CALIBRATION-AWARE TEMPORAL REPRESENTATION LEARNING UNDER FEATURE ESTIMATION NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal models typically treat the representation available at each time step as deterministic, although in practice intermediate features may be estimated under sensor noise, stochastic encoding, missing observations, or limited measurement precision. Unlike static prediction, representation error in a temporal model can alter the hidden state and subsequently propagate through all later transitions. We study this trajectory-level effect and its consequences for probabilistic prediction. Our analysis separates local uncertainty injection from downstream dynamical amplification: finite-perturbation bounds characterise worst-case error propagation, a local Jacobian analysis describes how feature covariance is transformed through the hidden trajectory, and a second-order expansion shows that even zero-mean feature noise can induce systematic hidden-state bias through nonlinear curvature. Motivated by this analysis, we propose Calibration-Aware Temporal Representation Learning (CaTRe), whose core mechanism is an uncertainty-conditioned gate that attenuates unreliable hidden-state updates before their perturbations propagate through time. We evaluate CaTRe on Room Occupancy Estimation and Occupancy Detection with GRU and liquid time-constant (LTC) backbones under clean evaluation, additive representation noise, channel missingness, and temporal thinning. The empirical effect is strongly regime dependent. When GRU baselines are already near saturation or have low probabilistic error, CaTRe produces limited gains and can trade off across metrics. In contrast, the largest improvements occur for LTC settings with high baseline probabilistic error: on Room Occupancy, CaTRe improves clean accuracy from 0.241 to 0.807, reduces NLL from 11.855 to 0.737, and reduces ECE from 0.753 to 0.203. Ablations further show that uncertainty gating is the dominant contributor, while auxiliary probabilistic and trajectory-stability losses are not always complementary. These results support a trajectory-level view of representation uncertainty: its effect depends not only on local noise magnitude, but also on how temporal dynamics transform that uncertainty before prediction.

1 INTRODUCTION

Temporal models typically treat the representation available at each time step as sufficiently accurate to be deterministic. In practice, however, the representation supplied to a recurrent or continuous-time model may itself be estimated under noise, stochastic encoding, missing information, or limited measurement precision. It is therefore useful to distinguish an underlying representation $\boldsymbol{\phi}_t$ from the representation actually available to the temporal model, $\widehat{\boldsymbol{\phi}}_t$.

This distinction is particularly consequential in temporal learning. For a generic residual update,

$$\mathbf{h}_{t+1} = \mathbf{h}_t + \Delta t_t f_\theta \left(\mathbf{h}_t, \widehat{\boldsymbol{\phi}}_t \right), \tag{1}$$

an error in $\widehat{\boldsymbol{\phi}}_t$ changes not only the current update but also the state from which future updates are computed. Its eventual effect therefore depends on both the magnitude of the local representation error and the sensitivity of subsequent hidden-state transitions. Existing work on irregular and continuous-time learning has established the importance of masks, elapsed time, and continuous hidden-state evolution (Che et al., 2018; Neil et al., 2016; Shukla & Marlin, 2021; Chen et al., 2018;

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.

Rubanova et al., 2019; Kidger et al., 2020; Hasani et al., 2021; 2022). Our focus is complementary: we study uncertainty in the *representation that drives the temporal dynamics*.

We model the estimated representation as

$$\widehat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_t + \boldsymbol{\epsilon}_t, \qquad \mathbb{E}[\boldsymbol{\epsilon}_t \mid \boldsymbol{\phi}_t] = \mathbf{0}, \qquad \mathrm{Cov}[\boldsymbol{\epsilon}_t \mid \boldsymbol{\phi}_t] = \boldsymbol{\Sigma}_t. \tag{2}$$

A central observation is that zero-mean local noise need not remain unbiased after nonlinear temporal composition. In general,

$$\mathbb{E}\left[f_\theta(\mathbf{h}_t, \boldsymbol{\phi}_t + \boldsymbol{\epsilon}_t)\right] \neq f_\theta(\mathbf{h}_t, \boldsymbol{\phi}_t),$$

so feature uncertainty can induce both trajectory dispersion and systematic bias. Moreover, the same local covariance can have very different downstream effects depending on where it enters the sequence. We therefore separate temporal feature uncertainty into *local uncertainty injection*, determined by $\boldsymbol{\Sigma}_t$, and *downstream dynamical amplification*, determined by the sensitivity of subsequent transitions.

This trajectory-level view is also relevant to probabilistic reliability. Modern neural networks can be accurate while producing poorly calibrated probabilities (Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021). When uncertain intermediate representations perturb the hidden trajectory, the final predictive distribution may change even without a change in the predicted class. Controlling only the final logits can therefore be insufficient when uncertainty has already propagated through the temporal state.

Motivated by this perspective, we propose *Calibration-Aware Temporal Representation Learning* (CaTRe). An overview of the resulting trajectory-level formulation and the CaTRe mechanism is shown in Figure 1. Its core mechanism is an uncertainty-conditioned temporal update,

$$\mathbf{h}_{t+1} = \mathbf{h}_t + \Delta t_t \mathbf{g}_t \odot f_\theta\left(\mathbf{h}_t, \widehat{\boldsymbol{\phi}}_t\right), \tag{3}$$

where $\mathbf{g}_t \in [0,1]^H$ depends on the hidden state, estimated representation, elapsed time, and representation uncertainty. The gate acts as a learned reliability controller that can attenuate uncertain updates before their perturbations are propagated through later temporal dynamics.

Our analysis characterises this mechanism at multiple levels. We first derive finite-perturbation bounds that separate local uncertainty from downstream temporal amplification. A local Jacobian analysis then describes how feature covariance propagates through the hidden trajectory, while a second-order expansion shows how zero-mean feature noise can induce nonzero trajectory bias through nonlinear curvature. We further analyse the uncertainty-conditioned gate itself, accounting explicitly for both update attenuation and the additional sensitivity introduced by the state-dependent gate.

We evaluate CaTRe on Room Occupancy Estimation and Occupancy Detection using GRU and liquid time-constant (LTC) backbones under clean evaluation, additive representation noise, channel missingness, and temporal thinning. The empirical effect is regime dependent. When the baseline already has low probabilistic error or is near saturation, CaTRe produces limited gains and can trade off across metrics. The largest improvements occur in the evaluated LTC settings with high baseline probabilistic error, and ablations identify uncertainty-conditioned gating as the dominant empirical mechanism. These observations are consistent with the central premise of the paper: the effect of representation uncertainty depends not only on its local magnitude, but also on how temporal dynamics transform it before prediction. Figure 1 summarises the central perspective of the paper: uncertain temporal representations can perturb hidden-state trajectories, these perturbations may be amplified by downstream dynamics, and CaTRe intervenes by conditioning the temporal update on representation uncertainty before the resulting error propagates further.

Our contributions are:

1. We formulate feature-estimation uncertainty as a trajectory-level problem, separating local representation uncertainty from its downstream amplification through temporal dynamics.
2. We develop a multi-level propagation analysis that combines finite perturbation bounds, local covariance propagation, curvature-induced trajectory bias, and stability conditions for uncertainty-conditioned temporal updates.

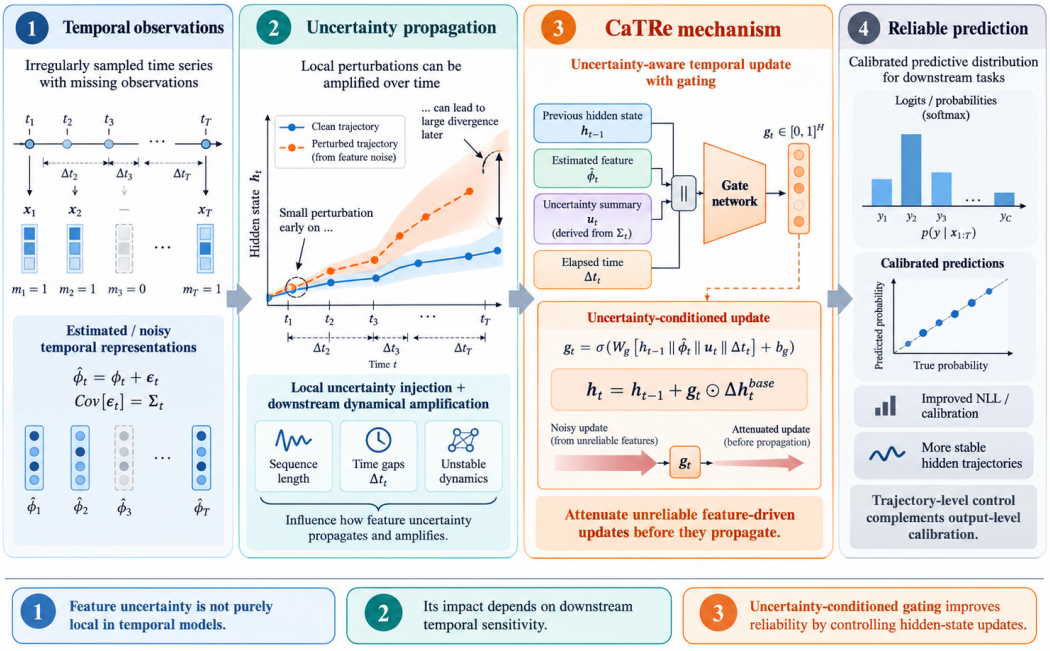

Figure 1: **Overview of Calibration-Aware Temporal Representation Learning (CaTRe).** Temporal observations may be irregular, partially missing, or represented by noisy feature estimates $\hat{\phi}_t = \phi_t + \epsilon_t$ with uncertainty Σ_t. Because the representation is repeatedly incorporated into the hidden trajectory, a local perturbation can influence later states through downstream temporal dynamics. CaTRe conditions a learned reliability gate on the current hidden state, estimated representation, uncertainty summary, and elapsed time, and uses this gate to attenuate unreliable feature-driven state increments before their perturbations propagate. The resulting trajectory is then mapped to the final predictive distribution. The figure illustrates the conceptual mechanism only and does not depict empirical results.

