Stable CDE Autoencoders with Acuity Regularization for Offline Reinforcement Learning in Sepsis Treatment

Yue Gao

Keebo.AI

yue@keebo.ai

Abstract

Effective reinforcement learning (RL) for sepsis treatment depends on learning stable, clinically meaningful state representations from irregular ICU time series. While previous works have explored representation learning for this task, the critical challenge of training instability in sequential representations and its detrimental impact on policy performance has been overlooked. This work demonstrates that Controlled Differential Equations (CDE) state representation can achieve strong RL policies when two key factors are met: (1) ensuring training stability through early stopping or stabilization methods, and (2) enforcing acuity-aware representations by correlation regularization with clinical scores (SOFA, SAPS-II, OASIS). Experiments on the MIMIC-III sepsis cohort reveal that stable CDE autoencoder produces representations strongly correlated with acuity scores and enables RL policies with superior performance (WIS return > 0.9). In contrast, unstable CDE representation leads to degraded representations and policy failure (WIS return ~ 0). Visualizations of the latent space show that stable CDEs not only separate survivor and non-survivor trajectories but also reveal clear acuity score gradients, whereas unstable training fails to capture either pattern. These findings highlight practical guidelines for using CDEs to encode irregular medical time series in clinical RL, emphasizing the need for training stability in sequential representation learning.

1 Introduction

Sequential decision-making is a cornerstone of modern healthcare, particularly in dynamic clinical scenarios, where timely diagnosis and adaptive treatment strategies are essential for patient survival [Shashikumar *et al.*, 2021; Solís-García *et al.*, 2023]. A representative case is sepsis management, which requires timely diagnosis and appropriate treatment strategies. Reinforcement learning (RL) offers a promising framework for modeling sequential decision-making in clinical settings, where treatment strategies must

adapt over time to a patient's evolving condition. In the context of sepsis management, RL has been employed to derive policies that optimize interventions including fluid resuscitation and vasopressor administration. For instance, the *AI Clinician* [Komorowski *et al.*, 2018], demonstrated RL's potential by training on the Medical Information Mart for Intensive Care III (MIMIC-III) dataset to recommend treatment strategies that outperformed clinician baselines. Subsequent advances in deep and distributional RL further refined policy robustness and flexibility, underscoring RL's role in sepsis management [Killian *et al.*, 2020; Jayaraman *et al.*, 2024; Böck *et al.*, 2022]. These advancements underscore the potential of RL to generate effective sequential treatment decisions in critical care environments.

In this work, we use MIMIC-III [Johnson *et al.*, 2016a; Johnson *et al.*, 2016b; Goldberger *et al.*, 2000] as experimental data, which contains clinical data from 40,000+ ICU patients, with vital signs, lab results, and treatments recorded at irregular time intervals. It provides a rich foundation for developing data-driven sepsis management tool. However, a key challenge lies in constructing informative state representations from this noisy, irregularly sampled data. Prior work has explored recurrent neural networks (RNNs) and autoencoders for this task, but their instability during training often leads to suboptimal representations and degraded policy performance [Böck *et al.*, 2022; Solís-García *et al.*, 2023; Killian *et al.*, 2020].

Neural Controlled Differential Equations (Neural CDEs) are state-of-the-art models for irregular time series data due to their ability to model continuous-time dynamics and handle irregular sampling [Killian *et al.*, 2020; Morrill *et al.*, 2021]. Unlike discrete-step RNNs, CDEs use differential equations to propagate hidden states, capturing latent physiological trends more accurately. This makes them particularly suited for sepsis management, where patient states evolve smoothly between sparse observations. However, CDEs are prone to training instability such as gradient explosion or collapse when unregularized or trained for excessive epochs [Agarwala *et al.*, 2022]. This instability stems fundamentally from numerical challenges in solving the underlying differential equations [Shoosmith, 2003], particularly when backpropagating through adaptive ODE solvers [Baker *et al.*, 2022].

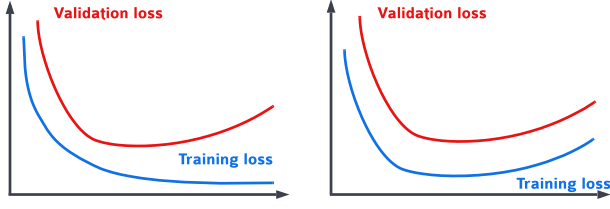A key challenge in applying Neural CDEs to clinical time

Figure 1: Distinguishing training failure modes: (Left) Classical overfitting; (Right) Numerical instability.

series is distinguishing between numerical instability and model overfitting, which are two fundamentally different failure modes. Overfitting appears when the training loss continues to decrease while the validation loss starts going upward; In contrast, CDEs exhibit unique instability patterns where both losses increase simultaneously (Figure 1). This instability occurs when numerical solvers fail to handle stiffness in the learned vector fields [Shoosmith, 2003]. For clinical time series modeling, this instability is particularly crucial, as trajectory smoothness and sudden acuity transitions must both be captured.

This limitation in CDE state representation was overlooked in prior studies. For instance, [Killian *et al.*, 2020] compared CDEs against other autoencoders for sepsis state representation and showed its superiority, but did not address their instability, leading to inconsistent results and an erroneous dismissal of clinical acuity score regularization.

In this work[1], we use the MIMIC-III database to train, evaluate, and compare CDE representations for sequential state encoding, with the goal of improving discrete Batch-Constrained Q-learning (dBCQ) models for offline sepsis treatment decision-making, with the following key contributions:

1. Identify and address the critical challenge in clinical RL: training instability in Neural CDEs for time-series state representation. Proposed a rigorous early stopping strategy that reliably produces robust representations.

2. Demonstrate that clinical acuity scores (SOFA/SAPS-II/OASIS) effectively regularize CDE autoencoder training when properly stabilized, contradicting prior claims about their ineffectiveness in RL representations.

3. Establish the consistency in MSE loss trend and acuity correlation loss trend during CDE autoencoder training process, validating acuity scores as meaningful proxies for patient state severity in CDE state representations.

