables use their specified no-event values; calendar and other fixed context retain observed values. This reference differs from the training-data median used in the numerical comparisons.

OhioT1DM and H&W selection. An initial feasibility screen found no eligible OhioT1DM case among 5,904 cases with a usable reference. Before inspecting attributions, the revised eligibility criteria required expected time in range between 30% and 70% of the forecast horizon, a meal recorded within 120 minutes, a stable reference prefix, and multiple history windows differing from the reference. Inflection and later high-glucose evidence were used as ranking preferences. Final selection used case identifiers and reference/full-input expected time in range, favoring larger decreases from the reference. Player contributions, observed future glucose, and forecast errors were inspected only after the two cases were selected.

Forecast and attribution budgets. TACTiS-2 generates 256 paired trajectories per coalition at five-minute spacing, over 60 minutes for OhioT1DM and 90 minutes for H&W. The duration payoff counts in-range coordinates along each trajectory; the terminal-event payoff uses only the last coordinate. The temporal game uses 64 reverse-paired orders, 510 distinct coalitions, and seed 729301. The variable game enumerates all 24 orders and 16 coalitions for OhioT1DM, or six orders and eight coalitions for H&W. Averaging contributions gives the scalar attribution for that game's expected payoff.

Replaying the stored forecast randomness checks trajectory reconstruction, summed-mean efficiency, and agreement with scalar attributions. The OhioT1DM run used 8,385 model calls and generated 134,144 trajectories in 737.3 seconds; H&W used 8,257 calls and 132,096 trajectories in 580.0 seconds. These totals span coalition forecasts and checks, rather than the 256 trajectories retained for one coalition.

Nested prefixes of 32, 64, 128, and 256 trajectories assess sensitivity without additional model calls. This analysis was specified after inspecting OhioT1DM contributions, but before inspecting prefix

results or H&W contributions. Appendix D.6 reports the diagnostics and the separate checks with fresh trajectory banks.

Outcome-aware Type 2 selection. The Type 2 study screened 512 held-out meal-aligned windows and selected eight eligible candidates before inspecting attributions. Final selection then used the attribution outputs: at least three temporal groups and two variable groups had to have absolute mean contributions of at least one minute, with temporal dominance below 0.8. One candidate qualified, with four temporal groups, two variable groups, and recorded dominance 0.7855. Observed future glucose was not used, and neither additional candidates nor substitutions were introduced after inspecting the outputs.

The reference and full-input forecasts yield 56.91 and 20.66 expected in-range minutes over the 60-minute horizon. The temporal game uses 256 trajectories, 64 reverse-paired orders, and 510 distinct coalitions; the three-group variable game uses the same trajectories and all six orders over eight coalitions. Both reconstruct the -36.25-minute change, with zero reported replay, efficiency, and pathwise reconstruction residuals. Figure 2 shows contribution means and empirical 5th–95th percentiles over trajectories and input orders. The percentiles describe contribution variation within the selected game.

D.3 Scalar recovery and computation time

Table 5 reports scalar recovery errors and computation times for all two-, four-, six-, and eight-group comparisons. The main summary in Table 3 uses the largest errors and average per-case attribution times from the six- and eight-group settings. Figure 3 plots the corresponding errors.

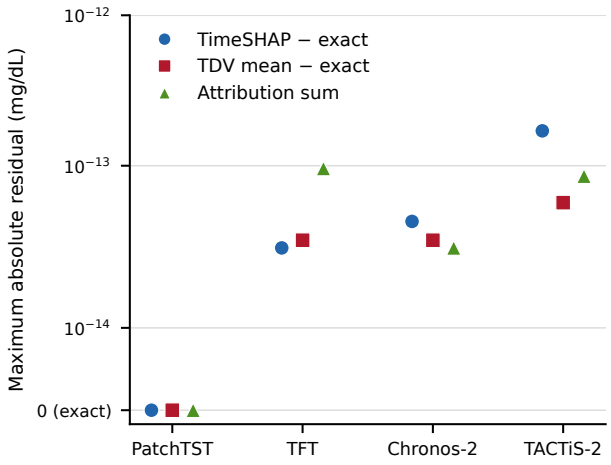


Figure 3: **Agreement with exact scalar Shapley values.** For the six- and eight-group comparisons, circles show the maximum absolute TimeSHAP error and squares show the maximum absolute error in the mean TDV contribution. Triangles show the maximum absolute difference between the summed mean contributions and the scalar forecast's change from its reference. The nonzero errors are consistent with floating-point rounding, supporting numerical agreement with scalar attribution. Exact zeros appear on a separate baseline below the logarithmic scale.

