ON THE DIVERGENCE OF ACCURACY AND MECHANISM CONSISTENCY IN TIME SERIES WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A time series world model (TSWM) takes the observed history of a controlled system and a plan of future actions and exogenous inputs, and predicts how the state will respond. Current approaches build such models as forecasters with the actions as covariates, trained and evaluated on prediction error under the executed plan. Yet a world model is used to compare plans that were never executed, so its response to a changed plan matters as much as its error, and this practice never tests it. Two questions are therefore open: which design choices matter, and whether an accurate forecaster responds to a changed plan the way the real system does. We address both with a formalization of TSWMs and a benchmark. The formalization separates state, actions and exogenous inputs, distinguishes continuous, mode and event actions, and introduces mechanism consistency, a metric built on declared mechanisms, relations between an action and a state variable whose direction is known, such as a vasopressor raising blood pressure: it checks whether shifting an action moves the forecast in the declared direction. The benchmark consolidates eight public datasets with real actions from engineered infrastructure and clinical care, and varies the prediction space, the plan fusion and the plan encoding across seven backbones and five seeds. First, a frozen latent prediction space lowers MAE by 9.9% over observation space and gated fusion at the output by 12.7% over concatenation at the input, both on average and on all eight datasets, whereas encoding the plan over time changes the average MAE by at most 2.2%. Second, prediction error and mechanism consistency diverge: the configuration with the lowest error is at or below chance in consistency on four of the five datasets with declared mechanisms, and no design choice avoids this. Finally, directional supervision, a loss term that penalizes the wrong-signed part of the response to a shifted action, raises consistency significantly on penalized mechanisms with no change in MAE. Together they give TSWMs a recipe: a frozen latent space and output-side fusion for accuracy, and a training objective for mechanism consistency.

1 INTRODUCTION

Many operational and clinical decisions depend on how a multivariate time series will evolve under different actions and exogenous inputs. For example, in an intensive-care unit, vital signs such as arterial blood pressure are held in an acceptable range by actions such as raising or lowering the infusion rate of a vasopressor, and are moved as well by uncontrollable exogenous inputs. A model that takes the observed history and a plan of future actions and exogenous inputs and predicts the response of the target is a world model in the sense used in reinforcement learning and video generation (Ha & Schmidhuber, 2018; Hafner et al., 2025; Agarwal et al., 2025); we call it a time series world model (TSWM). Current practice builds it as a forecasting model with actions and exoge-

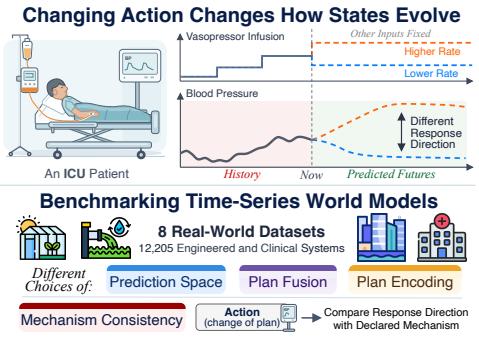


Figure 1: Two plans from the same history. We sweep three design choices and score mechanism consistency.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nous inputs as covariates (Das et al., 2023; Lim et al., 2021; Olivares et al., 2023; Arango et al., 2025; Wang et al., 2024), and recent work has begun to call such models world models for time series (Ding et al., 2025; Liu et al., 2026; Qi et al., 2026). Each of these works proposes one design and compares it against few alternatives inside its own architecture, so the design choices of a TSWM have not been investigated systematically. By contrast, prior work on control problems and video generation has examined world models along two such choices: the space in which the state is predicted, observational (Agarwal et al., 2025) or a pretrained feature space (Assran et al., 2025; Zhou et al., 2024), and the way the action enters the model, by concatenation (Hafner et al., 2019; Zhou et al., 2024), as interleaved tokens (Assran et al., 2025), or as an additive embedding (Bruce et al., 2024). Neither has been tested on time series with real actions.

In applications from multiple domains including retail demand, building control, and clinical prediction, models trained for prediction accuracy alone have been found to respond incorrectly to a change of action (Crasson et al., 2024; Kunz et al., 2023; Di Natale et al., 2022; Keogh & Van Geloven, 2024). A TSWM therefore needs a second property besides prediction accuracy, which we call mechanism consistency. A mechanism is a relation between an action and a state variable whose direction is known, and a mechanism-consistent model responds to a change of that action in that direction: a higher vasopressor infusion rate raises blood pressure, so raising the rate in the input should raise the predicted pressure. This property lets a world model answer what would happen under a plan that has not been taken, so a mechanism-consistent TSWM can compare candidate plans before one is carried out. However, existing benchmarks do not measure this. Forecasting benchmarks rarely contain action-conditional prediction (Godahewa et al., 2021; Aksu et al., 2024), and the benchmarks that score the reaction to an intervention usually test models on the dynamics of a simulator rather than of a real system (Hafner et al., 2019; Mu et al., 2026; Thumm et al., 2026).

Two desiderata are therefore unmet: a benchmark on time series with real actions that compares the design choices of a TSWM systematically, and a formalization of the TSWM with a metric of mechanism consistency, so that the comparison evaluates the reaction to a change of action next to prediction error. We formalize the TSWM problem and build a benchmark for it. The formalization assigns each channel a role, the target as the state, what the controller sets as the actions, and everything else that moves the state as an exogenous input; distinguishes three forms of action, a continuous level such as a dose, a discrete mode such as heating or cooling, and a one-off event such as a meal; and adds mechanism consistency as a metric next to prediction error (Section 3.1).

For dataset construction, we select eight public datasets with real actions, four from engineered infrastructure, namely horticulture, building climate, district heating, and water treatment, and four from clinical care, namely anesthesia, two glucose-monitoring cohorts, and intensive care. Together they cover 12,205 systems, sampling intervals from 2 seconds to 30 minutes, and 41 action channels of all three forms (Section 4.1). On these datasets, the benchmark sweeps the three design components of a TSWM one at a time with the other two fixed: the prediction space, the observation space or the latent space of a frozen autoencoder (AE), a frozen variational autoencoder (VAE) or a joint-embedding predictive architecture (JEPA) (Assran et al., 2023) trained with the forecaster; the plan fusion, at the backbone input by concatenation or at its output by adapters such as FiLM-0 (Perez et al., 2018) and Gate (Dauphin et al., 2017); and the plan encoding over time, with or without a memory of the preceding steps. Each option is run with seven forecasting backbones and five seeds (Section 4). Our contributions are as follows.

- We formalize the TSWM problem and build a benchmark for it from eight public datasets with real actions. Sweeping the three design choices shows that the frozen AE latent space is the best prediction space while JEPA is worse than the observation space, that FiLM-0 and Gate at the output are the best fusion methods and beat fusion at the input, and that a memory of the preceding steps in the plan encoding gives no consistent gain.
- We find a contradiction between prediction accuracy and mechanism consistency: the configuration selected for prediction accuracy is at or below chance in mechanism consistency on four of the five datasets with declared mechanisms, and no design choice escapes this.
- We propose directional supervision, a soft penalty on the sign of the response to a changed action, added to the training loss without changing the network, which raises mechanism consistency above 0.99 on engineered and clinical mechanisms alike without loss of prediction accuracy (Section 5).

2 Related work

Forecasting with exogenous inputs and actions. Covariate-aware forecasters differ in where the covariates enter the model and how they are combined with the target. Concatenation feeds them to the backbone input (Das et al., 2023); gating learns how much of them to pass (Lim et al., 2021; Han et al., 2026; Qiu et al., 2025b); addition keeps them in a separate path summed with the target (Olivares et al., 2023; Arango et al., 2025; Zhou et al., 2025); modulation scales the target representation instead (Qin et al., 2025); and cross-attention lets target tokens query covariate tokens (Wang et al., 2024; Yamaguchi et al., 2026), as it does for treatment sequences (Melnychuk et al., 2022; Bica et al., 2020). Foundation models that take covariates natively (Ansari et al., 2025; Hoo et al., 2025) predict in observation space and are outside our scope. How the effect of an input is spread over time is usually fixed by convention in each field: a carryover kernel in marketing (Jin et al., 2017), a learned decay in clinical series (Che et al., 2018). Each paper compares its design against two or three alternatives inside its own architecture, and none holds the backbone fixed while comparing the families. We hold the backbone fixed and split the plan path into a plan encoder and a fusion adapter that can be swapped independently.

World models in other domains. A world model learns the state transition kernel $p(x_{t+1} \mid x_{\leq t}, a_{\leq t})$ and is judged by rolling it forward. RL rolls it out in a latent space (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025), vision makes it a controllable video generator (Yang et al., 2023; Agarwal et al., 2025) or a predictor in a pretrained feature space (Assran et al., 2025; Zhou et al., 2024), and medicine has begun to use the name for treatment-conditioned disease trajectories (Ding et al., 2025; Liu et al., 2026). Two choices recur: whether the encoder is shaped by the dynamics, and whether the action enters by concatenation, interleaved tokens, or an additive embedding. Time series forecasting has only recently moved into a learned latent space (Yang et al., 2026), and its benchmarks do not score the response to a change of action. We study both choices on time series with real actions and add that score next to the forecast error.

Constraining learned dynamics with prior knowledge. Prior knowledge enters a learned dynamics model as the residual of a governing equation in the loss (Raissi et al., 2019; Karniadakis et al., 2021), as a monotonicity constraint built into the architecture (Nolte et al., 2023; Runje & Shankaranarayana, 2023), or as a loss on the ranking of treatment effects in a hybrid mechanistic–neural model (Zou et al., 2024). The treatment-effect models above instead recover the effect from observational data by correcting for confounding, and Crasson et al. (2024) show that forecast accuracy and causal accuracy can come apart. We impose the sign instead of learning it, because our actions come from a controller that reacts to the state, and the observational sign is often the reverse of the mechanism. Our penalty needs only the sign of a relation; it is a term in the loss, leaves the network unchanged, and is checked by a finite-difference intervention on the same held-out windows, reported alongside the forecast error.

3 Time Series World Model

3.1 Problem Formalization

We consider a controllable system observed at a fixed interval Δt as a discrete-time multivariate series whose channels play one of three roles. The action $\boldsymbol{u}_t \in \mathcal{U}$ collects the channels a controller sets directly, the only degrees of freedom in the system. The exogenous input $\boldsymbol{e}_t \in \mathcal{E} \subset \mathbb{R}^{d_e}$ collects channels that influence the system but that no controller sets. The state $\boldsymbol{x}_t \in \mathcal{X} \subset \mathbb{R}^{d_x}$ collects the regulated quantities, each scaled to $[-1, 1]$, and is what the model predicts. For instance, in a greenhouse, heating-pipe temperature and vent aperture are actions, outdoor temperature and irradiance are exogenous inputs, and indoor temperature and CO_2 concentration are the state.

An action has three components, continuous, mode and event, $\boldsymbol{u}_t = \left(\boldsymbol{u}_t^{\mathrm{c}},\ \boldsymbol{u}_t^{\mathrm{m}},\ u_t^{\mathrm{e}}\right)$. Subscripts give the time step and, where present, the channel; bold denotes all channels of a component; $[n] := \{1, \ldots, n\}$. **(i)** The continuous component holds d_c channels the controller may set anywhere in a bounded range, e.g. an infusion rate or a temperature setpoint. A channel whose declared range starts at an off position, e.g. a closed vent or a stopped infusion, is scaled to $[0, 1]$ so that zero is exactly off, and every other channel to $[-1, 1]$; thus $\boldsymbol{u}_t^{\mathrm{c}} \in [-1, 1]^{d_c}$. **(ii)** The mode component holds d_m channels of operating conditions the controller selects, e.g. a heat pump in heating or

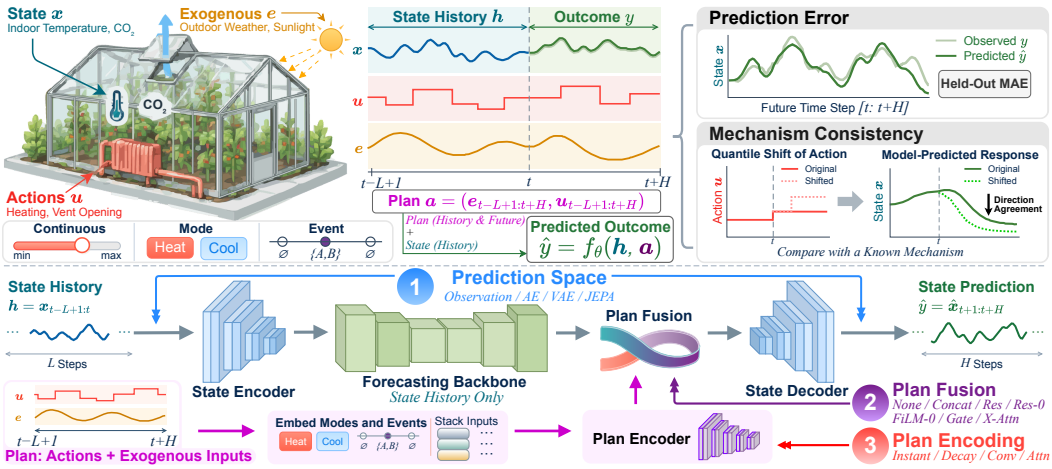


Figure 2: Overview of the formalization and the model. Top: the channel roles, the window of history, plan and outcome, and the two metrics (Section 3.1). Bottom: the model and its three design choices (Section 3.2); the concatenation baseline conditions the backbone instead (Appendix B).

in cooling. Channel $j \in [d_m]$ takes values in a finite unordered set $\mathcal{M}_j$ of size K_j, which we index by $[K_j]$, so $\boldsymbol{u}_t^{\mathrm{m}} \in \prod_{j=1}^{d_m}[K_j]$. **(iii)** The event component holds interventions that occur at an instant, e.g. a drug push. With $\mathcal{A}$ the finite set of event types in a dataset, the value at a step is the subset of $\mathcal{A}$ administered then, empty at most steps; it is a single channel $u_t^{\mathrm{e}} \in \mathcal{V}$, where $\mathcal{V} \subseteq 2^{\mathcal{A}}$ consists of $\varnothing$ and the subsets observed in the training set. The action space is the Cartesian product $\mathcal{U} = [-1,1]^{d_c} \times \prod_{j=1}^{d_m}[K_j] \times \mathcal{V}$.

Fix a context length L and a horizon H, and write $\boldsymbol{x}_{i:j}$ for the block of state values from step i to step j inclusive, likewise for $\boldsymbol{e}$ and $\boldsymbol{u}$. A window at time t is $W_t = (h, a, y)$ with history $h = \boldsymbol{x}_{t-L+1:t}$, plan $a = (\boldsymbol{e}_{t-L+1:t+H}, \boldsymbol{u}_{t-L+1:t+H})$ and outcome $y = \boldsymbol{x}_{t+1:t+H}$. Given training windows $\{(h_n, a_n, y_n)\}_{n=1}^{N}$ drawn from an unknown joint distribution P, the aim is to learn a map $f_\theta : \mathcal{X}^L \times (\mathcal{E} \times \mathcal{U})^{L+H} \to \mathcal{X}^H$ from the history and the plan to the outcome that minimizes the risk $\mathbb{E}_{(h,a,y)\sim P}\big[\|f_\theta(h,a) - y\|_2^2\big]$.

We evaluate a model on two metrics. Prediction error is measured on held-out data and reported as MAE in observation space. Mechanism consistency asks whether f_θ moves the state the way a known mechanism dictates once the plan is changed. A mechanism pairs a continuous action channel $i \in [d_c]$ with a state channel $c \in [d_x]$ under a sign $s_{i\to c} \in \{-1, +1\}$ declared from domain knowledge, $+1$ if raising channel i drives channel c up and -1 if it drives it down; opening a greenhouse vent lowers the air temperature, so $s_{i\to c} = -1$. To test the model, we shift that one action channel, which yields a new plan, and read off the direction in which the forecast moves. Let F_i be the empirical distribution function of $u_{t,i}^{\mathrm{c}}$ over all held-out windows and horizon steps. For $\delta \in (-1, 1) \setminus \{0\}$, the quantile shift $\tau_{i,\delta}$ sets $(\tau_{i,\delta}a)_{t+k,i}^{\mathrm{c}} = F_i^{-1}\big(\min\{\max\{F_i(u_{t+k,i}^{\mathrm{c}}) + \delta, 0\}, 1\}\big)$ at every horizon step $k \in [H]$ and leaves every other entry of h and a unchanged. Writing $[\,\cdot\,]_{k,c}$ for the entry of a forecast at horizon step k and state channel c, the predicted effect is $\Delta_\theta^{i\to c}(h, a; \delta) = \frac{1}{H}\sum_{k=1}^{H}[f_\theta(h, \tau_{i,\delta}a) - f_\theta(h, a)]_{k,c}$ and mechanism consistency is the rate at which its sign agrees with the mechanism, $\mathrm{MC}_\theta(i \to c; \delta) = \Pr_{(h,a)}[\operatorname{sign}\Delta_\theta^{i\to c}(h, a; \delta) = \operatorname{sign}(\delta)\, s_{i\to c}]$, the probability running over the held-out windows on which the shift changes the plan, $\tau_{i,\delta}a \neq a$, and the forecast moves, $|\Delta_\theta^{i\to c}| > 10^{-3}\sigma_c$ with σ_c the median absolute deviation of state channel c over the held-out outcomes; other windows are excluded from the count. A mechanism declared under a mode condition, e.g. only while a heat pump is heating, is scored over the windows that spend at least half of their steps in that mode. $\mathrm{MC}_\theta = 1$ is perfect agreement with the mechanism and $1/2$ is chance.