4. Compare three stabilization methods (gradient clipping, implicit solvers, stiffness regularization) for clinical time series representation training, and provide practical guidance for selecting methods based on impact on downstream RL policy performance.

Our findings revisit and clarify inconsistencies reported in prior works such as [Killian *et al.*, 2020], which concluded

---

[1]Code available at https://github.com/GAOYUEtianc/RL_mimic_CDE_stable

that acuity-based regularization fails to improve CDE representations. We show this claim likely arose from unstable training regimes that obscure representation quality. By addressing CDE instability, we show that acuity regularization improves both state representation quality and RL performance.

2 Background

2.1 MIMIC-III Data Irregular Properties

The MIMIC-III repository contains ICU patient data including physiological signals, laboratory results, medications, and interventions as multivariate time series [Johnson *et al.*, 2016a]. However, these measurements were recorded at irregular intervals based on clinical needs, resulting in uneven sampling patterns and substantial missing data [Johnson *et al.*, 2016b; Rubanova *et al.*, 2019]. The dataset contains inherent measurement noise and informative missing patterns, where missing observations could provide clinical insights. This creates a low signal-to-noise ratio that makes direct trajectory modeling challenging [Che *et al.*, 2018; Yoon *et al.*, 2018]. Different variables are recorded at varying frequencies, causing covariate shift and temporal aliasing that reduce model generalization [Zhang *et al.*, 2019]. Simple imputation techniques like forward-filling or mean substitution fail to capture complex temporal dependencies and often introduce bias [Yoon *et al.*, 2018].

To mitigate these challenges, researchers have developed several approaches. Time-aware recurrent models use masking and time interval embeddings to explicitly handle missingness and irregularity [Che *et al.*, 2018]. Continuous-time methods such as ODE-RNNs and latent ODEs learn differential equations that naturally adapt to varying time intervals [Rubanova *et al.*, 2019]. Among these, Neural Controlled Differential Equations offer a SOTA approach to encode MIMIC-III's high signal-to-noise trajectories. They produce well-represented states suited for downstream offline reinforcement learning in sepsis management [Rubanova *et al.*, 2019; Kidger *et al.*, 2021; Killian *et al.*, 2020].

2.2 Neural Controlled Differential Equations (Neural CDEs) as State Encoders

Neural Controlled Differential Equations are a class of continuous-time models that generalize recurrent neural networks by defining hidden state dynamics through differential equations [Kidger *et al.*, 2021]. We initialize the hidden state as $h(t_0) = Wo(t_0) + b$, using a learnable linear map $g : \mathbb{R}^d \to \mathbb{R}^h$ implemented as a single-layer neural network. Then for the continuous-time irregular observations $o_t \in \mathcal{O}$, CDE encodes a hidden state $h(t) \to \mathcal{H}$ in differential form via

$$\partial h(t) = f_\theta(h(t))\partial o_t \tag{1}$$

where f_θ is a neural network parameterizing the system dynamics, and the differential ∂o_t accounts for irregular sampling intervals [Chen *et al.*, 2018; Rubanova *et al.*, 2019]. The CDE acts as an encoder $\psi : \mathcal{O} \to \mathcal{H}$, mapping time-based observations to a latent space. A decoder $\phi : \mathcal{H} \to \mathcal{O}$
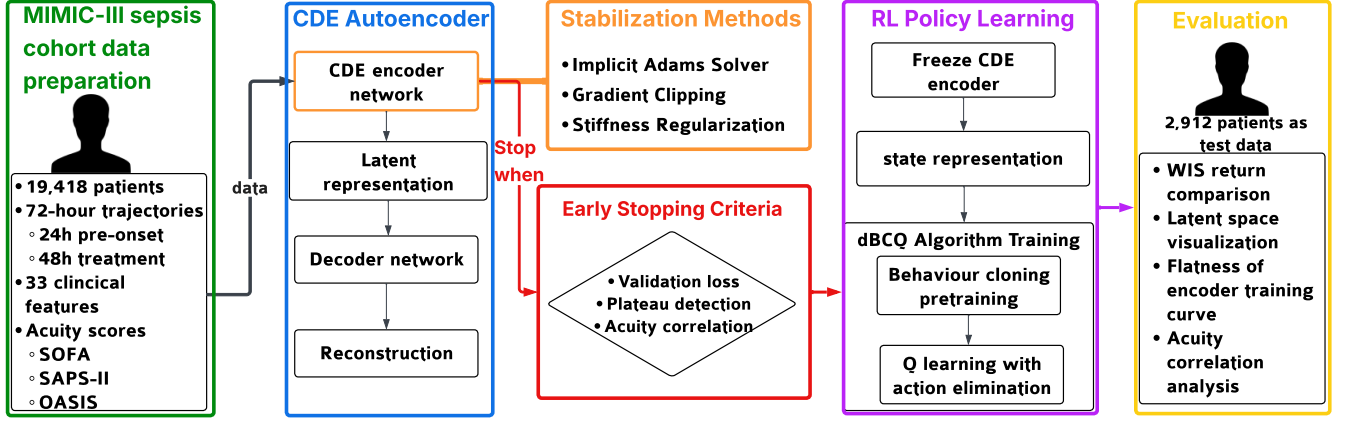
Figure 2: Overall architecture of the proposed framework.

reconstructing subsequent observations

$$\hat{o}_t = \phi(h(t)) \quad (2)$$

is trained to minimize the loss $\mathcal{L}_{\text{MSE}}(o_t, \hat{o}_t) = ||o_t - \hat{o}_t||^2$. This autoencoding framework ensures $h(t)$ retains clinically relevant information while discarding noise [Timothée *et al.*, 2018]. In this work, we use 4-th order Runge-Kutta (RK4) as the numerical solver for the CDE as a baseline [Hairer *et al.*, 2000].