3. We introduce CaTRe and show empirically that its effect is regime dependent, with uncertainty gating providing the strongest gains in the evaluated settings where baseline probabilistic error is large.

2 Related Work

2.1 Irregular and Continuous-Time Temporal Learning

Temporal models for irregularly sampled data explicitly account for missing observations, asynchronous events, and non-uniform time intervals. Missingness-aware recurrent models incorporate masks and elapsed times into recurrent updates, while event-driven and attention-based approaches use timing information to control hidden-state evolution (Che et al., 2018; Neil et al., 2016; Shukla & Marlin, 2021). Continuous-time approaches instead parameterise or control latent trajectories between observations, including Neural ODEs, Latent ODEs, neural controlled differential equations, and liquid or closed-form continuous-time models (Chen et al., 2018; Rubanova et al., 2019; Kidger et al., 2020; Hasani et al., 2021; 2022).

These methods primarily address *when* observations arrive and how hidden states should evolve under irregular sampling. Our focus is complementary: the representation driving the temporal update may itself be an uncertain estimate even when an observation is present. We therefore distinguish uncertainty in the current representation from its subsequent amplification through temporal dynamics.

2.2 CALIBRATION AND UNCERTAIN REPRESENTATIONS

Modern neural networks can achieve high predictive accuracy while producing unreliable probabilities (Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021). Probabilistic reliability is commonly assessed using proper scoring rules such as negative log-likelihood and the Brier score, together with empirical calibration diagnostics such as expected calibration error and reliability diagrams (Brier, 1950; Gneiting & Raftery, 2007; Guo et al., 2017). Post-hoc methods such as temperature scaling modify the final predictive distribution, whereas uncertainty in an intermediate representation can alter the hidden trajectory before the classifier is applied.

Intermediate features may themselves be stochastic or imprecise because of noisy sensing, stochastic encoding, compression, imputation, sampling, or finite measurement. In temporal models, such uncertainty is not restricted to a single predictive mapping: an error at one time step modifies the state from which later updates are computed. Our work therefore studies probabilistic reliability from a trajectory perspective, asking how local representation uncertainty is transformed by recurrent or continuous-time dynamics before reaching the final prediction.

2.3 POSITIONING

Recurrent gating, missingness-aware updates, and continuous-time dynamics already provide mechanisms for regulating temporal information flow (Che et al., 2018; Chen et al., 2018; Rubanova et al., 2019; Kidger et al., 2020; Hasani et al., 2021; 2022). However, the presence of a gate, mask, or elapsed-time variable does not by itself quantify the statistical precision of the representation supplied to the temporal model.

CaTRe explicitly conditions the temporal update on an estimate of representation uncertainty. Its central distinction is between *local uncertainty injection* and *downstream temporal amplification*: the effect of an uncertain representation depends not only on its local variance, but also on how later hidden-state transitions transform that perturbation. Accordingly, CaTRe is neither primarily an output-level calibration method nor a generic recurrent gating architecture; it is designed to control the pathway through which uncertain intermediate representations influence later states and predictive probabilities. Additional discussion of stochastic representations, uncertainty-aware gating, and measurement-budgeted features is provided in Appendix B.

3 METHODOLOGY

Figure 1 provides an overview of the relationship between feature-estimation uncertainty, temporal propagation, and the uncertainty-conditioned update studied in this section. We study temporal representation learning when the representation supplied to a recurrent or continuous-time model is itself estimated under uncertainty. Our analysis separates *local uncertainty injection*, which describes uncertainty in the current representation, from *temporal uncertainty amplification*, which describes how this uncertainty is transformed by subsequent hidden-state dynamics. Detailed derivations and proofs are provided in Appendix C.

3.1 PROBLEM FORMULATION

Let $\boldsymbol{\phi}_t \in \mathbb{R}^K$ denote the underlying representation at time t. We assume that the temporal model instead receives

$$\widehat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_t + \boldsymbol{\epsilon}_t, \qquad \mathbb{E}[\boldsymbol{\epsilon}_t \mid \boldsymbol{\phi}_t] = \mathbf{0}, \qquad \mathrm{Cov}[\boldsymbol{\epsilon}_t \mid \boldsymbol{\phi}_t] = \boldsymbol{\Sigma}_t. \tag{4}$$

The uncertainty $\boldsymbol{\Sigma}_t$ may be known, estimated, or replaced by an uncertainty proxy.

We write the clean temporal transition as

$$\mathbf{h}_t = F_t(\mathbf{h}_{t-1}, \boldsymbol{\phi}_t) := \mathbf{h}_{t-1} + \Delta t_t f_\theta(\mathbf{h}_{t-1}, \boldsymbol{\phi}_t), \tag{5}$$

and define the noisy trajectory analogously using $\widehat{\boldsymbol{\phi}}_t$. The final hidden state is mapped to $\mathbf{p} = \mathrm{softmax}(\mathbf{W}_o \mathbf{h}_T + \mathbf{b}_o)$. We denote the resulting hidden-state perturbation by

$$\boldsymbol{\delta}_t = \widehat{\mathbf{h}}_t - \mathbf{h}_t. \tag{6}$$

Our objective is to characterise its magnitude, covariance, and mean.

3.2 UNCERTAINTY PROPAGATION IN TEMPORAL DYNAMICS

Finite-perturbation propagation. Assume that the temporal vector field is locally Lipschitz in hidden state and representation, with constants $L_{h,t}$ and $L_{\phi,t}$. Then the terminal perturbation satisfies the following bound.

Proposition 1 (finite uncertainty propagation).

$$\|\boldsymbol{\delta}_T\| \leq \sum_{s=1}^{T} \Delta t_s L_{\phi,s} \|\boldsymbol{\epsilon}_s\| \prod_{k=s+1}^{T} (1 + \Delta t_k L_{h,k}) \,. \tag{7}$$

The bound separates the uncertainty injected at time s from its subsequent temporal amplification. Hence, two feature perturbations with the same local magnitude may have very different terminal effects depending on when they occur and on the sensitivity of later transitions.

Local covariance propagation. The worst-case bound does not exploit the covariance structure in equation 4. For a differentiable transition F_t, define the local Jacobians

$$\mathbf{A}_t = \frac{\partial F_t}{\partial \mathbf{h}}, \qquad \mathbf{B}_t = \frac{\partial F_t}{\partial \boldsymbol{\phi}}, \tag{8}$$

evaluated along the clean trajectory. A first-order expansion gives

$$\boldsymbol{\delta}_t \approx \mathbf{A}_t \boldsymbol{\delta}_{t-1} + \mathbf{B}_t \boldsymbol{\epsilon}_t. \tag{9}$$

Under conditionally independent feature-estimation errors, the hidden-state covariance $\mathbf{P}_t = \mathrm{Cov}[\boldsymbol{\delta}_t]$ therefore obeys

$$\mathbf{P}_t \approx \mathbf{A}_t \mathbf{P}_{t-1} \mathbf{A}_t^\top + \mathbf{B}_t \boldsymbol{\Sigma}_t \mathbf{B}_t^\top. \tag{10}$$

Defining $\boldsymbol{\Phi}_{T,s} = \mathbf{A}_T \cdots \mathbf{A}_{s+1}$, the terminal covariance is

$$\mathbf{P}_T \approx \sum_{s=1}^{T} \boldsymbol{\Phi}_{T,s} \mathbf{B}_s \boldsymbol{\Sigma}_s \mathbf{B}_s^\top \boldsymbol{\Phi}_{T,s}^\top. \tag{11}$$

Equation (11) makes the central distinction explicit: the contribution of uncertainty at time s depends jointly on its local covariance $\boldsymbol{\Sigma}_s$, its coupling to the hidden state through $\mathbf{B}_s$, and its downstream amplification through $\boldsymbol{\Phi}_{T,s}$.

Curvature-induced trajectory bias. Variance propagation alone is not sufficient to characterise nonlinear temporal models. Let $\mathbf{b}_t = \mathbb{E}[\boldsymbol{\delta}_t]$. A second-order expansion gives

$$\mathbf{b}_t \approx \mathbf{A}_t \mathbf{b}_{t-1} + \frac{1}{2} \mathcal{H}_{hh,t} : \mathbf{P}_{t-1} + \frac{1}{2} \mathcal{H}_{\phi\phi,t} : \boldsymbol{\Sigma}_t, \tag{12}$$

where $\mathcal{H}_{hh,t}$ and $\mathcal{H}_{\phi\phi,t}$ denote the corresponding Hessian tensors. Consequently,

$$\mathbb{E}[\boldsymbol{\epsilon}_t] = \mathbf{0} \quad \not\Rightarrow \quad \mathbb{E}[\widehat{\mathbf{h}}_T - \mathbf{h}_T] = \mathbf{0}. \tag{13}$$

Thus, nonlinear temporal composition can convert local estimator variance into systematic trajectory bias.