Table 3 averages the six- and eight-group attribution times per case: sum the corresponding totals and divide by 24 for PatchTST, TFT, and Chronos-2, or 48 for TACTiS-2 across both cohorts. The averages use unrounded timings. PatchTST forecast generation, including full-context replay, took 64.69 seconds across all four group counts; separate per-grid times are unavailable. TFT and Chronos-2 forecast times include setup and coalition inference but exclude serialization; TACTiS-2 times include inference and serialization after model loading and warm-up. TDV and TimeSHAP timings exclude model calls, while each first two-group TimeSHAP timing includes package loading.

Table 5: **Scalar recovery errors and computation time for all input-group counts.** Each row uses 12 forecast cases and all 2^n coalitions per case. TimeSHAP and TDV mean errors are maximum absolute deviations from exact scalar Shapley values; Sum measures the discrepancy between summed mean contributions and the full-input minus reference forecast. Times are total wall seconds across the 12 cases; attribution uses stored forecasts. Nonzero values below 0.01 use scientific notation, and a dash indicates an unavailable timer.

					Maximum error (mg/dL)			Total time (s)		
Model	Cohort	n	Cases	2^n	TimeSHAP	TDV mean	Sum	Forecasts	TDV	TimeSHAP
PatchTST	Type 2	2	12	4	0.00	0.00	0.00	–	2.84×10^{-5}	2.15
PatchTST	Type 2	4	12	16	0.00	0.00	0.00	–	2.14×10^{-4}	0.02
PatchTST	Type 2	6	12	64	0.00	0.00	0.00	–	8.67×10^{-3}	0.03
PatchTST	Type 2	8	12	256	0.00	0.00	0.00	–	0.65	0.06
TFT	Type 2	2	12	4	0.00	6.75×10^{-14}	7.82×10^{-14}	11.85	0.22	2.13
TFT	Type 2	4	12	16	7.11×10^{-15}	4.30×10^{-14}	4.44×10^{-14}	8.83	0.67	0.06
TFT	Type 2	6	12	64	2.13×10^{-14}	3.20×10^{-14}	9.41×10^{-14}	24.93	1.94	0.08
TFT	Type 2	8	12	256	2.84×10^{-14}	2.49×10^{-14}	9.59×10^{-14}	90.05	8.08	0.17
Chronos-2	H&W	2	12	4	0.00	2.84×10^{-14}	2.84×10^{-14}	15.28	0.21	1.98
Chronos-2	H&W	4	12	16	7.11×10^{-15}	1.89×10^{-14}	2.84×10^{-14}	21.60	0.43	0.05
Chronos-2	H&W	6	12	64	2.49×10^{-14}	1.30×10^{-14}	2.84×10^{-14}	80.99	0.94	0.08
Chronos-2	H&W	8	12	256	4.26×10^{-14}	3.20×10^{-14}	2.84×10^{-14}	309.52	2.72	0.16
TACTiS-2	H&W	2	12	4	0.00	1.44×10^{-14}	2.29×10^{-14}	43.08	3.27×10^{-3}	1.85
TACTiS-2	H&W	4	12	16	2.36×10^{-15}	1.35×10^{-14}	1.60×10^{-14}	155.57	0.02	0.02
TACTiS-2	H&W	6	12	64	7.11×10^{-15}	9.77×10^{-15}	1.60×10^{-14}	626.65	0.10	0.03
TACTiS-2	H&W	8	12	256	1.33×10^{-14}	7.99×10^{-15}	1.33×10^{-14}	2449.87	0.86	0.10
TACTiS-2	OhioT1DM	2	12	4	7.11×10^{-15}	2.86×10^{-14}	5.68×10^{-14}	56.43	3.29×10^{-3}	2.00
TACTiS-2	OhioT1DM	4	12	16	4.26×10^{-14}	2.30×10^{-14}	7.11×10^{-14}	220.94	0.02	0.02
TACTiS-2	OhioT1DM	6	12	64	1.14×10^{-13}	1.67×10^{-14}	3.55×10^{-14}	881.82	0.13	0.03
TACTiS-2	OhioT1DM	8	12	256	1.71×10^{-13}	5.68×10^{-14}	8.53×10^{-14}	3523.48	1.18	0.10