2.3 Instability in Neural CDE Training

During training, Neural CDEs can exhibit numerical instability due to the sensitivity of ODE solvers to the learned vector field f_θ. If f_θ has large Lipschitz constant or long time sequences, the backpropagating gradients can explode, leading to erratic updates and divergence [Shoosmith, 2003; Kim *et al.*, 2021]. Hence, picking a proper stabilization method that can balance the smoothness and sensitivity to rare high-acuity transitions is crucial.

Effective methodologies to stabilize Neural CDEs training include implicit solvers that handle stiffness by solving linearized equations at each step [Gebregiorgis and Gonfa, 2021; Baker *et al.*, 2022; Chen *et al.*, 2018; Fronk and Petzold, 2024], gradient clipping to prevent large updates [Nguyen *et al.*, 2022], and regularization techniques smoothing the learned vector field by penalizing high local error and stiffness [Pal *et al.*, 2022]. Building on these insights, we employ early stopping and stabilization techniques to ensure Neural CDEs training produces high-quality state representations that generalize well and enable effective downstream RL performance.

2.4 Reinforcement Learning with CDE State Representations

Sequential sepsis treatment can be modeled as a Partially Observable Markov Decision Process (POMDP):

- State : $s_t = h(t)$ (CDE encoded history of observations $o_{0:t}$).
- Action : $a_t \in \{1, ..., 25\}$ represents discrete combinations of intravenous fluids and vasopressor doses [Komorowski *et al.*, 2018]

- Reward :

$$r_t = \begin{cases} +1 & \text{if patient survives at trajectory end} \\ -1 & \text{if patient dies at trajectory end} \\ 0 & \text{otherwise (at intermediate steps)} \end{cases} \quad (3)$$

- Policy: $\pi(a_t|s_t)$ maps states to probability distribution of actions, aiming to optimize cumulative reward via offline RL.

2.5 Clinical Acuity Scores as Priors

We leverage three established clinical acuity scores (SOFA, SAPS-II, and OASIS) as semi-supervision to regularize the CDE latent space. Detailed score definitions are provided in Appendix B. We define the correlation loss:

$$\mathcal{L}_{\text{corr}}(\hat{s}_t) = -(\rho_{\text{SOFA}}(\hat{s}_t) + \rho_{\text{SAPS-II}}(\hat{s}_t) + \rho_{\text{OASIS}}(\hat{s}_t)) \quad (4)$$

The regularized total loss function is defined to be:

$$\mathcal{L}_{\text{total}}(o_t, \hat{o}_t) = \mathcal{L}_{\text{MSE}}(o_t, \hat{o}_t) + \lambda \cdot \mathcal{L}_{\text{corr}}(\hat{s}_t) \quad (5)$$

where $\rho(\hat{s}_t)$ denotes the Pearson correlation between the latent state representation $\hat{s}_t$ and acuity score. For simplicity, we denote the losses at epoch i as $\mathcal{L}_{\text{MSE}}(i), \mathcal{L}_{\text{corr}}(i), \mathcal{L}_{\text{total}}(i)$.

3 Methodology

We present a framework as shown in Figure 2 for sepsis treatment policy learning using stabilized CDE state representations, evaluated through offline RL and clinical interpretability metrics.

3.1 Overall Architecture

Our framework builds on prior work in sepsis treatment RL [Killian *et al.*, 2020; Böck *et al.*, 2022; Jayaraman *et al.*, 2024] but addresses a critical gap: the training instability of Neural CDEs for state representation. While existing approaches use CDEs to encode patient history into hidden states and decode future observations, they overlook how training instability affects both representation quality and downstream policy performance. We identify two key instability symptoms:

- Unpredictable fluctuations in observation prediction loss
- Weakened correlation between learned states and clinical acuity scores

To address these, we introduce:

- Training stabilization via early stopping and specialized techniques
- Acuity-aware regularization to maintain clinical relevance

The stabilized representations then feed into Batch Constrained Q-learning, where we evaluate their impact on final policy performance through both quantitative metrics and clinical interpretability measures.

3.2 Data Preparation

Our study utilizes the MIMIC-III v1.4 critical care database, processed according to the established sepsis cohort definition from prior work [Komorowski *et al.*, 2018]. After processing, the dataset comprises 19,418 adult sepsis patients, with each patient trajectory spanning a clinically relevant 72-hour window around sepsis onset : capturing 24 hours preceding identification through 48 hours of subsequent treatment. These trajectories reflect the real-world challenges of ICU care, exhibiting irregular sampling intervals and heterogeneous measurements across 33 time-varying physiological features. The action space follows prior work in discretizing clinical interventions into 25 distinct combinations of intravenous fluids and vasopressor doses, binned by percentile ranges to maintain clinically meaningful groupings while enabling reinforcement learning. Patient outcomes define trajectory termination, with mortality recorded for the 9.2% of patients who died within 48 hours of their final observation, while survivors comprise the remaining 90.8%. To ensure faithful evaluation while preserving outcome distributions, we split the cohort into training (70%, n=13,593), validation (15%, n=2,913), and test sets (15%, n=2,912), maintaining identical 9.2% mortality rates across splits. Clinical acuity scores (SOFA, SAPS-II, OASIS) for illness severity are calculated at each timestep using validated implementations that transform the 33 raw features into standardized risk metrics.

3.3 CDE Autoencoder Training with Rigorous Stopping Criteria

The CDE autoencoder is trained to learn clinically meaningful state representations through a carefully designed optimization process that addresses the inherent instability of continuous-time neural networks.

Continuous-Time Encoding Process Our CDE autoencoder learns continuous-time state representations through a neural controlled differential equation (Equation 1), implemented as a 3-layer network with ReLU activations and layer normalization. The encoder outputs a hidden state $h(t) \in \mathbb{R}^d$ where d is the representation dimension.

