CovNovBench: A Benchmark for Evaluating Time Series Foundation Models under Covariate Novelty

Wenhao Pan[*] University of Washington Seattle, USA

Ravi Kiran Selvam[†] Amazon Science Bellevue, USA

Michael W. Mahoney Amazon Science Berkeley, USA

Mayank Jauhari Amazon Science Bellevue, USA

Mengfei Cao Amazon Science Bellevue, USA

Dmitry Efimov Amazon Science New York City, USA

Boris N. Oreshkin Amazon Science Montreal, Canada

Abstract

Covariate-aware time-series foundation models (TSFMs) condition their forecasts on additional input variables, called covariates, that are supplied over the future horizon. This capability makes them appealing for interventional forecasting, in which a target is predicted under a planned action such as a new product pricing policy or a new patient treatment schedule. We refer to covariates that encode such actions as treatment covariates. Current covariate-aware TSFMs, however, are trained on observed patterns, and their behavior is unclear when future treatment covariates fall outside the range seen in the history. Examples include a numerical treatment that leaves its historical range and a categorical treatment that takes a previously unseen policy. We call this practically common setting covariate novelty, and note that it is not systematically represented in existing benchmarks. We introduce an operational definition of covariate novelty and use it to construct **CovNovBench**, a benchmark of 88,476 evaluation series generated or filtered from four simulator-based and real-world datasets. The benchmark spans categorical and numerical treatment covariates as well as stable and trending targets. We evaluate six covariate-aware TSFMs zero-shot using point and directional accuracy metrics. No model is consistently reliable in our benchmark: point error almost always increases under covariate novelty, directional performance changes non-uniformly, and the best observed model varies across datasets and metrics. `CovNovBench` provides a systematic testbed for developing TSFMs that forecast more reliably under covariate novelty.

1 Introduction

Early time series foundation models (TSFMs) predict a horizon-H future target path $Y_{T+1:T+H}$ from its length-T history $Y_{1:T}$ [Ansari et al., 2024, Das et al., 2024]. Recent TSFMs also condition on *covariates*, time-indexed inputs supplied over the history, forecast horizon, or both [Ansari et al., 2025, Potapczynski et al., 2026]. When future covariates encode planned actions such as prices or treatments, these models provide a zero-shot interface for assessing actions before they are taken. If such a path is actively set rather than observed, predicting the target is referred to as *interventional forecasting* [Schultz et al., 2023, Wu et al., 2026a, Xia et al., 2025], a world-modeling task that models a system's responses to prescribed inputs rather than extrapolating an observed trajectory.

[*]Work done during internship at Amazon Science

[†]Correspondence to ravisel@amazon.com.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Foundation Models for Temporal Systems: From Forecasting to World Modeling.

We call covariates encoding actions or exposures expected to causally affect the target *treatment covariates*, and all others *non-treatment covariates*. The term "covariate" denotes a model input and does not imply causal exogeneity [Pearl, 2009]. Let $\mathbf{Z}_t \in \mathbb{R}^{D_Z}$ and $\mathbf{X}_t \in \mathbb{R}^{D_X}$ collect the treatment and non-treatment covariates, and let $\mathcal{H} = (Y_{1:T}, \mathbf{X}_{1:T}, \mathbf{Z}_{1:T})$ denote the observed history. An interventional forecaster predicts

$$\mathbb{P}\big\{Y_{T+1:T+H} \;\big|\; \mathcal{H},\, \mathrm{do}(\mathbf{Z}_{T+1:T+H}),\, \mathbf{X}_{T+1:T+H}\big\}\,. \tag{1}$$

Current covariate-aware TSFMs instead target the corresponding *observational* conditional, without the do-operator, treating the future treatment path as passively observed; their forecasts therefore need not represent the interventional conditional [Wu et al., 2026a]. Nevertheless, such forecasts can be evaluated against intervention-generated outcomes without being interpreted as interventional estimates [Crasson et al., 2024, Schultz et al., 2023, Namazi and Shakeri, 2026]. Causal forecasters address this mismatch but typically require task-specific training, unlike the zero-shot TSFMs evaluated here [Bica et al., 2020, Melnychuk et al., 2022, Wu et al., 2026a].

A planned path exhibits *covariate novelty* when at least one treatment covariate's future value set is not contained in its historical value set. We ask: *do covariate-aware TSFMs accurately forecast targets under interventions involving covariate novelty?* Section 2.1 defines the IN-DISTRIBUTION reference and the OUT-OF-SUPPORT, POLICY-SWITCH, and POLICY-ONSET novelty scenarios.

To our knowledge, no existing benchmark systematically evaluates whether future treatment paths contain values or policies absent from the same series' history. `CancerTreatment` is widely used for forecasting under planned treatment sequences [Bica et al., 2020, Melnychuk et al., 2022, Wu et al., 2026a], while `fev-bench` evaluates general covariate-aware forecasting [Shchur et al., 2025]. Neither benchmark labels evaluation series by covariate novelty. Appendix B compares related benchmarks, causal and covariate-aware models, intervention generalization, and distribution shift.

We therefore define covariate novelty by comparing historical and future treatment-covariate value sets for each evaluation series and use it to construct `CovNovBench`, a benchmark of $88{,}476$ series from four sources spanning synthetic, semi-synthetic, and real-world data. The sources emphasize strong treatment effects and controlled or limited hidden confounding to make responses detectable and less ambiguous, without establishing causal identification. We evaluate six covariate-aware TSFMs zero-shot using point and directional metrics and find that no model is consistently reliable.

2 Evaluation framework

2.1 Covariate novelty

Covariate novelty compares historical and future values of each treatment covariate within a single series. For integers $a \le b$, let $[a:b] = \{a, \dots, b\}$ and $[b] = [1:b]$. For each treatment covariate $j \in [D_Z]$ and any nonempty time-index set $\mathcal{S}$, define

$$\mathcal{V}_j(\mathcal{S}) = \begin{cases} [\min_{t\in\mathcal{S}} Z_{j,t}, \max_{t\in\mathcal{S}} Z_{j,t}]\,, & \text{if numerical,} \\ \{Z_{j,t} : t \in \mathcal{S},\, Z_{j,t} \neq 0\}, & \text{if categorical.} \end{cases} \tag{2}$$

For categorical covariates, 0 denotes inactivity and nonzero values denote active policies. Let $\mathcal{V}_j^{\mathrm{hist}} = \mathcal{V}_j([T])$ and $\mathcal{V}_j^{\mathrm{fut}} = \mathcal{V}_j([T+1:T+H])$ be the historical and future value sets. A future treatment path is IN-DISTRIBUTION if $\mathcal{V}_j^{\mathrm{fut}} \subseteq \mathcal{V}_j^{\mathrm{hist}}$ for every $j \in [D_Z]$, and novel otherwise. Novelty is assessed per covariate and series, independently of other series or pretraining data.