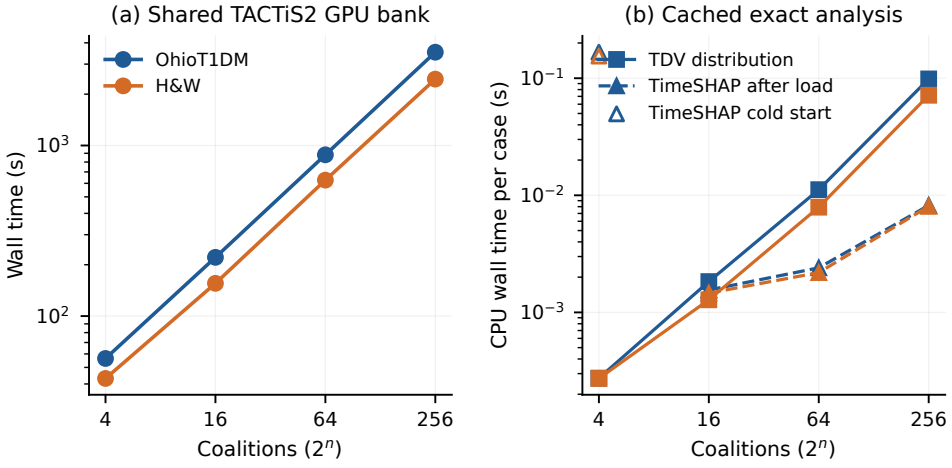


Figure 4: **Forecast-generation and attribution cost.** Left: total wall time to generate and store TACTiS-2 trajectory forecasts over 12 cases per cohort and all coalitions at each group count. Right: per-case CPU analysis time for TDV and TimeSHAP using those stored forecasts. The hollow four-coalition TimeSHAP points include initial package loading.

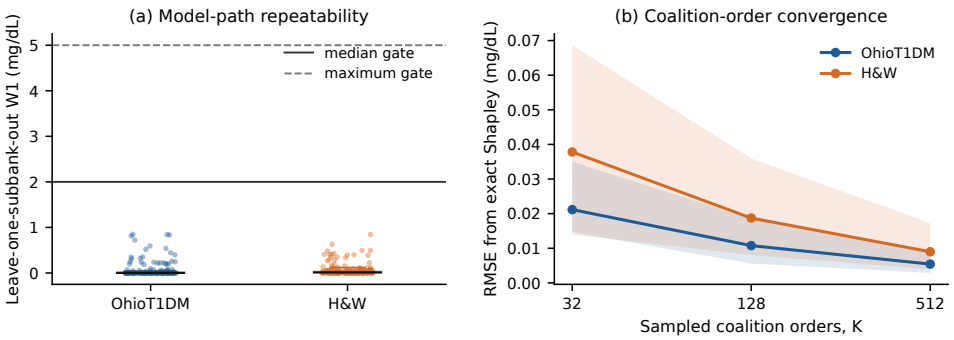


Figure 5: **Sensitivity to trajectory and input-order sampling.** Left: Wasserstein-1 distances between contribution estimates when one of four 64-trajectory groups is omitted. Right: root-mean-square error of estimated mean contributions against exact scalar Shapley values for $K = 32, 128, 512$ input orders; lines show cohort medians and bands show case ranges. The figure's *repeatability* and *convergence* headings refer to these checks within the fixed cases and trajectory banks.

D.4 Nine-window examples across forecasting models

These examples explain a forecast at 60 minutes using nine consecutive 10-minute input windows. Window W9 is the newest; code-style window labels in the figures use the same ordering. The upper panels show the observed history and forecast, while the lower panels show the represented endpoint distribution and signed probability changes allocated to input windows. For a glucose bin B_b at horizon t, write $p_{S,b} = \Pr(Y_{S,t} \in B_b)$ under the represented coalition forecast. Its signed allocation to input group i is

$$a_{i,b} = \mathbb{E}_{\Pi}\left[p_{\mathrm{Pre}_i(\Pi)\cup\{i\},b} - p_{\mathrm{Pre}_i(\Pi),b}\right], \qquad \sum_i a_{i,b} = p_{N,b} - p_{\varnothing,b}.$$

Finite-order estimates replace the expectation by an average over the evaluated orders. Thus these bars are scalar Shapley allocations of bin-probability changes, or equivalently mean contributions for the indicator payoff $\mathbf{1}\{Y_{S,t} \in B_b\}$. They differ from q_i, the distribution of paired glucose changes in mg/dL. Figure 1 provides the Chronos-2 example; the following figures provide PatchTST, TFT, and TACTiS-2 examples. The cases, references, and forecast representations differ across models.

Model	Primary change	Alternate change	Newest-window contribution
PatchTST	+156.39	+145.35	+159.51
TFT	−138.61	−138.91	−83.72
Chronos-2	+9.22	+41.27	+17.81
TACTiS-2	+5.39	+8.93	+9.47

Table 6: **Reference sensitivity in the nine-window examples.** Values are mg/dL at 60 minutes. The change is the full-input minus reference scalar forecast under the primary or alternate reference. The final column gives the newest window's mean contribution under the primary reference.

