MedWorld-Lite: A Missingness-Aware Personalized Framework for ICU Medical World Modeling

Nafiz Fahad[1,2][0000-0002-3207-6515]*

[1] Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia
[2] Elite Research Lab LLC, New York, NY, United States
nafiz.fahad@student.mmu.edu.my

Abstract. Medical world models should move beyond static risk scores by representing a patient's evolving state, modeling how that state changes, and exposing uncertainty when observations are incomplete. We present MedWorld-Lite, a compact framework for medical world modeling from irregular intensive-care data. It combines eight physiological variables with observation masks and time-since-last-measurement channels, personalizes the latent patient state using admission context, and specifies a recurrent transition with probabilistic future-state decoding. The temporal benchmark uses 24 h of history and targets 6/12/24 h future rollouts, but the recurrent transition and decoder are not empirically validated in this submission because raw sequence bytes were unavailable in the execution environment. We therefore report only an executed 4,000-patient processed Set-A state-representation probe. Across five stratified splits, adding personalization increased AUROC from 0.753 ± 0.022 to 0.794 ± 0.016; the complete representation reached 0.795 ± 0.018. On a held-out split, a random forest achieved 0.857 AUROC (95% bootstrap CI 0.823–0.889) and 0.469 AUPRC. Hiding 50% of available physiological state features reduced logistic-regression AUROC from 0.820 to 0.638. These results support missingness-aware personalized state construction only; they do not establish temporal forecasting performance, prospective mortality utility, or treatment effects.

Keywords: Medical world model · ICU · physiological forecasting · missing data · personalization · uncertainty

1 Introduction

Medical artificial intelligence has historically emphasized recognition: a model maps a snapshot or record to a diagnosis, score, or endpoint. A medical world model asks a broader question: what latent state best summarizes the patient now, how might that state evolve, and how uncertain are the predicted futures?

* Corresponding author.

Recent medical world-model work has begun to formalize this shift through treatment-conditioned tumor simulation, longitudinal EHR trajectory modeling, and medical time-series foundation models [13–15]. Contemporary reviews describe patient-state representation, temporal dynamics, intervention-conditioned simulation, and planning as a progression of capabilities rather than a single architecture [18, 19]. This makes irregular ICU time series a natural test bed: they are dynamic, partially observed, patient-specific, and clinically consequential.

World modeling in the ICU is difficult for three reasons. First, measurements are not synchronized: heart rate can be frequent while laboratory variables may appear only after clinical requests. Missingness can therefore encode care processes as well as physiology. Second, the same measured pattern may have different implications for different patients, motivating personalized state representations. Third, useful rollouts require calibrated uncertainty because long-horizon predictions can accumulate error. These challenges connect medical world modeling to earlier work on GRU-D, latent ODEs, irregular sensor graphs, and probabilistic multi-horizon forecasting [4–7].

We propose MedWorld-Lite, intentionally small enough for a single-GPU Colab workflow. The design follows the defining world-model pattern used in latent dynamics research: encode observations into a compact state, learn a transition in latent space, and decode predicted observations [2, 3]. The target temporal task uses the first 24 h of an ICU stay to roll out 6, 12, and 24 h future physiology. The framework does not claim causal treatment simulation: no intervention semantics are inferred from observational correlations, and predicted futures are not counterfactual treatment effects.

This paper makes four contributions. (1) It defines a missingness-aware, personalized ICU world-state representation that combines values, masks, time gaps, and static context. (2) It specifies, but does not empirically validate here, a lightweight recurrent latent transition with a heteroscedastic physiological decoder and auxiliary clinical-utility head. (3) It provides an evaluation protocol for multi-horizon forecasting, uncertainty, calibration, ablations, missingness stress, and patient-level rollouts. (4) It reports a fully executed 4,000-patient Set-A state-representation probe, including bootstrap uncertainty and subgroup diagnostics. All numerical results reported in this paper come from that probe, not from temporal rollouts.