Numerical treatment covariates. A future numerical range extending below the historical minimum or above the historical maximum is OUT-OF-SUPPORT. This setting tests forecasting under numerical treatment extrapolation. For example, with historical treatment doses 1 and 3, a future dose of 2 stays within range and 4 is Out-of-Support.

Categorical treatment covariates. An unseen active policy gives POLICY-ONSET after an inactive history ($\mathcal{V}_j^{\mathrm{hist}} = \emptyset$ and $\mathcal{V}_j^{\mathrm{fut}} \neq \emptyset$) or POLICY-SWITCH after an active history ($\mathcal{V}_j^{\mathrm{hist}} \neq \emptyset$ and $\mathcal{V}_j^{\mathrm{fut}} \not\subseteq \mathcal{V}_j^{\mathrm{hist}}$). Both cases test forecasting under a supplied unseen policy. Switch histories include active treatments that may inform the forecast.

Appendix D grounds this taxonomy with concrete "what-if" queries and representative `CancerTreatment` trajectories.

2.2 Point and directional forecast accuracy

We report RMSSE [Makridakis et al., 2022] for scaled point error, with 1 marking the stabilized historical one-step error scale. DA measures sign agreement between predicted and realized target changes over the horizon, which matters when decisions depend on whether the target will increase or decrease [Namazi and Shakeri, 2026]. CDA restricts DA to steps with a categorical policy change or a reversal in numerical treatment direction. Lower RMSSE and higher DA and CDA are better, with 0.5 as the random-direction reference for DA and CDA. See Appendix C for exact definitions.

3 CovNovBench

Table 1 summarizes supported dataset–scenario pairs. Where possible, ID and novel scenarios share the same historical treatment settings to ensure fair comparisons, as detailed in Appendix D.

For categorical datasets, we operationalize the policy-level definitions using one binary indicator per treatment type. A future policy is novel when it activates a treatment type absent from the history. The registry labels this transition POLICY-ONSET if no treatment type was active in the history and POLICY-SWITCH if at least one was active.

Table 1: CovNovBench datasets, totaling $88{,}476$ evaluation series. tr./st. denote trending/stable targets, and c/n denote categorical/numerical treatment covariates. ID, OOS, Sw, and On denote the four scenarios; observational datasets provide weaker causal guarantees.

				Scenarios					
Dataset	Type	Target (dyn.)	Treatment cov. (type)	ID	OOS	Sw	On	T/H	#series
CancerTreatment	Synth.	Log tumor volume (tr.)	Radio/chemo dose (c+n)	✓	✓	✓	✓	50/10	12,548
CausalRivers	Real	Discharge (tr.)	Upstream discharge (n)	✓	✓	✗	✗	100/20	38,992
MIMIC3-Real	Real	Blood pressure (st.)	Vasopressor/vent. (c)	✓	✗	✓	✓	30/10	7,936
MIMIC3-Synth	Semi-synth.	Synth. outcome (st.)	Synth. policy (c)	✓	✗	✓	✓	50/10	29,000

CancerTreatment. We adapt the PK/PD tumor-growth simulator of Geng et al. [2017] to forecast log tumor volume. Its treatment policy is no treatment, radiotherapy, chemotherapy, or both, with numerical dose covariates for active treatments. By varying both policy and dose paths, the simulator supports all four benchmark scenarios.

CausalRivers. CausalRivers [Stein et al., 2025] links 666 German gauging stations in a topology-based causal graph. We forecast downstream discharge from lagged upstream discharges and static inputs such as distance between gauges. Small departures from historical upstream ranges remain in ID. Larger departures in either direction define moderate and extreme OUT-OF-SUPPORT (Appendix D.2).

MIMIC3-Real and MIMIC3-Synth. Both use MIMIC-III [Johnson et al., 2016] through MIMIC-Extract [Wang et al., 2020], following Melnychuk et al. [2022]. MIMIC3-Real forecasts diastolic blood pressure from the vasopressor/ventilation policy; its treatment associations are suggestive, not causally identified. MIMIC3-Synth pairs real vitals with simulated treatments and one simulated outcome; both datasets use categorical treatment policies, so OUT-OF-SUPPORT is inapplicable.

4 Empirical evaluation

We evaluate six covariate-aware TSFMs zero-shot (three tabular and three non-tabular, Appendices E and F). Full dataset–scenario metric tables (RMSSE, DA, and CDA) are reported in Appendix A). Figure 1(a) gives a benchmark-wide summary.

Are TSFMs accurate under covariate novelty? No evaluated model is consistently accurate. All weighted RMSSE averages exceed 1, though some dataset–scenario means fall below 1. Many DA and CDA means are near 0.5 (Figure 1(a)).

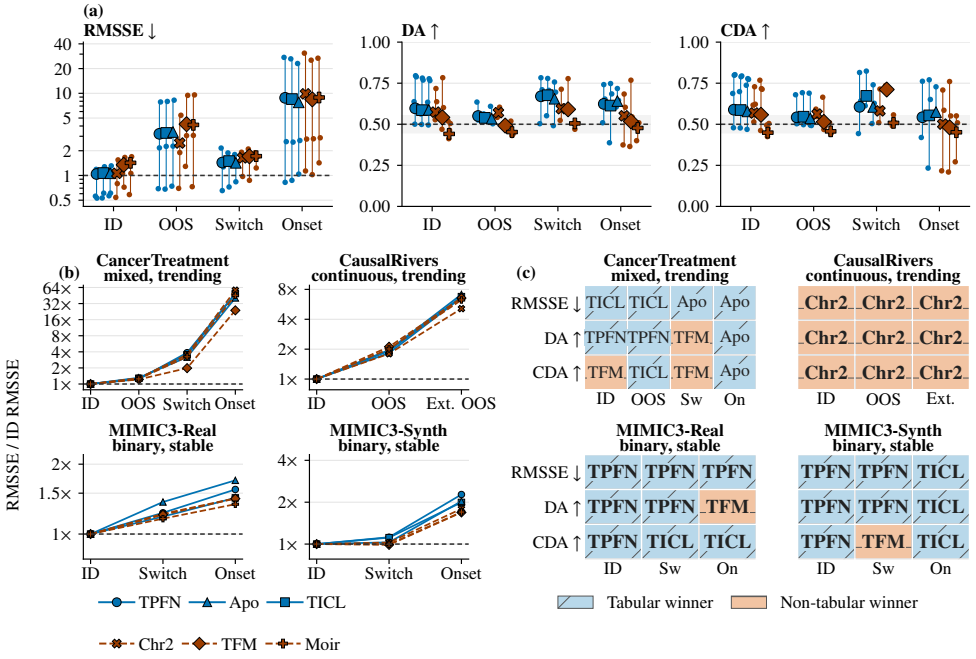


Figure 1: Main results. (a) Absolute metrics. Large markers show means weighted by series count. Small markers show dataset–scenario means, with stems spanning their range. For Onset, each model has three small markers, one each for `CancerTreatment`, `MIMIC3-Real`, and `MIMIC3-Synth`. CDA means use series with nonempty transition masks. (b) RMSSE relative to the same-model, same-dataset IN-DISTRIBUTION result. (c) Best observed model per dataset–scenario–metric cell; color and hatch indicate model family. Abbreviations: TPFN=`TabPFN-TS`, TICL=`TabICL-v2`, Apo=`ApolloPFN`, Chr2=`Chronos-2`, TFM=`TimesFM-2.5`, and Moir=`Moirai-1.1`.