3.2 Model Design Space

Figure 2 shows the model: a state encoder maps the state history into a prediction space, a backbone forecasts there from the history alone, and a fusion module adjusts that plan-blind forecast with

Table 1: The eight datasets, four engineered and four clinical. Systems are counted train/held-out; windows are held-out windows at $L = 256$, $H = 16$, stride 80. The action columns are the three factors of $\mathcal{U}$: d_c continuous channels, d_m mode channels with their cardinalities K_j, and the number of event types $|\mathcal{A}|$.

			scale		series dim.		actions		
Dataset	Domain	Δt	systems	windows	state	exog.	cont.	mode (K)	event
Greenhouse	horticulture	5 min	5/1	163	4	8	7	–	–
PleiaData	building HVAC	10 min	50/12	7,548	2	8	1	2 (2,6)	–
PREDIST	district heating	10 min	46/12	6,109	5	2	2	3 (2,3,2)	–
Wastewater	water treatment	2 min	1 plant	1,311	3	2	3	2 (2,2)	–
VitalDB	anesthesia	2 s	2,754/689	40,193	4	–	2	–	–
CGMacros	metabolic (CGM)	1 min	36/9	1,736	1	–	–	–	4
Shanghai	metabolic (CGM)	15 min	90/22	214	1	–	–	–	31
MIMIC-Cardio	critical care	30 min	6,782/1,696	8,370	6	–	16	–	2
eight datasets			*12,205*	*65,644*	*26*	*20*	*31*	*7*	*37*

the encoded plan, before a decoder returns to the state channels. The colored blocks are the three design choices. In the notation of Section 3.1, write $\boldsymbol{a}_s \in \mathbb{R}^C$ for step s of the plan as the model receives it, the exogenous input $\boldsymbol{e}_s$ and the continuous actions $\boldsymbol{u}^{\mathrm{c}}_s$ stacked with the embeddings of the mode indices $\boldsymbol{u}^{\mathrm{m}}_s$ and of the event value u^{e}_s. The plan encoder maps $\boldsymbol{a}_{t-L+1:t+H}$ causally to $\tilde{\boldsymbol{a}}_{t-L+1:t+H}$ of the same shape. The state encoder maps every state $\boldsymbol{x}_s$ to a latent $\boldsymbol{z}_s$ (the identity in observation space), and the backbone maps the latent history $\boldsymbol{z}_{t-L+1:t}$ to the plan-blind forecast $\hat{\boldsymbol{z}}_{t+1:t+H}$. The fusion method maps $\hat{\boldsymbol{z}}_{t+k}$ and $\tilde{\boldsymbol{a}}$ to the adjusted forecast $\boldsymbol{z}'_{t+k}$, which the decoder maps to the predicted state $\hat{\boldsymbol{x}}_{t+k}$. Tables 4 and 5 in Appendix B give the formula of every option. **(1) Prediction space**: the observation space, the frozen latent space of an autoencoder (AE) or variational autoencoder (VAE) pretrained on the state channels, or a JEPA trained jointly with the forecaster. **(2) Plan fusion**: NONE discards the plan, CONCAT feeds it into the backbone, and RES, RES-0, FILM-0, GATE and X-ATTN keep the backbone plan-blind and adjust its forecast at each horizon step by a residual, a zero-initialized residual, feature-wise scaling and shifting, a gated residual, or cross-attention over the plan. **(3) Plan encoding**: a causal transform of the plan before fusion, either the identity (INSTANT) or one that lets each step depend on the preceding steps through a learned exponential moving average (DECAY), dilated convolutions (CONV) or self-attention (ATTN), all three initialized as the identity map.

4 BENCHMARKING THE DESIGN CHOICES

4.1 DATASETS

Table 1 summarizes the eight public datasets. We selected them because they record actions as defined in Section 3.1. The four engineered datasets come from horticulture, building HVAC, district heating and wastewater treatment; the four clinical datasets come from anesthesia, two continuous-glucose-monitoring studies and critical care. The engineered datasets have exogenous inputs, and three of them also have mode actions; the clinical datasets have neither, but they are the only datasets with event actions. In total there are 31 continuous action channels, 7 mode channels and 3 event channels covering 37 event types, so each of the three action components is present in at least two datasets. Sampling intervals range from 2 s to 30 min, the number of systems from 1 to 8,478, and the state dimension is at most 6 in every dataset. Each channel is min–max scaled to the range given in Section 3.1. The scaling bounds are fixed per dataset and shared by all methods; they are taken from instrument or physiological limits where such limits exist and from percentiles of the data otherwise. Mode and event channels are represented as integer indices into an embedding table, one index per condition in $\mathcal{M}_j$ or per set in $\mathcal{V}$, plus one reserved index for missing codes and for sets not seen in training. Appendix A gives the channel roles, scaling ranges and split of each dataset.

Computing mechanism consistency requires a set of declared mechanisms. Each mechanism specifies a continuous action channel, a state channel that it affects, and the sign of the effect. Table 2 lists the twenty-one mechanisms used in this paper. All of them were fixed before any model was trained

Table 2: The twenty-one declared mechanisms. A plus sign means that raising the action raises the state, a minus sign that it lowers it; the PleiaData mechanisms hold in the named mode. BIS, bispectral index; MAP, mean arterial pressure; FiO_2, inspired oxygen fraction.

Dataset	Action	State	Sign
Greenhouse	heating-pipe temp.	air temp.	+
	leeward vent	air temp.	−
	windward vent	air temp.	−
	CO_2 dosing	CO_2 level	+
VitalDB	propofol	BIS	−
	propofol	MAP	−
	remifentanil	heart rate	−
	remifentanil	MAP	−
PleiaData	setpoint (heating)	indoor temp.	+
	setpoint (cooling)	indoor temp.	−
PREDIST	valve position	flow	+
	valve position	heat power	+
	supply setpoint	supply temp.	+
MIMIC-Cardio	norepinephrine	systolic pressure	+
	norepinephrine	mean pressure	+
	phenylephrine	systolic pressure	+
	phenylephrine	mean pressure	+
	vasopressin	mean pressure	+
	dobutamine	heart rate	+
	dopamine	heart rate	+
	FiO_2	oxygen saturation	+

or evaluated; in this section they are used only for evaluation, and in Section 5 a subset of them enters the training objective. For the engineered datasets, the signs are taken from the documented operation of each system: for example, a hotter heating pipe raises the greenhouse air temperature, opening a vent lowers it, opening the control valve increases flow and heat power, and the controlled temperature follows its setpoint. For the clinical datasets, the signs are taken from the pharmacology of each drug: propofol deepens hypnosis and lowers blood pressure, remifentanil lowers heart rate and blood pressure, vasopressors raise blood pressure, inotropes raise heart rate, and a higher inspired oxygen fraction raises oxygen saturation. The two PleiaData mechanisms depend on the operating mode: raising the setpoint increases the indoor temperature in heating mode and decreases it in cooling mode. Only five of the eight datasets are used for this metric. The two glucose-monitoring datasets have only event actions, for which the quantile shift of Section 3.1 is not defined, and the single mechanism of Wastewater is excluded for a reason given in Appendix E.1.

4.2 Experimental Setup

We run three experiments, each of which varies one of the design choices in Section 3.2 while fixing the other two. Experiment 1 compares the prediction spaces with CONCAT fusion and INSTANT plan encoding. Experiment 2 compares the fusion methods with the AE prediction space, which Experiment 1 selects, and INSTANT plan encoding. Experiment 3 compares the plan encodings with the AE prediction space and with each of GATE and FILM-0, the two best fusion methods in Experiment 2. Every configuration is trained with seven backbones, TimeXer (Wang et al., 2024), TiDE (Das et al., 2023), DUET (Qiu et al., 2025a), PatchTST (Nie et al., 2022), TimeKAN (Huang et al., 2025), CrossLinear (Zhou et al., 2025) and Amplifier (Fei et al., 2025), on all eight datasets with five random seeds, under a single training protocol (Table 6) and the fixed data split of Appendix A. For each run we keep the checkpoint with the lowest validation loss and report its validation MAE on the scaled state channels as the prediction error. The same checkpoints are then evaluated for mechanism consistency, without any retraining, using the MC metric of Section 3.1 with $|\delta| = 0.1$ in both directions on the twenty-one mechanisms in Table 2. Appendices B to H give the formulas and implementation of all options, hyperparameters, statistical tests, extended results, an auxiliary-loss and rollout diagnostic, and the complete result grids.

4.3 Results

Prediction space. The top block of Table 3 compares the four prediction spaces, pooling the same backbones and seeds for each dataset. The two frozen latent spaces perform best. AE has the lowest mean rank and beats observation-space prediction on all eight datasets. VAE is close to AE, within 3% on average (Appendix D). The jointly trained JEPA is worse than observation-space prediction. We therefore fix the prediction space to AE in the remaining experiments. A likely reason is the frozen, pretrained latent space itself: JEPA also predicts in a latent space, but its encoder is trained jointly, and it is worse than observation space.

Table 3: Validation MAE of the three sweeps. Mean ± std over the 35 backbone–seed runs of one dataset and one option (70 for encoding, which pools GATE and FiLM-0); the std spans backbones. The fixed choices of each sweep are given in Section 4.2. Bold is the lowest mean of each block, Rank is the mean rank over dataset–backbone cells, and the shaded row has the best rank of its block. Lower is better.

	Option	Greenhouse	PleiaData	PREDIST	Wastewater	VitalDB	CGMacros	Shanghai	MIMIC-Cardio	Rank
Space	OBSERVATION	$.0821_{\pm.0269}$	$.0476_{\pm.0306}$	$.1866_{\pm.0099}$	$.1718_{\pm.0116}$	$.0377_{\pm.0090}$	$.0131_{\pm.0054}$	$.1330_{\pm.0132}$	$.0562_{\pm.0029}$	2.52
	AE	$\mathbf{.0761}_{\pm.0200}$	$\mathbf{.0337}_{\pm.0252}$	$\mathbf{.1821}_{\pm.0047}$	$.1695_{\pm.0123}$	$\mathbf{.0329}_{\pm.0026}$	$.0114_{\pm.0049}$	$.1249_{\pm.0096}$	$.0552_{\pm.0021}$	**1.93**
	VAE	$.0956_{\pm.0134}$	$.0339_{\pm.0244}$	$.1836_{\pm.0054}$	$\mathbf{.1672}_{\pm.0106}$	$.0336_{\pm.0039}$	$\mathbf{.0105}_{\pm.0025}$	$\mathbf{.1206}_{\pm.0048}$	$\mathbf{.0552}_{\pm.0020}$	1.95
	JEPA	$.1041_{\pm.0271}$	$.0607_{\pm.0359}$	$.2143_{\pm.0327}$	$.1824_{\pm.0266}$	$.0403_{\pm.0112}$	$.0270_{\pm.0233}$	$.1418_{\pm.0208}$	$.0621_{\pm.0095}$	3.61
Fusion	NONE	$.0751_{\pm.0364}$	$.0292_{\pm.0172}$	$.1858_{\pm.0219}$	$.1928_{\pm.0085}$	$.0396_{\pm.0187}$	$.0174_{\pm.0187}$	$.1257_{\pm.0155}$	$.0608_{\pm.0174}$	4.89
	CONCAT	$.0761_{\pm.0200}$	$.0337_{\pm.0252}$	$.1821_{\pm.0047}$	$.1695_{\pm.0123}$	$.0329_{\pm.0026}$	$.0114_{\pm.0049}$	$.1249_{\pm.0096}$	$.0552_{\pm.0021}$	5.70
	RES	$.0651_{\pm.0041}$	$.0217_{\pm.0112}$	$.1644_{\pm.0019}$	$.1503_{\pm.0065}$	$.0314_{\pm.0008}$	$\mathbf{.0086}_{\pm.0003}$	$.1170_{\pm.0030}$	$.0544_{\pm.0007}$	3.43
	RES-0	$.0652_{\pm.0041}$	$.0216_{\pm.0111}$	$.1642_{\pm.0021}$	$.1509_{\pm.0067}$	$.0315_{\pm.0009}$	$.0087_{\pm.0004}$	$.1166_{\pm.0028}$	$.0544_{\pm.0007}$	3.66
	FiLM-0	$.0642_{\pm.0040}$	$\mathbf{.0203}_{\pm.0106}$	$.1645_{\pm.0023}$	$\mathbf{.1489}_{\pm.0053}$	$.0314_{\pm.0009}$	$.0087_{\pm.0004}$	$.1169_{\pm.0028}$	$.0545_{\pm.0007}$	2.91
	GATE	$\mathbf{.0632}_{\pm.0037}$	$.0205_{\pm.0104}$	$.1641_{\pm.0018}$	$.1500_{\pm.0064}$	$\mathbf{.0313}_{\pm.0009}$	$.0086_{\pm.0004}$	$\mathbf{.1165}_{\pm.0032}$	$\mathbf{.0543}_{\pm.0006}$	**2.52**
	X-ATTN	$.0764_{\pm.0229}$	$.0203_{\pm.0100}$	$\mathbf{.1567}_{\pm.0083}$	$.1617_{\pm.0122}$	$.0323_{\pm.0008}$	$.0107_{\pm.0010}$	$.1236_{\pm.0049}$	$.0548_{\pm.0011}$	4.89
Encoding	INSTANT	$\mathbf{.0637}_{\pm.0038}$	$.0204_{\pm.0104}$	$.1643_{\pm.0021}$	$.1495_{\pm.0059}$	$.0313_{\pm.0009}$	$\mathbf{.0087}_{\pm.0004}$	$.1167_{\pm.0030}$	$.0544_{\pm.0007}$	2.79
	DECAY	$.0638_{\pm.0041}$	$.0208_{\pm.0104}$	$.1597_{\pm.0021}$	$.1417_{\pm.0089}$	$\mathbf{.0313}_{\pm.0009}$	$.0087_{\pm.0005}$	$.1168_{\pm.0031}$	$\mathbf{.0544}_{\pm.0007}$	2.40
	CONV	$.0641_{\pm.0037}$	$.0187_{\pm.0086}$	$\mathbf{.1570}_{\pm.0018}$	$\mathbf{.1388}_{\pm.0061}$	$.0314_{\pm.0009}$	$.0087_{\pm.0004}$	$\mathbf{.1163}_{\pm.0030}$	$.0544_{\pm.0006}$	**2.06**
	ATTN	$.0638_{\pm.0039}$	$\mathbf{.0172}_{\pm.0068}$	$.1591_{\pm.0029}$	$.1471_{\pm.0063}$	$.0315_{\pm.0007}$	$.0093_{\pm.0005}$	$.1174_{\pm.0033}$	$.0545_{\pm.0006}$	2.74

Plan fusion. The middle block of Table 3 compares the seven fusion methods under the same AE representation. The two modulation methods work best: GATE and FiLM-0 have the best mean ranks, both beat CONCAT and NONE on all eight datasets, and they differ by less than 1% on average. The residual methods also do well, and cross-attention has no consistent advantage over these simpler adapters. Adjusting the predicted representation with the plan is therefore an effective way to bring in action and exogenous information. We keep both GATE and FiLM-0 for the encoding experiment; detailed comparisons are in Appendix D.

Plan encoding. The bottom block of Table 3 tests whether letting the encoding of each step depend on the preceding steps improves prediction. CONV has the best mean rank and a small average improvement over INSTANT, 2.2%, but the gain holds on only four of the eight datasets (Appendix D). DECAY and ATTN give no consistent improvement either. A memory of the preceding steps can therefore help on individual datasets, but we find no general benefit beyond the history summaries that the fusion module already receives.

Mechanism consistency. Figure 3 reports the MC metric of every design choice on the five datasets. No design choice does well: averaged over the five datasets, all choices score between 0.46 and 0.59, close to the chance level of 0.5. PREDIST is the easy case: its mechanisms are direct actuator effects, a valve on the flow it throttles and a setpoint on the temperature the controller tracks, and the training data already move in the declared direction, so fitting the training data is enough to get the sign right, and every choice is consistent there. On the other four datasets all choices are near or below chance, and on the two clinical datasets the models even learn the opposite sign. A likely reason is that the drugs are given in response to the state: a vasopressor is given when blood pressure is low and propofol when anesthesia is too shallow, so in the training data a higher dose coincides with the condition it is meant to correct, and a model that fits the training data may learn this association instead of the effect of the drug. The choice with the lowest MAE on each dataset, boxed in Figure 3, is no more consistent than the others, so training on forecasting error alone does not give a mechanism-consistent world model; Section 5 addresses this.

5 LEARNING MECHANISM-CONSISTENT RESPONSES

Section 4.3 shows that the forecasting objective supervises $f_\theta(h, a)$ only under the plan that was carried out and provides no target under an alternative plan. We add this supervision with a method we call directional supervision. Given the same history, we compare forecasts under the executed and the perturbed plan and penalize the part of their difference that opposes a declared mechanism. We apply this objective to the benchmark's AE–GATE–INSTANT configuration.

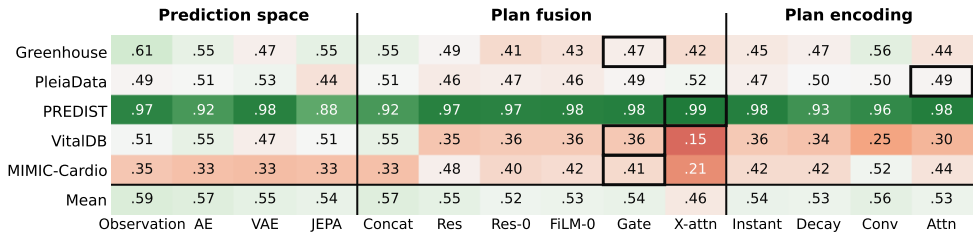


Figure 3: Mechanism consistency of the three sweeps. MC at $\delta = \pm 0.1$, averaged over both directions, the declared mechanisms and the runs of Table 3. White is chance, green agrees with the declared sign, red opposes it. Boxes mark the lowest-MAE choice of each row. NONE is omitted because its forecast does not move. Numbers are in Appendix E.

5.1 DIRECTIONAL SUPERVISION

Objective. The forecasting loss scores f_θ only under the plan that was carried out. We add a second term that scores its response to a change of plan. For a declared mechanism (i, c, s) with $s \in \{-1, +1\}$, raising action channel i should move the forecast of state channel c in direction s; the term compares two forecasts of the same window that differ only in the future values of channel i and penalizes the part of their difference that points against s. Let E and D be the frozen state encoder and decoder, B_θ the backbone and G_θ the gated fusion module, with the history-context arguments of G_θ omitted. The two forecasts are

$$\hat{z} = B_\theta(E(h)), \qquad \hat{y} = D(G_\theta(\hat{z}, a)), \qquad \hat{y}^+ = D(G_\theta(\hat{z}, \tau_{i,\delta} a)). \tag{1}$$

The backbone runs once, since it never sees the plan; history, exogenous inputs and the other action channels are identical in both branches. Gradients flow through D into G_θ and B_θ, while E and D stay frozen. Inference uses the single branch under the supplied plan.

Perturbation. The shifted plan is the quantile shift $\tau_{i,\delta} a$ of Section 3.1, with the quantiles of channel i estimated on the training actions, so the shifted values stay within the range the channel actually takes. For each batch we draw one $\delta \sim \mathcal{U}(0.05, 0.30)$, so the penalty covers a band of shift sizes. Training shifts only upward; downward shifts are never penalized and are reserved for evaluation.

Penalty. For a mechanism $p = (i, c, s)$, let $d^p_{k,c} = [\hat{y}^+ - \hat{y}]_{k,c}$ be the decoded response at horizon step k, let $\langle \cdot \rangle_p$ average over the windows of the batch and the horizon steps, restricted to the steps in the required mode for a conditional mechanism, and let $\mathcal{P}_b$ be the mechanisms with at least one eligible step in the batch. We minimize

$$\mathcal{L}_{\text{dir}} = \frac{1}{|\mathcal{P}_b|} \sum_{p \in \mathcal{P}_b} \frac{\left\langle \text{relu}(-s\, d^p_{k,c}) \right\rangle_p}{\text{sg}\left(\left\langle |d^p_{k,c}| \right\rangle_p \right) + \varepsilon}, \qquad \mathcal{L} = \mathcal{L}_{\text{forecast}} + \rho\, \mathcal{L}_{\text{dir}}, \tag{2}$$

where $\mathcal{L}_{\text{forecast}}$ is the unchanged MSE objective of the benchmark, including any backbone-specific auxiliary loss, sg stops gradients, $\varepsilon = 10^{-6}$, and $\mathcal{L}_{\text{dir}} = 0$ when $\mathcal{P}_b$ is empty. The numerator is the response mass in the wrong direction and the denominator the total response mass, so $\mathcal{L}_{\text{dir}}$ is roughly the fraction of the response with the wrong sign: it is zero when the response has the declared sign everywhere, it is dimensionless, so that one ρ serves datasets whose effect sizes differ, and the detached denominator lets the gradient act only on the wrong-signed part. The penalty is applied after decoding, because the declared sign refers to the state channels. We use $\rho = 0.1$, the smallest value in $\{0.1, 1, 10\}$ that met the consistency criterion of a Greenhouse pilot, fixed before the full grid; early stopping uses the forecasting loss alone. The constraint is soft and does not guarantee monotonicity.