2 Related Work and Medical World-Model Scope

World models and clinical trajectory simulation. World models learn internal dynamics that permit imagined future states rather than only reactive prediction [2, 3]. Medical applications have recently become explicit. MeWM simulates future tumor states conditioned on treatment plans [13]; EHRWorld targets long-horizon patient trajectories under sequential interventions [14]; and recent medical world-model surveys emphasize state construction, dynamics, action semantics, uncertainty, and staged clinical validation [17–19]. MedWorld-Lite addresses an earlier point on this capability ladder: physiological state construction

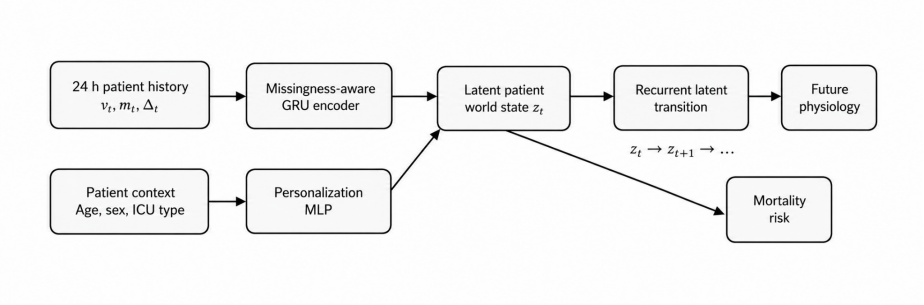

Fig. 1. MedWorld-Lite. The first 24 h of irregular observations are represented by values, observation masks, time gaps, and static patient context. The personalized world state is recurrently transitioned and probabilistically decoded into future physiology; a separate clinical head tests whether the learned state retains outcome-relevant information.

and predictive latent dynamics. It intentionally does not claim intervention-conditioned causal simulation.

Irregular medical time series. The PhysioNet/CinC Challenge 2012 established an ICU mortality benchmark from the first 48 h of patient data [1]. GRU-D demonstrated that masks and time intervals are informative for clinical sequences [4]. Latent ODEs model hidden continuous-time dynamics [5], while RAINDROP models relationships among irregular sensors [6]. MIRA further highlights continuous-time modeling, heterogeneous sampling, and missingness as core medical forecasting problems [15]. Standardized clinical benchmarks have also shown why patient-level splits and reproducible preprocessing are essential [8].

Personalization, uncertainty, and digital twins. Digital-twin research motivates patient-specific predictive states while also emphasizing the gap between statistical forecasts and mechanistic or causal simulation [12]. Heteroscedastic likelihoods provide observation-dependent uncertainty [9]; calibration metrics are needed for clinical probability outputs [10]; and distribution-free uncertainty methods motivate validation of interval coverage rather than reporting point error alone [11]. These ideas shape our evidence requirements even though the present model is not described as a validated clinical digital twin.

3 Problem Formulation

For patient i, let $x_t \in \mathbb{R}^D$ denote physiological variables in hour t, $m_t \in \{0,1\}^D$ indicate which variables were observed, and $\Delta_t \in \mathbb{R}^D$ denote time since the most recent observation of each variable. Static context p contains age, sex, ICU type, weight, and height. We observe $T_h = 24$ h and define future horizons $H \in \{6, 12, 24\}$.

A medical world model learns a latent patient state z_t satisfying three properties: it summarizes the observed history, a transition function evolves it forward, and a decoder maps future latent states to observable physiology. We distinguish three levels of evidence: state validity (does the representation retain clinically relevant information?), dynamics validity (do rollouts predict future observed physiology?), and intervention validity (do action-conditioned rollouts support causal or counterfactual claims?). This paper executes the first level and specifies a reproducible benchmark for the second; it makes no third-level claim.

4 MedWorld-Lite

4.1 Missingness-Aware Patient Encoder

Selected variables are HR, MAP, respiratory rate, temperature, SaO_2, glucose, GCS, and urine output. Continuous observations are robustly clipped using training-set quantiles and standardized using training statistics only. Missing values are filled with zero after the observation mask has been constructed. The encoder input is

$$u_t = [\tilde{x}_t; m_t; \Delta_t], \qquad h_t = \mathrm{GRU}_E(u_t, h_{t-1}). \tag{1}$$

Time gaps increase when a variable is unobserved and reset at observation, preserving information lost by ordinary mean imputation. The design does not assume synchronous or continuously available measurements: at inference time, only values already timestamped in the record are consumed; unavailable channels remain masked and their elapsed-time counters continue to increase.