Early Stopping Strategy To prevent training instability while preserving signal capture, we propose a multi-criteria early stopping method that selects the optimal checkpoint epoch e^* when all conditions are first met:

1. **Near-optimal validation loss:**

$$\mathcal{L}_{\text{val}}(e^*) \leq \min_{e \leq e^*} \mathcal{L}_{\text{val}}(e) + \epsilon_1 \tag{6}$$

2. **Stable training plateau:** For the last p epochs, the total loss variation remains within ϵ_2 fraction of the minimum loss:

$$\left| \max_{e^*-p \leq i \leq e^*} \mathcal{L}_{\text{total}}(i) - \min_{e^*-p \leq i \leq e^*} \mathcal{L}_{\text{total}}(i) \right| \leq$$
$$\epsilon_2 \cdot \min_{e^*-p \leq i \leq e^*} \mathcal{L}_{\text{total}}(i) \tag{7}$$

3. **Clinically meaningful representations over training:** Mean acuity score correlation on train set exceeds a threshold.

$$\rho(e^*) \geq \rho_{\text{threshold}} \tag{8}$$

We train CDE autoencoders across multiple random seeds through comprehensive hyperparameter tuning (learning rates, hidden sizes, ϵ_1, ϵ_2, p, $\rho_{\text{threshold}}$), then select the best performing configuration evaluated via the ultimate RL policy measured by WIS return [2].

3.4 Stabilization Methods for CDE Training

To improve training stability and facilitate model selection, we apply stabilization techniques to the CDE encoder, specifically on the vector field f_θ that governs hidden state dynamics. We implement three stabilization techniques:

1. **Gradient Clipping**: Constrains extreme gradient updates to prevent sudden spikes during backpropagation

2. **Implicit Adams Solver**: Uses an implicit numerical integration scheme to handle stiffness in the continuous-time dynamics

3. **Stiffness Regularization**: Directly penalizes high curvature in the learned vector field

Appendix D introduces those methodologies in detail.
Evaluation Protocol: We conduct comprehensive hyperparameter tuning for hyperparameters in Table 1 across

- Multiple random seeds : $25, 53, 1234, 2020$
- Learning rates : $1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}$
- Hidden dimensions : $4, 8, 16, 32, 64, 128$

Each configuration undergoes 100 epochs training until reaching early stopping criteria in Section 3.3, and is evaluated using two stability metrics:

1. **Plateau Length**: Number of consecutive epochs where the training loss remains on a plateau:

$$|\mathcal{L}_{\text{total}}(i) - \min \mathcal{L}_{\text{total}}| \leq \epsilon_2 \cdot \min \mathcal{L}_{\text{total}} \tag{9}$$

---

[2]Weighted Importance Sampling (WIS) is a technique used in offline reinforcement learning to estimate the expected return of a target policy using data collected from a different behavior policy. By normalizing importance weights across trajectories, WIS reduces variance compared to ordinary importance sampling, albeit introducing some bias. This trade-off often results in more stable and reliable policy evaluations [Mahmood *et al.*, 2014].

Method	Hyperparameter	Search Range
Gradient clipping	*max norm* : τ	$\{0.1, 0.5, 1.0, 1.5\}$
Implicit Adams solver	*step size* : Δt	$\{1/8, 1/4, 1/2\}$
Stiffness regularization	λ_{reg}	$\{0.005, 0.01, 0.015\}$

Table 1: Stabilization methods and hyperparameters to be tuned.

2. **Mean Absolute Slope on Plateau**:

$$S_1 = \frac{1}{T} \sum_i^T |\Delta\mathcal{L}_{\text{total}}(i)| \qquad (10)$$

where $\Delta\mathcal{L}_{\text{total}}(i) = \mathcal{L}_{\text{total}}(i+1) - \mathcal{L}_{\text{total}}(i)$ and T is the plateau length.

The final model evaluation considers both these stability metrics and the downstream RL policy performance (WIS return).

3.5 dBCQ Policy Learning and Evaluation

We employ the trained CDE encoder to transform raw patient trajectories into continuous state representations for offline reinforcement learning. Our policy learning approach builds on discrete Batch-Constrained Q-learning (dBCQ) [Killian *et al.*, 2020], which addresses key challenges in offline RL. The process is as follows:

Behavior Cloning Pre-training: We first train a policy to replicate the action distribution in the dataset, providing a conservative starting point.

Constrained Q-Learning: During policy optimization, the Q-function only considers actions that the behavior policy would likely take with probability $\geq \tau_{\text{BC}}$.

Policy Evaluation: We compute WIS returns of the trained Q-policy on the validation dataset to assess policy performance.

4 Experiments and Empirical Results

4.1 Effect of CDE Early Stopping on Final RL Policy Result

We conducted extensive hyperparameter tuning to identify optimal training configurations. We performed a grid search over hidden state sizes $\{4, 8, 16, 32, 64, 128\}$, learning rates $\{1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$, acuity correlation coefficients $\lambda \in \{0, 0.5, 1, 1.5\}$, and stopping criteria parameters : $\epsilon_1 \in \{0.05, 0.1, 0.15, 0.2\}, p \in \{20, 30, 40, 50\}, \epsilon_2 \in \{0.02, 0.03, 0.04, 0.05\}, \rho_{\text{threshold}} \in \{0.6, 0.65, 0.7, 0.75\}$.

The best configuration (hidden size=64, learning rate=2×10^{-4}, $\lambda = 1$, $\epsilon_1 = 0.1$, $p = 30$, $\epsilon_2 = 0.02$, $\rho_{\text{threshold}} = 0.7$) achieved the highest mean WIS return. Using this configuration, we trained the CDE autoencoder for 100 epoches and recorded its $\mathcal{L}_{\text{total}}$ (Equation 5) at each epoch.