The cases were selected without inspecting player attributions: PatchTST and TFT used input-based rules, Chronos-2 used a candidate roster and eligibility screen, and TACTiS-2 used input-based selection for the related H&W example. Table 6 shows that changing the reference changes the explained forecast difference, especially for Chronos-2. The newest window has the largest primary mean contribution in magnitude in each example. For TACTiS-2, this ordering also holds in each of four disjoint 64-trajectory subsets of the original 256-trajectory bank.

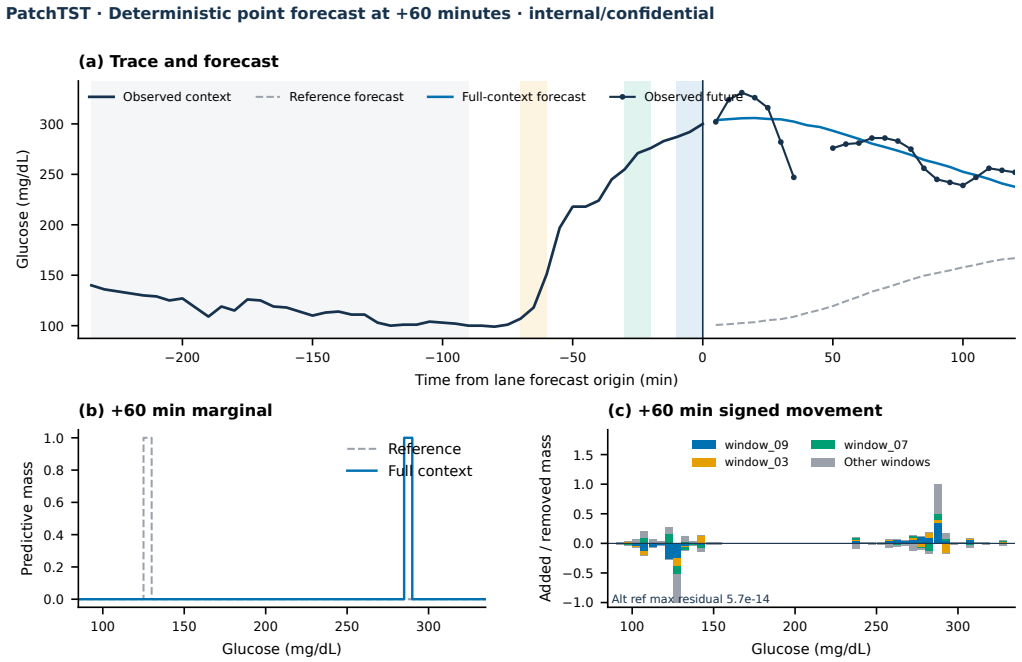


Figure 6: **PatchTST: contributions to a deterministic forecast.** The lower panels show the point forecasts as binned probability masses and their signed changes at 60 minutes. The newest window raises the prediction by 159.51 mg/dL, partly offset by the other windows. The observed-future line has gaps at +40 and +45 minutes because those stored targets were imputed.