4.2 Personalized World State

Static context is embedded by $e_p = g_\phi(p)$. The patient world state after the observed history is

$$z_t = \tanh\left(W_z[h_t; e_p] + b_z\right). \tag{2}$$

This construction allows identical observed vitals to map to different latent states when admission context differs.

4.3 Latent Transition and Probabilistic Decoder

Starting from z_t, the model recursively rolls the state forward:

$$\hat{z}_{t+k} = T_\theta(\hat{z}_{t+k-1}, e_k), \qquad k = 1, \ldots, H, \tag{3}$$

where e_k is a learned horizon embedding and T_θ is a GRU cell. A decoder predicts a mean and variance for each physiological variable,

$$[\mu_{t+k}, \log \sigma^2_{t+k}] = D_\psi(\hat{z}_{t+k}), \quad p(x_{t+k}|\hat{z}_{t+k}) = \mathcal{N}(\mu_{t+k}, \mathrm{diag}(\sigma^2_{t+k})). \tag{4}$$

Only observed future targets contribute to the forecasting loss, so sparse future measurements do not become artificial zeros.

4.4 Clinical Utility Head and Objective

A mortality head estimates $\hat{y} = \sigma(C(z_t))$. During raw-sequence training, representations obtained after future observations can be used as detached consistency targets z^*_{t+k}. The proposed objective is

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda_m \mathcal{L}_{\text{BCE}} + \lambda_z H^{-1} \sum_{k=1}^{H} \|\hat{z}_{t+k} - z^*_{t+k}\|_2^2, \tag{5}$$

with starting values $\lambda_m = 0.5$ and $\lambda_z = 0.1$. These are tunable hyperparameters, not claimed universal constants.

5 Data and Preprocessing

The PhysioNet/CinC Challenge 2012 includes 12,000 adult ICU stays, while Set A provides 4,000 labeled training records with outcomes [1]. The challenge records up to 42 descriptors/measurements during the first 48 h with substantial irregularity. For the temporal benchmark, measurements are binned by elapsed ICU hour. Multiple continuous measurements within one hour use the median; urine is summed within the hour. Invasive MAP is preferred, with NIMAP used when invasive MAP is absent. Static personalization uses age, gender, ICU type, weight, and height.

Patients are split at the patient level into 70% train, 15% validation, and 15% test partitions with mortality stratification and seed 42. Clipping limits, scaling statistics, and imputation medians are computed from training patients only. The first 24 h form model history and hours 24–47 form future targets. No outcome information is used in preprocessing.

The executed state-validation experiment uses a public processed Set-A table with 4,000 rows and 55 columns. The table contains admission descriptors, aggregate physiological values, selected variability statistics, length of stay, and in-hospital death. We exclude length of stay and all outcome-derived fields from predictors. The executed representation uses 19 physiological/static/variability features plus explicit missingness indicators.

6 Experimental Framework

6.1 Executed State-Representation Probe

The mortality classification experiment is used only as a representation probe: it tests whether the constructed state features retain outcome-relevant information, not whether MedWorld-Lite is prospectively useful for mortality prediction. For interpretable component ablation, class-balanced logistic regression is evaluated over five stratified 80/20 splits (seeds 11, 22, 33, 44, 55). Feature groups are added cumulatively: physiological state variables, static personalization, variability features, and explicit missingness indicators. On a fixed seed-42 split,

Table 1. Executed state-representation probe: feature ablation over five stratified splits (mean ± standard deviation).

State features	AUROC	AUPRC	Brier	F1	Bal. Acc.
Vitals only	.753±.022	.340±.040	.202±.003	.384±.023	.688±.024
+ Personalization	.794±.016	.378±.048	.188±.006	.405±.011	.713±.013
+ Variability	.790±.017	.389±.042	.186±.004	.404±.014	.709±.017
+ Missingness mask	.795±.018	.388±.041	.183±.004	.406±.010	.710±.012

logistic regression, random forest, Extra Trees, and histogram gradient boosting are compared. We report AUROC, AUPRC, Brier score, F1, and balanced accuracy; random-forest uncertainty is estimated using 2,000 patient-level bootstrap resamples.