Implications for probabilistic prediction. If the prediction head is locally Lipschitz, then

$$\|\widehat{\mathbf{p}} - \mathbf{p}\|_1 \leq L_p \|\boldsymbol{\delta}_T\|. \tag{14}$$

Trajectory perturbations therefore induce corresponding changes in the final predictive distribution and, consequently, in proper scoring rules such as NLL and the Brier score. Appendix C gives explicit bounds for these quantities. We do not claim an analogous deterministic bound for binned ECE, which is treated as an empirical calibration diagnostic.

3.3 TRAINING OBJECTIVE

We treat uncertainty-conditioned gating as the core CaTRe mechanism and use probabilistic and trajectory-consistency terms as auxiliary regularisers.

For a minibatch of N labelled sequences, the classification loss is

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{n=1}^{N} \log p_{y_n}^{(n)}. \tag{15}$$

We additionally use the Brier score as a proper probabilistic scoring rule,

$$\mathcal{L}_{\mathrm{Brier}} = \frac{1}{N}\sum_{n=1}^{N} \left\| \mathbf{p}^{(n)} - \mathbf{e}_{y_n} \right\|_2^2, \tag{16}$$

where $\mathbf{e}_{y_n}$ is the one-hot target vector.

To encourage the hidden trajectory to remain insensitive to repeated realisations of the same representation uncertainty, we construct two independently perturbed views of the same input sequence and denote their hidden trajectories by $\{\mathbf{h}_t^{(n,a)}\}_{t=1}^{T}$ and $\{\mathbf{h}_t^{(n,b)}\}_{t=1}^{T}$. The trajectory-stability loss is

$$\mathcal{L}_{\mathrm{stab}} = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \left\| \mathbf{h}_t^{(n,a)} - \mathbf{h}_t^{(n,b)} \right\|_2^2. \tag{17}$$

The full objective is

$$\mathcal{L}_{\mathrm{CaTRe}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{Brier}}\mathcal{L}_{\mathrm{Brier}} + \lambda_{\mathrm{stab}}\mathcal{L}_{\mathrm{stab}}. \tag{18}$$

The ablation study separately evaluates the gate, Brier term, and stability term. Throughout the main Base–CaTRe comparisons, "CaTRe" denotes the full variant in Eq. equation 18; "Gate only" is reserved for the corresponding ablation without the two auxiliary regularisers.

4 EXPERIMENTS

Our experiments examine when uncertainty-aware temporal updating is useful rather than assuming uniform improvement across temporal models. We compare CaTRe with standard temporal backbones under clean evaluation, additive representation noise, channel missingness, and temporal thinning. The study evaluates both predictive performance and probabilistic reliability, with particular attention to whether the effect of CaTRe depends on the underlying backbone and corruption regime. Full preprocessing, perturbation, and reporting details are provided in Appendix D.

4.1 EXPERIMENTAL SETUP

We evaluate on two temporal sensor-classification datasets from the UCI Machine Learning Repository: Room Occupancy Estimation and Occupancy Detection (Singh & Chaudhari, 2018; Candanedo, 2016). Room Occupancy Estimation is a multi-class prediction task, whereas Occupancy Detection is binary. Both datasets are chronologically ordered and split into training, validation, and test partitions; normalisation statistics are computed using the training split only.

We consider two temporal backbones: a gated recurrent unit (GRU) and a liquid time-constant network (LTC) (Hasani et al., 2021). For each backbone, we compare a standard **Base** model with its uncertainty-aware **CaTRe** counterpart. Base and CaTRe use the same backbone dimensionality, classifier structure, optimiser, batch size, and training schedule. CaTRe adds the uncertainty-conditioned update; the full variant additionally uses the probabilistic and trajectory-stability regularisers.

We evaluate three forms of test-time degradation:

$$\sigma \in \{0, .05, .10, .20, .40\}, \qquad \rho_c \in \{0, .1, .3, .5\}, \qquad \rho_t \in \{0, .2, .4, .6\}, \tag{19}$$

Table 1: **Clean test performance.** Results are mean ± standard deviation over three seeds. Higher is better for Accuracy and Macro-F1; lower is better for NLL and ECE.

Dataset	Backbone	Method	Accuracy	Macro-F1	NLL	ECE
Occupancy Detection	GRU	Base	**0.993 ± 0.001**	**0.991 ± 0.001**	**0.031 ± 0.005**	**0.011 ± 0.004**
Occupancy Detection	GRU	CaTRe	0.992 ± 0.001	0.989 ± 0.001	0.034 ± 0.012	0.013 ± 0.007
Occupancy Detection	LTC	Base	0.598 ± 0.130	0.482 ± 0.010	5.174 ± 4.170	0.370 ± 0.183
Occupancy Detection	LTC	CaTRe	**0.841 ± 0.258**	**0.827 ± 0.277**	**2.506 ± 4.125**	**0.212 ± 0.213**
Room Occupancy	GRU	Base	**0.915 ± 0.006**	**0.507 ± 0.010**	**0.143 ± 0.010**	0.035 ± 0.003
Room Occupancy	GRU	CaTRe	0.913 ± 0.015	0.502 ± 0.039	0.155 ± 0.021	**0.033 ± 0.010**
Room Occupancy	LTC	Base	0.241 ± 0.068	0.137 ± 0.026	11.855 ± 0.680	0.753 ± 0.058
Room Occupancy	LTC	CaTRe	**0.807 ± 0.131**	**0.309 ± 0.066**	**0.737 ± 0.558**	**0.203 ± 0.126**

where σ denotes additive Gaussian representation noise, ρ_c the channel-missingness rate, and ρ_t the temporal-thinning rate. Gaussian noise directly instantiates the feature-estimation model in equation 4; channel missingness provides structured feature degradation; and temporal thinning changes the density and spacing of available observations. Within each robustness sweep, the selected checkpoint is fixed across corruption levels.

Unless otherwise stated, experiments are repeated with three random seeds and results are reported as mean and standard deviation. Stress-test results with fewer available seeds are explicitly identified and interpreted descriptively.

4.2 Evaluation protocol

Predictive performance is evaluated primarily using accuracy and macro-F1. Probabilistic quality is evaluated using negative log-likelihood (NLL) and the Brier score, together with expected calibration error (ECE) as an empirical calibration diagnostic. Because ECE depends on confidence binning, we interpret it jointly with proper scoring rules rather than in isolation.

To identify the source of any improvement, we additionally compare five ablation variants: **Base**, **Gate only**, **Brier only**, **Stability only**, and **Full CaTRe**. This distinguishes architectural control of the temporal update from improvements caused only by auxiliary training objectives.

Model selection and early stopping use validation NLL, and all reported metrics are computed on the held-out test split. Robustness curves are interpreted as sensitivity profiles of fixed trained models rather than as independently optimised models at each corruption level.

5 Results

We organise the results around a regime-dependent question: when does uncertainty-conditioned temporal updating materially affect predictive and probabilistic behaviour? The experiments do not directly measure the Jacobian, covariance, or curvature quantities in Section 3.2; instead, we examine whether the observed behaviour is qualitatively consistent with the proposed propagation mechanism.

5.1 Clean behaviour depends strongly on the backbone

Table 1 reports clean-test performance.

The GRU results provide little evidence of a uniform CaTRe advantage. On Occupancy Detection, the baseline is already near saturation, achieving 0.993 ± 0.001 accuracy and 0.011 ± 0.004 ECE; CaTRe produces nearly identical predictive performance with slightly worse NLL and ECE. Room Occupancy shows a similarly mixed pattern, with a small reduction in ECE but slightly lower accuracy and higher NLL.

The LTC settings are substantially different. On Occupancy Detection, CaTRe increases accuracy from 0.598 ± 0.130 to 0.841 ± 0.258 while reducing NLL and ECE. On Room Occupancy, accuracy increases from 0.241 ± 0.068 to 0.807 ± 0.131, NLL decreases from 11.855 ± 0.680 to $0.737 \pm$

Table 2: **Robustness at the strongest tested corruption level.** Gaussian noise uses $\sigma = 0.40$, channel missingness $\rho_c = 0.5$, and temporal thinning $\rho_t = 0.6$. Values are CaTRe–Base. Positive ΔAcc and negative ΔNLL/ΔECE are favourable.

Dataset	Backbone	Corruption	ΔAcc	ΔNLL	ΔECE	n
Occupancy Detection	GRU	Gaussian	+0.001	+0.002	+0.004	3
Occupancy Detection	GRU	Channel	+0.026	-0.532	-0.071	3
Occupancy Detection	GRU	Temporal	+0.077	-0.291	-0.045	3
Occupancy Detection	LTC	Gaussian	+0.235	-7.580	-0.312	3
Occupancy Detection	LTC	Channel	+0.182	-6.528	-0.324	3
Occupancy Detection	LTC	Temporal	-0.044	+0.707	+0.044	3
Room Occupancy	GRU	Gaussian	+0.004	+0.024	-0.003	3
Room Occupancy	GRU	Channel	-0.004	-0.015	-0.011	3
Room Occupancy	GRU	Temporal	+0.000	-0.033	+0.001	3
Room Occupancy	LTC	Gaussian	+0.083	-9.154	-0.383	3
Room Occupancy	LTC	Channel	+0.240	-10.295	-0.529	1
Room Occupancy	LTC	Temporal	-0.056	+0.893	+0.056	1

0.558, and ECE decreases from 0.753 ± 0.058 to 0.203 ± 0.126. Thus, the largest observed effects occur in settings with high baseline predictive and probabilistic error, rather than uniformly across backbones.