5.2 RESULTS

We train on the six datasets with continuous actions, with the backbones, seeds, split and protocol of Section 4.2, and pair each run with the GATE run of Experiment 2 that shares its dataset, backbone and seed. The penalty covers twelve mechanisms of Table 2, three on Greenhouse, one on PleiaData,

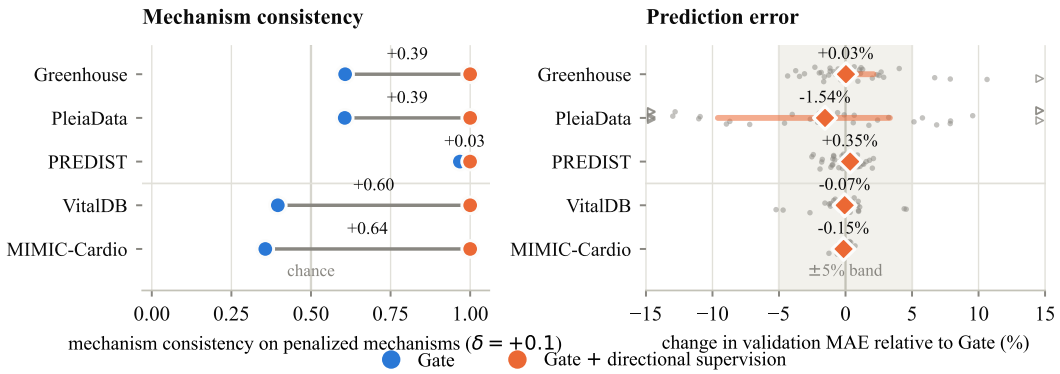


Figure 4: Directional supervision against the GATE control. Left: mechanism consistency (MC at $\delta = +0.1$) on the penalized mechanisms, mean over the 35 backbone–seed runs of each dataset, with the paired gain. Right: change in validation MAE relative to GATE per run, with the median (diamond), a 90% interval (bar) and the $\pm 5\%$ band (shaded).

two on PREDIST, two on VitalDB and four on MIMIC-Cardio; the remaining nine are held out, and Wastewater, which has no penalized mechanism, serves as a null control. Figure 4 reports MC at $\delta = +0.1$ on held-out windows and validation MAE, with the paired change per run; Appendix F gives every number. MC on the penalized mechanisms rises on every dataset to a mean above 0.99 over the 35 runs on each mechanism, including the windward vent, on which GATE scores 0.196, and four of the six clinical mechanisms, on which GATE is below chance. The downward shift, which the penalty never sees, is as consistent as the upward one, so the learned response holds on held-out windows and in the untrained direction. Prediction error is unchanged: the median change in validation MAE, pooled over datasets, is -0.05%, and the per-dataset medians lie between -1.54% and $+0.35\%$, inside the $\pm 5\%$ band; MIMIC-Cardio shows a small improvement, and PleiaData has the widest spread across runs. The gain is confined to the supplied signs: the held-out drug mechanisms stay below chance except dopamine, and three held-out mechanisms move away from their sign, the PleiaData cooling-mode setpoint, vasopressin and remifentanil on mean pressure (Appendix F). Directional supervision therefore enforces the signs it is given and only those: it does not recover undeclared mechanisms and does not constrain effect sizes. Within that scope, it gives a forecaster that keeps the accuracy of the best adapter and responds to a changed plan in the declared direction, which no choice in the design space achieved.

6 CONCLUSION

We formalized the time series world model, a forecaster that takes the observed history and a plan of actions and exogenous inputs and predicts the response of the state, and asked two questions that current practice leaves open: which design choices matter, and whether an accurate forecaster responds to a changed plan the way the real system does. A benchmark of eight public datasets with real actions, swept one design choice at a time, answers the first: a frozen latent prediction space and gated fusion at the output lower the error across datasets and backbones, whereas a memory in the plan encoding brings no consistent gain. Mechanism consistency, a metric that scores the direction of the response to a shifted action, answers the second in the negative: the most accurate configurations are at or below chance on most datasets, and no design choice escapes this, because the sign in the training data is observational and an architecture does not change the data. Directional supervision supplies the sign through the loss and raises consistency to near its maximum at no cost in error, so a mechanism-consistent world model is a property of the training objective rather than of the architecture. Future work includes extending the supervision to event actions, clinical validation of the declared pharmacological signs, and evaluation under closed-loop control.

AI USE STATEMENT

In this work, we used generative AI tools to check the manuscript for grammatical and spelling errors, to refine language expression, to prepare the tables and figures from the stored metric files, and to organize the code. We did not use generative AI tools to write the manuscript, to write the experiment code, to generate or label data, to declare the mechanism signs of Table 2, or to produce or select the experimental results. The authors are responsible for the final manuscript, code and other research artifacts, including AI-assisted content.

ETHICS STATEMENT

This work studies action-conditioned forecasting on existing public datasets and pretrained models; no new data were collected from human subjects. VitalDB and MIMIC-IV are de-identified and used under their access agreements; MIMIC-IV requires credentialed access, so we release the extraction code for MIMIC-Cardio rather than the data. The declared mechanisms are simplified statements of direction, not clinical guidance, and a model that satisfies mechanism consistency is not thereby validated for treatment decisions: the metric checks the sign of a response and says nothing about its size or its safety. Directional supervision enforces the signs it is given, so a wrongly declared sign would be enforced as well; the signs should be reviewed by domain experts before the method is used in a new setting. Any deployment on patient data requires evaluation in the intended clinical setting and appropriate privacy protections.

REPRODUCIBILITY STATEMENT

The problem formalization, the three action forms and the mechanism-consistency metric are defined in Section 3.1, and the design space in Section 3.2, with the formula of every fusion method and action encoder in Tables 4 and 5. Section 4.1 and Appendix A give the channel roles, scaling ranges, split and windowing of each dataset, and Table 2 the declared mechanisms. Section 4.2 states what each experiment fixes and varies; Table 6 lists every hyperparameter of the forecasting protocol, the codecs, the fusion and encoding modules and the mechanism-consistency probe, and Appendix F those of the directional-supervision term; none is tuned per dataset or backbone. Appendix C gives the aggregation and the bootstrap intervals behind the reported effects, Appendix D the paired effects and their interaction with the backbone, Appendix H the complete validation grids of Experiments 1–3, and Appendix E the numbers behind Figure 3, the per-mechanism scores and the mechanisms that were set aside. The objective of Section 5 is given in full in Equations 1 and 2, and Appendix F gives its hyperparameters, the penalized set and the score of every mechanism and dataset. The preprocessing code, the MIMIC-Cardio extraction script, the declared mechanisms, the training and evaluation code, the mechanism-consistency probe, the configuration of every run and the launch scripts of every experiment are available at `https://anonymous.4open.science/r/TSWM-Official-Repo-4016/`.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

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. *arXiv preprint arXiv:2410.10393*, 2024.

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. *arXiv preprint arXiv:2510.15821*, 2025.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Senetaire, Caner Turkmen, Oleksandr Shchur, Danielle C Maddix, et al. Chronosx: Adapting pretrained time series models with exogenous variables. *arXiv preprint arXiv:2503.12107*, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela Van Der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Thomas Crasson, Yacine Nabet, and Mathias Lécuyer. Training and evaluating causal forecasting models for time-series. *arXiv preprint arXiv:2411.00126*, 2024.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Anurag Das, David Kerr, Namino Glantz, Wendy Bevier, Rony Santiago, Ricardo Gutierrez-Osuna, and Bobak J Mortazavi. Cgmacros: a pilot scientific dataset for personalized nutrition and diet monitoring. *Scientific Data*, 12(1):1557, 2025.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Loris Di Natale, Bratislav Svetozarevic, Philipp Heer, and Colin N Jones. Physically consistent neural networks for building thermal modeling: theory and analysis. *Applied Energy*, 325:119806, 2022.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. Clarity: Medical world model for guiding treatment decisions by modeling context-aware disease trajectories in latent space. *arXiv preprint arXiv:2512.08029*, 2025.

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: Bringing attention to neglected low-energy components in time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11645–11653, 2025.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Yinbo Sun, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, et al. Unica: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 6403–6443, 2026.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending tabpfn-v2 to time series forecasting. *arXiv preprint arXiv:2501.02945*, 2025.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 28514–28529, 2025.

Antonio Martínez Ibarra, Aurora González-Vidal, and Antonio Skarmeta. Pleiadata: consumption, hvac, temperature, weather and motion sensor data for smart buildings applications. *Scientific Data*, 10(1):118, 2023.

Yuxue Jin, Yueqing Wang, Yunting Sun, David Chan, and Jim Koehler. Bayesian methods for media mix modeling with carryover and shape effects. Technical report, Google Inc., 2017.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Ruth H Keogh and Nan Van Geloven. Prediction under interventions: evaluation of counterfactual performance using longitudinal observational data. *Epidemiology (Cambridge, Mass.)*, 35(3): 329–339, 2024.

Manuel Kunz, Stefan Birr, Mones Raslan, Lei Ma, and Tim Januschowski. Deep learning based forecasting: a case study from the online fashion industry. In *Forecasting with artificial intelligence: theory and applications*, pp. 279–311. Springer, 2023.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

Ke Liu, Mengxuan Li, Yanyi Bao, Tianyun Zhang, Chong Chu, Jiajun Bu, and Haishuai Wang. Medical world models: representing medical states, modelling clinical dynamics and guiding intervention policies. *arXiv preprint arXiv:2606.16721*, 2026.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*, pp. 15293–15329. PMLR, 2022.

Esmaeel Mohammadi, Anju Rani, Mikkel Stokholm-Bjerregaard, Daniel Ortiz-Arroyo, and Petar Durdevic. Wastewater treatment plant data for nutrient removal system. *arXiv preprint arXiv:2407.05346*, 2024.

Wenhao Mu, Facundo Yan, Anik Mumssen, Marisa Eisenberg, and Alexander Rodríguez. Benchmarking counterfactual prediction in epidemic time series with time-varying interventions. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 9593–9603, 2026.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Niklas Nolte, Ouail Kitouni, and Mike Williams. Expressive monotonic neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Kin G Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Anna Petropoulou, Bart van Marrewijk, Silke Hemming, Feije de Zwart, Anne Elings, and Monique Bijlaard. 3rd autonomous greenhouse challenge: Time-series data on realized climate with annotated crop lettuce-images. 2023.

Xiaoqian Qi, Haoye Chai, Yue Wang, and Yong Li. Beyond static forecasting: Unleashing the power of world models for mobile traffic extrapolation. *arXiv preprint arXiv:2604.08199*, 2026.

Guo Qin, Zhi Chen, Yong Liu, Zhiyuan Shi, Haixuan Liu, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Cora: Covariate-aware adaptation of time series foundation models. *arXiv preprint arXiv:2510.12681*, 2025.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 1185–1196, 2025a.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. Dag: A dual correlation network for time series forecasting with exogenous variables. *arXiv preprint arXiv:2509.14933*, 2025b.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Cyriana MA Roelofs, Edison Guevara Bastidas, Thomas Hugo, Stefan Faulstich, and Anna Cadenbach. Enabling predictive maintenance in district heating substations: A labelled dataset and fault detection evaluation framework based on service data. *Energy*, pp. 141178, 2026.

Davor Runje and Sharath M Shankaranarayana. Constrained monotonic neural networks. In *International Conference on Machine Learning*, pp. 29338–29353. PMLR, 2023.

Dennis Thumm, Billy Tim Anthony, and Ying Chen. Dotime: A synthetic benchmark generator for interventional and counterfactual time series. *arXiv preprint arXiv:2607.27263*, 2026.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in neural information processing systems*, 37:469–498, 2024.

Yosuke Yamaguchi, Issei Suemitsu, and Wenpeng Wei. Citras: Covariate-informed transformer for time series forecasting. *IEEE Access*, 2026.

Jie Yang, Yifan Hu, Yuante Li, Kexin Zhang, Kaize Ding, and Philip S Yu. From observations to states: Latent time series forecasting. *arXiv preprint arXiv:2602.00297*, 2026.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Qinpei Zhao, Jinhao Zhu, Xuan Shen, Chuwen Lin, Yinjia Zhang, Yuxiang Liang, Baige Cao, Jiangfeng Li, Xiang Liu, Weixiong Rao, et al. Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, 10(1):35, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4120–4131, 2025.

Bob Junyi Zou, Matthew E Levine, Dessi P Zaharieva, Ramesh Johari, and Emily B Fox. Hybrid^2 neural ode causal modeling and an application to glycemic response. *arXiv preprint arXiv:2402.17233*, 2024.

A Datasets and Preprocessing

This appendix supports Section 4.1. It describes each dataset of Table 1 under the three roles of Section 3.1, state, action and exogenous input, and then the two preprocessing steps shared by all methods, the scaling of every channel and the split into training and held-out windows.

A.1 The eight datasets

Greenhouse (Petropoulou et al., 2023) contains the climate records of six greenhouse compartments from the third Autonomous Greenhouse Challenge, in which six teams each grew a lettuce crop under their own control strategy over about six months, at 5-minute resolution. The state consists of the indoor temperature, relative humidity, CO_2 concentration and photosynthetically active radiation. The seven continuous actions are actuator intensities: heating-pipe temperature, leeward and windward vent apertures, artificial lighting, energy screen, blackout screen and CO_2 dosing. The exogenous inputs are eight outdoor weather variables, including temperature, humidity, irradiance, wind and rain. **PleiaData** (Ibarra et al., 2023) contains one year of operation data from a university building at 10-minute resolution; a system pairs an indoor temperature sensor with the HVAC unit serving its room. The state consists of the indoor temperature and the energy consumed per step. The continuous action is the temperature setpoint entered by the occupants, and two mode channels record the on/off status and the operating mode of the unit (off, heat, cool, dry, fan, auto). The exogenous inputs are eight outdoor weather variables. **PREDIST** (Roelofs et al., 2026) contains operational records of 58 district-heating substations at 10-minute resolution. The state consists of the secondary supply temperature, the primary and network return temperatures, heat power and flow. The continuous actions are the control-valve position and the supply-temperature setpoint issued by the control unit; the mode channels are the pump status, the control mode and the position of the domestic hot-water three-way valve. The exogenous inputs are the outdoor temperature and the network supply temperature. We exclude maintenance events, which are manual repairs rather than control signals. **Wastewater** (Mohammadi et al., 2024) contains sensor and dosing records from a biological reactor of the Agtrup wastewater treatment plant in Kolding, Denmark, at 2-minute resolution. The state consists of the dissolved oxygen, ammonium and phosphate concentrations in the tank. The continuous actions are two metal-salt dosing rates for phosphate precipitation and a chemical dosing factor; the mode channels are the process phases at the inlet and the outlet. The exogenous inputs are the influent flow and the wastewater temperature. Since the dataset covers a single plant, we split it in time rather than by system. **VitalDB** (Lee et al., 2022) contains intraoperative vital signs of 3,443 surgical cases at 2-second resolution. The state consists of mean arterial pressure, heart rate, peripheral oxygen saturation and the bispectral index. The continuous actions are the infusion rates of propofol and remifentanil, the hypnotic and the opioid used to maintain anesthesia. There are no exogenous inputs. **CGMacros** (Das et al., 2025) contains continuous glucose monitoring traces of 45 participants at 1-minute resolution together with their meal logs. The state is the glucose concentration. The actions are events of four types (breakfast, lunch, dinner and snack). **Shanghai** (Zhao et al., 2023) contains continuous glucose monitoring traces of 112 patients with type 1 or type 2 diabetes at 15-minute resolution together with their meal, injection and oral-medication records. The state is the glucose concentration. The actions are events of 31 types: meals, 16 injected agents (insulin preparations and GLP-1 agonists) and 14 oral hypoglycemic drugs, several of which may be administered at the same step. **MIMIC-Cardio** is our extraction from MIMIC-IV (Johnson et al., 2023) of the intensive-care stays of 8,478 patients, resampled to 30-minute resolution. The state consists of the heart rate, respiratory rate, oxygen saturation, and systolic, diastolic and mean non-invasive blood pressure. The 16 continuous actions are the rates of nine continuous infusions (norepinephrine, epinephrine, phenylephrine, vasopressin, dopamine, dobutamine, milrinone, furosemide and intravenous fluid), the amounts of two boluses (furosemide and fluid), and five ventilator settings (FiO_2, PEEP, oxygen flow, set tidal volume and set respiratory rate). The event channel records the two bolus types as pushes. We exclude the invasive arterial-pressure channels, which are missing for 52.6% of stays, and 89 epinephrine pushes whose recorded amounts are unit-entry errors.

A.2 Scaling ranges

The scaling range of a channel is fixed once per dataset before training. Greenhouse, PleiaData, VitalDB, CGMacros, Shanghai and the ventilator channels of MIMIC-Cardio use constants declared from instrument or physiological limits, e.g. 40 to 400 mg/dL for glucose. Wastewater uses the 1st and 99th percentiles of the training segment, PREDIST uses the 1st and 99th percentiles of each substation's own series, and the drug channels of MIMIC-Cardio use the 99.5th percentile of the non-zero doses over the cohort. The continuous actions of Greenhouse, VitalDB and MIMIC-Cardio have declared ranges whose lower end is an off position and map to $[0, 1]$. All other channels map to $[-1, 1]$, including the PREDIST valve position and the Wastewater dosing rates, whose ranges are percentiles of the data, so their zero, although it is an off position, is not mapped to a fixed point.

A.3 Splits and windows

Seven datasets contain several systems of one kind, and we split them by system: all windows of one system go either to the training set or to the held-out set, so that a model is tested on systems it has never seen and the N training windows, which overlap in time within a system, are not independent samples. The system is the team for Greenhouse, the sensor–HVAC pair for PleiaData, the substation for PREDIST, the surgical case for VitalDB, the subject for CGMacros and Shanghai, and the patient for MIMIC-Cardio; the train/held-out counts are in Table 1. Wastewater covers a single plant and is split in time, the first 80% for training and the rest held out. Within a system, the series is cut wherever two consecutive samples are further apart than a tolerance, and every window lies entirely inside one of the resulting segments. The tolerance is 30 minutes for Greenhouse, PleiaData, PREDIST, CGMacros and Shanghai, 10 minutes for Wastewater and 2 minutes for VitalDB; a MIMIC-Cardio stay is a single contiguous segment on its 30-minute grid, so it is never cut.