Robustness is tested by randomly hiding 10%, 30%, and 50% of available physiological/variability features at test time while updating the corresponding missingness indicators. This is deliberately more severe than natural missingness and measures graceful degradation. Subgroup results are descriptive diagnostics rather than fairness certification because the test set is modest and subgroup confidence intervals are not powered for clinical equivalence claims.

6.2 Specified Temporal Benchmark (Not Executed)

The released notebook specifies a temporal benchmark comparing MedWorld-Lite against last observation carried forward (LOCF), direct GRU, direct LSTM, and Transformer forecasting. All neural baselines receive matched value/mask/time-gap/static inputs. No metric described in this subsection is reported as an executed temporal result in this paper. Forecasting is evaluated at 6, 12, and 24 h using normalized macro MAE in standardized units plus per-variable MAE/RMSE in original clinical units. Mortality utility is evaluated with AUROC, AUPRC, Brier, F1, sensitivity, specificity, and balanced accuracy.

For the Gaussian decoder, a nominal 90% interval is $\mu \pm 1.645\sigma$. We report prediction-interval coverage probability (PICP) and normalized mean interval width. Mortality reliability curves assess calibration [10]. Ablations remove personalization, mask/time-gap channels, the recurrent transition, and probabilistic decoding. Additional observation-loss stress is applied to the raw history sequence. Finally, patient-level plots show observed history, ground-truth future measurements, predicted mean trajectory, and uncertainty bands. These visualizations are essential because acceptable average error can conceal implausible individual rollouts.

7 Results

Set A contains 554 deaths among 4,000 rows in the processed table (13.85%). All numerical results in this section concern the executed state-representation

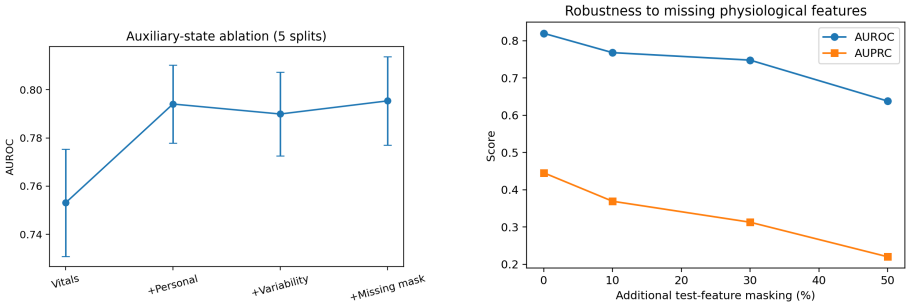

Fig. 2. Executed state-representation probe. Left: cumulative representation ablation. Right: degradation when additional physiological features are hidden at test time.

Table 2. Executed representation-probe model comparison on the fixed seed-42 held-out split.

Model	AUROC	AUPRC	Brier	F1	Bal. Acc.
Random Forest	.858	.474	.099	.325	.600
Hist. Gradient Boosting	.840	.461	.094	.276	.580
Extra Trees	.829	.459	.119	.433	.669
Logistic Regression	.820	.446	.195	.428	.748

probe; none validate the recurrent transition, probabilistic decoder, consistency term, or 6/12/24 h forecasting loss. Table 1 shows the five-split state-feature ablation. Personalization provides the largest AUROC improvement relative to vitals alone. Variability increases mean AUPRC but slightly reduces mean AUROC relative to the personalized state, while explicit missingness indicators recover the best mean AUROC and Brier score. The result supports representing patient context and observation structure explicitly, although it does not establish superiority of the unexecuted temporal transition.

On the fixed held-out split, random forest obtains the highest AUROC (0.858 in the primary rerun; 0.857 with the extended 500-tree validation rerun) and AUPRC, while histogram gradient boosting has the lowest Brier score (Table 2). The extended random-forest run yields AUROC 0.857 with 95% bootstrap CI 0.823–0.889, AUPRC 0.469 (0.383–0.573), and Brier 0.099 (0.089–0.109). The width of these intervals is a useful reminder that a single test split should not be overinterpreted.