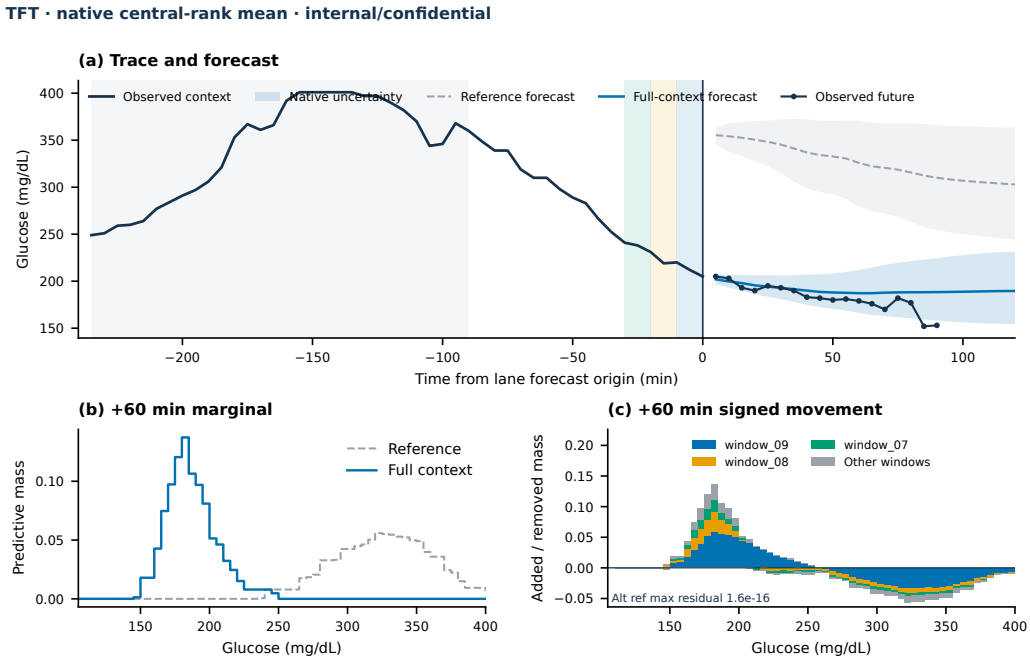


Figure 7: **TFT: contributions to a marginal forecast.** At 60 minutes, the represented mean decreases by 138.61 mg/dL from the primary reference; the newest window contributes −83.72 mg/dL. The lower panels show the represented probability masses and signed changes over glucose bins. These distributions use the supported quantile ranks and do not identify omitted predictive tails.

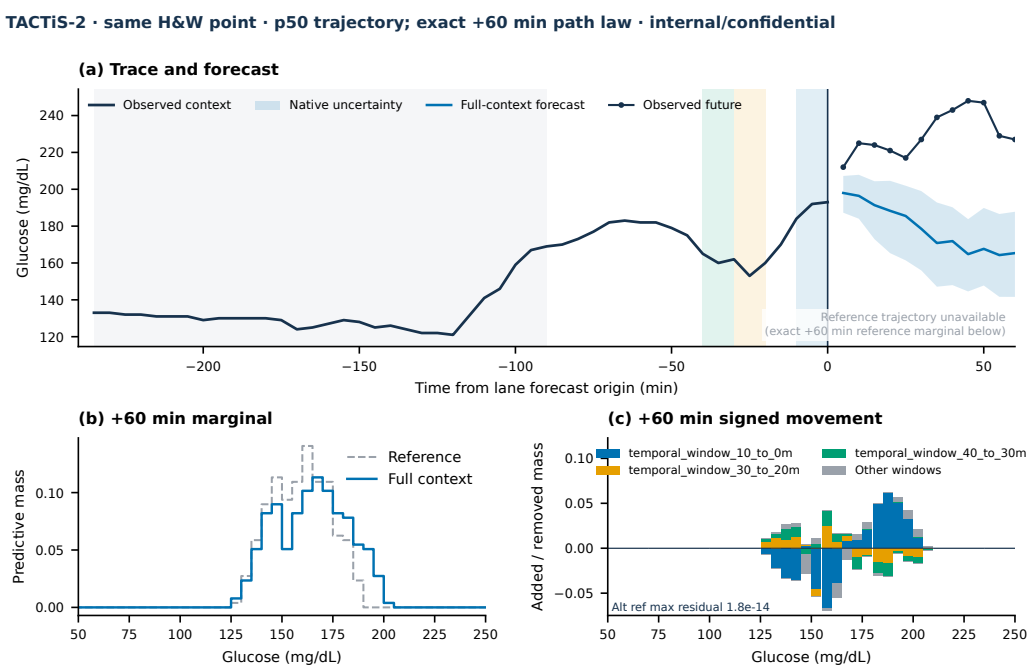


Figure 8: **TACTiS-2: a marginal view of paired trajectory forecasts.** This H&W example is related to the Chronos-2 case in Figure 1, with its forecast origin 30 minutes later. At 60 minutes, the mean rises by 5.39 mg/dL, with +9.47 mg/dL attributed to the newest window. The reference trajectory is unavailable in the upper panel; the lower panel includes its endpoint distribution.

D.5 Time-in-range companion cases

The OhioT1DM and H&W examples explain predicted minutes in $[70, 180]$ mg/dL over 60 and 90 minutes, respectively, using the protocol in Appendix D.2. Both use 256 trajectories and 64 temporal orders; their variable-group calculations enumerate all orders. Means allocate the change in expected duration. The empirical 5th–95th percentiles describe variation across trajectories and orders, rather than uncertainty in an estimated population mean.

For OhioT1DM, expected time in range falls from 60.00 to 30.06 minutes. The newest window contributes -29.70 minutes, with the other eleven summing to -0.24 minutes. The separate variable allocation assigns -29.87 minutes to glucose history and -0.07 minutes to meals.

For H&W, expected time in range falls from 90.00 to 45.92 minutes. The newest two windows contribute -33.36 and -6.00 minutes; the variable allocation assigns -44.77 minutes to glucose history and $+0.68$ minutes to meals. Together with the Type 2 example in Figure 2, these cases motivate checking how the forecaster uses meal records and information already present in glucose history.