B Model Design Space: Formulas and Implementation

This appendix supports Section 3.2. It gives the formula of every option of the three design choices in the notation of Section 3.1 and the implementation details that the formulas leave open.

B.1 Formulas

Tables 4 and 5 give the formula of every fusion method and every plan encoder in the notation of Section 3.1, for the window W_t defined there: the history covers steps $t - L + 1$ to t and the horizon covers steps $t + k$ for $k \in [H]$. We write $\boldsymbol{a}_s \in \mathbb{R}^C$ for step s of the plan a as the model receives it, i.e., the exogenous input $\boldsymbol{e}_s$ and the continuous actions $\boldsymbol{u}^{\mathrm{c}}_s$ stacked with the embeddings of the mode indices $\boldsymbol{u}^{\mathrm{m}}_s$ and of the event value u^{e}_s. The action encoder maps $\boldsymbol{a}_{t-L+1:t+H}$ to $\tilde{\boldsymbol{a}}_{t-L+1:t+H}$ of the same shape, causally, so that $\tilde{\boldsymbol{a}}_s$ depends on $\boldsymbol{a}_{t-L+1:s}$ only. The state encoder maps every state $\boldsymbol{x}_s$ to a latent $\boldsymbol{z}_s \in \mathbb{R}^{32}$; in observation space $\boldsymbol{z}_s = \boldsymbol{x}_s$. The backbone maps the latent history $\boldsymbol{z}_{t-L+1:t}$ to the plan-blind forecast $\hat{\boldsymbol{z}}_{t+1:t+H}$, the fusion method maps $\hat{\boldsymbol{z}}_{t+k}$ to the adjusted forecast $\boldsymbol{z}'_{t+k}$, and the decoder maps $\boldsymbol{z}'_{t+k}$ to the predicted state $\hat{\boldsymbol{x}}_{t+k}$. The stepwise fusion methods share a history summary $\boldsymbol{r} = \mathrm{ReLU}\big(W_r[\boldsymbol{z}_t; \bar{\boldsymbol{z}}; \tilde{\boldsymbol{a}}_t; \bar{\tilde{\boldsymbol{a}}}]\big)$, where a bar denotes the mean over the L history steps, and a per-step feature $\boldsymbol{\phi}_k = [\boldsymbol{r}; \tilde{\boldsymbol{a}}_{t+k}; \boldsymbol{p}_k]$ with a learned position embedding $\boldsymbol{p}_k$; g is a two-layer MLP with a ReLU hidden layer of width 128, and $\odot$ is the elementwise product. A zero-initialized layer makes its method the identity at the start of training, i.e., $\boldsymbol{z}'_{t+k} = \hat{\boldsymbol{z}}_{t+k}$ or $\tilde{\boldsymbol{a}}_s = \boldsymbol{a}_s$.

B.2 Implementation

The AE pretraining objective is reconstruction MSE. The VAE adds a KL penalty with weight 10^{-3} and uses the posterior mean when forecasting. The pretrained codec is frozen and shared across the five forecasting seeds. Our JEPA variant is trained jointly from scratch with observation-space forecasting MSE plus a latent MSE term of weight 0.02. The latter compares the predicted latent, after plan fusion, with an exponential-moving-average encoder of the true future states; its momentum rises from 0.996 to 1 on a cosine schedule. Built-in backbone regularizers are retained, including DUET's expert load-balancing term. All runs share the protocol of Table 6; the held-out set of the fixed split selects the checkpoint and provides the reported MAE. The plan carries actions and exogenous inputs together, and the three experiments do not separate their contributions.

Table 4: Plan-fusion methods. $\hat{\boldsymbol{z}}_{t+k}$ is the plan-blind forecast of the backbone at horizon step $t+k$ and $\boldsymbol{z}'_{t+k}$ the adjusted forecast that is decoded; $\tilde{\boldsymbol{a}}$ is the encoded plan. CONCAT is the only method whose plan enters the backbone itself.

Method	Adjusted forecast $\boldsymbol{z}'_{t+k}$	Details
NONE	$\hat{\boldsymbol{z}}_{t+k}$	Plan discarded.
CONCAT	$\hat{\boldsymbol{z}}_{t+k}$, backbone input $[\boldsymbol{z}_s; \tilde{\boldsymbol{a}}_s]$	The history plan enters with the latent history; the future plan $\tilde{\boldsymbol{a}}_{t+k}$ enters through the native covariate interface of TimeXer and TiDE, and through $\mathrm{MLP}([\hat{\boldsymbol{z}}_{t+k}; \tilde{\boldsymbol{a}}_{t+k}])$ otherwise.
RES	$\hat{\boldsymbol{z}}_{t+k} + g(\boldsymbol{\phi}_k)$	Standard initialization.
RES-0	$\hat{\boldsymbol{z}}_{t+k} + g(\boldsymbol{\phi}_k)$	Output layer of g zero-initialized.
FILM-0	$(1+\boldsymbol{\gamma}_k) \odot \hat{\boldsymbol{z}}_{t+k} + \boldsymbol{\beta}_k$	$[\boldsymbol{\gamma}_k; \boldsymbol{\beta}_k] = g(\boldsymbol{\phi}_k)$, output layer zero-initialized.
GATE	$\hat{\boldsymbol{z}}_{t+k} + \boldsymbol{\alpha}_k \odot \sigma(\boldsymbol{b}_k)$	$[\boldsymbol{\alpha}_k; \boldsymbol{b}_k] = g([\boldsymbol{\phi}_k; \hat{\boldsymbol{z}}_{t+k}])$, σ the logistic function.
X-ATTN	$\hat{\boldsymbol{z}}_{t+k} + W_o\,\mathrm{MHA}(\boldsymbol{q}_k,\, K,\, K)$	Query $\boldsymbol{q}_k = W_q \hat{\boldsymbol{z}}_{t+k} + \boldsymbol{p}_k$; keys and values $K = W_{kv}\tilde{\boldsymbol{a}}_{t-L+1:t+H} + P'$ over the whole plan, with position embeddings P'; four heads, width 128.

Table 5: Plan encoders. Each maps the plan $\boldsymbol{a}_{t-L+1:t+H}$ to $\tilde{\boldsymbol{a}}_{t-L+1:t+H}$ of the same width, causally and per channel unless stated otherwise; all parameters are per-channel vectors in $\mathbb{R}^C$ except for ATTN.

Encoder	Encoded plan $\tilde{\boldsymbol{a}}_s$	Details
INSTANT	$\boldsymbol{a}_s$	Identity.
DECAY	$\boldsymbol{a}_s + \boldsymbol{\eta} \odot (\boldsymbol{m}_s - \boldsymbol{a}_s)$	$\boldsymbol{m}_s = (1-\boldsymbol{\alpha}) \odot \boldsymbol{a}_s + \boldsymbol{\alpha} \odot \boldsymbol{m}_{s-1}$ with $\boldsymbol{\alpha} = \sigma(\boldsymbol{\theta})$ learned per channel; the gate $\boldsymbol{\eta}$ is zero-initialized.
CONV	$\boldsymbol{v}_s^{(6)}$, with $\boldsymbol{v}^{(0)} = \boldsymbol{a}$	Six residual blocks $\boldsymbol{v}_s^{(l)} = \boldsymbol{v}_s^{(l-1)} + \mathrm{GELU}(\boldsymbol{c}_s^{(l)})$ with $\boldsymbol{c}_s^{(l)} = \sum_{j=0}^{7} \boldsymbol{w}_j^{(l)} \odot \boldsymbol{v}_{s-2^{l-1}j}^{(l-1)}$: depthwise kernels of eight taps with dilation 2^{l-1}, zero-initialized; receptive field 442 steps.
ATTN	$\boldsymbol{a}_s + W_o\,\mathrm{MHA}(\boldsymbol{o}_s,\, \boldsymbol{o}_{\le s},\, \boldsymbol{o}_{\le s})$	$\boldsymbol{o}_s = W_i \boldsymbol{a}_s + \boldsymbol{p}'_s$ with position embeddings $\boldsymbol{p}'_s$; causal mask; four heads, width 64; W_o zero-initialized.

CONCAT concatenates the plan history with the state representation. The future plan enters TimeXer and TiDE through their native covariate interfaces, and the other five backbones through a shared output MLP. The other fusion methods use a backbone without action or exogenous inputs. RES, RES-0, FILM-0, and GATE also receive last-value and mean summaries of state and plan history. The additional DECAY encoder computes a learned per-channel exponential moving average and mixes it with the current input through a learned gate, initialized at zero. CONV uses six residual, dilated depthwise convolution layers with GELU activations. Both it and ATTN are causal over the concatenated history and future plan.

C TRAINING AND EVALUATION PROTOCOL

This appendix supports Section 4.2. In all three main experiments, the data split and forecast windows are fixed across methods. The AE entries in Experiment 1 are reused as CONCAT in Experiment 2. Experiment 3 reuses the GATE and FILM-0 entries as its INSTANT baselines, and the diagnostic of Appendix G reuses GATE as its $\lambda = 0$ baseline. Their stored prediction metrics match exactly. These 1,120 shared entries explain the difference between 6,160 grid entries and 5,040 distinct forecasting runs; the counts exclude codec pretraining and rollout evaluation.

C.1 HYPERPARAMETERS

Table 6 lists every hyperparameter of the benchmark, and the mechanism-consistency probe of Sections 4.3 and 5. None was tuned per dataset or per backbone: the forecasting protocol is shared by all 5,040 runs, the codecs are pretrained once per dataset and frozen, and the backbones use the

default configuration of their reference implementation in the Time series Forecasting Benchmark (TFB) codebase with instance normalization enabled.

Table 6: Hyperparameters. Shared by every dataset, backbone, design choice and seed unless a row says otherwise. L, H and d_c are as in Section 3.1; d is the latent width; MAD is the median absolute deviation.

Forecasting protocol (Experiments 1–3, the diagnostic of Appendix G and Section 5, whose added term is in Appendix F)	
Window	context $L = 256$, horizon $H = 16$, stride 80
Optimizer	Adam, learning rate 10^{-3}, no weight decay, batch size 128
Schedule	halve the learning rate after 5 epochs without held-out improvement
Stopping	at most 100 epochs; stop after 20 epochs without held-out improvement; the checkpoint with the lowest held-out loss is kept
Loss	MSE between the decoded forecast and the observed future states; gradients clipped at norm 1
Seeds	training seeds 0–4 (initialization, minibatch order, dropout); the split is fixed at seed 0 with a held-out fraction of 0.2
Backbones	TFB reference implementations with their default configuration; instance normalization on; built-in regularizers kept (e.g. DUET's load balancing)
Prediction space	
AE, VAE	step-wise MLP codec, $d_x \to 64 \to d$ encoder and $d \to 64 \to d_x$ decoder with ReLU, $d = 32$; pretrained on the state channels with reconstruction MSE (plus $10^{-3}\times$KL for the VAE, posterior mean used at forecast time); Adam 10^{-3}, batch 128, at most 100 epochs, patience 15, same windows and split as above; frozen and shared across seeds
JEPA	same codec shape, $d = 32$, trained jointly from scratch; latent prediction loss weight 0.02; target encoder is an EMA of the online encoder with momentum 0.996 ramped to 1 on a cosine schedule; the latent term is excluded from early stopping
Plan fusion	
RES, RES-0, FILM-0, GATE	adapter MLP g with one ReLU hidden layer of width 128; position embedding of width 32; history summary from the last value and the mean of the state and plan histories; output layer zero-initialized for RES-0 and FILM-0
X-ATTN	four heads, width 128, keys and values over the whole plan with position embeddings
CONCAT	native covariate interface of TimeXer and TiDE; for the other backbones an output MLP with one hidden layer of width 256
Plan encoding	
DECAY	per-channel geometric kernel with $\boldsymbol{\alpha} = \sigma(\boldsymbol{\theta})$, $\boldsymbol{\theta}$ and the mixing gate initialized at 0
CONV	six residual depthwise causal blocks, eight taps, dilation 2^{l-1}, GELU, zero-initialized; receptive field 442 steps
ATTN	causal self-attention, four heads, width 64, output projection zero-initialized
Auxiliary loss (diagnostic of Appendix G)	
λ	$\{0.1, 0.3\}$ on the pre-fusion forecast, decoded to observation space; $\lambda = 0$ is the GATE baseline; the auxiliary term is excluded from early stopping
Mechanism-consistency probe (Sections 4.3 and 5)	
Shift	$\lvert\delta\rvert = 0.1$ along the empirical quantiles of the action channel, both directions
Counted windows	the shift changes the plan, and $\lvert\Delta_\theta^{i\to c}\rvert > 10^{-3}\,\mathrm{MAD}_c$ with MAD_c floored at 10^{-3}; a mode condition must hold on at least half of the window's steps

C.2 Result aggregation and uncertainty

Table 3 pools individual run scores before calculating their mean and sample SD. For scores $m_1, \ldots, m_n$, we report $\bar{m} = n^{-1} \sum_i m_i$ and $s = \sqrt{\sum_i (m_i - \bar{m})^2/(n-1)}$, with $n = 35$ for prediction space and fusion and $n = 70$ for encoding. These SDs include differences between backbones and, for encoding, fusion methods. They are distinct from both within-configuration seed SDs and the dataset-level uncertainty estimates below. Mean ranks are computed after averaging seeds within each dataset–backbone configuration, and within each fusion method in the encoding experiment.

Paired effects. We average the five training seeds before comparing methods within each configuration. A relative effect is the geometric mean of the paired MAE ratios over dataset–backbone cells (Experiment 3 also fixes the fusion method), minus one. Its 95% interval is a percentile bootstrap over the eight datasets, which resamples whole datasets with all their cells, so that an interval excluding zero marks an effect that survives a change in the set of datasets. Next to it we count the datasets on which the effect has the same sign.

D Extended Results of the Design-Choice Sweeps

This appendix supports Section 4.3. It reports, for the three sweeps of Table 3, the paired effects with their uncertainty, computed as described in Appendix C.2, and how the effects interact with the backbone.

D.1 Paired effects and uncertainty

Table 7: Additional summaries of the three sweeps. References are observation space, CONCAT, and INSTANT, respectively. Relative MAE and ranks use seed-averaged paired cells. Parameter counts are dataset-averaged fusion-adapter or temporal-encoder sizes, excluding the backbone, codec, and shared embeddings.

Sweep	Method	Mean rank	ΔMAE vs. ref.	Params.
Prediction space	Observation	2.52	–	–
	AE	1.93	-9.87%	–
	VAE	1.95	-7.54%	–
	JEPA	3.61	$+20.21\%$	–
Plan fusion	NONE	4.89	$+5.73\%$	0
	CONCAT	5.70	–	–
	RES	3.43	-11.73%	38k
	RES-0	3.66	-11.62%	38k
	FILM-0	2.91	-12.58%	43k
	GATE	2.52	-12.73%	47k
	X-ATTN	4.89	-7.27%	113k
Plan encoding	INSTANT	2.79	–	0
	DECAY	2.40	-0.81%	26
	CONV	2.06	-2.20%	612
	ATTN	2.74	-1.15%	35.8k

Table 8: Paired effects of the three sweeps. ΔMAE is the geometric-mean change in validation MAE of the first method against the second, with its 95% bootstrap interval over datasets; Datasets counts the datasets on which the first method is better. Negative values favor the first method.

Exp.	Contrast	ΔMAE	95% CI (bootstrap over datasets)	Datasets
1	AE vs. observation	-9.87%	$[-17.71, -3.93]\%$	8/8
1	VAE vs. observation	-7.54%	$[-16.35, +1.96]\%$	7/8
1	JEPA vs. observation	$+20.21\%$	$[+9.28, +37.40]\%$	0/8
1	VAE vs. AE	$+2.59\%$	$[-1.93, +9.75]\%$	3/8
2	GATE vs. CONCAT	-12.73%	$[-19.03, -7.12]\%$	8/8
2	GATE vs. NONE	-17.47%	$[-24.22, -10.90]\%$	8/8
2	FILM-0 vs. CONCAT	-12.58%	$[-19.19, -6.75]\%$	8/8
2	FILM-0 vs. NONE	-17.32%	$[-24.31, -10.36]\%$	8/8
2	GATE vs. FILM-0	-0.17%	$[-0.84, +0.61]\%$	6/8
3	DECAY vs. INSTANT	-0.81%	$[-2.37, +0.44]\%$	4/8
3	CONV vs. INSTANT	-2.20%	$[-4.33, -0.13]\%$	4/8
3	ATTN vs. INSTANT	-1.15%	$[-4.80, +1.99]\%$	3/8

Close pairs. The two pairs that Section 4.3 calls close are close on every column: VAE differs from AE by $+2.59\%$ with an interval of $[-1.93, +9.75]\%$ and is better on three of the eight datasets, and GATE differs from FILM-0 by -0.17% with an interval of $[-0.84, +0.61]\%$.

D.2 INTERACTION WITH THE BACKBONE

The complete grids of Appendix H list the std over the five training seeds of every cell. It is largest for JEPA, whose jointly trained encoder makes some cells vary by more than their mean, and smaller for the output adapters than for CONCAT; the frozen AE and VAE codecs are shared across seeds, so their variation excludes codec retraining.

Table 9: How design choices and backbone rankings interact. Panel (a) ranks design choices within each backbone, then averages over datasets. Panel (b) ranks backbones within each choice. Lower is better; bold marks the best entry in the relevant comparison. An output adapter ranks first for each backbone, but the ordering of backbones changes with fusion. For example, the NONE and CONCAT backbone rankings have Kendall $\tau = 0.05$.