5.2 Robustness is regime and corruption dependent

Table 2 summarises the CaTRe–Base difference at the strongest tested corruption level for additive noise, channel missingness, and temporal thinning. Full curves over all corruption strengths are provided in Appendix D.8.

Under additive representation noise, the GRU differences remain small, whereas both LTC settings show substantially lower NLL and ECE for CaTRe at $\sigma = 0.40$. Channel missingness produces a similar pattern: CaTRe substantially reduces probabilistic error for Occupancy Detection/LTC, while the available Room Occupancy/LTC result shows the same direction but is based on a single run.

Temporal thinning behaves differently. For Occupancy Detection/LTC, CaTRe has lower accuracy and higher NLL and ECE than the baseline at $\rho_t = 0.6$; the available single-run Room Occupancy/LTC result exhibits the same direction. This distinction is consistent with the scope of the current mechanism: gating can attenuate an unreliable *observed feature-driven update*, whereas temporal thinning changes the observation process itself and may leave no update to attenuate at removed time steps.

5.3 Ablation identifies the uncertainty gate as the primary mechanism

The strongest ablation evidence arises in the LTC settings, where the main performance differences are observed. Table 3 therefore reports the Base, Gate-only, and Full CaTRe variants for LTC; the complete ablation over both backbones and all auxiliary-loss variants appears in Appendix D.9.

In both LTC settings, the Gate-only variant outperforms the full objective on the reported metrics. By contrast, the complete ablation in Appendix D.9 shows that the Brier-only and stability-only variants do not reproduce the same improvement. The strongest empirical evidence therefore supports uncertainty-conditioned control of the temporal update as the principal mechanism, while the auxiliary losses are better viewed as optional regularisers rather than necessary components.

Overall, the experiments support a deliberately limited conclusion. CaTRe is not uniformly beneficial: its effect is small or mixed in several GRU settings, substantially larger in the evaluated

Table 3: **Key LTC ablation.** Results are mean ± standard deviation over three seeds.

Dataset	Variant	Accuracy	NLL	ECE
Occupancy Detection	Base	0.598 ± 0.130	5.174 ± 4.170	0.370 ± 0.183
Occupancy Detection	Gate only	$\mathbf{0.990 \pm 0.002}$	$\mathbf{0.045 \pm 0.011}$	$\mathbf{0.020 \pm 0.010}$
Occupancy Detection	Full CaTRe	0.841 ± 0.258	2.506 ± 4.125	0.212 ± 0.213
Room Occupancy	Base	0.241 ± 0.068	11.855 ± 0.680	0.753 ± 0.058
Room Occupancy	Gate only	$\mathbf{0.921 \pm 0.002}$	$\mathbf{0.245 \pm 0.049}$	$\mathbf{0.067 \pm 0.010}$
Room Occupancy	Full CaTRe	0.891 ± 0.021	0.754 ± 0.586	0.165 ± 0.078

high-error LTC regimes, and weaker under severe temporal thinning. This pattern is qualitatively consistent with the view that representation uncertainty matters through its interaction with temporal dynamics, while not constituting a direct empirical verification of the Jacobian or covariance quantities in the theory.

6 Conclusion

We studied temporal representation learning when the representation available at each time step is an uncertain estimate rather than an exact feature. Our analysis separates *local uncertainty injection* from *downstream temporal amplification*. Finite-perturbation bounds characterise worst-case error propagation, a local Jacobian analysis describes how feature covariance is transformed through the hidden trajectory, and a second-order expansion shows that even zero-mean feature noise can induce systematic trajectory bias through nonlinear curvature. These results motivate CaTRe, whose core mechanism is an uncertainty-conditioned temporal update that attenuates unreliable feature-driven state changes before their perturbations propagate through later dynamics.

The empirical results support a regime-dependent rather than universal benefit. CaTRe produces limited or mixed changes in several GRU settings where baseline predictive and probabilistic error is already low, whereas substantially larger improvements appear in the evaluated LTC settings with high baseline NLL and ECE, particularly under additive representation noise and channel missingness. The ablation study further identifies uncertainty-conditioned gating as the dominant empirical mechanism: the auxiliary Brier and trajectory-stability terms do not reproduce the same gains on their own and are not always complementary to the gate. Performance under temporal thinning is less consistent, highlighting an important distinction between attenuating an unreliable observed update and modelling sparse temporal evolution.

The present study is limited to two related sensor-classification datasets, and the experiments do not directly measure the Jacobian, covariance, or curvature quantities appearing in the theoretical analysis. Some stress-test settings also contain fewer repeated runs and should therefore be interpreted descriptively. Accordingly, the empirical results should be viewed as qualitatively consistent with the proposed propagation mechanism rather than as a direct identification of its theoretical quantities. Overall, our findings suggest that representation uncertainty in temporal models should be considered jointly with the dynamics through which it propagates, motivating reliability-aware control of hidden-state updates when intermediate representations are estimated, noisy, or otherwise uncertain.

AI use statement

In this work, we used generative AI tools for proofreading the manuscript. We have not used generative AI tools for any other tasks. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Luis Candanedo. Occupancy detection, 2016. URL https://doi.org/10.24432/C5X01N. Dataset.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8: 6085, 2018. doi: 10.1038/s41598-018-24271-9. URL https://doi.org/10.1038/s41598-018-24271-9.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7657–7666, 2021. doi: 10.1609/aaai.v35i9.16936. URL https://doi.org/10.1609/aaai.v35i9.16936.

Ramin Hasani, Mathias Lechner, Alexander Amini, Lucas Liebenwein, Aaron Ray, Max Tschaikowski, Gerald Teschl, and Daniela Rus. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, 4:992–1003, 2022. doi: 10.1038/s42256-022-00556-7. URL https://doi.org/10.1038/s42256-022-00556-7.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html.

Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased LSTM: Accelerating recurrent network training for long or event-based sequences. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/5bce843dd76db8c939d5323dd3e54ec9-Abstract.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5320–5330, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

Adarsh Pal Singh and Sachin Chaudhari. Room occupancy estimation, 2018. URL https://doi.org/10.24432/C5P605. Dataset.

APPENDIX

A EXTENDED MOTIVATION AND SCOPE

A.1 SOURCES OF REPRESENTATION UNCERTAINTY

The feature supplied to a temporal model can be uncertain for several distinct reasons. Sensor measurements may be corrupted by acquisition noise; learned encoders may produce stochastic outputs; dropout or sampling-based representations may vary across repeated forward passes; compressed representations may lose information; and intermediate quantities estimated under finite computational or measurement budgets may have non-negligible sampling variance. In each case, the downstream temporal model operates on an estimated representation rather than an exact latent feature.

The formulation

$$\widehat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_t + \boldsymbol{\epsilon}_t \tag{20}$$

provides a common abstraction when the estimation error can reasonably be treated as an additive perturbation. The covariance $\boldsymbol{\Sigma}_t$ may be known analytically, estimated from repeated stochastic evaluations, or approximated by a lower-dimensional uncertainty proxy.

The zero-mean model used in the main theoretical development is not intended to imply that all forms of representation degradation are unbiased. Missing-value imputation, systematic sensor corruption, or distribution shift may introduce nonzero bias. We use the zero-mean setting because it isolates an important phenomenon: even when the local feature estimator is unbiased, nonlinear temporal composition can itself generate trajectory-level bias. Extensions to biased or temporally correlated feature errors are discussed as natural generalisations of the same propagation framework.

A.2 STATIC PERTURBATION VERSUS TEMPORAL PROPAGATION

The distinction between uncertain representations in static and temporal models is structural. In a static classifier, a perturbation in the feature vector affects one predictive mapping. In a temporal model, the perturbed representation modifies the hidden state, and that modified state becomes an input to all later transitions.

Consequently, two feature perturbations with the same local variance need not have comparable sequence-level effects. An early perturbation followed by sensitive hidden-state transitions may produce a large terminal deviation, whereas an equally uncertain feature introduced near the end of a stable trajectory may have little effect. This motivates the distinction used throughout the paper between *local uncertainty injection* and *downstream dynamical amplification*.

This distinction also clarifies why uncertainty magnitude alone is insufficient as a temporal reliability signal. A useful temporal mechanism should account not only for whether a feature is uncertain, but also for how strongly the current update can influence subsequent hidden dynamics.

A.3 RELATION TO PROBABILISTIC CALIBRATION

Our primary theoretical quantities concern hidden-state perturbation, covariance propagation, trajectory bias, and changes in the final predictive distribution. These quantities are related to, but should not be identified with, empirical calibration metrics.

Proper scoring rules such as negative log-likelihood and the Brier score change continuously with the predictive distribution and can therefore be related directly to probability perturbations. By contrast, binned expected calibration error depends on finite-sample bin assignments and is not

globally smooth. We therefore use ECE as an empirical calibration diagnostic rather than claiming a deterministic trajectory-to-ECE bound.