Figure 3 shows the training-epoch-wise curves for the total loss, MSE loss, and correlation loss, with shaded bands indicating ± 1 standard deviation across random seeds. To
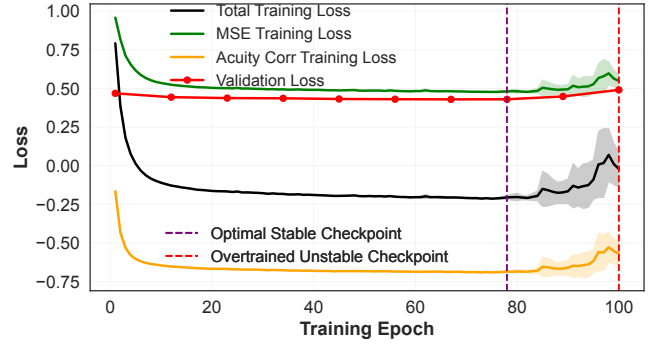

Figure 3: CDE autoencoder loss curves ($mean \pm std$ across multiple random seeds) versus training epoch with *optimal and stable* & *overtrained and unstable* checkpoints.
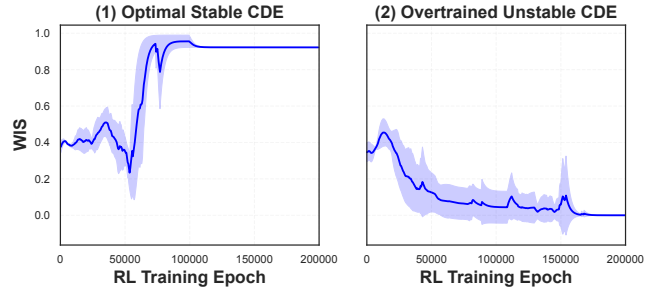

Figure 4: Evaluated WIS on validation set ($mean \pm std$ across random seeds) for dBCQ policies trained using *optimal and stable* and *overtrained and unstable* CDE state representations.

illustrate the impact of early stopping strategy, we highlight and compare those two checkpoints:

1. *optimal and stable*: meets all stopping criteria in Section 3.3 with hyperparameters above, achieves a low, stable loss and high acuity correlation

2. *overtrained and unstable*: shows unstable loss patterns and degraded generalization

To evaluate how early stoppping affects downstream RL, we freeze each CDE encoder at the two checkpoints above and train a dBCQ RL policy for them respectively using identical setting: policy learning rate $= 1 \times 10^{-5}$, dBCQ action elimination threshold $\tau_{\text{BC}} = 0.3$, and epochs $= 2 \times 10^5$. Figure 4 shows the WIS returns on validation patients trajectories versus RL training epoch, with shaded bands representing ± 1 standard deviation across random seeds. The policy initialized from the *optimal and stable* checkpoint consistently achieves high final return 0.9195. In contrast, policies derived from the *overtrained and unstable* CDE representations exhibits lower, more erratic RL performance, collapsing to a WIS return 3.2×10^{-7}, indicating complete value estimation failure. Collectively, these results demonstrate that selecting the right stopping epoch is essential for obtaining stable, generalizable representations from the CDE autoencoder, which has a critical impact on downstream RL policy quality. And our proposed early stopping strategy (Section 3.3) effectively optimized the state representation quality, yielding
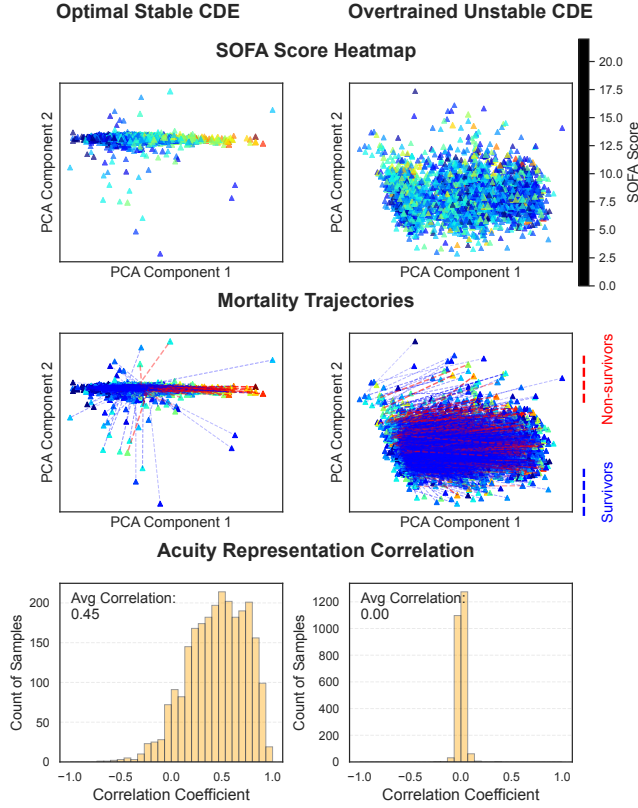
Figure 5: Comparative visualization of clinical alignment of *optimal and stable* and *overtrained and unstable* CDE representations on validation dataset trajectories

Pearson corr. coefficient r	significant p-value	mutual information
0.9578 ± 0.0127	$< 10^{-45}$	1.5296 ± 0.1847 nats

Table 2: Correlation between MSE training loss $\mathcal{L}_{\text{MSE}}$ and acuity correlation training loss $\mathcal{L}_{\text{corr}}$ during CDE autoencoder training, averaged across random seeds.

	All Epochs	**Plateau Epochs**
Pearson Corr.	0.6241 ± 0.1934	0.9284 ± 0.0295
p-value	< 0.05	< 0.05

Table 3: Correlation between training loss $\mathcal{L}_{\text{total}}$ and validation loss $\mathcal{L}_{\text{val}}$ over CDE autoencoder training epochs, averaged across random seeds, compared on plateau V.S. whole training epochs.

superior RL performance (WIS$\sim > 0.9$) compared to prior works [Killian *et al.*, 2020; Huang *et al.*, 2022] which reported WIS~ 0.775.

4.2 Clinical Alignment Through Early Stopping

Building on our policy results, we analyze how early stopping affects the clinical interpretability of CDE representations. We compare latent spaces from *optimal and stable* checkpoint and *overtrained and unstable* checkpoint. All experiments in this section use the best hyperparameters from Section 4.1, we use principal component analysis (PCA) to project those representations on validation trajectories (with first observation and end observation) into a lower dimensional space. Figure 5 shows three facets of the CDE representations, illustrates how stopping early affects clinical alignment on validation trajectories. In this section, we only show the result for SOFA score, but similar results hold for SAPS-II and OASIS as well (see Appendix B.1 for details).