(a) Mean rank of the arms within each backbone (1 = that backbone's best arm)

	Exp. 1: prediction space				Exp. 2: fusion							Exp. 3: encoder			
Backbone	Observation	AE	VAE	JEPA	NONE	CONCAT	RES	RES-0	FILM-0	GATE	X-ATTN	INSTANT	DECAY	CONV	ATTN
TimeXer	2.00	**1.88**	2.50	3.62	4.50	6.12	2.88	3.88	3.25	**2.00**	5.38	2.50	2.31	**2.19**	3.00
TiDE	**1.75**	2.00	2.25	4.00	5.00	5.50	3.50	3.38	**3.00**	3.00	4.62	2.88	2.75	**1.88**	2.50
DUET	2.75	1.75	**1.50**	4.00	5.00	6.88	3.00	3.25	**1.88**	3.12	4.88	2.50	2.44	**1.94**	3.12
PatchTST	3.12	1.88	**1.50**	3.50	3.88	6.62	4.25	3.38	2.88	**2.25**	4.75	3.12	2.25	**1.88**	2.75
TimeKAN	2.25	**1.62**	2.12	4.00	4.88	5.38	3.62	3.12	3.12	**2.75**	5.12	2.75	2.50	**2.06**	2.69
CrossLinear	3.62	2.25	**2.00**	2.12	7.00	3.50	3.12	4.50	2.38	**2.00**	5.50	2.69	2.44	**2.12**	2.75
Amplifier	2.12	2.12	**1.75**	4.00	4.00	5.88	3.62	4.12	3.88	**2.50**	4.00	3.12	**2.12**	2.38	2.38

(b) Mean rank of the backbones within each arm (1 = that arm's best backbone)

	Exp. 1: prediction space				Exp. 2: fusion							Exp. 3: encoder			
Backbone	Observation	AE	VAE	JEPA	NONE	CONCAT	RES	RES-0	FILM-0	GATE	X-ATTN	INSTANT	DECAY	CONV	ATTN
TimeXer	3.00	3.88	5.12	3.38	3.62	3.88	3.12	3.62	3.62	3.50	3.88	3.38	3.38	3.75	3.88
TiDE	**2.00**	2.88	3.50	5.62	3.38	2.88	3.88	3.62	4.12	3.88	2.88	4.25	4.12	3.75	3.50
DUET	6.38	6.50	6.25	6.62	4.50	6.50	5.12	5.38	4.62	5.50	4.44	5.12	5.25	5.50	5.00
PatchTST	4.62	4.25	3.75	3.25	**2.00**	4.25	3.75	3.00	**2.88**	**2.38**	**2.75**	**2.62**	**2.62**	**2.38**	**2.38**
TimeKAN	2.75	3.12	3.12	3.38	3.75	3.12	**2.62**	**2.88**	3.38	3.25	3.12	3.12	3.25	3.12	3.75
CrossLinear	5.50	**2.62**	**2.12**	**1.00**	6.62	**2.62**	5.25	5.25	4.62	4.75	6.25	4.62	4.62	4.88	4.75
Amplifier	3.75	4.75	4.12	4.75	4.12	4.75	4.25	4.25	4.75	4.75	4.69	4.88	4.75	4.62	4.75

In Experiment 3, CONV ranks first or second for each backbone. Its overall gain is similar under GATE and FILM-0. To assess that interaction, we compute each encoder's log MAE ratio to INSTANT under each fusion method, subtract the two ratios, and average over backbones within each

dataset. The two ratios differ by less than 2% on average for DECAY, CONV and ATTN, so the average encoding effect is similar under the two retained fusion methods, while the dataset-specific gains remain heterogeneous.

E MECHANISM CONSISTENCY

This appendix supports Table 2 of Section 4.1 and the mechanism-consistency paragraph of Section 4.3. It states which declared relations enter the metric and why the others are set aside, which windows are counted, and gives the numbers behind Figure 3 together with the score of every relation.

E.1 DECLARED MECHANISMS AND EXCLUSIONS

The six datasets with a continuous action declare twenty-four relations, read from the documented process for the engineered datasets and from the pharmacology of the drug for the clinical ones. Two are set aside before any model is evaluated. The greenhouse assimilation lamp runs at full output for 98.8% of the held-out steps, so the upward quantile shift $\tau_{i,\delta}$ changes the plan in almost no window and the conditioning event of MC nearly never holds. The PleiaData on/off relation is categorical, and $\tau_{i,\delta}$ is defined for continuous channels only; Appendix F probes it by switching the unit on and reports it separately. The remaining twenty-two are listed in Table 11. One of them, the single relation of Wastewater, is set aside after evaluation and is reported in the table but not counted. It states that metal-salt dosing lowers the phosphate concentration in the tank by precipitation, a mechanism that acts through the reaction and the residence of the water in the tank, on a timescale beyond the 32-minute horizon ($H = 16$ steps at 2 min) over which MC averages the response, so the declared sign is not expected to appear within the window the metric scores. Its scores, 0.03 to 0.21 across the design choices, are the lowest in the grid; we report them so that this exclusion, made after the numbers were seen, stays visible. Section 4.3 therefore averages the twenty-one relations of the remaining five datasets.

E.2 COUNTED WINDOWS AND REPRODUCIBILITY CHECK

MC conditions on the perturbation changing the plan and on the forecast moving. Both conditions bite. A channel already at its maximum cannot be shifted up, and a model that routes the plan into a path with no influence on the output produces $|\Delta_\theta^{i\to c}|$ below the threshold in every window; those cells have no defined consistency and are reported as such rather than counted as chance. NONE is the extreme case, since it never receives the plan, but it is not the only one: 85 of the 3150 cells that do receive the plan also fail to move, and every one of them is TiDE under a prediction space that routes the plan through the backbone rather than an adapter.

Every number of Section 4.3 comes from the checkpoints of the three experiments; no model is retrained. Reloading a checkpoint and recomputing its validation MAE reproduces the stored value to 3.6×10^{-6} for every backbone but DUET, whose mixture-of-experts router samples at inference and drifts by at most 1.1×10^{-3}. Within one probed cell the router is held to a single sample across the factual and the perturbed forecast, which is what a counterfactual requires: the same unit, the same history, and the same draw of any internal randomness, with only the plan changed. The factual forecast is verified to be bitwise identical when the plan is passed twice.

E.3 RESULTS PER DESIGN CHOICE AND PER MECHANISM

Table 10 gives the numbers drawn in Figure 3, and Table 11 the score of every declared relation under the three fusion methods that Section 4.3 discusses.

Table 10: Mechanism consistency of the three sweeps, the numbers behind Figure 3. The MC metric of Section 3.1 at $\delta = \pm 0.1$, averaged over both directions and over the mechanisms a dataset declares, as mean $\pm$ std over the 35 backbone–seed runs of one dataset and one option, or 70 for encoding, which pools GATE and FILM-0. Chance is 0.5 and 1 is perfect agreement with the declared mechanism. NONE discards the plan, so its forecast does not move and it is omitted. Bold marks the highest unrounded mean of each block in each column; Mean averages the five datasets. Wastewater is set aside, see Appendix E.1 and Table 11. Higher is better.

	Option	Greenhouse	PleiaData	PREDIST	VitalDB	MIMIC-Cardio	Mean
Space	OBSERVATION	**.61**$_{\pm.13}$	.49$_{\pm.10}$	.97$_{\pm.03}$	.51$_{\pm.10}$	**.35**$_{\pm.14}$	**.59**
	AE	.55$_{\pm.17}$	.51$_{\pm.11}$	.92$_{\pm.15}$	**.55**$_{\pm.10}$	.33$_{\pm.17}$	.57
	VAE	.47$_{\pm.26}$	**.53**$_{\pm.15}$	**.98**$_{\pm.03}$	.47$_{\pm.05}$	.33$_{\pm.08}$	.55
	JEPA	.55$_{\pm.15}$	.44$_{\pm.16}$	.88$_{\pm.13}$	.51$_{\pm.11}$	.33$_{\pm.10}$	.54
Fusion	CONCAT	**.55**$_{\pm.17}$	.51$_{\pm.11}$	.92$_{\pm.15}$	**.55**$_{\pm.10}$	.33$_{\pm.17}$	**.57**
	RES	.49$_{\pm.11}$	.46$_{\pm.10}$	.97$_{\pm.02}$	.35$_{\pm.09}$	**.48**$_{\pm.09}$	.55
	RES-0	.41$_{\pm.15}$	.47$_{\pm.13}$	.97$_{\pm.02}$	.36$_{\pm.11}$	.40$_{\pm.12}$	.52
	FILM-0	.43$_{\pm.14}$	.46$_{\pm.07}$	.98$_{\pm.02}$	.36$_{\pm.11}$	.42$_{\pm.08}$	.53
	GATE	.47$_{\pm.11}$	.49$_{\pm.12}$	.98$_{\pm.02}$	.36$_{\pm.08}$	.41$_{\pm.09}$	.54
	X-ATTN	.42$_{\pm.13}$	**.52**$_{\pm.06}$	**.99**$_{\pm.01}$	.15$_{\pm.19}$	.21$_{\pm.11}$	.46
Encoding	INSTANT	.45$_{\pm.12}$	.47$_{\pm.10}$	.98$_{\pm.02}$	**.36**$_{\pm.10}$	.42$_{\pm.09}$	.54
	DECAY	.47$_{\pm.12}$	**.50**$_{\pm.10}$	.93$_{\pm.03}$	.34$_{\pm.10}$	.42$_{\pm.09}$	.53
	CONV	**.56**$_{\pm.15}$	.50$_{\pm.10}$	.96$_{\pm.02}$	.25$_{\pm.11}$	**.52**$_{\pm.10}$	**.56**
	ATTN	.44$_{\pm.12}$	.49$_{\pm.06}$	**.98**$_{\pm.01}$	.30$_{\pm.14}$	.44$_{\pm.10}$	.53

Table 11: Mechanism consistency per declared relation under the three fusion methods that Section 4.3 discusses. Mean over 35 backbone–seed cells; the four engineered datasets first, then the two clinical ones. Sign is the declared direction. Seed range is the median over backbones of the spread of the five seeds of one cell configuration on that relation.

Dataset	Mechanism	Sign	CONCAT	GATE	FILM-0	seed range
Greenhouse	Tpipe→Tair	+	.83	.97	.99	.03
	CO2dosing→CO2air	+	.48	.08	.17	.29
	VentLee→Tair	−	.61	.64	.36	.77
	VentWind→Tair	−	.28	.19	.22	.66
PleiaData	setpoint_norm@mode=1→indoor_temp	+	.80	.60	.50	.75
	setpoint_norm@mode=2→indoor_temp	−	.21	.37	.42	.57
PREDIST	valve_pos→flow	+	.91	.97	.96	.07
	valve_pos→heat_power	+	.92	.97	.97	.05
	supply_setpoint→s_hc1_supply_temp	+	.92	1.00	1.00	.01
Wastewater	metal_dosing→po4	−	.15	.03	.07	.10
VitalDB	propofol→BIS	−	.63	.24	.34	.31
	propofol→ART_MBP	−	.38	.56	.52	.49
	remifentanil→HR	−	.54	.35	.36	.33
	remifentanil→ART_MBP	−	.65	.27	.23	.26
MIMIC-Cardio	norepinephrine→NBPs	+	.21	.20	.14	.24
	norepinephrine→NBPm	+	.15	.14	.13	.18
	phenylephrine→NBPs	+	.29	.15	.27	.29
	phenylephrine→NBPm	+	.29	.11	.16	.26
	vasopressin→NBPm	+	.34	.62	.53	.42
	dobutamine→HR	+	.55	.86	.92	.33
	dopamine→HR	+	.42	.99	.92	.05
	FiO2→SpO2	+	.26	.22	.30	.41

Two relations on one actuator. Averaging a dataset's relations hides a sharper failure. PleiaData declares opposite signs for the same setpoint channel in the two modes of the heat pump, so a model that has learned the mechanism must get both. Across the cells of that dataset only 21 of 506 do,

and the two relations are anticorrelated across cells at $\rho = -0.86$: a cell that gets one right gets the other wrong, which is the signature of a single mode-independent sign rather than a mode-dependent mechanism. The two greenhouse vents, whose apertures are nearly collinear in the training data, are also rarely both right, 49 of 518 cells, but there the two relations are uncorrelated across cells rather than anticorrelated, so that pair is a weaker instance of the same point.

F Directional Supervision: Details and Extended Results

This appendix supports Section 5. It gives the hyperparameters of the added term and how ρ and the penalized set were fixed, the score of every mechanism under both shift directions, and the prediction error of every dataset including the null control.

F.1 Hyperparameters and the penalized set

Directional supervision adds the term of Equation 2 to the forecasting protocol of Table 6 and changes nothing else: the same AE–GATE–INSTANT model, backbones, seeds, split, windows, optimizer and stopping rule; each mechanism costs one extra pass through G_θ and D per batch. The term uses $\rho = 0.1$, one shift size per batch drawn from $\delta \sim \mathcal{U}(0.05, 0.30)$ and applied upward only, $\varepsilon = 10^{-6}$, and the quantiles of $\tau_{i,\delta}$ estimated on the training actions; it is excluded from early stopping, which uses the forecasting loss alone. The value of ρ follows a rule fixed before the grid was run: the smallest value in $\{0.1, 1, 10\}$ whose consistency on the penalized mechanisms of the training windows reached 0.95 in a Greenhouse pilot with three backbones. All three values reached 1.000, so the rule selects $\rho = 0.1$.

The penalized set holds twelve of the twenty-one mechanisms of Table 2 and was fixed before any supervised model was trained. Mechanisms were assigned by action channel, so that the two targets of the PREDIST valve and the two targets of norepinephrine are on the same side and no held-out mechanism shares its actuator with a penalized one; the single exception is the PleiaData cooling-mode setpoint, which shares its channel with the penalized heating-mode mechanism and differs only in the mode condition, so it tests whether the learned response follows the mode. On the clinical datasets the held-out side was chosen to contain both a mechanism that GATE already gets right, dopamine, and mechanisms it gets wrong, phenylephrine and remifentanil. Wastewater has no penalized mechanism: its single relation is set aside in Appendix E.1, so the added term is identically zero there and the dataset serves as a null control. The PleiaData on/off relation is listed with the penalized set in the code but the term is defined for continuous channels only and is zero for it, so Figure 4 leaves it out of the PleiaData mean. It is probed by switching the unit from off to on instead of by the quantile shift, and Table 12 reports it with Wastewater in a separate block.

F.2 Results per mechanism

Table 12 gives the numbers behind the left panel of Figure 4 and adds the downward shift and the held-out mechanisms. On every penalized mechanism the mean of +DS rounds to 1.000 in both directions, the heating pipe included, which GATE already gets right at 0.973; the lowest single run is 0.991, on the leeward vent. The downward shift, which the term never sees, is as consistent as the upward one. On the held-out mechanisms the picture is mixed and is reported in full. Two stay where GATE left them, the PREDIST supply setpoint and dopamine, both already consistent. Four move by less than 0.08 under either shift and stay well below chance: CO_2 dosing and both phenylephrine relations rise slightly, and remifentanil on heart rate falls slightly. Three move away from the declared sign: the PleiaData cooling-mode setpoint falls from 0.373 to 0.000, remifentanil on mean pressure from 0.284 to 0.156, and vasopressin on mean pressure from 0.531 to 0.182. The cooling-mode case is the conditional mechanism above: the learned heating-mode response extends to the cooling mode, so the model has a single mode-independent response, the failure that Appendix E already found in the unsupervised models, and supervising one mode does not correct it. Remifentanil and vasopressin on mean pressure each share their state channel with a penalized mechanism, propofol and norepinephrine respectively, but not its action channel; we have no evidence for why their scores fall, and note only that the supervision does not constrain them. Directional supervision therefore enforces the signs it is given and should not be expected to transfer to a mechanism it is not given, including one that shares a channel with a penalized mechanism.

Table 12: Mechanism consistency of every declared relation with and without directional supervision. Mean over the 35 backbone–seed runs of GATE and of GATE + directional supervision (+DS) under the upward and the downward shift; Pen. marks the twelve penalized mechanisms, the training shift is upward only. The last column is the lowest of the 35 runs of +DS under the upward shift. The last block holds the two relations that Section 4.1 sets aside; the on/off relation is categorical and has no downward shift.

				$\delta = +0.1$		$\delta = -0.1$		lowest cell
Dataset	Action → State	Sign	Pen.	GATE	+DS	GATE	+DS	+DS, $\delta = +0.1$
Greenhouse	heating-pipe temp. → air temp.	+	✓	0.973	1.000	0.974	1.000	1.000
Greenhouse	leeward vent → air temp.	−	✓	0.653	1.000	0.634	1.000	0.991
Greenhouse	windward vent → air temp.	−	✓	0.196	1.000	0.181	1.000	1.000
Greenhouse	CO_2 dosing → CO_2 level	+		0.097	0.154	0.068	0.126	0.000
PleiaData	setpoint (heating) → indoor temp.	+	✓	0.606	1.000	0.601	1.000	1.000
PleiaData	setpoint (cooling) → indoor temp.	−		0.373	0.000	0.376	0.000	0.000
PREDIST	valve position → flow	+	✓	0.966	1.000	0.965	1.000	0.999
PREDIST	valve position → heat power	+	✓	0.969	1.000	0.970	1.000	0.997
PREDIST	supply setpoint → supply temp.	+		0.999	0.998	0.999	0.997	0.989
VitalDB	propofol → BIS	−	✓	0.218	1.000	0.258	1.000	1.000
VitalDB	propofol → MAP	−	✓	0.574	1.000	0.551	1.000	1.000
VitalDB	remifentanil → heart rate	−		0.370	0.306	0.330	0.270	0.083
VitalDB	remifentanil → MAP	−		0.284	0.156	0.258	0.119	0.004
MIMIC-Cardio	norepinephrine → systolic pressure	+	✓	0.218	1.000	0.188	1.000	1.000
MIMIC-Cardio	norepinephrine → mean pressure	+	✓	0.150	1.000	0.127	1.000	1.000
MIMIC-Cardio	phenylephrine → systolic pressure	+		0.180	0.236	0.127	0.198	0.004
MIMIC-Cardio	phenylephrine → mean pressure	+		0.103	0.128	0.109	0.155	0.000
MIMIC-Cardio	vasopressin → mean pressure	+		0.531	0.182	0.702	0.274	0.003
MIMIC-Cardio	dobutamine → heart rate	+	✓	0.864	1.000	0.852	1.000	1.000
MIMIC-Cardio	dopamine → heart rate	+		0.999	0.978	0.987	0.992	0.230
MIMIC-Cardio	FiO_2 → oxygen saturation	+	✓	0.194	1.000	0.240	1.000	1.000
Set aside in Section 4 (Appendix E)								
Wastewater	metal-salt dosing → phosphate	−		0.031	0.033	0.028	0.029	0.000
PleiaData	unit on/off → energy per step	+		0.579	0.500	—	—	0.000

Table 13: Validation MAE with and without directional supervision. Mean over the 35 backbone–seed runs of each dataset; Pen. is the number of penalized mechanisms. ΔMAE is the median of the paired relative change and the interval is a t-based 90% interval on the paired log ratios, as in Figure 4. The pooled row pairs all 210 runs. Wastewater has no penalized mechanism and is the null control.

Dataset	Pen.	GATE	+DS	ΔMAE (median)	90% interval
Greenhouse	3	0.0632	0.0639	+0.03%	$[-0.24, +2.11]\%$
PleiaData	1	0.0205	0.0198	−1.54%	$[-9.60, +3.34]\%$
PREDIST	2	0.1641	0.1641	+0.35%	$[-0.34, +0.38]\%$
VitalDB	2	0.0313	0.0312	−0.07%	$[-0.70, +0.33]\%$
MIMIC-Cardio	4	0.0543	0.0542	−0.15%	$[-0.26, -0.03]\%$
Wastewater	0	0.1500	0.1498	+0.00%	$[-0.59, +0.34]\%$
Pooled	12	0.0806	0.0805	−0.05%	$[-1.58, +0.63]\%$

F.3 PREDICTION ERROR AND THE NULL CONTROL

Table 13 gives the numbers behind the right panel of Figure 4 and adds Wastewater. On the null control the two arms are the same model trained twice under the same seed, since the added term is identically zero, and their MAE agrees to 0.0002 with a median paired change of 0.00%; this verifies that the training path of the supervised arm matches GATE apart from the term itself. On the five datasets with penalized mechanisms the median change lies between −1.54% and +0.35%, and every interval lies within ±5% except on PleiaData, whose spread across runs is the widest; MIMIC-Cardio shows a small decrease.