RMSSE was higher than the IN-DISTRIBUTION value in 52 of 54 comparisons within the same model and dataset (Figure 1(b)). The two exceptions are `TimesFM-2.5` and `Moirai-1.1` on `MIMIC3-Synth` POLICY-SWITCH, at or slightly below their IN-DISTRIBUTION values. The RMSSE ratios relative to IN-DISTRIBUTION are smaller on `MIMIC3-Real`, reach roughly 24–57× on `CancerTreatment` POLICY-ONSET, and increase from moderate to extreme OUT-OF-SUPPORT for every `CausalRivers` model.

Do winners vary with dataset characteristics? Model winners vary by dataset, and neither family has the best observed mean across all dataset types (Figure 1(c)). `Chronos-2` has the best observed mean in all nine numerical-covariate, trending-target `CausalRivers` cells; tabular models do so in 16 of 18 binary-covariate, stable-target MIMIC cells, except for `TimesFM-2.5` on `MIMIC3-Real` POLICY-ONSET DA and `MIMIC3-Synth` POLICY-SWITCH CDA. On `CancerTreatment` (categorical and numerical treatments; trending target), tabular models have the lowest observed RMSSE mean in all four scenarios, while directional winners vary. These are descriptive associations because dataset characteristics vary jointly.

5 Conclusion

We introduced `CovNovBench`, a benchmark of 88,476 series drawn from both synthetic and real-world sources for evaluating covariate-aware TSFMs when future treatments contain values or policies absent from the history. Our zero-shot evaluation of six TSFMs shows that point error increases under novelty in nearly all matched comparisons, directional accuracy frequently approaches the random-direction reference, and no model or model family is consistently best across datasets, leaving substantial room for improvement in forecasting under novel interventions. Promising directions include retrieval-augmented few-shot forecasting over series where the novel treatment was observed, and conditioning on textual descriptions of each treatment and its expected effect.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. In *NeurIPS Workshop on Time Series in the Age of Large Models (TSALM)*, 2024. URL https://openreview.net/forum?id=Z2cMOOANFX.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In *Advances in Neural Information Processing Systems 38*, 2025. URL https://openreview.net/forum?id=RblaNJGx8C.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg866NFvB.

Thomas Crasson, Yacine Nabet, and Mathias Lécuyer. Training and evaluating causal forecasting models for time-series, 2024. URL https://arxiv.org/abs/2411.00126.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a biomathematical model. *Scientific reports*, 7(1):13542, 2017.

Rakshitha W. Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/eddea82ad2755b24c4e168c5fc2ebd40-Abstract-round2.html.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. IGC-Net for conditional average potential outcome estimation over time. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZmhpqpKzAT.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending TabPFN-v2 to time series forecasting, 2025. URL https://arxiv.org/abs/2501.02945.

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

Jesús Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. doi: 10.1016/j.apenergy.2021.116983.

Bryan Lim, Ahmed M. Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 7483–7493, 2018.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=d2L1ndOKjq.

Spyros Makridakis and Michele Hibon. The m3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, 2000.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4): 1325–1336, 2022.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 15293–15329, 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

Alireza Namazi and Heman Shakeri. From prediction to practice: A task-aware evaluation framework for blood glucose forecasting, 2026. URL https://arxiv.org/abs/2605.00645.

Hamed Nilforoshan, Michael Moor, Yusuf Roohani, Yining Chen, Anja Šurina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. Zero-shot causal learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 6862–6901. Curran Associates, Inc., 2023. doi: 10.52202/075280-0301. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/15ddb1773510075ef44981cdb204330b-Abstract-Conference.html.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Andres Potapczynski, Ravi Kiran Selvam, Tatiana Konstantinova, Malcolm Wolff, Kin G. Olivares, Ruijun Ma, Michael W. Mahoney, Andrew Gordon Wilson, Boris N. Oreshkin, and Dmitry Efimov. Time-aware prior fitted networks for zero-shot forecasting with exogenous variables. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=nJARpxp3cF.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://icml.cc/virtual/2026/poster/63874.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 174811–174848. Curran Associates, Inc., 2025. doi: 10.52202/085713-5814. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ff997469ac66cf893c4183efeb22212a-Paper-Conference.pdf.

Sorawit James Saengkyongam, Elan Rosenfeld, Pradeep K Ravikumar, Niklas Pfister, and Jonas Peters. Identifying representations for intervention extrapolation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3cuJwmPxXj.

Douglas Schultz, Johannes Stephan, Julian Sieber, Trudie Yeh, Manuel Kunz, Patrick Doupe, and Tim Januschowski. Causal forecasting for pricing, 2023. URL https://arxiv.org/abs/2312.15282.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting, 2025. URL https://arxiv.org/abs/2509.26468.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3): 653–677, 2025.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. CausalRivers – scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=wmV4cIbgl6.

Dennis Thumm and Ying Chen. Interventional time series priors for causal foundation models. In *ICLR 2026 Workshop on Time Series in the Age of Large Models*, 2026. URL https://openreview.net/forum?id=JbTgx2L9Z2.

Leena Chennuru Vankadara, Philipp Michael Faller, Michaela Hardt, Lenon Minorics, Debarghya Ghoshdastidar, and Dominik Janzing. Causal forecasting: generalization bounds for autoregressive models. In *Uncertainty in Artificial Intelligence*, pages 2002–2012. PMLR, 2022.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. MIMIC-Extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Dongze Wu, Feng Qiu, and Yao Xie. DoFlow: Flow-based generative models for interventional and counterfactual forecasting on time series. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=4IPIhOgVqz.

Xin Wu, Fei Teng, Xingwang Li, Ji Zhang, Qiang Duan, and Tianrui Li. Out-of-distribution generalization in time series: A survey. *Information Fusion*, page 104336, 2026b.

Yutong Xia, Chang Xu, Yuxuan Liang, Li Zhao, Qingsong Wen, Roger Zimmermann, and Jiang Bian. Causal time series generation via diffusion models, 2025. URL https://arxiv.org/abs/2509.20846.