The term "calibration-aware" in CaTRe reflects this broader objective of preserving reliable predictive probabilities under uncertain temporal representations. The core mechanism nevertheless operates before the final classifier: it controls the hidden-state update through which representation uncertainty enters the trajectory.

A.4 Missingness and temporal sparsity

Channel missingness and temporal thinning are included in the empirical study as structured forms of degraded temporal information, but they are not identical to the zero-mean additive-noise model used in the main analysis.

Channel missingness removes selected components of the observed representation. When a missingness mask is available, it provides an explicit signal about which feature components are unreliable or absent, and can therefore be used as an uncertainty proxy for the temporal update.

Temporal thinning is conceptually different. Removing an observation changes not only the reliability of a feature but also the temporal observation process itself. The elapsed time between available observations increases, and in some cases no observation-driven update is available at the removed time step. For this reason, the additive feature-noise theory should not be interpreted as a complete model of temporal thinning. We include thinning as a stress test of whether uncertainty-conditioned temporal updates remain effective when temporal information becomes sparse.

This distinction is also useful for interpreting the empirical results: the current CaTRe mechanism is naturally suited to attenuating unreliable observed updates, whereas sparse temporal evolution may require additional modelling of latent dynamics between observations.

A.5 Measurement-budgeted representations

Finite-measurement representations provide a concrete example in which feature uncertainty is directly controlled by a resource budget. In quantum machine learning, for example, expectation values produced by parameterised quantum circuits are estimated from a finite number of repeated measurements. An estimator based on S measurements has nonzero sampling variance even when it is unbiased, and increasing the measurement budget reduces that variance.

For an observable with expectation $q_{t,j} \in [-1, 1]$, a standard finite-shot estimator satisfies

$$\mathrm{Var}[\hat{q}_{t,j}^{(S)}] = \frac{1 - q_{t,j}^2}{S}. \tag{21}$$

The number of measurements therefore directly determines the precision of the representation supplied to the temporal model.

We use this setting only as a motivating example of measurement-budgeted representation uncertainty. The empirical study in this paper does not include a quantum experiment, and CaTRe does not rely on any quantum-specific structure. The same temporal propagation problem arises whenever intermediate features are estimated from finite samples, stochastic simulations, Monte Carlo procedures, or other resource-limited computations.

B Additional Related Work

B.1 Noisy and stochastic representations

Intermediate representations are often estimated rather than observed exactly. Examples include stochastic encoder outputs, dropout-based representations, noisy sensor embeddings, compressed features, imputed observations, and sampled latent variables. In such settings, the downstream model receives an estimator $\widehat{\phi}_t$ of an underlying representation ϕ_t, with uncertainty that may vary across examples and time steps.

The temporal setting introduces an additional mechanism beyond uncertainty in a single predictive mapping. An error in $\widehat{\phi}_t$ first changes the current hidden state and can then affect all future states that depend on it. Representation uncertainty is therefore transformed by the temporal transition dynamics rather than acting only at the final classifier.

Our formulation makes this effect explicit by associating the estimated representation with a time-dependent uncertainty $\mathbf{\Sigma}_t$ and analysing how this local uncertainty interacts with subsequent dynamical sensitivity. The objective is therefore not only to make individual representations robust, but also to characterise when uncertainty introduced at a particular time step can have a large downstream effect.

B.2 Gating and uncertainty-aware temporal updates

Gating mechanisms are fundamental to recurrent neural architectures because they regulate how much new information enters or modifies the hidden state. Missingness-aware temporal models extend this idea by conditioning recurrent behaviour on observation masks or elapsed time (Che et al., 2018). Continuous-time approaches similarly allow hidden-state evolution to depend explicitly on time and current state (Chen et al., 2018; Rubanova et al., 2019; Kidger et al., 2020; Hasani et al., 2021; 2022).

A binary observation mask and a time-gap variable, however, represent different information from feature-estimation precision. Two features can both be observed at the same temporal spacing while having substantially different uncertainty. Similarly, stochastic representations can have comparable means but different variances.

CaTRe therefore conditions its temporal update explicitly on an estimate of representation uncertainty. The gate is intended to act as a reliability-dependent controller of the pathway through which uncertain features affect the hidden trajectory. This distinction also motivates our theoretical treatment: because the gate depends on the current state and representation, its effect on temporal stability depends both on the amount of attenuation it introduces and on the sensitivity of the gate itself.

B.3 Calibration and trajectory-level reliability

Probabilistic calibration concerns the agreement between predictive confidence and empirical correctness. Modern neural networks can remain poorly calibrated despite high predictive accuracy (Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021). Negative log-likelihood and the Brier score provide proper scoring rules for assessing probabilistic predictions, whereas expected calibration error and reliability diagrams provide empirical summaries of confidence alignment (Brier, 1950; Gneiting & Raftery, 2007; Guo et al., 2017).

Calibration can become more difficult under distribution shift or corrupted inputs because the predictive distribution may change in ways that are not reflected by accuracy alone (Ovadia et al., 2019; Minderer et al., 2021). For temporal models, uncertainty may enter repeatedly throughout a sequence, meaning that the final probability vector can be affected by perturbations that occurred many steps earlier.

Our work therefore differs from approaches that operate only on the final predictive distribution. We study how representation uncertainty is transformed before the prediction head is applied. This motivates analysing trajectory perturbation and proper-scoring degradation while treating empirical calibration metrics such as ECE as complementary evaluation quantities rather than as direct dynamical objects.

B.4 Measurement-budgeted representations

Finite-measurement representations provide a concrete example of feature-estimation uncertainty. In quantum machine learning, parameterised quantum circuits can be used as feature maps or differentiable components of hybrid models (Biamonte et al., 2017; Cerezo et al., 2021; Benedetti et al., 2019; Havlíček et al., 2019; Schuld & Killoran, 2019). Expectation values produced by such circuits are typically estimated from a finite number of measurements rather than observed exactly, so their

estimation variance depends on the available measurement budget (Schuld et al., 2019; Sweke et al., 2020).

This setting illustrates the distinction between feature value and feature reliability. Two estimated representations can have similar expected values while having different statistical precision, and these differences can produce different levels of trajectory perturbation once the representations are incorporated into a temporal model.

We use finite-measurement representations only as a motivating instance of the broader problem considered in this paper. The proposed framework is not specific to quantum models, and the empirical study does not include a quantum experiment. The same reasoning applies whenever temporal models depend on features estimated from finite samples, stochastic simulations, compressed measurements, or other resource-limited procedures.

C Theoretical Details and Proofs

C.1 Full problem setup and assumptions

Consider a temporal sequence

$$\mathcal{S} = \{(\mathbf{x}_t, \mathbf{m}_t, \Delta t_t)\}_{t=1}^{T}, \tag{22}$$

where $\mathbf{x}_t \in \mathbb{R}^d$ denotes the observation, $\mathbf{m}_t$ is an optional observation mask, and $\Delta t_t \geq 0$ denotes the elapsed time since the preceding observation. The goal is to predict $y \in \{1, \ldots, C\}$.

The clean and noisy temporal trajectories are

$$\mathbf{h}_t = F_t(\mathbf{h}_{t-1}, \boldsymbol{\phi}_t) = \mathbf{h}_{t-1} + \Delta t_t f_\theta(\mathbf{h}_{t-1}, \boldsymbol{\phi}_t), \tag{23}$$

$$\widehat{\mathbf{h}}_t = F_t(\widehat{\mathbf{h}}_{t-1}, \widehat{\boldsymbol{\phi}}_t) = \widehat{\mathbf{h}}_{t-1} + \Delta t_t f_\theta(\widehat{\mathbf{h}}_{t-1}, \widehat{\boldsymbol{\phi}}_t), \tag{24}$$

with $\mathbf{h}_0 = \widehat{\mathbf{h}}_0 = \mathbf{0}$.

The residual form includes an explicit Euler discretisation of continuous-time dynamics and can also represent recurrent transitions after expressing the state change as an increment.

The final prediction is

$$\mathbf{o} = \mathbf{W}_o \mathbf{h}_T + \mathbf{b}_o, \qquad \mathbf{p} = \mathrm{softmax}(\mathbf{o}), \tag{25}$$

with analogous quantities for the noisy trajectory.

We assume that feature-estimation noise satisfies

$$\widehat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_t + \boldsymbol{\epsilon}_t, \qquad \mathbb{E}[\boldsymbol{\epsilon}_t \mid \boldsymbol{\phi}_t] = \mathbf{0}, \qquad \mathrm{Cov}[\boldsymbol{\epsilon}_t \mid \boldsymbol{\phi}_t] = \boldsymbol{\Sigma}_t. \tag{26}$$

The zero-mean assumption isolates uncertainty propagation from systematic local estimator bias. The empirical stress tests involving missingness and temporal thinning are broader structured perturbations and need not satisfy this model exactly.