Table 14: Auxiliary-loss results by dataset. The first two blocks report 16-step full-model and pre-fusion MAE over seven backbones and five seeds. Rollout error is for steps 49–64, not the cumulative horizon. It uses 273 matched triples: 35 per dataset except VitalDB (28; all TimeXer runs and two TimeKAN seeds missing). Subscripts are SDs over five seed means of available backbones; VitalDB coverage varies by seed. Bold marks the lowest unrounded mean of each row. Relative changes use seed-averaged cells in the first two blocks and matched backbone–seed triples for rollout.

Dataset	$\lambda=0$	$\lambda=0.1$	$\lambda=0.3$
MAE (fused, $H=16$)			
Greenhouse	**0.0632**$_{\pm.0005}$	0.0651$_{\pm.0004}$	0.0658$_{\pm.0009}$
VitalDB	0.0313$_{\pm.0002}$	0.0312$_{\pm.0001}$	**0.0312**$_{\pm.0001}$
CGMacros	**0.0086**$_{\pm.0001}$	0.0095$_{\pm.0001}$	0.0098$_{\pm.0002}$
Shanghai	**0.1165**$_{\pm.0010}$	0.1174$_{\pm.0010}$	0.1177$_{\pm.0007}$
MIMIC-Cardio	**0.0543**$_{\pm.0001}$	0.0545$_{\pm.0001}$	0.0545$_{\pm.0001}$
PleiaData	0.0205$_{\pm.0011}$	**0.0204**$_{\pm.0007}$	0.0212$_{\pm.0005}$
PREDIST	0.1641$_{\pm.0006}$	0.1640$_{\pm.0013}$	**0.1639**$_{\pm.0002}$
Wastewater	0.1500$_{\pm.0011}$	0.1505$_{\pm.0009}$	**0.1499**$_{\pm.0011}$
Δ vs. $\lambda=0$	–	$+1.75\%$	$+2.78\%$
MAE$_{\text{bypass}}$ (pre-fusion, $H=16$)			
Greenhouse	0.1637$_{\pm.0027}$	0.0859$_{\pm.0015}$	**0.0814**$_{\pm.0026}$
VitalDB	0.0961$_{\pm.0002}$	0.0430$_{\pm.0003}$	**0.0411**$_{\pm.0003}$
CGMacros	0.1397$_{\pm.0062}$	0.0253$_{\pm.0030}$	**0.0203**$_{\pm.0024}$
Shanghai	0.1852$_{\pm.0022}$	0.1435$_{\pm.0042}$	**0.1322**$_{\pm.0013}$
MIMIC-Cardio	0.1147$_{\pm.0003}$	0.0626$_{\pm.0002}$	**0.0614**$_{\pm.0001}$
PleiaData	0.1735$_{\pm.0095}$	0.0361$_{\pm.0010}$	**0.0325**$_{\pm.0007}$
PREDIST	0.2333$_{\pm.0009}$	0.1953$_{\pm.0008}$	**0.1906**$_{\pm.0011}$
Wastewater	0.2832$_{\pm.0044}$	0.2122$_{\pm.0018}$	**0.2047**$_{\pm.0014}$
Δ vs. $\lambda=0$	–	-44.73%	-47.83%
Rollout MAE (steps 49–64)			
Greenhouse	**0.1160**$_{\pm.0013}$	0.1195$_{\pm.0017}$	0.1204$_{\pm.0031}$
VitalDB	**0.0587**$_{\pm.0005}$	0.0592$_{\pm.0005}$	0.0590$_{\pm.0004}$
CGMacros	**0.0733**$_{\pm.0003}$	0.0755$_{\pm.0009}$	0.0756$_{\pm.0011}$
Shanghai	**0.1260**$_{\pm.0006}$	0.1278$_{\pm.0015}$	0.1295$_{\pm.0009}$
MIMIC-Cardio	**0.0730**$_{\pm.0002}$	0.0732$_{\pm.0001}$	0.0735$_{\pm.0002}$
PleiaData	**0.1203**$_{\pm.0015}$	0.1209$_{\pm.0007}$	0.1240$_{\pm.0006}$
PREDIST	**0.1989**$_{\pm.0013}$	0.1998$_{\pm.0021}$	0.1997$_{\pm.0012}$
Wastewater	**0.1939**$_{\pm.0022}$	0.1953$_{\pm.0017}$	0.1951$_{\pm.0013}$
Δ vs. $\lambda=0$	–	$+1.19\%$	$+1.71\%$

G Auxiliary Prediction Loss and Rollout Diagnostic

This appendix reports the diagnostic that Section 4.2 mentions. As a separate diagnostic, we fix AE, GATE, and INSTANT and add supervision to the pre-fusion state forecast:

$$\mathcal{L}_{\text{prediction}} = \text{MSE}(\hat{y}, y) + \lambda\,\text{MSE}\big(D(z_{\text{pre}}), y\big), \qquad \lambda \in \{0, 0.1, 0.3\}. \tag{3}$$

Here z_{pre} is the backbone forecast before fusion, D is the frozen decoder, and $\hat{y}$ is the full model's output. Both terms are in observation space; backbone-specific training regularizers are retained. This is an additional objective comparison, not a fourth architectural sweep.

Equation 3 adds observation-space supervision to the decoded pre-fusion forecast. The validation criterion used for checkpoint selection excludes this added term, retaining the full-model prediction criterion. The bypass metric decodes the pre-fusion forecast alone for a single 16-step block. It is not a forecast of a no-action intervention: its training target is still the observed outcome under the recorded plan. At $\lambda = 0.1$, the bypass MAE improves by 44.73% while the full-model MAE changes by $+1.75\%$. The latter has a dataset-level 90% interval of $[-0.20, +3.74]\%$, within $\pm 5\%$. At $\lambda = 0.3$, full-model MAE changes by $+2.78\%$, with an interval of $[-0.19, +5.84]\%$ that reaches beyond 5%.

Table 15: Sensitivity to backbone inclusion. Each row removes one backbone and recomputes the geometric-mean change relative to $\lambda = 0$. Negative values indicate lower MAE. Removing CrossLinear changes the average full-model effect to approximately zero while retaining a large bypass gain, showing where the observed cost is concentrated.

Backbone dropped	ΔMAE		$\Delta\text{MAE}_{\text{bypass}}$	
	$\lambda=0.1$	$\lambda=0.3$	$\lambda=0.1$	$\lambda=0.3$
none (full set)	$+1.75\%$	$+2.78\%$	-44.73%	-47.83%
TimeXer	$+2.35\%$	$+3.73\%$	-45.24%	-48.53%
TiDE	$+2.16\%$	$+3.37\%$	-45.90%	-49.20%
DUET	$+2.05\%$	$+2.90\%$	-45.85%	-48.87%
PatchTST	$+1.95\%$	$+3.26\%$	-45.70%	-48.63%
TimeKAN	$+1.93\%$	$+3.03\%$	-46.06%	-49.26%
CrossLinear	-0.22%	-0.02%	-37.92%	-40.85%
Amplifier	$+2.05\%$	$+3.26\%$	-45.97%	-48.96%

Rollout feeds the full model's predictions back as state history, while teacher forcing restores observed history; both retain the recorded plan. The first block is identical. At each later block, we subtract the relative change in teacher-forced MAE from the relative change in rollout MAE, both against $\lambda = 0$. Negative differences indicate a rollout-specific gain. Neither weight shows a consistent rollout-specific advantage at the blocks ending at steps 32, 48 and 64; $\lambda = 0$ also has the lowest final-block mean on all eight datasets. These results apply to the available matched runs.

H Complete Validation Grids

Tables 16–19 show every configuration behind Table 3: five-seed means with seed SDs, with the lowest unrounded mean in each row in bold.

Table 16: Experiment 1, prediction space: the complete grid. Validation MAE, mean and SD over five training seeds.

Dataset	Backbone	Observation	AE	VAE	JEPA
Greenhouse	TimeXer	$0.0627_{\pm.0008}$	$\mathbf{0.0604}_{\pm.0006}$	$0.1051_{\pm.0009}$	$0.0882_{\pm.0206}$
	TiDE	$\mathbf{0.0635}_{\pm.0026}$	$0.0658_{\pm.0027}$	$0.1094_{\pm.0007}$	$0.1158_{\pm.0152}$
	DUET	$0.0920_{\pm.0084}$	$\mathbf{0.0709}_{\pm.0016}$	$0.0962_{\pm.0134}$	$0.1453_{\pm.0261}$
	PatchTST	$0.0986_{\pm.0334}$	$0.0924_{\pm.0334}$	$\mathbf{0.0921}_{\pm.0025}$	$0.1019_{\pm.0152}$
	TimeKAN	$\mathbf{0.0620}_{\pm.0007}$	$0.0642_{\pm.0011}$	$0.0770_{\pm.0022}$	$0.0894_{\pm.0282}$
	CrossLinear	$0.1147_{\pm.0327}$	$0.1066_{\pm.0024}$	$0.0996_{\pm.0142}$	$\mathbf{0.0827}_{\pm.0065}$
	Amplifier	$0.0814_{\pm.0265}$	$\mathbf{0.0723}_{\pm.0092}$	$0.0898_{\pm.0164}$	$0.1051_{\pm.0196}$
VitalDB	TimeXer	$0.0357_{\pm.0009}$	$\mathbf{0.0326}_{\pm.0014}$	$0.0356_{\pm.0012}$	$0.0339_{\pm.0043}$
	TiDE	$0.0303_{\pm.0002}$	$0.0310_{\pm.0001}$	$\mathbf{0.0303}_{\pm.0002}$	$0.0444_{\pm.0096}$
	DUET	$0.0446_{\pm.0043}$	$\mathbf{0.0384}_{\pm.0013}$	$0.0417_{\pm.0020}$	$0.0631_{\pm.0059}$
	PatchTST	$0.0334_{\pm.0009}$	$\mathbf{0.0318}_{\pm.0004}$	$0.0331_{\pm.0003}$	$0.0338_{\pm.0022}$
	TimeKAN	$0.0310_{\pm.0003}$	$\mathbf{0.0307}_{\pm.0001}$	$0.0309_{\pm.0002}$	$0.0339_{\pm.0031}$
	CrossLinear	$0.0562_{\pm.0005}$	$0.0321_{\pm.0005}$	$\mathbf{0.0310}_{\pm.0002}$	$0.0326_{\pm.0004}$
	Amplifier	$0.0326_{\pm.0004}$	$0.0335_{\pm.0008}$	$\mathbf{0.0323}_{\pm.0005}$	$0.0403_{\pm.0035}$
CGMacros	TimeXer	$0.0152_{\pm.0006}$	$\mathbf{0.0116}_{\pm.0002}$	$0.0119_{\pm.0005}$	$0.0206_{\pm.0115}$
	TiDE	$0.0108_{\pm.0010}$	$\mathbf{0.0093}_{\pm.0002}$	$0.0105_{\pm.0006}$	$0.0538_{\pm.0247}$
	DUET	$0.0233_{\pm.0054}$	$0.0219_{\pm.0058}$	$\mathbf{0.0152}_{\pm.0029}$	$0.0414_{\pm.0161}$
	PatchTST	$0.0142_{\pm.0037}$	$0.0094_{\pm.0005}$	$\mathbf{0.0091}_{\pm.0003}$	$0.0164_{\pm.0058}$
	TimeKAN	$0.0109_{\pm.0012}$	$0.0101_{\pm.0016}$	$\mathbf{0.0095}_{\pm.0012}$	$0.0117_{\pm.0020}$
	CrossLinear	$\mathbf{0.0086}_{\pm.0001}$	$0.0090_{\pm.0001}$	$0.0088_{\pm.0001}$	$0.0088_{\pm.0001}$
	Amplifier	$0.0087_{\pm.0001}$	$0.0084_{\pm.0002}$	$\mathbf{0.0084}_{\pm.0001}$	$0.0366_{\pm.0374}$
Shanghai	TimeXer	$0.1279_{\pm.0024}$	$\mathbf{0.1236}_{\pm.0029}$	$0.1270_{\pm.0025}$	$0.1625_{\pm.0238}$
	TiDE	$0.1220_{\pm.0010}$	$0.1205_{\pm.0025}$	$\mathbf{0.1197}_{\pm.0011}$	$0.1324_{\pm.0038}$
	DUET	$0.1426_{\pm.0153}$	$0.1403_{\pm.0158}$	$\mathbf{0.1263}_{\pm.0037}$	$0.1630_{\pm.0067}$
	PatchTST	$0.1556_{\pm.0041}$	$0.1235_{\pm.0070}$	$\mathbf{0.1172}_{\pm.0018}$	$0.1487_{\pm.0224}$
	TimeKAN	$0.1187_{\pm.0015}$	$0.1201_{\pm.0006}$	$\mathbf{0.1182}_{\pm.0018}$	$0.1300_{\pm.0211}$
	CrossLinear	$0.1330_{\pm.0034}$	$0.1175_{\pm.0046}$	$\mathbf{0.1145}_{\pm.0018}$	$0.1223_{\pm.0084}$
	Amplifier	$0.1310_{\pm.0029}$	$0.1286_{\pm.0031}$	$\mathbf{0.1211}_{\pm.0025}$	$0.1335_{\pm.0057}$
MIMIC-Cardio	TimeXer	$\mathbf{0.0542}_{\pm.0007}$	$0.0543_{\pm.0006}$	$0.0548_{\pm.0003}$	$0.0570_{\pm.0039}$
	TiDE	$\mathbf{0.0534}_{\pm.0001}$	$0.0537_{\pm.0002}$	$0.0541_{\pm.0002}$	$0.0631_{\pm.0041}$
	DUET	$0.0587_{\pm.0019}$	$0.0596_{\pm.0023}$	$\mathbf{0.0586}_{\pm.0039}$	$0.0754_{\pm.0110}$
	PatchTST	$0.0554_{\pm.0004}$	$\mathbf{0.0543}_{\pm.0003}$	$0.0546_{\pm.0005}$	$0.0582_{\pm.0031}$
	TimeKAN	$\mathbf{0.0541}_{\pm.0002}$	$0.0542_{\pm.0002}$	$0.0543_{\pm.0000}$	$0.0608_{\pm.0124}$
	CrossLinear	$0.0618_{\pm.0004}$	$0.0543_{\pm.0002}$	$\mathbf{0.0543}_{\pm.0003}$	$0.0558_{\pm.0012}$
	Amplifier	$0.0557_{\pm.0001}$	$0.0558_{\pm.0002}$	$\mathbf{0.0555}_{\pm.0002}$	$0.0645_{\pm.0106}$
PleiaData	TimeXer	$\mathbf{0.0281}_{\pm.0108}$	$0.0337_{\pm.0099}$	$0.0312_{\pm.0070}$	$0.0564_{\pm.0125}$
	TiDE	$0.0121_{\pm.0021}$	$\mathbf{0.0115}_{\pm.0009}$	$0.0123_{\pm.0007}$	$0.0866_{\pm.0260}$
	DUET	$0.0923_{\pm.0057}$	$0.0795_{\pm.0058}$	$\mathbf{0.0763}_{\pm.0101}$	$0.0936_{\pm.0065}$
	PatchTST	$0.0448_{\pm.0232}$	$0.0318_{\pm.0140}$	$\mathbf{0.0313}_{\pm.0162}$	$0.0663_{\pm.0186}$
	TimeKAN	$0.0643_{\pm.0007}$	$\mathbf{0.0552}_{\pm.0015}$	$0.0580_{\pm.0007}$	$0.0874_{\pm.0329}$
	CrossLinear	$0.0766_{\pm.0023}$	$0.0075_{\pm.0005}$	$0.0081_{\pm.0004}$	$\mathbf{0.0070}_{\pm.0001}$
	Amplifier	$\mathbf{0.0155}_{\pm.0004}$	$0.0166_{\pm.0004}$	$0.0198_{\pm.0009}$	$0.0273_{\pm.0227}$
PREDIST	TimeXer	$\mathbf{0.1737}_{\pm.0057}$	$0.1799_{\pm.0021}$	$0.1799_{\pm.0051}$	$0.1864_{\pm.0038}$
	TiDE	$\mathbf{0.1773}_{\pm.0039}$	$0.1775_{\pm.0006}$	$0.1784_{\pm.0047}$	$0.2593_{\pm.0193}$
	DUET	$0.2013_{\pm.0076}$	$\mathbf{0.1878}_{\pm.0029}$	$0.1909_{\pm.0025}$	$0.2257_{\pm.0063}$
	PatchTST	$0.1884_{\pm.0027}$	$\mathbf{0.1816}_{\pm.0048}$	$0.1853_{\pm.0017}$	$0.1999_{\pm.0073}$
	TimeKAN	$0.1837_{\pm.0028}$	$\mathbf{0.1827}_{\pm.0029}$	$0.1842_{\pm.0017}$	$0.2050_{\pm.0389}$
	CrossLinear	$0.1955_{\pm.0009}$	$\mathbf{0.1774}_{\pm.0024}$	$0.1784_{\pm.0013}$	$0.1864_{\pm.0052}$
	Amplifier	$\mathbf{0.1863}_{\pm.0022}$	$0.1875_{\pm.0019}$	$0.1881_{\pm.0010}$	$0.2377_{\pm.0373}$
Wastewater	TimeXer	$\mathbf{0.1626}_{\pm.0050}$	$0.1681_{\pm.0027}$	$0.1674_{\pm.0065}$	$0.1814_{\pm.0107}$
	TiDE	$\mathbf{0.1657}_{\pm.0076}$	$0.1809_{\pm.0044}$	$0.1806_{\pm.0088}$	$0.2109_{\pm.0347}$
	DUET	$0.1824_{\pm.0237}$	$0.1801_{\pm.0152}$	$\mathbf{0.1722}_{\pm.0061}$	$0.2126_{\pm.0134}$
	PatchTST	$0.1651_{\pm.0015}$	$0.1699_{\pm.0051}$	$0.1615_{\pm.0049}$	$\mathbf{0.1581}_{\pm.0053}$
	TimeKAN	$0.1702_{\pm.0020}$	$\mathbf{0.1574}_{\pm.0050}$	$0.1605_{\pm.0027}$	$0.1801_{\pm.0063}$
	CrossLinear	$0.1754_{\pm.0016}$	$0.1520_{\pm.0026}$	$0.1520_{\pm.0031}$	$\mathbf{0.1478}_{\pm.0028}$
	Amplifier	$0.1811_{\pm.0027}$	$0.1783_{\pm.0018}$	$\mathbf{0.1762}_{\pm.0041}$	$0.1861_{\pm.0067}$

Table 17: Experiment 2, plan fusion: the complete grid. The CONCAT entries reuse the AE results from Table 16.