Jialin Yu, Andreas Koukorinis, Nicolò Colombo, Yuchen Zhu, and Ricardo Silva. Structured learning of compositional sequential interventions. In *Advances in Neural Information Processing Systems*, volume 37, pages 115409–115439. Curran Associates, Inc., 2024. doi: 10.52202/079017-3665. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d10c7e24c96db4b222688efd11b02940-Abstract-Conference.html.

A Full numerical results

Tables 2, 3, and 4 list the scores underlying Figure 1. For `CausalRivers`, the unmarked out-of-support row is moderate and "Ext." denotes extreme. Appendix E.3 describes aggregation.

Table 2: Full RMSSE (↓) results across scenarios and datasets. Best per benchmark-cell row is in **bold**; *Average* rows are sample-count weighted over the displayed cells in each scenario block.

Scenario	Dataset	TabPFN-TS	TabICL-v2	TimesFM-2.5	Moirai-1.1	Chronos-2	ApolloPFN
IN-DISTRIBUTION	CancerTreatment	0.56	**0.53**	1.06	0.59	0.54	0.57
	CausalRivers	1.19	1.18	1.44	1.53	**1.06**	1.16
	MIMIC3-Real	**0.53**	0.61	0.72	1.06	0.80	0.61
	MIMIC3-Synth	**1.14**	1.28	1.66	1.69	1.55	1.31
	Average	**1.04**	1.07	1.35	1.42	1.06	1.06
OUT-OF-SUPPORT	CancerTreatment	0.69	**0.68**	1.29	0.73	0.70	0.74
	CausalRivers	2.17	2.26	3.05	3.07	**1.90**	2.27
	CausalRivers (Ext.)	7.85	7.97	9.44	9.56	**5.41**	8.25
	Average	3.21	3.29	4.21	4.13	**2.48**	3.38
POLICY-SWITCH	CancerTreatment	2.16	1.88	2.10	1.89	1.92	**1.80**
	MIMIC3-Real	**0.65**	0.72	0.87	1.23	0.97	0.84
	MIMIC3-Synth	**1.27**	1.42	1.64	1.69	1.59	1.35
	Average	1.44	1.50	1.71	1.72	1.64	**1.43**
POLICY-ONSET	CancerTreatment	27.39	26.28	25.31	26.76	30.83	**23.05**
	MIMIC3-Real	**0.82**	0.87	1.02	1.42	1.13	1.04
	MIMIC3-Synth	2.58	**2.55**	2.80	2.88	2.77	2.66
	Average	8.81	8.51	8.44	8.89	9.85	**7.77**

Table 3: Full Directional Accuracy (DA, ↑) results across scenarios and datasets. Best per benchmark-cell row is in **bold**; *Average* rows are sample-count weighted over the displayed cells in each scenario block.

Scenario	Dataset	TabPFN-TS	TabICL-v2	TimesFM-2.5	Moirai-1.1	Chronos-2	ApolloPFN
IN-DISTRIBUTION	CancerTreatment	**.638**	.563	.568	.520	.559	.634
	CausalRivers	.501	.499	.470	.412	**.528**	.496
	MIMIC3-Real	**.796**	.781	.783	.464	.718	.777
	MIMIC3-Synth	**.789**	.769	.601	.505	.637	.766
	Average	**.597**	.586	.542	.441	.574	.587
OUT-OF-SUPPORT	CancerTreatment	**.635**	.563	.508	.519	.542	.609
	CausalRivers	.526	.521	.478	.431	**.567**	.519
	CausalRivers (Ext.)	.529	.556	.510	.441	**.604**	.503
	Average	.550	.539	.492	.453	**.570**	.535
POLICY-SWITCH	CancerTreatment	.504	.552	**.581**	.505	.523	.491
	MIMIC3-Real	**.783**	.779	.778	.470	.713	.757
	MIMIC3-Synth	**.710**	.706	.587	.506	.609	.695
	Average	.670	**.677**	.591	.505	.595	.655
POLICY-ONSET	CancerTreatment	.509	.387	.365	.400	.374	**.605**
	MIMIC3-Real	.741	.748	**.768**	.470	.684	.718
	MIMIC3-Synth	.653	**.685**	.545	.509	.603	.645
	Average	.624	.614	.518	.478	.551	**.641**

Table 4: Full Conditional Directional Accuracy (CDA, ↑) results across scenarios and datasets. Best per benchmark-cell row is in **bold**; *Average* rows are sample-count weighted over the displayed cells in each scenario block.

Scenario	Dataset	TabPFN-TS	TabICL-v2	TimesFM-2.5	Moirai-1.1	Chronos-2	ApolloPFN
IN-DISTRIBUTION	CancerTreatment	.688	.699	**.768**	.495	.617	.737
	CausalRivers	.479	.478	.462	.425	**.514**	.469
	MIMIC3-Real	**.798**	.789	.717	.474	.728	.778
	MIMIC3-Synth	**.802**	.794	.713	.505	.622	.773
	Average	**.589**	.586	.558	.449	.569	.579
OUT-OF-SUPPORT	CancerTreatment	.680	**.692**	.665	.510	.592	.689
	CausalRivers	.503	.497	.468	.439	**.547**	.495
	CausalRivers (Ext.)	.508	.525	.485	.451	**.570**	.492
	Average	.542	.546	.515	.457	**.562**	.536
POLICY-SWITCH	CancerTreatment	.443	.612	**.736**	.490	.510	.504
	MIMIC3-Real	.815	**.824**	.734	.557	.714	.740
	MIMIC3-Synth	.642	.683	**.704**	.511	.596	.628
	Average	.606	.673	**.712**	.508	.581	.605
POLICY-ONSET	CancerTreatment	.420	.233	.209	.271	.217	**.571**
	MIMIC3-Real	.762	**.771**	.760	.551	.708	.728
	MIMIC3-Synth	.562	**.651**	.551	.509	.585	.555
	Average	.543	.554	.481	.451	.501	**.574**

B Additional related work

This appendix expands the compact positioning discussion in the main paper.

Forecasting benchmarks. M3 [Makridakis and Hibon, 2000] and M4 [Makridakis et al., 2020] mainly forecast targets from their own histories, while M5 [Makridakis et al., 2022] adds a large retail panel with price and calendar inputs; the Monash [Godahewa et al., 2021] and GIFT-Eval [Aksu et al., 2024] broaden domain, frequency, and scale coverage. General benchmarks do not define treatment covariates, while resources with known future inputs, including M5, electricity-price forecasting with day-ahead load and generation forecasts [Lago et al., 2021], and fev-bench's covariate-aware tasks [Shchur et al., 2025], treat those inputs as observational continuations rather than distinguish passively observed from actively set paths. None annotates whether a future treatment path is novel relative to its own series history, the axis targeted by CovNovBench.