1. **SOFA Score Distribution**: Each point is a validation observation colored by its SOFA score. At the *optimal and stable* CDE representation, patient states occupy a compact, well-populated manifold, where samples with similar SOFA scores cluster together, revealing a smooth severity gradient. By contrast, the *overtrained and unstable* CDE representation's observations are sparse and unstructured, without clear pattern for SOFA score distribution.

2. **Mortality Trajectories**: We overlay each validation patient's first-to-last latent jump, with survivors in blue and non-survivors in red. In the *optimal and stable* representation, non-survivor paths cluster roughly in one zone, suggesting clear prognostic separation. In the *overtrained and unstable* representation, both red and blue trajectories crisscross indiscriminately, offering no predictive signal.

3. **Acuity Representation Correlation**: We use histograms to show per-validation-trajectory mean Pearson correlation between latent features and acuity scores. The *optimal and stable* representation's distribution skews to the right ($mean\ correlation \sim 0.45$). The *overtrained and unstable* model's distribution is centered around zero, with a tight unimodal peak, indicating that its latent features no longer capture any meaningful severity signal.

In summary, early stopping strategy yields a CDE representation with better clinical interpretability.

Also, this comparison indicates that acuity correlation serves as a reliable proxy for evaluating the clinical alignment of learned representations. The representation with clearer survivor separation and a structured SOFA score gradient also yields superior downstream RL policy performance. This is verified by Figure 3 that the trend of MSE loss, acuity correlation loss, and validation loss are consistent. Quantitatively, we calculate the Pearson correlation coefficient and mutual information of $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{corr}}$ over training epochs. As shown in Table 2, the high Pearson correlation coefficient with significant p-value indicates a strong linear relationship; The mutual information of 1.62 nats further supports this relationship, suggesting that $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{corr}}$ are capturing the same underlying latent quality [Young *et al.*, 2023]. See Appendix B.2 for plot visualizing the correlation.

Preventing overfitting. As shown in Table 3, training-validation loss correlation improves from moderate (~ 0.62) on global to near-perfect (~ 0.93) on the plateau. This

Method	Plateau Length	S_1	$\overline{\text{WIS}}$
Baseline	42 ± 8.6	0.0028 ± 0.0005	0.9195
Gradient clipping	45.8 ± 4.5	0.0031 ± 0.0021	0.6547
Implicit Adams solver	46.5 ± 5.7	0.0027 ± 0.0006	0.9206
Stiffness regularization	43 ± 9.4	0.0030 ± 0.0008	0.9189

Table 4: Flatness metrics and downstream RL WIS return for each stabilization method (under best hyperparameters, across random seeds).

demonstrates our stopping criteria not only prevent instability but also avoid overfitting and ensure reliable generalization.

Contradicting prior work. Notably, this observation runs counter to claims in prior work [Killian *et al.*, 2020], which concluded clinical separability does not strongly correlate with downstream RL performance. They did not account for training stability of the CDE autoencoder, and trained for a fixed epoch length. In fact, when the CDE training is properly stabilized, both the MSE loss and the acuity correlation loss govern the same underlying latent quality, leading to representations that yield high RL policy returns and capture clinically meaningful trajectories. Appendix B.3 further verifies this conclusion by comparing the RL policies trained on CDE representations with and without acuity correlation regularization.

4.3 Stabilization Method Comparison

We evaluate three stabilization methods using our early stopping criteria, with all methods using $\lambda = 1$ for acuity correlation. Each method was trained for 100 epochs with optimal hyperparameters from Table 1. For each method, we compute **flatness metrics** mentioned in Section 3.4 on the loss curve, respectively **plateau length** and **mean absolute slope on plateau (S_1)**. Then we freeze each encoder at its *optimal and stable* checkpoint, train a downstream dBCQ RL policy using identical setting: policy learning rate $= 1 \times 10^{-5}$, dBCQ action elimination threshold $\tau_{\text{BC}} = 0.3$, and epochs $= 2 \times 10^5$, and get their mean evaluated WIS returns on validation trajectories.

As shown in Table 4, the implicit Adam solver produces the longest plateau, lowest mean absolute slope, and achieves the highest mean RL return ($\overline{WIS} = 0.9206$). Both implicit Adam solver and stiffness regularization yield comparable RL performance w.r.t baseline, while gradient clipping weakened both the downstream RL policy performance and flatness of training curve.

Figure 6 demonstrates that when training RL on the CDE representations from baseline and implicit Adam solver, the downstream RL policy converges fast to an optimal WIS return; The stiffness regularized representation converges
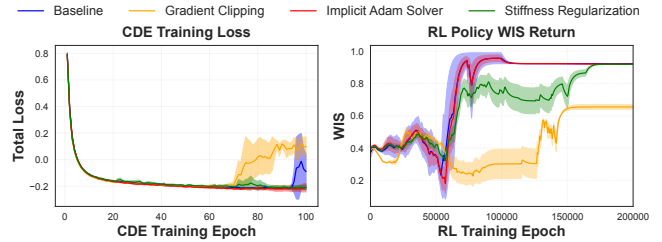

Figure 6: The CDE autoencoder training loss ($mean \pm std$) and WIS return ($mean \pm std$) for downstream RL on validation trajectories across random seeds for each stabilization method.

slower, but achieves a comparable WIS return finally. The gradient clipping representation, however, converges to a sub-optimal WIS return, and has the most erratic CDE training curvature. As a result, gradient clipping is not suitable for this clinical application, as it may overly dampen the dynamics of the learned vector field; Implicit Adam solver is the best method to balance training stability and representation quality.