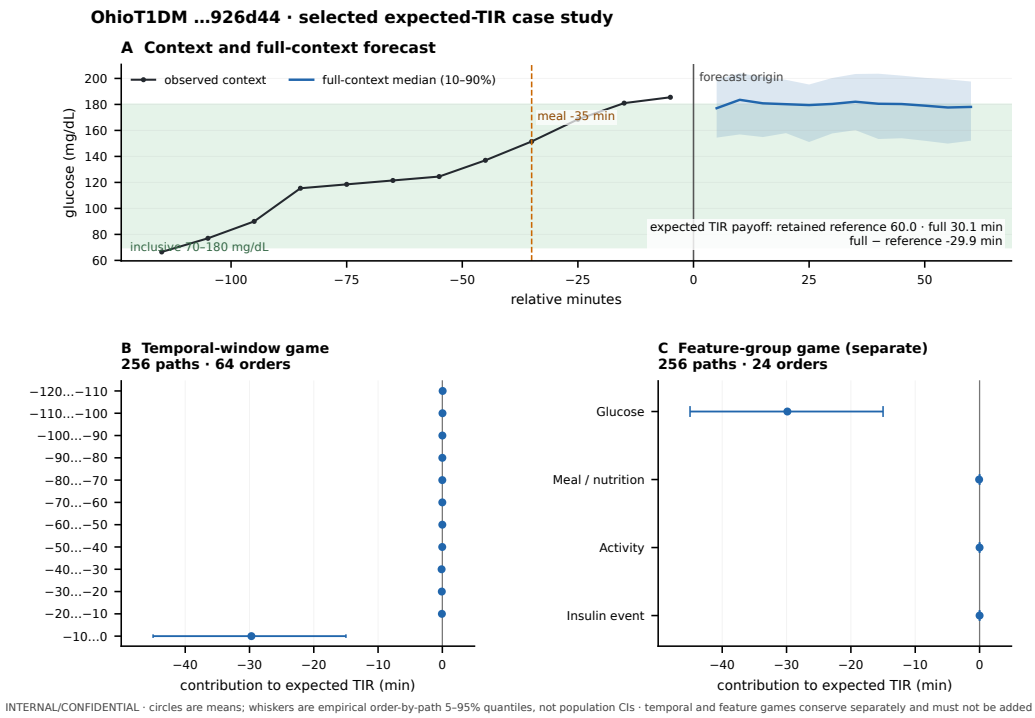


Figure 9: **OhioT1DM time-in-range example.** (A) Observed history and full-input forecast over 60 minutes, with a 10–90% predictive band. Expected in-range duration falls from 60.00 reference minutes to 30.06 full-input minutes. (B,C) Separate temporal and variable-group explanations; circles show means and whiskers empirical 5th–95th percentiles over trajectories and orders. The newest window and glucose history account for most of the decrease.

D.6 Contribution distributions and sampling diagnostics

Figure 10 extends the scalar comparison by showing temporal contribution distributions for TFT and Chronos-2 at three forecast horizons. The horizontal axis records signed contribution size in mg/dL, and color records nonnegative probability mass, averaged equally over the 12 cases. The mass near zero and the less frequent larger changes describe features of the contribution distribution that its mean alone does not specify.

The terminal-event analysis asks whether adding each group switches the last predicted glucose value into or out of range. Figure 11 separates the probabilities of negative, zero, and positive switches. The positive-switch probability minus the negative-switch probability gives the mean event contribution, while the separate probabilities distinguish unchanged outcomes from opposing switches.

Figure 12 separates variation in contributions from sensitivity of their estimated means to the trajectory count. Using 128 trajectories instead of the full 256 changes the largest player mean by 0.62 minutes for OhioT1DM and 0.70 minutes for H&W. The final 256-trajectory comparison is zero by construction because it uses the same bank as its reference.

Shuffling the alignment of contribution records preserves their separate distributions but changes the reconstructed endpoint distribution (Figure 13, left). This illustrates why reconstruction retains dependence among contributions and with the baseline. Three independent 256-trajectory banks provide a separate check of sampling variation for the same cases: each retains negative mean glucose contributions and the original sign of the smaller meal contribution. Across those banks, glucose means range from -30.34 to -29.83 minutes for OhioT1DM and from -45.05 to -42.88 minutes for H&W. Meal means range from -0.060 to -0.028 and from $+0.576$ to $+0.732$ minutes, respectively.