C.2 Proof of the finite-perturbation bound

Assume that, along the relevant trajectories,

$$\|f_\theta(\mathbf{h}, \boldsymbol{\phi}) - f_\theta(\mathbf{h}', \boldsymbol{\phi}')\| \leq L_{h,t}\|\mathbf{h} - \mathbf{h}'\| + L_{\phi,t}\|\boldsymbol{\phi} - \boldsymbol{\phi}'\|. \tag{27}$$

Using $\boldsymbol{\delta}_t = \widehat{\mathbf{h}}_t - \mathbf{h}_t$, we obtain

$$\|\boldsymbol{\delta}_t\| \leq \|\boldsymbol{\delta}_{t-1}\| + \Delta t_t \|f_\theta(\widehat{\mathbf{h}}_{t-1}, \widehat{\boldsymbol{\phi}}_t) - f_\theta(\mathbf{h}_{t-1}, \boldsymbol{\phi}_t)\| \tag{28}$$

$$\leq (1 + \Delta t_t L_{h,t})\, \|\boldsymbol{\delta}_{t-1}\| + \Delta t_t L_{\phi,t} \|\boldsymbol{\epsilon}_t\|. \tag{29}$$

Repeated substitution with $\boldsymbol{\delta}_0 = \mathbf{0}$ gives

$$\|\boldsymbol{\delta}_T\| \leq \sum_{s=1}^{T} \Delta t_s L_{\phi,s} \|\boldsymbol{\epsilon}_s\| \prod_{k=s+1}^{T} (1 + \Delta t_k L_{h,k}), \tag{30}$$

which proves Proposition 1.

The factor

$$\Delta t_s L_{\phi,s} \|\boldsymbol{\epsilon}_s\| \tag{31}$$

describes local uncertainty injection, whereas

$$\prod_{k=s+1}^{T} (1 + \Delta t_k L_{h,k}) \tag{32}$$

describes worst-case downstream amplification.

C.3 First-order covariance propagation

Assume that F_t is twice continuously differentiable in a neighbourhood of the clean trajectory. Define

$$\mathbf{A}_t = \left.\frac{\partial F_t}{\partial \mathbf{h}}\right|_{(\mathbf{h}_{t-1}, \boldsymbol{\phi}_t)} = \mathbf{I} + \Delta t_t \mathbf{J}_{h,t}, \tag{33}$$

and

$$\mathbf{B}_t = \left.\frac{\partial F_t}{\partial \boldsymbol{\phi}}\right|_{(\mathbf{h}_{t-1}, \boldsymbol{\phi}_t)} = \Delta t_t \mathbf{J}_{\phi,t}. \tag{34}$$

Taylor expansion around the clean trajectory yields

$$\boldsymbol{\delta}_t = \mathbf{A}_t \boldsymbol{\delta}_{t-1} + \mathbf{B}_t \boldsymbol{\epsilon}_t + \mathbf{r}_t, \tag{35}$$

where

$$\|\mathbf{r}_t\| = \mathcal{O}\left(\|\boldsymbol{\delta}_{t-1}\|^2 + \|\boldsymbol{\epsilon}_t\|^2\right). \tag{36}$$

Suppose that feature-estimation errors are conditionally independent across time and independent of the preceding first-order hidden-state perturbation. Ignoring higher-order terms gives

$$\mathbf{P}_t \approx \mathbf{A}_t \mathbf{P}_{t-1} \mathbf{A}_t^\top + \mathbf{B}_t \boldsymbol{\Sigma}_t \mathbf{B}_t^\top. \tag{37}$$

Define

$$\boldsymbol{\Phi}_{T,s} = \mathbf{A}_T \mathbf{A}_{T-1} \cdots \mathbf{A}_{s+1}, \qquad \boldsymbol{\Phi}_{T,T} = \mathbf{I}. \tag{38}$$

Then, with $\mathbf{P}_0 = \mathbf{0}$,

$$\mathbf{P}_T \approx \sum_{s=1}^{T} \boldsymbol{\Phi}_{T,s} \mathbf{B}_s \boldsymbol{\Sigma}_s \mathbf{B}_s^\top \boldsymbol{\Phi}_{T,s}^\top. \tag{39}$$

For interpretation, the contribution associated with time s can be summarised as

$$\mathcal{A}_s^{(h)} = \mathrm{Tr}\left(\boldsymbol{\Phi}_{T,s} \mathbf{B}_s \boldsymbol{\Sigma}_s \mathbf{B}_s^\top \boldsymbol{\Phi}_{T,s}^\top\right). \tag{40}$$

This quantity is large either when the local representation is uncertain or when downstream dynamics strongly amplify a modest perturbation.

C.4 Second-order curvature-induced bias

Let

$$\mathbf{b}_t = \mathbb{E}[\boldsymbol{\delta}_t]. \tag{41}$$

A second-order Taylor expansion of F_t around the clean trajectory yields

$$\mathbf{b}_t \approx \mathbf{A}_t \mathbf{b}_{t-1} + \frac{1}{2}\mathcal{H}_{hh,t} : \mathbf{P}_{t-1} + \frac{1}{2}\mathcal{H}_{\phi\phi,t} : \boldsymbol{\Sigma}_t. \tag{42}$$

Here, $\mathcal{H}_{hh,t}$ and $\mathcal{H}_{\phi\phi,t}$ are the output-wise Hessian tensors of F_t. For a tensor $\mathcal{H} \in \mathbb{R}^{H \times d \times d}$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$, the contraction is defined component-wise by

$$(\mathcal{H} : \mathbf{C})_i = \sum_{j,k} \mathcal{H}_{i,j,k} \mathbf{C}_{j,k}. \tag{43}$$

Therefore,

$$\mathbb{E}[\boldsymbol{\epsilon}_t] = \mathbf{0} \not\Rightarrow \mathbb{E}[\widehat{\mathbf{h}}_T - \mathbf{h}_T] = \mathbf{0}. \tag{44}$$

The bias disappears under a globally affine temporal map because the Hessian terms vanish. It is therefore a consequence of nonlinear temporal composition rather than biased local feature estimation.

C.5 Predictive-probability and proper-scoring perturbations

The composition of the affine prediction head and softmax map is locally Lipschitz, so

$$\|\widehat{\mathbf{p}} - \mathbf{p}\|_1 \le L_p \|\boldsymbol{\delta}_T\|. \tag{45}$$

For the Brier score

$$\mathcal{B}(\mathbf{p}, y) = \|\mathbf{p} - \mathbf{e}_y\|_2^2, \tag{46}$$

we have

$$|\mathcal{B}(\widehat{\mathbf{p}}, y) - \mathcal{B}(\mathbf{p}, y)| \le 2\sqrt{2}\|\widehat{\mathbf{p}} - \mathbf{p}\|_2 + \|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2. \tag{47}$$

For negative log-likelihood, suppose

$$p_y \ge \eta, \qquad \widehat{p}_y \ge \eta \tag{48}$$

for some $\eta > 0$. Then the mean-value theorem gives

$$|-\log \widehat{p}_y + \log p_y| \le \frac{1}{\eta}|\widehat{p}_y - p_y| \le \frac{1}{\eta}\|\widehat{\mathbf{p}} - \mathbf{p}\|_1. \tag{49}$$

Together, these results establish the pathway

feature uncertainty $\longrightarrow$ trajectory variance and bias $\longrightarrow$ predictive-probability perturbation $\longrightarrow$ proper-scoring degradation.

Binned expected calibration error does not admit the same direct smooth bound because changes in confidence may alter bin assignments. We therefore use ECE as an empirical calibration diagnostic rather than as a quantity directly controlled by the theoretical inequalities.

D Experimental Details

D.1 Datasets and temporal preprocessing

We evaluate on two temporal sensor-classification datasets from the UCI Machine Learning Repository: Room Occupancy Estimation and Occupancy Detection (Singh & Chaudhari, 2018; Candanedo, 2016). Both datasets are used in their complete available form after preprocessing; no artificial dataset subsampling is introduced.

Room Occupancy Estimation is a multi-class environmental sensing task in which temporal measurements are used to predict room occupancy level. **Occupancy Detection** is a binary environmental sensing task that predicts whether a room is occupied.

For each dataset, observations are first ordered chronologically. Numerical features are normalised using statistics computed exclusively from the training split, and the same transformation is applied to validation and test data. The ordered sequence is converted into fixed-length temporal windows

$$\mathcal{S}^{(n)} = \left\{ (\mathbf{x}_t^{(n)}, \mathbf{m}_t^{(n)}, \Delta t_t^{(n)}) \right\}_{t=1}^{L}, \tag{50}$$

where $\mathbf{x}_t$ denotes the observed sensor vector, $\mathbf{m}_t$ the observation mask, and Δt_t the elapsed temporal interval associated with the current observation.

Training, validation, and test partitions are constructed chronologically to prevent information from future observations entering the training set. Hyperparameter selection and early stopping use only the validation partition, and final metrics are computed on the held-out test partition.

D.2 TEMPORAL BACKBONES AND TRAINING

We use two temporal backbones.

GRU. The gated recurrent unit provides a standard discrete-time recurrent baseline whose internal gating regulates information flow.

LTC. The liquid time-constant network provides a continuous-time-inspired temporal backbone with adaptive hidden dynamics (Hasani et al., 2021).

For each backbone, the Base and CaTRe variants use the same backbone dimensionality, classifier structure, batch size, optimiser, and training schedule. Baseline models are trained using cross-entropy. Full CaTRe uses

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{cal}}\mathcal{L}_{\mathrm{Brier}} + \lambda_{\mathrm{stab}}\mathcal{L}_{\mathrm{stab}}, \tag{51}$$

as defined in Section 3.3.

All models are optimised with AdamW. Early stopping is based on validation negative log-likelihood. Unless otherwise stated, experiments are repeated using three random seeds and reported as mean and standard deviation. Where fewer runs are available, the number of seeds is reported explicitly and the result is treated descriptively.