Covariate-aware foundation models. Early TSFMs such as Chronos [Ansari et al., 2024] and TimesFM [Das et al., 2024] mainly studied univariate zero-shot forecasting; Chronos-2 [Ansari et al., 2025], TimesFM-2.5 [Das et al., 2024], and Moirai-1.1 [Woo et al., 2024] accept covariates, as do tabular models adapted by representing forecast steps as supervised examples: ApolloPFN [Potapczynski et al., 2026], TabPFN-TS [Hoo et al., 2025], and TabICL-v2 [Qu et al., 2026]. These interfaces make natural baselines, but their observational forecasts need not equal the interventional conditional under actively set treatment paths; we therefore evaluate them as frozen observational baselines, not causal forecasters.

Causal forecasting and causal foundation models. Models for specified treatment regimes include task-specific RMSN [Lim et al., 2018], CRN [Bica et al., 2020], CT [Melnychuk et al., 2022], and IGC-Net [Hess et al., 2026], as well as DoFlow [Wu et al., 2026a] and CaTSG [Xia et al., 2025] for interventional or counterfactual trajectory distributions; these approaches require task-specific training, causal assumptions, or structural information. Across tasks, Do-PFN [Robertson et al., 2025], CausalPFN [Balazadeh et al., 2025], and CausalFM [Ma et al., 2026] apply foundation-model or in-context ideas mainly to tabular effect estimation, while CausalTimePrior [Thumm and Chen, 2026] provides a temporal causal prior and a proof-of-concept PFN trained on synthetic temporal structural

causal models; none provides a general covariate-aware TSFM for zero-shot multi-step interventional forecasting. Theory distinguishes statistical from causal forecast risk [Vankadara et al., 2022], and empirical work studies changed pricing policies, actions outside the observational distribution, and paired factual and counterfactual insulin plans [Schultz et al., 2023, Crasson et al., 2024, Namazi and Shakeri, 2026]. These works motivate intervention evaluation but provide neither our future-path taxonomy nor a cross-dataset benchmark of frozen covariate-aware TSFMs; `CovNovBench` instead evaluates supplied novel paths against simulator-generated or observationally filtered outcomes, depending on the dataset.

Novel interventions. Zero-shot causal learning also studies unseen interventions: `CaML` [Nilforoshan et al., 2023] uses semantic and structured intervention representations for personalized effects and combinations, while compositional sequential-intervention methods exploit structure to handle sparse or unseen sequences [Yu et al., 2024]. These settings resemble POLICY-ONSET and POLICY-SWITCH, but they develop causal or structured intervention models, whereas `CovNovBench` evaluates whether existing frozen forecasters can use a supplied novel treatment path.

Distribution shift and extrapolation. Classical covariate shift assumes a fixed conditional under input-distribution change [Shimodaira, 2000]; time-series OOD work studies broader shifts across time, domains, and generating regimes [Wu et al., 2026b]; and regression or intervention extrapolation addresses prediction beyond observed support [Shen and Meinshausen, 2025, Saengkyongam et al., 2024]. These notions define shift relative to a population, dataset, domain, or action distribution, whereas our label compares a supplied horizon with the same series' history, separating treatment-path novelty from broader target non-stationarity or dataset shift.

C Metric details

RMSSE. For one evaluation series with target history $Y_{1:T}$ and horizon targets $Y_{T+1:T+H}$, $\hat{Y}_{T+h}$ is the median predicted quantile and $\varepsilon = 1 \times 10^{-2}$ stabilizes denominators in all three metrics. RMSSE [Makridakis et al., 2022] scales forecast mean squared error by historical one-step variation,

$$\mathrm{RMSSE} = \sqrt{\frac{\frac{1}{H}\sum_{h=1}^{H}\left(\hat{Y}_{T+h} - Y_{T+h}\right)^2}{\frac{1}{T-1}\sum_{t=2}^{T}\left(Y_t - Y_{t-1}\right)^2 + \varepsilon}}. \tag{3}$$

The denominator is the mean squared historical one-step error plus ε. Lower RMSSE is better, and values above 1 indicate that the forecast mean squared error exceeds this stabilized reference.

DA and CDA. DA measures directional agreement, which matters when an incorrect direction can misstate whether an outcome improves or worsens [Namazi and Shakeri, 2026]. Set $\hat{Y}_T = Y_T$, and define $r_h = (Y_{T+h} - Y_{T+h-1})/(|Y_{T+h-1}| + \varepsilon)$ and $\hat{r}_h$ analogously. Let $\mathcal{M} \subseteq [H]$ be the union of treatment-covariate transition steps—categorical policy changes or numerical direction reversals—counting a step once even if several covariates transition. DA averages sign agreement over the horizon, whereas CDA restricts it to $\mathcal{M}$:

$$\mathrm{DA} = \frac{1}{H}\sum_{h=1}^{H}\mathbf{1}[\mathrm{sign}(\hat{r}_h) = \mathrm{sign}(r_h)], \qquad \mathrm{CDA} = \frac{1}{|\mathcal{M}|}\sum_{h\in\mathcal{M}}\mathbf{1}[\mathrm{sign}(\hat{r}_h) = \mathrm{sign}(r_h)]. \tag{4}$$

Both lie in $[0, 1]$ and higher is better; the mask controls only where CDA is evaluated and is distinct from the novelty criterion.

Aggregation. All three metrics are computed per evaluation series. RMSSE and DA are averaged arithmetically over all series assigned to each dataset–scenario result. CDA is computed only when $\mathcal{M} \neq \emptyset$; series with no designated treatment transition are excluded from the dataset–scenario CDA mean.

D Benchmark interpretation, construction, and scenario-subset registries

Table 5 translates the taxonomy into concrete supply-chain and `CancerTreatment` *what-if* questions, and Figure 2 visualizes the `CancerTreatment` cases.

Table 5: Illustrative questions for the non-novel IN-DISTRIBUTION reference and three novelty scenarios, ordered nominally from easy to hard. BOGO denotes a buy-one-get-one offer; the `CancerTreatment` cases appear in Figure 2.

Scenario (easy ↓ hard)	Scenario pattern	Query: *What if we . . . ?*	
		Supply chain (demand)	`CancerTreatment` (tumor volume)
IN-DISTRIBUTION	Future values seen in history	Repeat a previously used discount?	Continue a prior radiotherapy dose?
OUT-OF-SUPPORT	Future range outside history	Offer 40% after only $<$ 20% discounts?	Use a chemo dose outside history?
POLICY-SWITCH	Active history $\rightarrow$ new policy	Replace a discount with BOGO?	Switch from chemo to radiotherapy?
POLICY-ONSET	Inactive $\rightarrow$ active policy	Launch the product's first discount?	Start radiotherapy for a new patient?

Within a dataset–scenario pair, an *evaluation subset* fixes the treatment covariates and their exact history-to-horizon pattern. Each dataset's scenario-subset registry is fixed before generation or observational filtering; registered subsets are pooled for dataset–scenario reporting, except that moderate and extreme `CausalRivers` OUT-OF-SUPPORT remain separate results. Categorical registries exclude inactive-to-inactive horizons and de-escalation-only transitions because neither introduces a novel active value.