Dataset	Backbone	NONE	CONCAT	RES	RES-0	FILM-0	GATE	X-ATTN
Greenhouse	TimeXer	**0.0587**$_{\pm.0006}$	0.0604$_{\pm.0006}$	0.0625$_{\pm.0019}$	0.0621$_{\pm.0013}$	0.0617$_{\pm.0008}$	0.0608$_{\pm.0012}$	0.0651$_{\pm.0021}$
	TiDE	**0.0609**$_{\pm.0005}$	0.0658$_{\pm.0027}$	0.0636$_{\pm.0017}$	0.0635$_{\pm.0020}$	0.0629$_{\pm.0012}$	0.0615$_{\pm.0006}$	0.0641$_{\pm.0011}$
	DUET	0.0657$_{\pm.0006}$	0.0709$_{\pm.0016}$	0.0670$_{\pm.0013}$	0.0678$_{\pm.0013}$	**0.0642**$_{\pm.0006}$	0.0649$_{\pm.0017}$	0.0694$_{\pm.0014}$
	PatchTST	**0.0563**$_{\pm.0008}$	0.0924$_{\pm.0334}$	0.0606$_{\pm.0013}$	0.0612$_{\pm.0013}$	0.0605$_{\pm.0026}$	0.0590$_{\pm.0014}$	0.0694$_{\pm.0039}$
	TimeKAN	**0.0587**$_{\pm.0004}$	0.0642$_{\pm.0011}$	0.0656$_{\pm.0020}$	0.0650$_{\pm.0022}$	0.0645$_{\pm.0014}$	0.0636$_{\pm.0017}$	0.0659$_{\pm.0026}$
	CrossLinear	0.1626$_{\pm.0056}$	0.1066$_{\pm.0024}$	0.0731$_{\pm.0009}$	0.0732$_{\pm.0016}$	0.0721$_{\pm.0039}$	**0.0703**$_{\pm.0019}$	0.1310$_{\pm.0060}$
	Amplifier	0.0627$_{\pm.0006}$	0.0723$_{\pm.0092}$	0.0631$_{\pm.0008}$	0.0638$_{\pm.0007}$	0.0636$_{\pm.0010}$	**0.0625**$_{\pm.0008}$	0.0697$_{\pm.0039}$
VitalDB	TimeXer	0.0326$_{\pm.0013}$	0.0326$_{\pm.0014}$	0.0310$_{\pm.0001}$	0.0311$_{\pm.0002}$	0.0311$_{\pm.0004}$	**0.0308**$_{\pm.0002}$	0.0328$_{\pm.0007}$
	TiDE	0.0312$_{\pm.0002}$	0.0310$_{\pm.0001}$	**0.0308**$_{\pm.0002}$	0.0309$_{\pm.0003}$	0.0309$_{\pm.0003}$	0.0310$_{\pm.0001}$	0.0317$_{\pm.0004}$
	DUET	0.0342$_{\pm.0016}$	0.0384$_{\pm.0013}$	0.0312$_{\pm.0001}$	0.0312$_{\pm.0001}$	0.0312$_{\pm.0001}$	**0.0311**$_{\pm.0001}$	0.0323$_{\pm.0001}$
	PatchTST	0.0316$_{\pm.0003}$	0.0318$_{\pm.0004}$	0.0310$_{\pm.0003}$	0.0309$_{\pm.0002}$	0.0309$_{\pm.0003}$	**0.0307**$_{\pm.0002}$	0.0322$_{\pm.0005}$
	TimeKAN	0.0308$_{\pm.0002}$	0.0307$_{\pm.0001}$	0.0308$_{\pm.0001}$	0.0308$_{\pm.0002}$	0.0306$_{\pm.0001}$	**0.0304**$_{\pm.0001}$	0.0312$_{\pm.0002}$
	CrossLinear	0.0846$_{\pm.0005}$	**0.0321**$_{\pm.0005}$	0.0328$_{\pm.0005}$	0.0331$_{\pm.0006}$	0.0329$_{\pm.0006}$	0.0329$_{\pm.0008}$	0.0335$_{\pm.0006}$
	Amplifier	0.0321$_{\pm.0002}$	0.0335$_{\pm.0008}$	0.0324$_{\pm.0005}$	0.0324$_{\pm.0002}$	0.0323$_{\pm.0004}$	**0.0319**$_{\pm.0004}$	0.0323$_{\pm.0004}$
CGMacros	TimeXer	0.0109$_{\pm.0007}$	0.0116$_{\pm.0002}$	**0.0085**$_{\pm.0001}$	0.0086$_{\pm.0002}$	0.0086$_{\pm.0002}$	0.0085$_{\pm.0002}$	0.0108$_{\pm.0006}$
	TiDE	0.0095$_{\pm.0004}$	0.0093$_{\pm.0002}$	0.0093$_{\pm.0002}$	0.0093$_{\pm.0000}$	**0.0092**$_{\pm.0002}$	0.0092$_{\pm.0002}$	0.0104$_{\pm.0002}$
	DUET	0.0116$_{\pm.0007}$	0.0219$_{\pm.0058}$	**0.0085**$_{\pm.0001}$	0.0086$_{\pm.0001}$	0.0087$_{\pm.0003}$	0.0088$_{\pm.0002}$	0.0118$_{\pm.0005}$
	PatchTST	0.0092$_{\pm.0002}$	0.0094$_{\pm.0005}$	0.0086$_{\pm.0002}$	0.0085$_{\pm.0003}$	0.0084$_{\pm.0001}$	**0.0084**$_{\pm.0002}$	0.0098$_{\pm.0002}$
	TimeKAN	0.0099$_{\pm.0013}$	0.0101$_{\pm.0016}$	**0.0085**$_{\pm.0004}$	0.0090$_{\pm.0005}$	0.0094$_{\pm.0005}$	0.0087$_{\pm.0008}$	0.0105$_{\pm.0009}$
	CrossLinear	0.0624$_{\pm.0011}$	0.0090$_{\pm.0001}$	0.0085$_{\pm.0001}$	0.0085$_{\pm.0001}$	0.0084$_{\pm.0001}$	**0.0084**$_{\pm.0000}$	0.0121$_{\pm.0009}$
	Amplifier	0.0084$_{\pm.0001}$	0.0084$_{\pm.0002}$	**0.0084**$_{\pm.0001}$	0.0085$_{\pm.0001}$	0.0085$_{\pm.0001}$	0.0085$_{\pm.0000}$	0.0095$_{\pm.0003}$
Shanghai	TimeXer	0.1204$_{\pm.0013}$	0.1236$_{\pm.0029}$	0.1197$_{\pm.0011}$	0.1199$_{\pm.0021}$	0.1195$_{\pm.0015}$	**0.1174**$_{\pm.0030}$	0.1211$_{\pm.0018}$
	TiDE	0.1175$_{\pm.0012}$	0.1205$_{\pm.0025}$	0.1151$_{\pm.0004}$	0.1144$_{\pm.0008}$	0.1148$_{\pm.0007}$	**0.1142**$_{\pm.0031}$	0.1199$_{\pm.0008}$
	DUET	0.1260$_{\pm.0023}$	0.1403$_{\pm.0158}$	0.1220$_{\pm.0021}$	**0.1201**$_{\pm.0023}$	0.1214$_{\pm.0013}$	0.1221$_{\pm.0008}$	0.1250$_{\pm.0023}$
	PatchTST	**0.1164**$_{\pm.0012}$	0.1235$_{\pm.0070}$	0.1172$_{\pm.0022}$	0.1171$_{\pm.0021}$	0.1164$_{\pm.0010}$	0.1169$_{\pm.0023}$	0.1179$_{\pm.0022}$
	TimeKAN	0.1208$_{\pm.0017}$	0.1201$_{\pm.0006}$	0.1148$_{\pm.0017}$	**0.1138**$_{\pm.0015}$	0.1142$_{\pm.0026}$	0.1141$_{\pm.0014}$	0.1268$_{\pm.0021}$
	CrossLinear	0.1623$_{\pm.0008}$	0.1175$_{\pm.0046}$	**0.1149**$_{\pm.0013}$	0.1156$_{\pm.0005}$	0.1153$_{\pm.0009}$	0.1155$_{\pm.0014}$	0.1292$_{\pm.0069}$
	Amplifier	0.1162$_{\pm.0005}$	0.1286$_{\pm.0031}$	0.1151$_{\pm.0017}$	0.1155$_{\pm.0012}$	0.1169$_{\pm.0009}$	**0.1151**$_{\pm.0012}$	0.1254$_{\pm.0038}$
MIMIC-Cardio	TimeXer	**0.0535**$_{\pm.0001}$	0.0543$_{\pm.0006}$	0.0538$_{\pm.0001}$	0.0539$_{\pm.0001}$	0.0538$_{\pm.0001}$	0.0535$_{\pm.0001}$	0.0542$_{\pm.0001}$
	TiDE	0.0540$_{\pm.0003}$	**0.0537**$_{\pm.0002}$	0.0542$_{\pm.0001}$	0.0543$_{\pm.0004}$	0.0544$_{\pm.0003}$	0.0543$_{\pm.0002}$	0.0542$_{\pm.0002}$
	DUET	**0.0540**$_{\pm.0003}$	0.0596$_{\pm.0023}$	0.0542$_{\pm.0001}$	0.0543$_{\pm.0001}$	0.0542$_{\pm.0001}$	0.0541$_{\pm.0001}$	0.0548$_{\pm.0002}$
	PatchTST	**0.0533**$_{\pm.0002}$	0.0543$_{\pm.0003}$	0.0541$_{\pm.0003}$	0.0540$_{\pm.0003}$	0.0541$_{\pm.0001}$	0.0539$_{\pm.0003}$	0.0541$_{\pm.0002}$
	TimeKAN	**0.0534**$_{\pm.0001}$	0.0542$_{\pm.0002}$	0.0538$_{\pm.0002}$	0.0537$_{\pm.0001}$	0.0539$_{\pm.0003}$	0.0539$_{\pm.0001}$	0.0540$_{\pm.0002}$
	CrossLinear	0.1029$_{\pm.0004}$	**0.0543**$_{\pm.0002}$	0.0557$_{\pm.0005}$	0.0558$_{\pm.0003}$	0.0558$_{\pm.0007}$	0.0553$_{\pm.0001}$	0.0573$_{\pm.0007}$
	Amplifier	**0.0546**$_{\pm.0002}$	0.0558$_{\pm.0002}$	0.0551$_{\pm.0001}$	0.0550$_{\pm.0001}$	0.0551$_{\pm.0002}$	0.0551$_{\pm.0002}$	0.0550$_{\pm.0001}$
PleiaData	TimeXer	0.0284$_{\pm.0010}$	0.0337$_{\pm.0099}$	0.0263$_{\pm.0008}$	0.0270$_{\pm.0019}$	**0.0251**$_{\pm.0009}$	0.0251$_{\pm.0068}$	0.0294$_{\pm.0046}$
	TiDE	0.0116$_{\pm.0013}$	0.0115$_{\pm.0009}$	0.0105$_{\pm.0003}$	0.0108$_{\pm.0008}$	**0.0103**$_{\pm.0001}$	0.0110$_{\pm.0009}$	0.0104$_{\pm.0007}$
	DUET	0.0250$_{\pm.0030}$	0.0795$_{\pm.0058}$	0.0221$_{\pm.0031}$	0.0227$_{\pm.0030}$	**0.0220**$_{\pm.0028}$	0.0230$_{\pm.0051}$	0.0236$_{\pm.0032}$
	PatchTST	0.0226$_{\pm.0053}$	0.0318$_{\pm.0140}$	0.0238$_{\pm.0030}$	0.0207$_{\pm.0011}$	0.0203$_{\pm.0011}$	0.0219$_{\pm.0025}$	**0.0196**$_{\pm.0037}$
	TimeKAN	0.0685$_{\pm.0006}$	0.0552$_{\pm.0015}$	0.0435$_{\pm.0016}$	0.0434$_{\pm.0019}$	0.0411$_{\pm.0012}$	0.0392$_{\pm.0019}$	**0.0364**$_{\pm.0021}$
	CrossLinear	0.0254$_{\pm.0014}$	0.0075$_{\pm.0005}$	0.0076$_{\pm.0002}$	0.0076$_{\pm.0001}$	**0.0070**$_{\pm.0002}$	0.0071$_{\pm.0002}$	0.0074$_{\pm.0002}$
	Amplifier	0.0229$_{\pm.0003}$	0.0166$_{\pm.0004}$	0.0182$_{\pm.0004}$	0.0187$_{\pm.0008}$	0.0160$_{\pm.0004}$	0.0159$_{\pm.0006}$	**0.0154**$_{\pm.0002}$
PREDIST	TimeXer	0.1756$_{\pm.0014}$	0.1799$_{\pm.0021}$	0.1633$_{\pm.0018}$	0.1630$_{\pm.0018}$	0.1650$_{\pm.0030}$	0.1646$_{\pm.0005}$	**0.1533**$_{\pm.0013}$
	TiDE	0.1773$_{\pm.0012}$	0.1775$_{\pm.0006}$	0.1649$_{\pm.0015}$	0.1641$_{\pm.0009}$	0.1663$_{\pm.0011}$	0.1636$_{\pm.0018}$	**0.1545**$_{\pm.0009}$
	DUET	0.1752$_{\pm.0006}$	0.1878$_{\pm.0029}$	0.1651$_{\pm.0013}$	0.1658$_{\pm.0013}$	0.1642$_{\pm.0027}$	0.1654$_{\pm.0007}$	**0.1522**$_{\pm.0016}$
	PatchTST	0.1758$_{\pm.0010}$	0.1816$_{\pm.0048}$	0.1635$_{\pm.0018}$	0.1637$_{\pm.0010}$	0.1643$_{\pm.0027}$	0.1628$_{\pm.0029}$	**0.1530**$_{\pm.0015}$
	TimeKAN	0.1785$_{\pm.0002}$	0.1827$_{\pm.0029}$	0.1625$_{\pm.0005}$	0.1616$_{\pm.0010}$	0.1618$_{\pm.0009}$	0.1632$_{\pm.0006}$	**0.1510**$_{\pm.0026}$
	CrossLinear	0.2384$_{\pm.0015}$	0.1774$_{\pm.0024}$	0.1657$_{\pm.0023}$	0.1664$_{\pm.0025}$	0.1644$_{\pm.0024}$	**0.1632**$_{\pm.0015}$	0.1759$_{\pm.0026}$
	Amplifier	0.1799$_{\pm.0004}$	0.1875$_{\pm.0019}$	0.1656$_{\pm.0012}$	0.1652$_{\pm.0011}$	0.1654$_{\pm.0010}$	0.1655$_{\pm.0020}$	**0.1568**$_{\pm.0009}$
Wastewater	TimeXer	0.1772$_{\pm.0013}$	0.1681$_{\pm.0027}$	0.1434$_{\pm.0018}$	0.1449$_{\pm.0028}$	0.1436$_{\pm.0014}$	**0.1417**$_{\pm.0028}$	0.1474$_{\pm.0052}$
	TiDE	0.1991$_{\pm.0034}$	0.1809$_{\pm.0044}$	0.1549$_{\pm.0029}$	0.1570$_{\pm.0061}$	**0.1534**$_{\pm.0019}$	0.1541$_{\pm.0029}$	0.1680$_{\pm.0046}$
	DUET	0.1890$_{\pm.0025}$	0.1801$_{\pm.0152}$	0.1506$_{\pm.0031}$	0.1503$_{\pm.0021}$	**0.1485**$_{\pm.0010}$	0.1549$_{\pm.0035}$	0.1657$_{\pm.0062}$
	PatchTST	0.1875$_{\pm.0027}$	0.1699$_{\pm.0051}$	0.1451$_{\pm.0031}$	0.1447$_{\pm.0028}$	**0.1440**$_{\pm.0012}$	0.1447$_{\pm.0016}$	0.1515$_{\pm.0060}$
	TimeKAN	0.1942$_{\pm.0023}$	0.1574$_{\pm.0050}$	**0.1430**$_{\pm.0020}$	0.1440$_{\pm.0011}$	0.1437$_{\pm.0014}$	0.1440$_{\pm.0029}$	0.1493$_{\pm.0049}$
	CrossLinear	0.2005$_{\pm.0015}$	**0.1520**$_{\pm.0026}$	0.1575$_{\pm.0025}$	0.1580$_{\pm.0004}$	0.1566$_{\pm.0034}$	0.1556$_{\pm.0029}$	0.1767$_{\pm.0031}$
	Amplifier	0.2019$_{\pm.0005}$	0.1783$_{\pm.0018}$	0.1572$_{\pm.0016}$	0.1574$_{\pm.0026}$	**0.1527**$_{\pm.0022}$	0.1553$_{\pm.0036}$	0.1732$_{\pm.0046}$

Table 18: Experiment 3, plan encoding with GATE: the complete grid. The INSTANT entries reuse the GATE results from Table 17.