5 Conclusion

Our work establishes practical guidelines for training Neural CDEs in clinical RL, addressing critical gaps in representation training stability and clinical alignment that were overlooked in prior works. We demonstrate that stabilizing CDE autoencoder training through early stopping and stabilization methodologies is essential for obtaining clinical meaningful and robust state representations, and could also significantly improve downstream RL policy performance. Our proposed multi-criteria early stopping strategy enables downstream RL policies to achieve superior performance (WIS > 0.9 compared to ~ 0.775 in [Killian *et al.*, 2020]), while stabilization methods like implicit Adam solver provide additional robustness by extending stable training plateaus.

We reconcile the controversy around acuity scores by showing they improve both RL policy returns and clinical interpretability when paired with stabilized CDE.

Our correlation analysis reveals a deeper insight that the state representation quality and acuity correlation are fundamentally linked by a shared latent structure. This shared latent space property explains why our stabilized training simultaneously achieves both high policy returns and clinical interpretability.

References

[Agarwala *et al.*, 2022] Susama Agarwala, Ben Dees, and Corey Lowman. Geometric instability of out of distribution data across autoencoder architecture, 01 2022.

[Alain and Bengio, 2014] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data generating distribution, 2014.

[Baker *et al.*, 2022] Justin Baker, Hedi Xia, Yiwei Wang, Elena Cherkaev, Akil Narayan, Long Chen, Jack Xin, Andrea L. Bertozzi, Stanley J. Osher, and Bao Wang. Proximal implicit ode solvers for accelerating learning neural odes, 2022.

[Brouwer and Krishnan, 2023] Edward De Brouwer and Rahul G. Krishnan. Anamnesic neural differential equations with orthogonal polynomial projections, 2023.

[Böck *et al.*, 2022] Markus Böck, Julien Malle, Daniel Pasterk, Hrvoje Kukina, Ramin Hasani, and Clemens Heitzinger. Superhuman performance on sepsis mimic-iii data by distributional reinforcement learning. *PLOS ONE*, 17(11):e0275358, 2022.

[Che *et al.*, 2018] 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.

[Chen *et al.*, 2018] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[Choo and Chang, 2022] Yoo Jin Choo and Min Cheol Chang. Use of machine learning in stroke rehabilitation: A narrative review. *Brain & NeuroRehabilitation*, 15(3):e26, Nov 2022. eCollection 2022 Nov. PMID: 36742082.

[Coelho *et al.*, 2024] C. Coelho, M. Fernanda P. Costa, and L.L. Ferrás. Enhancing continuous time series modelling with a latent ode-lstm approach. *Applied Mathematics and Computation*, 475:128727, 2024.

[Fronk and Petzold, 2024] Colby Fronk and Linda Petzold. Training stiff neural ordinary differential equations with implicit single-step methods, 2024.

[Gall *et al.*, 1993] J. R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24):2957–2963, Dec 1993.

[Gebregiorgis and Gonfa, 2021] Solomon Gebregiorgis and Genanew Gonfa. The comparison of runge-kutta and adams-bashforh-moulton methods for the first order ordinary differential equations. *International Journal of Current Research*, 8:27356–27360, 06 2021.

[Giannini *et al.*, 2019] Heather M. Giannini, Jennifer C. Ginestra, Corey Chivers, Michael Draugelis, Asaf Hanish, William D. Schweickert, Barry D. Fuchs, Laurie Meadows, Michael Lynch, Patrick J. Donnelly, Kimberly Pavan, Neil O. Fishman, C. William Hanson 3rd, and Craig A. Umscheid. A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice. *Critical Care Medicine*, 47(11):1485–1492, Nov 2019. Research supported by NIH (UL1 RR024134, UL1 TR000003). PMID: 31389839, PMCID: PMC8635476.

[Goldberger *et al.*, 2000] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[Hairer *et al.*, 2000] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I Nonstiff problems*. Springer, Berlin, second edition, 2000.

[Hassan *et al.*, 2023] Abbas M. Hassan, Aashish Rajesh, Malke Asaad, Jonas A. Nelson, J. Henk Coert, Babak J. Mehrara, and Charles E. Butler. Artificial intelligence and machine learning in prediction of surgical complications: Current state, applications, and implications. *The American Surgeon*, 89(1):25–30, Jan 2023. Epub 2022 May 13. Research supported by NIH grant P30 CA008748.

[Huang *et al.*, 2022] Yong Huang, Rui Cao, and Amir Rahmani. Reinforcement learning for sepsis treatment: A continuous action space solution. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 631–647. PMLR, 05–06 Aug 2022.

[Jayaraman *et al.*, 2024] Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N. Nadkarni, and Ankit Sakhuja. A primer on reinforcement learning in medicine for clinicians. *npj Digital Medicine*, 7(337), 2024.

[Johnson *et al.*, 2013] Alistair E. W. Johnson, Andrew A. Kramer, and Gari D. Clifford. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical Care Medicine*, 41(7):1711–1718, 2013.

[Johnson *et al.*, 2016a] Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database (version 1.4). https://doi.org/10.13026/C2XW26, 2016. PhysioNet.

[Johnson *et al.*, 2016b] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[Kidger *et al.*, 2021] Patrick Kidger, James Foster, Xuechen Li, and Terry J Lyons. Neural sdes as infinite-dimensional gans. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5453–5463. PMLR, 18–24 Jul 2021.

[Killian *et al.*, 2020] Taylor W. Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 139–160. PMLR, 11 Dec 2020.

[Kim *et al.*, 2021] Suyong Kim, Weiqi Ji, Sili Deng, Yingbo Ma, and Christopher Rackauckas. Stiff neural ordinary differential equations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(9), September 2021.

[Komorowski *et al.*, 2018] M. Komorowski, L.A. Celi, O. Badawi, A.C. Gordon, and A.A Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

[Le *et al.*, 2018] Lei Le, Andrew Patterson, and Martha White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[Mahmood *et al.*, 2014] A. Rupam Mahmood, Hado van Hasselt, and Richard S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[Morrill *et al.*, 2021] James Morrill, Patrick Kidger, Lingyi Yang, and Terry Lyons. Neural controlled differential equations for online prediction tasks. *arXiv preprint arXiv:2106.11028*, 2021.