D.3 EVALUATION-TIME PERTURBATIONS

We evaluate three forms of representation or observation degradation.

Additive representation noise. To instantiate the feature-estimation model directly, we perturb the representation according to

$$\widehat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{52}$$

with

$$\sigma \in \{0.00, 0.05, 0.10, 0.20, 0.40\}. \tag{53}$$

For this perturbation, $\boldsymbol{\Sigma}_t = \sigma^2\mathbf{I}$ is known by construction.

Channel missingness. Sensor channels are randomly masked with missingness rate

$$\rho_c \in \{0.0, 0.1, 0.3, 0.5\}. \tag{54}$$

The observation mask remains available to the temporal model, allowing missing values to be distinguished from observed numerical values. For CaTRe, the mask also acts as a structured uncertainty signal.

Temporal thinning. Intermediate observations are randomly removed with rate

$$\rho_t \in \{0.0, 0.2, 0.4, 0.6\}. \tag{55}$$

The resulting elapsed time Δt_t is supplied to the model.

Temporal thinning differs from feature corruption because it changes the observation process itself rather than merely degrading an observed representation. It is therefore treated as a structured stress test rather than as a direct instance of the additive noise model.

For all robustness experiments, the trained checkpoint is held fixed across perturbation strengths. Differences across σ, ρ_c, and ρ_t therefore measure evaluation-time sensitivity rather than retraining effects.

D.4 EVALUATION METRICS

We distinguish predictive performance from probabilistic reliability.

For predictive performance, we report accuracy and macro-F1 as the primary metrics. Balanced accuracy and weighted-F1 are additionally used where appropriate to characterise class-imbalanced performance.

For probabilistic quality, we report negative log-likelihood (NLL) and Brier score. Both depend on the full predictive distribution and therefore complement the trajectory-level probability perturbation analysis.

Expected calibration error (ECE) and maximum calibration error (MCE) are reported as empirical calibration diagnostics. Given confidence bins $\{B_m\}_{m=1}^{M}$,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \tag{56}$$

Lower NLL, Brier score, ECE, and MCE indicate better performance according to the corresponding probabilistic metric. Because ECE depends on confidence binning and is not a smooth proper scoring rule, we interpret it jointly with NLL, Brier score, and reliability diagrams rather than in isolation.

D.5 Ablation and reporting protocol

We evaluate five variants:

- **Base**: the underlying temporal backbone without CaTRe components;
- **Gate only**: uncertainty-conditioned temporal updating without auxiliary probabilistic or trajectory-stability terms;
- **Brier only**: the Base backbone trained with the additional Brier-score term but without uncertainty-conditioned gating;
- **Stability only**: the Base backbone trained with trajectory-stability regularisation but without the uncertainty gate;
- **Full CaTRe**: uncertainty-conditioned gating together with both auxiliary terms.

The ablation is evaluated for both GRU and LTC backbones on both datasets. Its purpose is to distinguish improvements caused by direct control of the hidden-state update from those caused by probabilistic or consistency regularisation alone.

For comparisons against the corresponding baseline, we define

$$\Delta_{\mathrm{metric}} = \mathrm{metric}_{\mathrm{CaTRe}} - \mathrm{metric}_{\mathrm{Base}}. \tag{57}$$

Positive values indicate improvement for accuracy and F1-based metrics, whereas negative values indicate improvement for NLL, Brier score, ECE, and MCE.

Given the limited number of random seeds, reported means and standard deviations are treated as descriptive measures of variability. Small numerical differences are not interpreted as evidence of a systematic effect in isolation; emphasis is instead placed on qualitative behaviour across corruption levels, backbones, and metrics.

D.6 Additional clean-result visualisations

Figure 2 visualises the relationship between clean accuracy and expected calibration error across datasets, backbones, and methods.

D.7 Complete additive-noise results

We evaluate Gaussian representation noise over

$$\sigma \in \{0, 0.05, 0.10, 0.20, 0.40\}.$$

The main paper reports the strongest corruption level; here we provide the additional visualisations and metric comparisons over the full sweep.

D.8 Complete channel-missingness and temporal-thinning results

Table 5 reports the strongest tested structured perturbations, including macro-F1 and the number of available seeds.

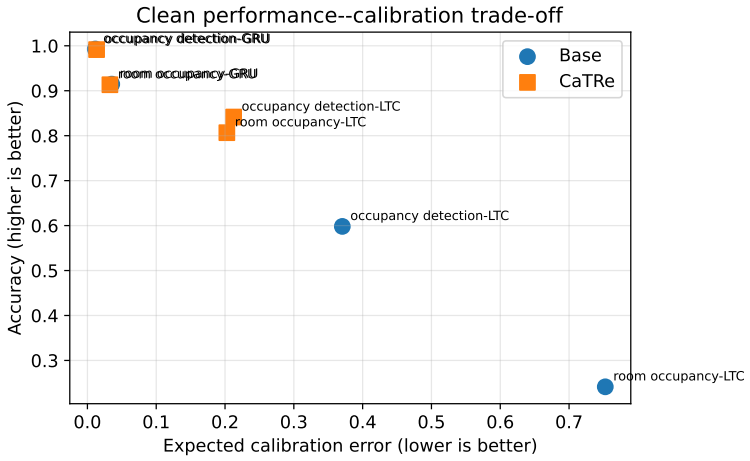

Figure 2: **Clean accuracy–calibration behaviour.** Each point denotes one dataset/backbone/method combination averaged over seeds. The largest separation between Base and CaTRe occurs in the LTC settings with high baseline calibration error, whereas the GRU points remain close.

Table 4: **CaTRe–Base difference under additive feature-estimation noise.** Positive values are favourable for Accuracy and Macro-F1; negative values are favourable for NLL and ECE.

Dataset	Backbone	$\Delta\text{Acc}_{0.0}$	$\Delta\text{F1}_{0.0}$	$\Delta\text{NLL}_{0.0}$	$\Delta\text{ECE}_{0.0}$	$\Delta\text{Acc}_{0.4}$	$\Delta\text{F1}_{0.4}$	$\Delta\text{NLL}_{0.4}$	$\Delta\text{ECE}_{0.4}$
Occupancy Detection	GRU	-0.001	-0.002	+0.004	+0.003	+0.001	+0.001	+0.002	+0.004
Occupancy Detection	LTC	+0.196	-0.029	-6.338	-0.175	+0.235	-0.029	-7.580	-0.312
Room Occupancy	GRU	+0.008	-0.015	+0.023	-0.006	+0.004	-0.021	+0.024	-0.003
Room Occupancy	LTC	-0.104	-0.073	-4.415	-0.246	+0.083	+0.005	-9.154	-0.383

D.9 Complete ablation study

Table 6 reports all ablation variants for both datasets and both temporal backbones.

D.10 Additional observations

The complete results reinforce three observations.

First, GRU behaviour is conservative across most conditions. The Occupancy Detection GRU baseline is already near saturation, and neither clean evaluation nor additive representation noise produces a consistent CaTRe advantage. Room Occupancy with GRU shows small metric-dependent trade-offs rather than uniform improvement.

Second, the largest CaTRe differences occur in the evaluated LTC regimes with high baseline probabilistic error. Under additive noise and channel missingness, these differences are particularly large in NLL and ECE. Because the experiments do not directly estimate local Jacobians or propagated covariance, these results should be interpreted as qualitatively consistent with the theoretical mechanism rather than as direct validation of the propagation equations.

Third, temporal thinning differs from feature corruption. The LTC results under severe thinning are weaker or negative even where channel masking is favourable. This suggests that the current uncertainty gate is better suited to controlling unreliable observed updates than to compensating for missing temporal evolution.

Finally, the complete ablation confirms that the uncertainty gate is the dominant component in the LTC settings. The Brier-only and stability-only variants do not reproduce the same improvements, and the full objective can underperform the gate-only model. We therefore treat the auxiliary objectives as optional regularisers rather than as necessary components of the core CaTRe mechanism.

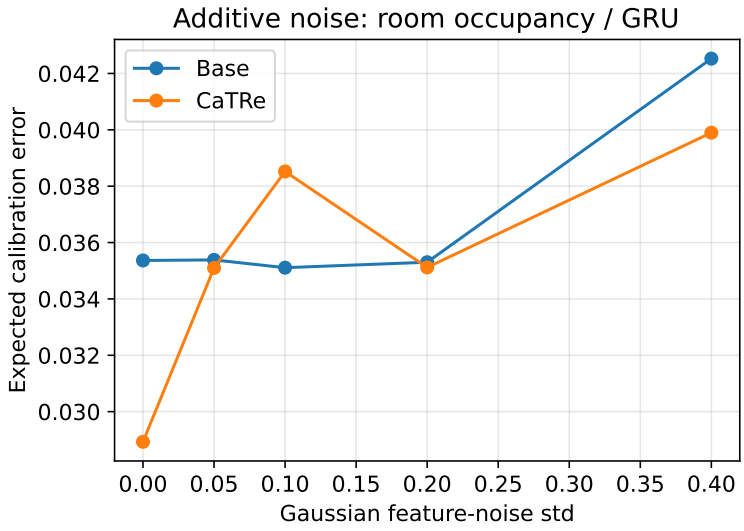

Figure 3: **ECE under additive representation noise.** Expected calibration error on Room Occupancy with the GRU backbone under increasing Gaussian representation noise.