History comparability. In `CancerTreatment`, OUT-OF-SUPPORT and POLICY-SWITCH subsets have IN-DISTRIBUTION counterparts with the same allowed historical treatment types and dose ranges (Table 6). For example, radiotherapy histories in IN-DISTRIBUTION and OUT-OF-SUPPORT both use doses in $[1, 2]$. The POLICY-SWITCH subsets in `MIMIC3-Real` and `MIMIC3-Synth` share historical active treatment sets with their IN-DISTRIBUTION counterparts (Tables 8 and 9). `CausalRivers` uses the same graph types and windowing procedure for IN-DISTRIBUTION and OUT-OF-SUPPORT (Table 7). POLICY-ONSET starts from untreated histories and has no corresponding IN-DISTRIBUTION subset. The reported scenario means pool separately generated or selected windows.

D.1 `CancerTreatment`

Construction and assumptions. This synthetic dataset uses the PK/PD tumor-growth model of Geng et al. [2017], adapted by Bica et al. [2020], Melnychuk et al. [2022], to forecast log tumor volume from binary radiotherapy/chemotherapy actions, continuous doses, and a static patient type. Radiotherapy acts immediately, chemotherapy decays over time, and treatment assignment depends on recent tumor diameter, creating controlled time-varying confounding; the specified-policy outcomes are simulator semantics rather than clinical claims, and very strong treatment can approach the simulator's boundary near zero tumor volume. For each registry row, split-window controls constrain history and horizon treatment types and doses, disallowed assignment probabilities are zeroed, and 1,000 patient-specific trajectories of $L = 60$ daily steps are sampled. After removing incomplete or infeasible trajectories and shifting treatments by one step, the benchmark uses $T = 50$, $H = 10$, and retains 12,548 series: 2,969 IN-DISTRIBUTION, 2,868 OUT-OF-

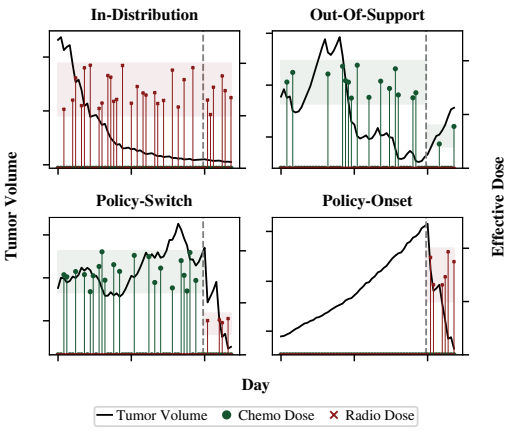


Figure 2: Representative `CancerTreatment` trajectories for Table 5. Black curves show tumor volume, green and red stems show chemotherapy and radiotherapy doses, dashed lines mark the forecast boundary, and shading marks registered active-dose ranges.

SUPPORT, 2,749 POLICY-ONSET, and 3,962 POLICY-SWITCH.

Registry. Table 6 lists all 13 subsets with their history and horizon policies. In-range radiotherapy and chemotherapy doses are $[1, 2]$ and $[3, 5]$, while their OUT-OF-SUPPORT ranges are $[0, 1] \cup [2, 3]$ and $[0, 3] \cup [5, 6]$, respectively.

Table 6: `CancerTreatment` evaluation-subset registry. Radio and chemo denote radiotherapy and chemotherapy.

Scenario	Subset pattern	History policy	Horizon policy
IN-DISTRIBUTION	radio	radio, dose $[1, 2]$	radio, dose $[1, 2]$
IN-DISTRIBUTION	chemo	chemo, dose $[3, 5]$	chemo, dose $[3, 5]$
IN-DISTRIBUTION	both	radio + chemo, in-range doses	radio + chemo, in-range doses
OUT-OF-SUPPORT	radio dose	radio, dose $[1, 2]$	radio, dose $[0, 1] \cup [2, 3]$
OUT-OF-SUPPORT	chemo dose	chemo, dose $[3, 5]$	chemo, dose $[0, 3] \cup [5, 6]$
OUT-OF-SUPPORT	both doses	radio + chemo, in-range doses	radio + chemo, out-of-range doses
POLICY-ONSET	to radio	no treatment	radio, dose $[1, 2]$
POLICY-ONSET	to chemo	no treatment	chemo, dose $[3, 5]$
POLICY-ONSET	to both	no treatment	radio + chemo, in-range doses
POLICY-SWITCH	radio to chemo	radio, dose $[1, 2]$	chemo, dose $[3, 5]$
POLICY-SWITCH	chemo to radio	chemo, dose $[3, 5]$	radio, dose $[1, 2]$
POLICY-SWITCH	radio to both	radio, dose $[1, 2]$	radio + chemo, in-range doses
POLICY-SWITCH	chemo to both	chemo, dose $[3, 5]$	radio + chemo, in-range doses

D.2 `CausalRivers`

Construction and assumptions. We use RiversEastGermany from `CausalRivers` [Stein et al., 2025], containing 15-minute discharge from 666 gauges and 651 directed edges during 2019–2023. For each pair, chain, or confluence subgraph, the downstream discharge is the target, lagged upstream discharges are treatment covariates, and distance, elevation difference, and same-river status are static inputs. The topology gives a physically motivated direction but not an identified intervention effect: unobserved rainfall, varying transport lags, and omitted tributaries can confound these observational associations. We retain gauges with more than 95% valid positive observations, aggregate to 6-hour means, restrict chain/confluence targets to modeled direct parents, and form windows with $T = 100$, $H = 20$, and stride H. After removing invalid or near-constant histories, capping windows per subgraph, and shifting upstream covariates by one step, 38,992 series remain: 28,532 IN-DISTRIBUTION, 7,258 moderate OUT-OF-SUPPORT, and 3,202 extreme OUT-OF-SUPPORT.

Registry. For upstream covariates $\mathbf{Z}$, the deviation score is the maximum normalized departure from each covariate's own history range:

$$d(\mathbf{Z}) = \max_j \frac{\max\{0, Z^{\text{fut}}_{j,\max} - Z^{\text{hist}}_{j,\max}, Z^{\text{hist}}_{j,\min} - Z^{\text{fut}}_{j,\min}\}}{Z^{\text{hist}}_{j,\max} - Z^{\text{hist}}_{j,\min}}. \tag{5}$$

Every structure in Table 7 is crossed with $d < 0.1$ (IN-DISTRIBUTION), $0.1 \leq d < 1$ (moderate OUT-OF-SUPPORT), and $d \geq 1$ (extreme OUT-OF-SUPPORT); thus small positive departures remain in the reference group. Continuous discharge has no natural inactive state, so POLICY-ONSET and categorical POLICY-SWITCH are undefined.