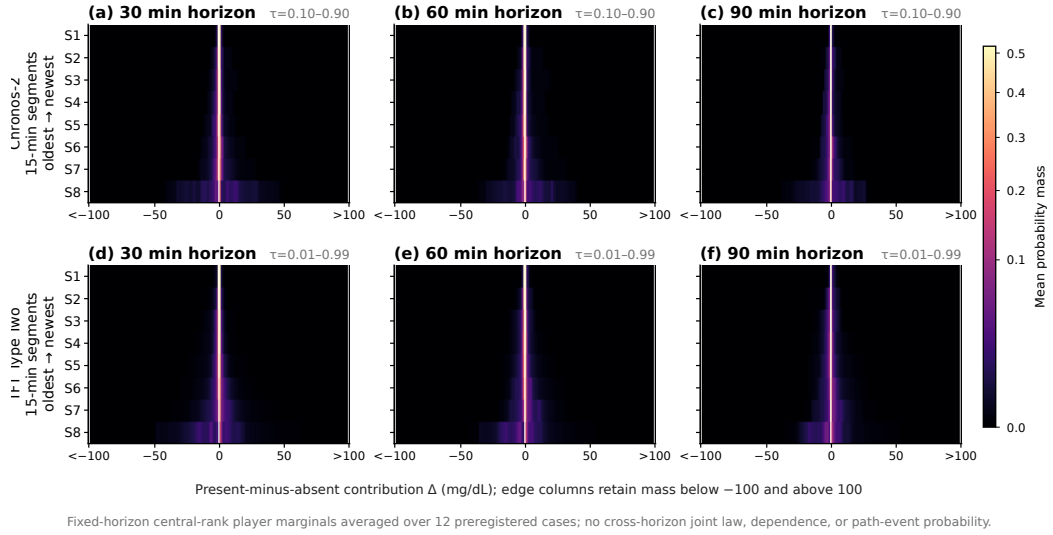


Figure 10: **Temporal contribution distributions for marginal forecasts.** Each row S1–S8 corresponds to a 15-minute input window, ordered oldest to newest. Columns bin contribution sizes in mg/dL, including values below -100 and above 100; color uses a square-root scale for probability mass. Panels average equally across 12 cases and show 30-, 60-, and 90-minute horizons separately. Chronos-2 represents ranks $[0.10, 0.90]$ and TFT $[0.01, 0.99]$. These are marginal distributions at each horizon.

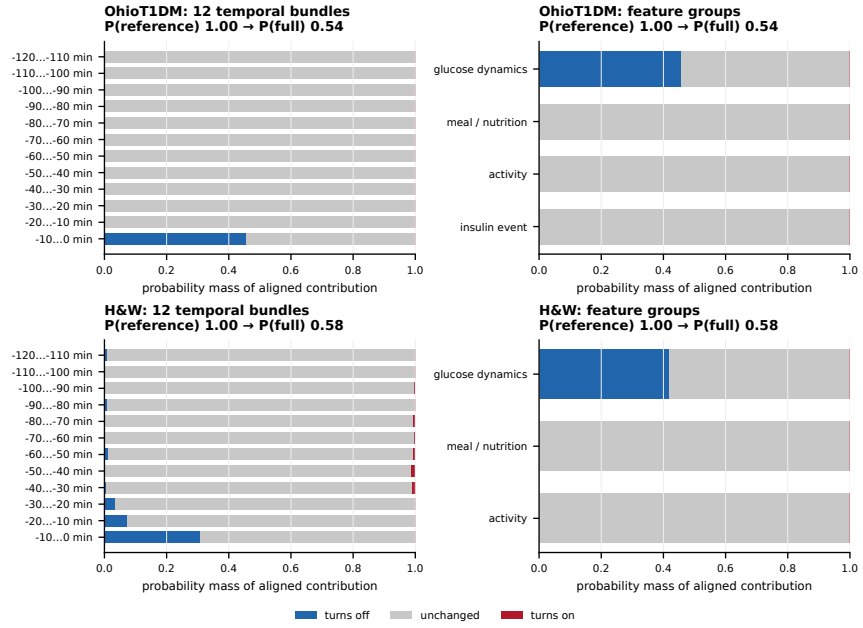
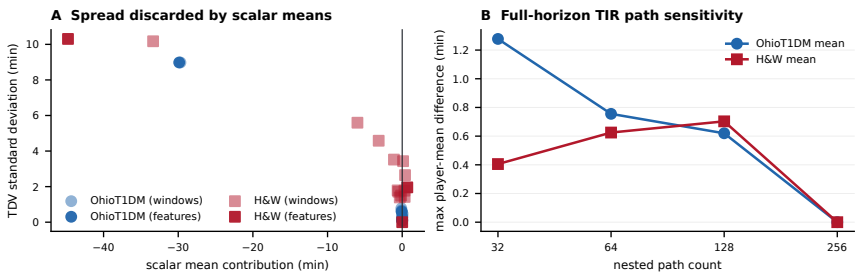