Observation loss is substantial: logistic-regression AUROC changes from 0.820 with no additional hiding to 0.768, 0.748, and 0.638 after hiding 10%, 30%, and 50% of available physiological/variability features. AUPRC decreases from 0.446 to 0.369, 0.313, and 0.220. This supports treating missingness as both an input signal and a stress-test dimension instead of assuming complete measurements.

Descriptive subgroup evaluation shows AUROC 0.871 for age < 65 versus 0.834 for age ≥ 65, and 0.847/0.863 for the two coded gender groups. ICU-type

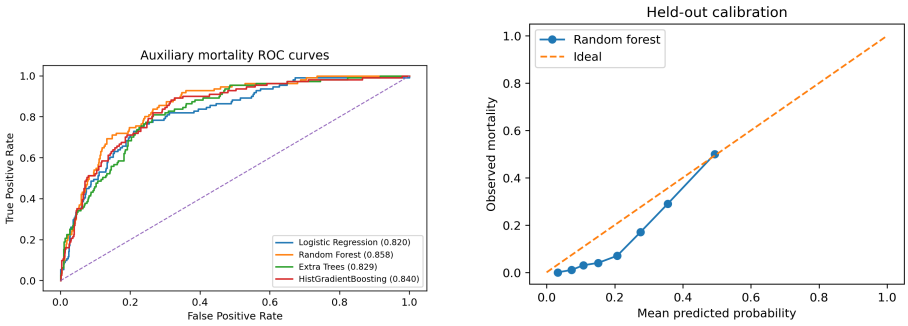

Fig. 3. Executed held-out diagnostics. Left: ROC curves for state-based clinical utility models. Right: random-forest reliability curve. Discrimination and calibration are reported separately because a high AUROC does not imply calibrated risk.

AUROCs range from 0.796 to 0.938, but some groups contain few deaths; these differences therefore motivate uncertainty-aware subgroup validation rather than claims of fairness or superiority. In particular, ICU type 2 has only six deaths in the held-out split.

8 Discussion

The executed representation probe supports two design choices. First, static context materially improves information retained by the constructed patient state. Second, performance degrades sharply as observed physiological information is removed. These findings concern state construction only. They do not validate the recurrent transition, probabilistic decoder, consistency term, or future-physiology forecasting objective; a transition model cannot compensate for an invalid initial state.

The study also illustrates why the term "world model" should be used carefully in medicine. MedWorld-Lite contains a latent state, transition, and observation decoder, but the present evidence validates only state construction and its ability to retain outcome-relevant information. The temporal code is an implementation and evaluation specification, not empirical evidence of dynamics validity. Even a successful 6/12/24 h rollout study would not establish intervention validity: observational ICU data contain treatment effects, clinician behavior, informative sampling, and selection processes, so predicting what usually happens next is not equivalent to predicting what would happen under a changed treatment.

Compared with large foundation approaches such as MIRA or intervention-conditioned EHRWorld [14,15], MedWorld-Lite trades scale for transparent, low-cost experimentation. The limitations are equally clear: hourly binning loses within-hour timing, a diagonal Gaussian can under-represent multimodal futures, and recurrent rollouts may accumulate error.

Intended use and user-facing uncertainty. MedWorld-Lite is an early-stage research framework for studying ICU state representations and future-trajectory models; it is not a diagnostic system, mortality decision aid, or treatment recommender. A prospective interface should expose observation masks and time gaps alongside predictions, display uncertainty intervals, and issue a reduced-confidence or insufficient-data warning when coverage is sparse, gaps are long, or inputs appear shifted from the development distribution. The present study does not validate such an interface or any bedside workflow.

External validity and observation-process bias. The benchmark is an older public cohort and monitoring patterns can reflect hospital protocols, clinician behavior, resource availability, and historical care inequities. Missingness may therefore encode care processes rather than physiology alone. Before clinical use, the evidence ladder should include executed temporal validation, external cohorts from multiple hospitals and care protocols, subgroup calibration, distribution-shift testing, and prospective silent evaluation. Only measurements available and timestamped by the inference time should be used. If interventions are later modeled as actions, causal assumptions and action semantics must be explicit [12].