D.3 `MIMIC3-Real`

Construction and assumptions. We use the hourly MIMIC-Extract [Wang et al., 2020] representation of deidentified MIMIC-III [Johnson et al., 2016], following Melnychuk et al. [2022]. The target is diastolic blood pressure; vasopressor administration and mechanical ventilation are binary treatment covariates, with 25 hourly clinical variables and static gender, ethnicity, and age as other inputs after the source imputation and normalization. Because treatments respond to patient

Table 7: `CausalRivers` registry: three structures crossed with ID, moderate-OOS, and extreme-OOS bands (nine subsets).

Structure	Treatment covariates	Role
Pair ($A \to B$)	one upstream discharge	direct upstream influence
Chain ($A \to B \to C$)	grandparent and parent discharges	mediated upstream influence
Confluence ($A \to C \leftarrow B$)	two upstream tributary discharges	simultaneous upstream influence

state and counterfactual outcomes are unavailable, confounding by indication prevents interpreting these observational associations as intervention effects. Eligible ICU stays have at least 40 hourly observations and are capped at 250 hours. POLICY-ONSET and POLICY-SWITCH boundaries use the first activation of a horizon-novel treatment type, with the preceding history checked against the registry; IN-DISTRIBUTION uses a persistent active policy, and one stay may supply distinct qualifying transitions. After a one-step treatment shift, the $T = 30$, $H = 10$ windows total $7{,}936$: $6{,}487$ IN-DISTRIBUTION, 934 POLICY-ONSET, and 515 POLICY-SWITCH.

Registry. The two indicators define none, vasopressor-only, ventilation-only, and joint policies. A vasopressor-to-ventilation transition is POLICY-ONSET for the ventilation indicator and is registered as POLICY-SWITCH for the overall policy because vasopressors were active in the history. Table 8 preserves all ten registered transitions; all-off horizons and de-escalation-only changes are excluded.

Table 8: `MIMIC3-Real` scenario-subset registry. Vaso and Vent denote vasopressor and ventilation.

Scenario	# subsets	Registered history → horizon patterns
IN-DISTRIBUTION	3	Vaso→Vaso, Vent→Vent, Both→Both
POLICY-ONSET	3	None→Vaso, None→Vent, None→Both
POLICY-SWITCH	4	Vaso→Vent, Vent→Vaso, Vaso→Both, Vent→Both

D.4 `MIMIC3-Synth`

Construction and assumptions. Following Melnychuk et al. [2022], this semi-synthetic dataset combines 25 normalized hourly MIMIC-Extract variables and static gender, ethnicity, and age with three simulated binary treatments and one simulated outcome. The untreated outcome includes smooth temporal structure, covariate dependence, patient heterogeneity, and noise; treatment propensities depend on recent outcomes and covariates, creating controlled time-varying confounding, while effects are additive over finite windows. The mechanism supports specified-policy outcomes, conditional on the simulator, and reconstruction requires authorized MIMIC-III access. Eligible stays have at least 60 hours and are capped at 100; for each registry row, a history–horizon allow mask constrains assignment and a separately seeded generator samples $1{,}000$ trajectories of $L = 60$. After a one-step treatment shift, the $T = 50$, $H = 10$ windows total exactly $29{,}000$: $7{,}000$ IN-DISTRIBUTION, $7{,}000$ POLICY-ONSET, and $15{,}000$ POLICY-SWITCH.

Registry. For active sets formed from t_1, t_2, t_3, Table 9 compactly specifies all 29 subsets.

Table 9: `MIMIC3-Synth` registry (0: inactive; distinct i, j, k).

Scenario	# subsets	Registered history → horizon patterns
IN-DISTRIBUTION	7	$S \to S$ for every nonempty $S \subseteq \{t_1, t_2, t_3\}$
POLICY-ONSET	7	$0 \to S$ for every nonempty $S \subseteq \{t_1, t_2, t_3\}$
POLICY-SWITCH	15	six pure replacements $t_i \to t_j$; three single-to-complementary-pair replacements $t_i \to t_j t_k$; three pair-to-complement replacements $t_i t_j \to t_k$; three pair-to-triple escalations $t_i t_j \to t_1 t_2 t_3$

E Experimental setup

E.1 Models and protocol

All six frozen checkpoints are evaluated once in zero-shot mode, independently on each of the 88,476 evaluation series, using its target history, observed covariates, and supplied future covariates; the median probabilistic forecast is the point forecast. No reference series, benchmark-specific tuning or fine-tuning, supervised adaptation, or test-outcome feedback is used. Benchmark construction and stochastic inference use fixed seeds; inference uses seed 1 except for `Chronos-2` and `Moirai-1.1`, which use seed 42, with deterministic backend settings enabled where supported. The results therefore characterize one fixed benchmark realization rather than repeated-run or alternative-draw uncertainty. Table 10 consolidates the model-specific settings.

Table 10: Checkpoints, interfaces, and configurations used in zero-shot evaluation.

Model	Checkpoint / interface	Configuration / preprocessing
`ApolloPFN`	Closed, owner-authorized checkpoint; forecast-step tabular examples; MQLoss head outputs nine quantiles.	12-layer PFNT, 192 hidden dimensions, 6 heads, MLP factor 4, no bias or dropout; feature grouping 2, input size 512, `norm_y=False`. The checkpoint is unavailable publicly.
`TabPFN-TS`	`tabpfn` with `tabpfn-v2-regressor.ckpt` (Prior-Labs, HuggingFace); forecast-step tabular examples; nine quantiles from the inverse CDF of a BarDistribution.	12-layer PFNT, 192 hidden dimensions, 6 heads, no dropout; 5,000 bins from `borders.npy`, feature grouping 2, per-series z-normalization (`norm_y=True`), per-batch uniform QuantileTransformer, StandardScaler, and rank-1 SVD augmentation. NaNs receive indicators; zero-valid-context series receive zero predictions.
`TabICL-v2`	`tabicl`; TabICL-v2 checkpoint auto-downloaded from HuggingFace; `TabICLRegressor.fit()` uses historical covariates and targets, then `.predict()` uses future covariates.	Eight ensemble estimators; most recent 1,024 history steps retained when needed; no per-series optimization beyond in-context learning.
`Chronos-2`	`chronos-forecasting` with `amazon/chronos-2`; named past/future-covariate dictionaries passed to the native covariate-aware API.	Context determined from target history up to 2,048 tokens; quantiles returned directly.
`TimesFM-2.5`	`timesfm` with the public TimesFM-2.5 200M checkpoint through `TimesFM_2p5_200M_torch`; `dynamic_numerical_covariates` passed to `forecast_with_covariates()`.	Context 1,024, horizon 64; `normalize_inputs`, `use_continuous_quantile_head`, `force_flip_invariance`, `infer_is_positive`, and `fix_quantile_crossing` set true. Padding is stripped before each call.
`Moirai-1.1`	`uni2ts` with `Salesforce/moirai-1.1-R-large`, loaded by `MoiraiModule.from_pretrained` and `MoiraiForecast`; `feat_dynamic_real` spans context and horizon.	Context 512, patch size 16, and 100 forecast samples converted to quantiles. NaN covariates are zeroed with an observed-mask indicator; seed 42 makes sampling repeatable.