Figure 11: **Contributions to terminal in-range events.** Each row partitions probability into switching off (-1), leaving unchanged (0), or switching on ($+1$) the event that terminal glucose lies in $[70, 180]$ mg/dL. Terminal horizons are 60 minutes for OhioT1DM and 90 minutes for H&W. Temporal and variable-group panels are separate explanations.



cohort	wall s	CPU s	calls	paths	RSS MiB
OhioT1DM	737.3	737.6	8385	134144	2510
H&W	580.0	580.4	8257	132096	2625

cohort	max residual	replay	gate
OhioT1DM	0.0e+00	0.0e+00	pass
H&W	2.3e-13	0.0e+00	pass

Figure 12: **Contribution spread and trajectory-count sensitivity.** (A) Contribution standard deviation versus mean for the selected games. (B) Largest change in a player mean relative to the 256-trajectory result, using nested prefixes of the same bank. (C) Total resource use across coalition forecasts and checks; *paths* denotes all generated trajectories, not the per-coalition budget. (D) Reconstruction, replay, and scalar-identity checks.

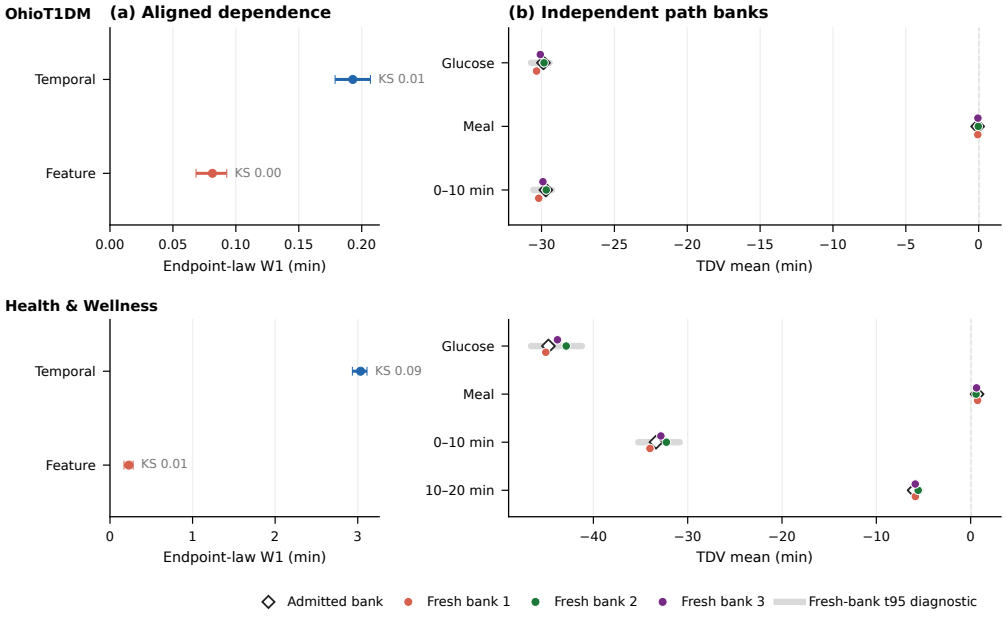


Figure 13: **Alignment and fresh-trajectory checks.** Left: Wasserstein-1 distance in minutes after shuffling alignment; markers and whiskers show medians and 5th–95th percentiles of the recorded shuffle results. KS denotes the Kolmogorov–Smirnov distance; a displayed 0.00 is rounded, not necessarily zero. Right: mean contributions from the original bank and three independent 256-trajectory banks. Pale intervals are t intervals over the three bank means, describing Monte Carlo variation conditional on the selected case and game. The *0–10* and *10–20 min* labels denote the newest and second-newest history windows before the forecast origin.

E REPRODUCIBILITY

Appendices A–C specify the mathematical construction, forecast representations, and computation. Appendix D describes the numerical comparisons, case selection, and reported results. Code is not released with this submission; sensitive source data and trained checkpoints are not redistributed.

Reproducing a numerical explanation requires the trained model, input preprocessing, case and player definitions, reference values, forecast representation, and shared-randomness and sampling settings. To check reconstruction from stored results, the baseline and contributions must retain their common order and shared-state identifiers, as described in Appendix C. Aggregate tables and rendered figures alone do not contain these records.