Conclusion. MedWorld-Lite separates patient-state, dynamics, and intervention evidence. The executed 4,000-patient representation probe supports personalized, missingness-aware state construction and demonstrates sensitivity to observation loss. It does not establish temporal forecasting performance or prospective clinical utility; direct 6/12/24 h rollout evaluation remains the next evidence step.

Disclosure of Interests

The author declares no competing interests.

References

1. Silva, I., Moody, G.B., Scott, D.J., Celi, L.A., Mark, R.G.: Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in Cardiology Challenge 2012. Computing in Cardiology 39, 245–248 (2012)
2. Ha, D., Schmidhuber, J.: World Models. arXiv:1803.10122 (2018)
3. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. In: ICML. PMLR 97, 2555–2565 (2019)
4. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Scientific Reports 8, 6085 (2018)
5. Rubanova, Y., Chen, R.T.Q., Duvenaud, D.: Latent ODEs for irregularly-sampled time series. In: NeurIPS 32 (2019)
6. Zhang, X., Zeman, M., Tsiligkaridis, T., Zitnik, M.: Graph-guided network for irregularly sampled multivariate time series. In: ICLR (2022)
7. Lim, B., Arik, S.O., Loeff, N., Pfister, T.: Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting 37(4), 1748–1764 (2021)

8. Harutyunyan, H., Khachatrian, H., Kale, D.C., Ver Steeg, G., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. Scientific Data 6, 96 (2019)
9. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: NeurIPS 30 (2017)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. PMLR 70, 1321–1330 (2017)
11. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning 16(4), 494–591 (2023)
12. Corral-Acero, J., et al.: The "Digital Twin" to enable the vision of precision cardiology. European Heart Journal 41(48), 4556–4564 (2020)
13. Yang, Y., Wang, Z.-Y., Liu, Q., Sun, S., Wang, K., Chellappa, R., Zhou, Z., Yuille, A., Zhu, L., Zhang, Y.-D., Chen, J.: Medical World Model: Generative Simulation of Tumor Evolution for Treatment Planning. arXiv:2506.02327 (2025)
14. Mu, L., Huang, Z., Gu, Y., Qin, S., Zhang, S., Zhang, X.: EHRWorld: A Patient-Centric Medical World Model for Long-Horizon Clinical Trajectories. arXiv:2602.03569 (2026)
15. Li, H., Deng, B., Xu, C., Feng, Z., Schlegel, V., Huang, Y.-H., Sun, Y., Sun, J., Yang, K., Yu, Y., Bian, J.: MIRA: Medical Time Series Foundation Model for Real-World Health Data. arXiv:2506.07584 (2025)
16. Guo, W.: DT-ICU: Towards Explainable Digital Twins for ICU Patient Monitoring via Multi-Modal and Multi-Task Iterative Inference. arXiv:2601.07778 (2026)
17. Qazi, M.A., Nadeem, M., Yaqub, M.: Beyond Generative AI: World Models for Clinical Prediction, Counterfactuals, and Planning. arXiv:2511.16333 (2025)
18. Chen, Z., Cong, Z., Jin, Z., Fan, W., Zhou, D., Ai, Q., Gong, H., Liao, C., Liu, X., Wang, C.: Medical world models in healthcare: foundations, applications, and challenges for trustworthy clinical translation. arXiv:2607.25242 (2026)
19. Liu, K., Li, M., Bao, Y., Zhang, T., Chu, C., Bu, J., Wang, H.: Medical world models: representing medical states, modelling clinical dynamics and guiding intervention policies. arXiv:2606.16721 (2026)
20. Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmarking deep learning models on large healthcare datasets. Journal of Biomedical Informatics 83, 112–134 (2018)
21. Johnson, A.E.W., Pollard, T.J., Shen, L., et al.: MIMIC-III, a freely accessible critical care database. Scientific Data 3, 160035 (2016)
22. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In: ICLR (2023)