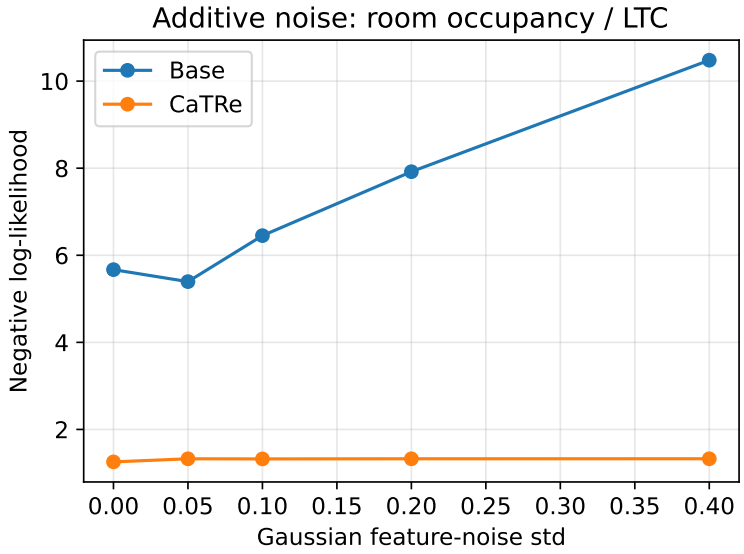

Figure 4: **NLL under additive representation noise.** Negative log-likelihood on Room Occupancy with the LTC backbone under increasing Gaussian representation noise.

Appendix References

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019. doi: 10.1088/2058-9565/ab4eb5. URL https://doi.org/10.1088/2058-9565/ab4eb5.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017. doi: 10.1038/nature23474. URL https://doi.org/10.1038/nature23474.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Table 5: **Structured degradation at the strongest tested level.** Channel masking uses $\rho_c = 0.5$ and temporal thinning uses $\rho_t = 0.6$. Values are CaTRe–Base. Rows with $n = 1$ are descriptive.

Dataset	Backbone	Type	ΔAcc	ΔF1	ΔNLL	ΔECE	n
Occupancy Detection	GRU	Channel	+0.026	+0.069	-0.532	-0.071	3
Occupancy Detection	GRU	Temporal	+0.077	+0.109	-0.291	-0.045	3
Occupancy Detection	LTC	Channel	+0.182	-0.004	-6.528	-0.324	3
Occupancy Detection	LTC	Temporal	-0.044	-0.031	+0.707	+0.044	3
Room Occupancy	GRU	Channel	-0.004	-0.051	-0.015	-0.011	3
Room Occupancy	GRU	Temporal	+0.000	-0.028	-0.033	+0.001	3
Room Occupancy	LTC	Channel	+0.240	+0.042	-10.295	-0.529	1
Room Occupancy	LTC	Temporal	-0.056	-0.047	+0.893	+0.056	1

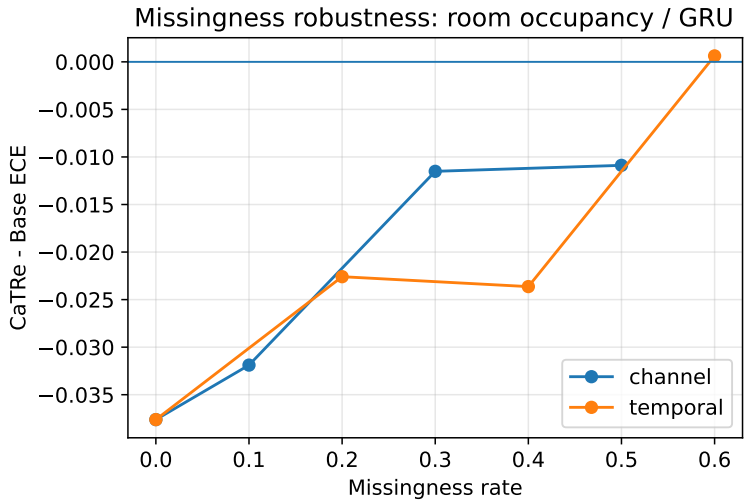

Figure 5: **Structured-degradation calibration behaviour.** CaTRe–Base ECE difference on Room Occupancy with the GRU backbone under channel missingness and temporal thinning. Negative values indicate lower ECE for CaTRe.

Luis Candanedo. Occupancy detection, 2016. URL https://doi.org/10.24432/C5X01N. Dataset.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Łukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3:625–644, 2021. doi: 10.1038/s42254-021-00348-9. URL https://doi.org/10.1038/s42254-021-00348-9.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8: 6085, 2018. doi: 10.1038/s41598-018-24271-9. URL https://doi.org/10.1038/s41598-018-24271-9.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Table 6: **Complete ablation study.** Results are mean ± standard deviation over three seeds.

Dataset	Backbone	Variant	Accuracy	Macro-F1	NLL	ECE
Occupancy Detection	GRU	Base	0.993 ± 0.001	0.991 ± 0.001	0.031 ± 0.005	0.011 ± 0.004
Occupancy Detection	GRU	Gate only	0.992 ± 0.001	0.989 ± 0.001	0.035 ± 0.012	0.014 ± 0.007
Occupancy Detection	GRU	Full CaTRe	0.992 ± 0.001	0.989 ± 0.001	0.034 ± 0.012	0.013 ± 0.007
Occupancy Detection	GRU	Brier only	0.992 ± 0.001	0.989 ± 0.001	0.037 ± 0.008	0.014 ± 0.003
Occupancy Detection	GRU	Stability only	0.992 ± 0.000	0.990 ± 0.000	0.031 ± 0.005	0.012 ± 0.004
Occupancy Detection	LTC	Base	0.598 ± 0.130	0.482 ± 0.010	5.174 ± 4.170	0.370 ± 0.183
Occupancy Detection	LTC	Gate only	**0.990 ± 0.002**	**0.987 ± 0.003**	**0.045 ± 0.011**	**0.020 ± 0.010**
Occupancy Detection	LTC	Full CaTRe	0.841 ± 0.258	0.827 ± 0.277	2.506 ± 4.125	0.212 ± 0.213
Occupancy Detection	LTC	Brier only	0.551 ± 0.169	0.428 ± 0.054	5.126 ± 3.915	0.441 ± 0.141
Occupancy Detection	LTC	Stability only	0.829 ± 0.267	0.815 ± 0.284	2.592 ± 4.374	0.169 ± 0.269
Room Occupancy	GRU	Base	0.915 ± 0.006	0.507 ± 0.010	0.143 ± 0.010	0.035 ± 0.003
Room Occupancy	GRU	Gate only	**0.923 ± 0.002**	0.492 ± 0.057	0.167 ± 0.037	**0.029 ± 0.006**
Room Occupancy	GRU	Full CaTRe	0.913 ± 0.015	0.502 ± 0.039	0.155 ± 0.021	0.033 ± 0.010
Room Occupancy	GRU	Brier only	0.908 ± 0.008	0.460 ± 0.066	0.170 ± 0.054	0.039 ± 0.007
Room Occupancy	GRU	Stability only	0.907 ± 0.006	0.457 ± 0.058	0.170 ± 0.055	0.042 ± 0.003
Room Occupancy	LTC	Base	0.241 ± 0.068	0.137 ± 0.026	11.855 ± 0.680	0.753 ± 0.058
Room Occupancy	LTC	Gate only	**0.921 ± 0.002**	**0.448 ± 0.048**	**0.245 ± 0.049**	**0.067 ± 0.010**
Room Occupancy	LTC	Full CaTRe	0.891 ± 0.021	0.355 ± 0.087	0.754 ± 0.586	0.165 ± 0.078
Room Occupancy	LTC	Brier only	0.297 ± 0.032	0.167 ± 0.037	11.202 ± 0.511	0.703 ± 0.032
Room Occupancy	LTC	Stability only	0.272 ± 0.136	0.149 ± 0.034	5.779 ± 5.398	0.501 ± 0.355

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7657–7666, 2021. doi: 10.1609/aaai.v35i9.16936. URL https://doi.org/10.1609/aaai.v35i9.16936.

Ramin Hasani, Mathias Lechner, Alexander Amini, Lucas Liebenwein, Aaron Ray, Max Tschaikowski, Gerald Teschl, and Daniela Rus. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, 4:992–1003, 2022. doi: 10.1038/s42256-022-00556-7. URL https://doi.org/10.1038/s42256-022-00556-7.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. doi: 10.1038/s41586-019-0980-2. URL https://doi.org/10.1038/s41586-019-0980-2.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*,

volume 32, pp. 5320–5330, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019. doi: 10.1103/PhysRevLett.122.040504. URL https://doi.org/10.1103/PhysRevLett.122.040504.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019. doi: 10.1103/PhysRevA.99.032331. URL https://doi.org/10.1103/PhysRevA.99.032331.

Adarsh Pal Singh and Sachin Chaudhari. Room occupancy estimation, 2018. URL https://doi.org/10.24432/C5P605. Dataset.

Ryan Sweke, Frederik Wilde, Johannes Meyer, Maria Schuld, Paul K. Faehrmann, Barthélémy Meynard-Piganeau, and Jens Eisert. Stochastic gradient descent for hybrid quantum-classical optimization. *Quantum*, 4:314, 2020. doi: 10.22331/q-2020-08-31-314. URL https://doi.org/10.22331/q-2020-08-31-314.