Implementation. All wrappers share a PyTorch Lightning harness driven by a dataset-registry YAML with one combined-Parquet entry per scenario; the runner generates scenario configurations and writes per-sample and aggregate metric CSVs. LLM-based assistance was limited to code implementation of the author-defined scenario-subset registries and filtering pipeline.

E.2 Compute resources

All 78 model–dataset–scenario evaluations ran on one AWS `p4d.24xlarge` instance with 96 vCPUs, 1,152 GiB memory, eight 40-GiB NVIDIA A100 GPUs, and eight 1-TB NVMe SSDs. Table 11 reports the recorded elapsed forward-pass time for each evaluation. The underlying unrounded durations total 13,604.2 seconds (3.78 hours); model loading, preprocessing, and broader project compute are excluded.

Table 11: Forward-pass inference times (s); totals sum unrounded entries and are rounded to 0.1 s.

Scenario	Dataset	TabPFN-TS	TabICL-v2	TimesFM-2.5	Moirai-1.1	Chronos-2	Apollo PFN
IN-DISTRIBUTION	CancerTreatment	23.5	805.6	98.7	30.3	20.1	17.3
	CausalRivers	370.7	1986.2	521.6	49.9	31.4	88.1
	MIMIC3-Real	31.0	584.6	192.9	74.7	23.1	18.8
	MIMIC3-Synth	26.2	557.9	232.8	76.6	28.4	27.7
OUT-OF-SUPPORT	CancerTreatment	22.9	691.2	94.3	29.6	19.5	17.4
	CausalRivers	92.9	506.9	149.4	36.2	23.7	36.5
	CausalRivers (Ext.)	48.7	522.0	81.9	29.1	27.0	34.3
POLICY-SWITCH	CancerTreatment	22.3	627.4	94.3	29.9	21.3	18.4
	MIMIC3-Real	12.0	276.2	26.0	36.3	14.8	11.4
	MIMIC3-Synth	35.7	1051.2	380.1	118.9	30.5	29.0
POLICY-ONSET	CancerTreatment	19.8	671.7	92.0	30.1	18.7	18.1
	MIMIC3-Real	14.7	511.1	35.2	54.2	16.0	11.9
	MIMIC3-Synth	27.6	538.9	197.5	75.8	27.2	26.3
	Total	748.3	9330.8	2196.5	671.7	301.7	355.3

E.3 Result aggregation

Each dataset–scenario RMSSE and DA value is the arithmetic mean over all of its per-series values; CDA is averaged over the series with nonempty transition masks. The benchmark-wide averages in Figure 1 weight each dataset–scenario result by its evaluation-series count; moderate and extreme CausalRivers OUT-OF-SUPPORT results contribute separately. All benchmark-wide averages are descriptive; matched same-model, same-dataset comparisons against IN-DISTRIBUTION are the primary evidence.

F Access and reproducibility boundaries

Data and licenses. The MIT-licensed Causal Transformer code supports both generated datasets, and MIMIC-Extract is MIT-licensed; CausalRivers data use the German Data Licence–Attribution–Version 2.0. Under the PhysioNet Credentialed Health Data License 1.5.0, MIMIC-III v1.4 requires credentialing, human-subjects training, and a data-use agreement, so MIMIC3-Real and MIMIC3-Synth's MIMIC-based covariates cannot be regenerated without authorized access. These terms do not establish redistribution rights for every derived benchmark artifact.

Models and artifacts. Apache-2.0 covers TabPFN-TS and Uni2TS code, Chronos-2 and TimesFM-2.5 code and checkpoints, and TabICL-v2's TabPFN-TS component; the TabPFN-v2 checkpoint uses Prior Labs License 1.1 (Apache-2.0 plus attribution), the rest of TabICL-v2 uses BSD-3-Clause, and Moirai-1.1-R-Large weights use CC BY-NC 4.0. The five public baselines are available through the named repositories or model hubs. The owner-authorized ApolloPFN checkpoint remains closed and unavailable, so its predictions cannot be publicly regenerated; no code, data, model, or supplementary artifact accompanies this submission.

Participation and evaluation boundaries. Secondary, deidentified MIMIC-III data involve no participant recruitment, interaction, or intervention. Pre-evaluation registries and frozen models use no benchmark-specific training or adaptation, reference retrieval, or test-outcome feedback. Without a formal pretraining-contamination or deduplication audit, incompletely documented corpora may include related sources; novelty applies only to each evaluation history, not pretraining.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sections 1–5 state the benchmark scope, evaluation setting, findings, and limitations consistently with the reported results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 notes that dataset characteristics vary jointly, and Appendix D explains that scenario means pool separately generated or selected windows. Section 5 identifies shared-history evaluation as future work. Appendix F documents access restrictions and possible pretraining overlap.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents no theoretical results or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendices C–F document the metrics, dataset and scenario-subset construction, model settings, aggregation, and access limits. MIMIC-based reconstruction requires credentialed access, and the ApolloPFN checkpoint is not publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No code, data, model, or supplementary artifact is released with this submission due to the company policy. MIMIC-based data require credentialed access, and the evaluated ApolloPFN checkpoint is closed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 2–4 and Appendices C–E specify scenario construction, metrics, dataset generation or filtering, seeds, model settings, and aggregation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The reported scores summarize one fixed benchmark realization (Appendix E). Sampling uncertainty and repeated-run variability remain unestimated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.2 reports the AWS hardware and forward-pass inference time for every model–dataset–scenario evaluation, totaling 13,604.2 seconds (3.78 hours), with stated exclusions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors reviewed the NeurIPS Code of Ethics and confirm compliance. Data-access and reproducibility boundaries are documented in Appendix F.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This preliminary work introduces an evaluation benchmark and does not deploy a decision-making system or release a model; we identify no direct societal impact requiring a dedicated discussion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The submission releases no data or model requiring misuse safeguards; MIMIC-III is not redistributed and remains subject to PhysioNet access controls.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendices D–F credit the data, software, and model assets and state their licenses and access terms. The closed ApolloPFN checkpoint was used with authorization and is not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new code, dataset, model, or supplementary artifact is released with this submission.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study uses secondary analysis of deidentified MIMIC-III data and involves no crowdsourcing, participant recruitment, interaction, or intervention.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The study analyzes previously collected, deidentified MIMIC-III data under its access requirements and involves no participant recruitment, interaction, or intervention.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM assistance was limited to code implementation of the author-defined scenario-subset registries and filtering pipeline, as disclosed in Appendix E.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.