Dataset	Backbone	INSTANT	DECAY	CONV	ATTN
Greenhouse	TimeXer	**0.0608**$_{\pm.0012}$	0.0610$_{\pm.0017}$	0.0611$_{\pm.0018}$	0.0608$_{\pm.0010}$
	TiDE	0.0615$_{\pm.0006}$	0.0614$_{\pm.0007}$	0.0616$_{\pm.0009}$	**0.0614**$_{\pm.0011}$
	DUET	0.0649$_{\pm.0017}$	0.0657$_{\pm.0011}$	0.0656$_{\pm.0017}$	**0.0646**$_{\pm.0009}$
	PatchTST	0.0590$_{\pm.0014}$	**0.0586**$_{\pm.0016}$	0.0596$_{\pm.0009}$	0.0589$_{\pm.0018}$
	TimeKAN	0.0636$_{\pm.0017}$	0.0637$_{\pm.0016}$	0.0635$_{\pm.0018}$	**0.0634**$_{\pm.0015}$
	CrossLinear	**0.0703**$_{\pm.0019}$	0.0706$_{\pm.0028}$	0.0705$_{\pm.0013}$	0.0713$_{\pm.0007}$
	Amplifier	**0.0625**$_{\pm.0008}$	0.0626$_{\pm.0007}$	0.0629$_{\pm.0008}$	0.0628$_{\pm.0014}$
VitalDB	TimeXer	**0.0308**$_{\pm.0002}$	0.0310$_{\pm.0002}$	0.0309$_{\pm.0002}$	0.0315$_{\pm.0003}$
	TiDE	0.0310$_{\pm.0001}$	**0.0307**$_{\pm.0004}$	0.0309$_{\pm.0003}$	0.0309$_{\pm.0002}$
	DUET	**0.0311**$_{\pm.0001}$	0.0311$_{\pm.0001}$	0.0312$_{\pm.0002}$	0.0318$_{\pm.0005}$
	PatchTST	0.0307$_{\pm.0002}$	**0.0306**$_{\pm.0001}$	0.0307$_{\pm.0002}$	0.0311$_{\pm.0003}$
	TimeKAN	**0.0304**$_{\pm.0001}$	0.0305$_{\pm.0002}$	0.0305$_{\pm.0002}$	0.0307$_{\pm.0002}$
	CrossLinear	0.0329$_{\pm.0008}$	0.0332$_{\pm.0004}$	0.0334$_{\pm.0012}$	**0.0325**$_{\pm.0009}$
	Amplifier	0.0319$_{\pm.0004}$	**0.0318**$_{\pm.0005}$	0.0318$_{\pm.0004}$	0.0318$_{\pm.0005}$
CGMacros	TimeXer	0.0085$_{\pm.0002}$	0.0085$_{\pm.0002}$	**0.0084**$_{\pm.0002}$	0.0092$_{\pm.0002}$
	TiDE	**0.0092**$_{\pm.0002}$	0.0093$_{\pm.0004}$	0.0092$_{\pm.0002}$	0.0097$_{\pm.0006}$
	DUET	0.0088$_{\pm.0002}$	0.0088$_{\pm.0001}$	**0.0087**$_{\pm.0001}$	0.0091$_{\pm.0002}$
	PatchTST	**0.0084**$_{\pm.0002}$	0.0084$_{\pm.0002}$	0.0084$_{\pm.0001}$	0.0090$_{\pm.0003}$
	TimeKAN	**0.0087**$_{\pm.0008}$	0.0090$_{\pm.0010}$	0.0088$_{\pm.0009}$	0.0097$_{\pm.0007}$
	CrossLinear	0.0084$_{\pm.0000}$	**0.0084**$_{\pm.0001}$	0.0084$_{\pm.0002}$	0.0086$_{\pm.0004}$
	Amplifier	**0.0085**$_{\pm.0000}$	0.0085$_{\pm.0001}$	0.0086$_{\pm.0002}$	0.0093$_{\pm.0003}$
Shanghai	TimeXer	**0.1174**$_{\pm.0030}$	0.1196$_{\pm.0025}$	0.1185$_{\pm.0031}$	0.1185$_{\pm.0008}$
	TiDE	**0.1142**$_{\pm.0031}$	0.1147$_{\pm.0029}$	0.1147$_{\pm.0032}$	0.1157$_{\pm.0012}$
	DUET	0.1221$_{\pm.0008}$	0.1217$_{\pm.0018}$	**0.1211**$_{\pm.0010}$	0.1235$_{\pm.0019}$
	PatchTST	0.1169$_{\pm.0023}$	0.1177$_{\pm.0028}$	0.1167$_{\pm.0027}$	**0.1165**$_{\pm.0021}$
	TimeKAN	0.1141$_{\pm.0014}$	0.1139$_{\pm.0010}$	**0.1138**$_{\pm.0012}$	0.1187$_{\pm.0027}$
	CrossLinear	0.1155$_{\pm.0014}$	**0.1153**$_{\pm.0015}$	0.1156$_{\pm.0013}$	0.1153$_{\pm.0013}$
	Amplifier	0.1151$_{\pm.0012}$	0.1147$_{\pm.0009}$	0.1147$_{\pm.0021}$	**0.1142**$_{\pm.0009}$
MIMIC-Cardio	TimeXer	**0.0535**$_{\pm.0001}$	0.0536$_{\pm.0001}$	0.0538$_{\pm.0002}$	0.0539$_{\pm.0001}$
	TiDE	0.0543$_{\pm.0002}$	0.0543$_{\pm.0003}$	**0.0541**$_{\pm.0004}$	0.0542$_{\pm.0003}$
	DUET	0.0541$_{\pm.0001}$	**0.0540**$_{\pm.0001}$	0.0547$_{\pm.0005}$	0.0546$_{\pm.0002}$
	PatchTST	0.0539$_{\pm.0003}$	**0.0536**$_{\pm.0002}$	0.0537$_{\pm.0001}$	0.0539$_{\pm.0002}$
	TimeKAN	0.0539$_{\pm.0001}$	**0.0537**$_{\pm.0002}$	0.0540$_{\pm.0000}$	0.0539$_{\pm.0001}$
	CrossLinear	0.0553$_{\pm.0001}$	0.0552$_{\pm.0002}$	**0.0551**$_{\pm.0002}$	0.0553$_{\pm.0004}$
	Amplifier	0.0551$_{\pm.0002}$	**0.0550**$_{\pm.0002}$	0.0550$_{\pm.0002}$	0.0551$_{\pm.0002}$
PleiaData	TimeXer	0.0251$_{\pm.0068}$	0.0290$_{\pm.0013}$	**0.0217**$_{\pm.0074}$	0.0217$_{\pm.0052}$
	TiDE	0.0110$_{\pm.0009}$	0.0110$_{\pm.0009}$	0.0109$_{\pm.0008}$	**0.0104**$_{\pm.0003}$
	DUET	0.0230$_{\pm.0051}$	0.0241$_{\pm.0020}$	**0.0206**$_{\pm.0020}$	0.0215$_{\pm.0035}$
	PatchTST	0.0219$_{\pm.0025}$	0.0222$_{\pm.0013}$	0.0215$_{\pm.0021}$	**0.0183**$_{\pm.0012}$
	TimeKAN	0.0392$_{\pm.0019}$	0.0392$_{\pm.0012}$	0.0329$_{\pm.0018}$	**0.0253**$_{\pm.0005}$
	CrossLinear	0.0071$_{\pm.0002}$	0.0070$_{\pm.0000}$	0.0067$_{\pm.0000}$	**0.0067**$_{\pm.0001}$
	Amplifier	0.0159$_{\pm.0006}$	0.0156$_{\pm.0003}$	0.0156$_{\pm.0004}$	**0.0145**$_{\pm.0003}$
PREDIST	TimeXer	0.1646$_{\pm.0005}$	0.1584$_{\pm.0012}$	**0.1579**$_{\pm.0004}$	0.1587$_{\pm.0021}$
	TiDE	0.1636$_{\pm.0018}$	0.1577$_{\pm.0011}$	**0.1568**$_{\pm.0011}$	0.1586$_{\pm.0047}$
	DUET	0.1654$_{\pm.0007}$	0.1615$_{\pm.0027}$	**0.1583**$_{\pm.0003}$	0.1592$_{\pm.0015}$
	PatchTST	0.1628$_{\pm.0029}$	0.1581$_{\pm.0016}$	0.1562$_{\pm.0016}$	**0.1560**$_{\pm.0013}$
	TimeKAN	0.1632$_{\pm.0006}$	0.1583$_{\pm.0016}$	**0.1548**$_{\pm.0010}$	0.1576$_{\pm.0024}$
	CrossLinear	0.1632$_{\pm.0015}$	0.1590$_{\pm.0015}$	**0.1563**$_{\pm.0017}$	0.1613$_{\pm.0025}$
	Amplifier	0.1655$_{\pm.0020}$	0.1604$_{\pm.0015}$	**0.1573**$_{\pm.0009}$	0.1604$_{\pm.0025}$
Wastewater	TimeXer	0.1417$_{\pm.0028}$	**0.1321**$_{\pm.0029}$	0.1333$_{\pm.0022}$	0.1438$_{\pm.0014}$
	TiDE	0.1541$_{\pm.0029}$	0.1528$_{\pm.0038}$	**0.1414**$_{\pm.0032}$	0.1515$_{\pm.0031}$
	DUET	0.1549$_{\pm.0035}$	0.1424$_{\pm.0045}$	**0.1401**$_{\pm.0033}$	0.1455$_{\pm.0018}$
	PatchTST	0.1447$_{\pm.0016}$	0.1345$_{\pm.0024}$	**0.1339**$_{\pm.0020}$	0.1414$_{\pm.0059}$
	TimeKAN	0.1440$_{\pm.0029}$	0.1331$_{\pm.0008}$	**0.1321**$_{\pm.0015}$	0.1412$_{\pm.0036}$
	CrossLinear	0.1556$_{\pm.0029}$	0.1546$_{\pm.0027}$	**0.1501**$_{\pm.0030}$	0.1566$_{\pm.0033}$
	Amplifier	0.1553$_{\pm.0036}$	0.1495$_{\pm.0025}$	**0.1416**$_{\pm.0022}$	0.1520$_{\pm.0021}$

Table 19: Experiment 3, plan encoding with FiLM-0: the complete grid. The INSTANT entries reuse the FiLM-0 results from Table 17; Table 18 gives the same grid with GATE.

Dataset	Backbone	INSTANT	DECAY	CONV	ATTN
Greenhouse	TimeXer	$0.0617_{\pm.0008}$	$\mathbf{0.0613}_{\pm.0007}$	$0.0623_{\pm.0013}$	$0.0618_{\pm.0011}$
	TiDE	$0.0629_{\pm.0012}$	$0.0629_{\pm.0012}$	$0.0631_{\pm.0013}$	$\mathbf{0.0627}_{\pm.0015}$
	DUET	$\mathbf{0.0642}_{\pm.0006}$	$0.0642_{\pm.0014}$	$0.0649_{\pm.0005}$	$0.0655_{\pm.0014}$
	PatchTST	$0.0605_{\pm.0026}$	$\mathbf{0.0596}_{\pm.0009}$	$0.0609_{\pm.0024}$	$0.0598_{\pm.0010}$
	TimeKAN	$0.0645_{\pm.0014}$	$0.0654_{\pm.0026}$	$0.0648_{\pm.0015}$	$\mathbf{0.0637}_{\pm.0013}$
	CrossLinear	$0.0721_{\pm.0039}$	$0.0724_{\pm.0038}$	$0.0722_{\pm.0033}$	$\mathbf{0.0720}_{\pm.0029}$
	Amplifier	$0.0636_{\pm.0010}$	$\mathbf{0.0636}_{\pm.0008}$	$0.0643_{\pm.0010}$	$0.0641_{\pm.0006}$
VitalDB	TimeXer	$0.0311_{\pm.0004}$	$\mathbf{0.0309}_{\pm.0001}$	$0.0309_{\pm.0002}$	$0.0316_{\pm.0001}$
	TiDE	$\mathbf{0.0309}_{\pm.0003}$	$0.0309_{\pm.0002}$	$0.0309_{\pm.0001}$	$0.0309_{\pm.0001}$
	DUET	$\mathbf{0.0312}_{\pm.0001}$	$0.0313_{\pm.0001}$	$0.0312_{\pm.0001}$	$0.0316_{\pm.0004}$
	PatchTST	$0.0309_{\pm.0003}$	$0.0309_{\pm.0001}$	$\mathbf{0.0309}_{\pm.0001}$	$0.0315_{\pm.0003}$
	TimeKAN	$\mathbf{0.0306}_{\pm.0001}$	$0.0306_{\pm.0002}$	$0.0307_{\pm.0001}$	$0.0307_{\pm.0002}$
	CrossLinear	$0.0329_{\pm.0006}$	$0.0327_{\pm.0007}$	$\mathbf{0.0327}_{\pm.0006}$	$0.0328_{\pm.0004}$
	Amplifier	$0.0323_{\pm.0004}$	$\mathbf{0.0322}_{\pm.0004}$	$0.0323_{\pm.0004}$	$0.0323_{\pm.0004}$
CGMacros	TimeXer	$0.0086_{\pm.0002}$	$\mathbf{0.0085}_{\pm.0003}$	$0.0086_{\pm.0004}$	$0.0091_{\pm.0002}$
	TiDE	$\mathbf{0.0092}_{\pm.0002}$	$0.0092_{\pm.0002}$	$0.0092_{\pm.0001}$	$0.0097_{\pm.0002}$
	DUET	$0.0087_{\pm.0003}$	$0.0087_{\pm.0003}$	$\mathbf{0.0087}_{\pm.0001}$	$0.0090_{\pm.0002}$
	PatchTST	$0.0084_{\pm.0001}$	$0.0084_{\pm.0001}$	$\mathbf{0.0084}_{\pm.0002}$	$0.0090_{\pm.0002}$
	TimeKAN	$0.0094_{\pm.0005}$	$\mathbf{0.0091}_{\pm.0006}$	$0.0093_{\pm.0005}$	$0.0104_{\pm.0002}$
	CrossLinear	$\mathbf{0.0084}_{\pm.0001}$	$0.0084_{\pm.0001}$	$0.0085_{\pm.0001}$	$0.0088_{\pm.0002}$
	Amplifier	$\mathbf{0.0085}_{\pm.0001}$	$0.0085_{\pm.0002}$	$0.0086_{\pm.0002}$	$0.0092_{\pm.0003}$
Shanghai	TimeXer	$0.1195_{\pm.0015}$	$0.1188_{\pm.0020}$	$\mathbf{0.1182}_{\pm.0019}$	$0.1186_{\pm.0020}$
	TiDE	$0.1148_{\pm.0007}$	$0.1151_{\pm.0006}$	$\mathbf{0.1142}_{\pm.0011}$	$0.1154_{\pm.0005}$
	DUET	$0.1214_{\pm.0013}$	$0.1215_{\pm.0015}$	$\mathbf{0.1205}_{\pm.0010}$	$0.1238_{\pm.0018}$
	PatchTST	$0.1164_{\pm.0010}$	$0.1162_{\pm.0012}$	$\mathbf{0.1157}_{\pm.0020}$	$0.1162_{\pm.0016}$
	TimeKAN	$0.1142_{\pm.0026}$	$0.1137_{\pm.0025}$	$\mathbf{0.1127}_{\pm.0020}$	$0.1161_{\pm.0019}$
	CrossLinear	$0.1153_{\pm.0009}$	$0.1155_{\pm.0014}$	$0.1154_{\pm.0015}$	$\mathbf{0.1149}_{\pm.0020}$
	Amplifier	$0.1169_{\pm.0009}$	$0.1166_{\pm.0010}$	$\mathbf{0.1161}_{\pm.0010}$	$0.1161_{\pm.0008}$
MIMIC-Cardio	TimeXer	$\mathbf{0.0538}_{\pm.0001}$	$0.0539_{\pm.0001}$	$0.0540_{\pm.0001}$	$0.0540_{\pm.0002}$
	TiDE	$0.0544_{\pm.0003}$	$0.0543_{\pm.0003}$	$\mathbf{0.0542}_{\pm.0002}$	$0.0544_{\pm.0003}$
	DUET	$\mathbf{0.0542}_{\pm.0001}$	$0.0543_{\pm.0002}$	$0.0544_{\pm.0002}$	$0.0550_{\pm.0003}$
	PatchTST	$0.0541_{\pm.0001}$	$\mathbf{0.0539}_{\pm.0002}$	$0.0540_{\pm.0002}$	$0.0544_{\pm.0004}$
	TimeKAN	$\mathbf{0.0539}_{\pm.0003}$	$0.0541_{\pm.0003}$	$0.0541_{\pm.0002}$	$0.0540_{\pm.0001}$
	CrossLinear	$0.0558_{\pm.0007}$	$0.0559_{\pm.0007}$	$\mathbf{0.0554}_{\pm.0004}$	$0.0557_{\pm.0004}$
	Amplifier	$0.0551_{\pm.0002}$	$0.0551_{\pm.0002}$	$0.0551_{\pm.0001}$	$\mathbf{0.0550}_{\pm.0004}$
PleiaData	TimeXer	$0.0251_{\pm.0009}$	$0.0261_{\pm.0024}$	$0.0240_{\pm.0005}$	$\mathbf{0.0239}_{\pm.0010}$
	TiDE	$0.0103_{\pm.0001}$	$0.0104_{\pm.0005}$	$0.0099_{\pm.0004}$	$\mathbf{0.0099}_{\pm.0002}$
	DUET	$0.0220_{\pm.0028}$	$0.0202_{\pm.0029}$	$0.0208_{\pm.0034}$	$\mathbf{0.0187}_{\pm.0038}$
	PatchTST	$0.0203_{\pm.0011}$	$0.0224_{\pm.0020}$	$\mathbf{0.0196}_{\pm.0016}$	$0.0216_{\pm.0029}$
	TimeKAN	$0.0411_{\pm.0012}$	$0.0403_{\pm.0015}$	$0.0350_{\pm.0013}$	$\mathbf{0.0265}_{\pm.0017}$
	CrossLinear	$\mathbf{0.0070}_{\pm.0002}$	$0.0071_{\pm.0003}$	$0.0071_{\pm.0002}$	$0.0072_{\pm.0001}$
	Amplifier	$0.0160_{\pm.0004}$	$0.0161_{\pm.0007}$	$0.0158_{\pm.0002}$	$\mathbf{0.0145}_{\pm.0003}$
PREDIST	TimeXer	$0.1650_{\pm.0030}$	$0.1608_{\pm.0014}$	$\mathbf{0.1570}_{\pm.0017}$	$0.1588_{\pm.0014}$
	TiDE	$0.1663_{\pm.0011}$	$0.1616_{\pm.0010}$	$\mathbf{0.1565}_{\pm.0012}$	$0.1577_{\pm.0014}$
	DUET	$0.1642_{\pm.0027}$	$0.1623_{\pm.0006}$	$\mathbf{0.1582}_{\pm.0008}$	$0.1595_{\pm.0023}$
	PatchTST	$0.1643_{\pm.0027}$	$0.1596_{\pm.0014}$	$\mathbf{0.1569}_{\pm.0012}$	$0.1576_{\pm.0015}$
	TimeKAN	$0.1618_{\pm.0009}$	$0.1567_{\pm.0016}$	$\mathbf{0.1540}_{\pm.0013}$	$0.1585_{\pm.0040}$
	CrossLinear	$0.1644_{\pm.0024}$	$0.1594_{\pm.0013}$	$\mathbf{0.1584}_{\pm.0010}$	$0.1639_{\pm.0008}$
	Amplifier	$0.1654_{\pm.0010}$	$0.1614_{\pm.0006}$	$\mathbf{0.1592}_{\pm.0016}$	$0.1597_{\pm.0022}$
Wastewater	TimeXer	$0.1436_{\pm.0014}$	$\mathbf{0.1327}_{\pm.0008}$	$0.1335_{\pm.0024}$	$0.1428_{\pm.0041}$
	TiDE	$0.1534_{\pm.0019}$	$0.1470_{\pm.0015}$	$\mathbf{0.1426}_{\pm.0037}$	$0.1480_{\pm.0039}$
	DUET	$0.1485_{\pm.0010}$	$\mathbf{0.1376}_{\pm.0041}$	$0.1399_{\pm.0034}$	$0.1505_{\pm.0040}$
	PatchTST	$0.1440_{\pm.0012}$	$0.1338_{\pm.0020}$	$\mathbf{0.1332}_{\pm.0010}$	$0.1405_{\pm.0046}$
	TimeKAN	$0.1437_{\pm.0014}$	$0.1334_{\pm.0012}$	$\mathbf{0.1333}_{\pm.0008}$	$0.1408_{\pm.0037}$
	CrossLinear	$0.1566_{\pm.0034}$	$0.1550_{\pm.0031}$	$\mathbf{0.1480}_{\pm.0024}$	$0.1557_{\pm.0033}$
	Amplifier	$0.1527_{\pm.0022}$	$0.1459_{\pm.0012}$	$\mathbf{0.1406}_{\pm.0020}$	$0.1486_{\pm.0033}$