[Nguyen *et al.*, 2022] Ho Huu Nghia Nguyen, Tan Nguyen, Huyen Vo, Stanley Osher, and Thieu Vo. Improving neural ordinary differential equations with nesterov's accelerated gradient method. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7712–7726. Curran Associates, Inc., 2022.

[Pal *et al.*, 2022] Avik Pal, Yingbo Ma, Viral Shah, and Christopher Rackauckas. Opening the blackbox: Accelerating neural differential equations by regularizing internal solver heuristics, 2022.

[Rubanova *et al.*, 2019] Yulia Rubanova, Ricky T. Q. Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Shashikumar *et al.*, 2021] Supreeth P. Shashikumar, Christopher S. Josef, Ashish Sharma, and Shamim Nemati. Deepaise – an interpretable and recurrent neural survival model for early prediction of sepsis. *Artificial Intelligence in Medicine*, 113:102036, 2021.

[Shoosmith, 2003] John N. Shoosmith. Numerical analysis. In Robert A. Meyers, editor, *Encyclopedia of Physical Science and Technology (Third Edition)*, pages 39–70. Academic Press, New York, third edition edition, 2003.

[Solís-García *et al.*, 2023] Javier Solís-García, Belén Vega-Márquez, Juan A. Nepomuceno, José C. Riquelme-Santos, and Isabel A. Nepomuceno-Chamorro. Comparing artificial intelligence strategies for early sepsis detection in the icu: an experimental study. *Applied Intelligence*, 53(24):30691–30705, 2023.

[Timothée *et al.*, 2018] Lesort Timothée, Díaz-Rodríguez Natalia, Goudou Jean-Franois, and Filliat David. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018.

[Vincent *et al.*, 1996] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. on behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intensive Care Medicine*, 22(7):707–710, Jul 1996.

[Yoon *et al.*, 2018] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, 2018.

[Young *et al.*, 2023] Alexander L. Young, Willem van den Boom, Rebecca A. Schroeder, Vijay Krishnamoorthy, Karthik Raghunathan, Hau-Tieng Wu, and David B. Dunson. Mutual information: Measuring nonlinear dependence in longitudinal epidemiological data. *PLoS One*, 18(4):e0284904, 2023. eCollection 2023.

[Zhang *et al.*, 2019] Kun Zhang, Yuan Xue, Gerardo Flores, Alvin Rajkomar, and Andrew M. Dai. Modelling ehr time-series by restricting feature interaction. arXiv preprint arXiv:1911.06410, 2019.

Appendix

A   Details about Patient Cohort

We follow the approach from [Komorowski *et al.*, 2018] to process the sepsis cohort from the MIMIC-III v1.4 critical care database. The original extracted data contains 48 variables including demographics, elixhauser status, laboratory results, vital signs, fluids and vasopressors, and fluid balance. There are missing and irregulaly sampled data in original MIMIC-III dataset, following the method by [Komorowski *et al.*, 2018], 19,418 adult sepsis patients were selected from the MIMIC-III database by applying the criteria:

- Patients are aged 18 or older;
- An acute increase in the Sequential Organ Failure Assessment (SOFA) score of 2 or more;
- Exclude admissions where treatment was withdrawn or mortality was not documented.

Then the data was filled using a time-limited approach based on clinically relevant periods for each variable, then further imputed using nearest neighbour algorithm. We then select 33 features (Table 5) that are most relevant to sepsis treatment.

For reinforcement learning in this task, actions are defined as combinations of intravenous (IV) fluids and vasopressors, discretized into 25 clinically meaningful bins based on percentile ranges. As shown in Table 6, these bins span from no administration to higher dose quartiles, forming a 5×5 action space, comprising 25 distinct treatment actions. Different types of vasopressors are converted to noradrenaline-equivalent, and the unit is mcg/kg/min. IV fluids are corrected for tonicity and converted to a standard unit.

B   Acuity score alignment

Patient acuity scores are used to measure the severity of a patient's condition, and are crucial for clinical decision-making. To constrain the learning of state representations, we extract three acuity scores from the full patient observations from each 4 hour time step (in MIMIC-III dataset):

- **SOFA (Sequential Organ Failure Assessment)** [Vincent *et al.*, 1996]: Assesses dysfunction across respiratory, coagulation, liver, cardiovascular, CNS, and renal systems. Scores range from 0 (normal) to 24 (severe failure).
- **SAPS-II (Simplified Acute Physiology Score)** [Gall *et al.*, 1993]: Predicts ICU mortality using 17 physiological measurements. Scores range from 0 to 163, with higher values indicating greater mortality risk.
- **OASIS (Oxford Acute Severity of Illness Score)** [Johnson *et al.*, 2013]: Estimates mortality risk from 10 clinical variables. Scores range from 10 to 83, where higher scores correspond to worse prognosis.

These scores are used to regularize the CDE latent space by the Pearson correlation between the state representation and the acuity socres. Equation 11 defines the acuity correlation regularized loss, where we choose the hyperparameter λ to be
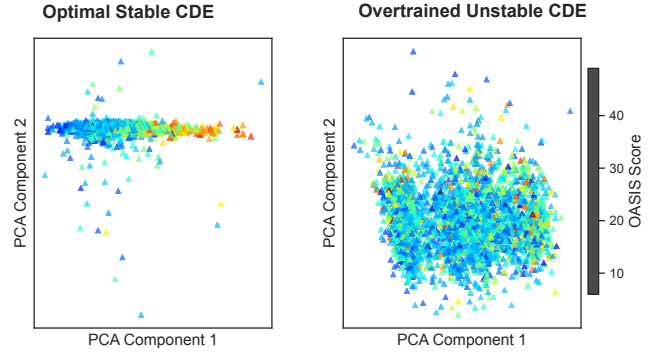

Figure 7: Heatmap of OASIS score distribution at *optimal and stable* and *overtrained and unstable* CDE representations on validation dataset
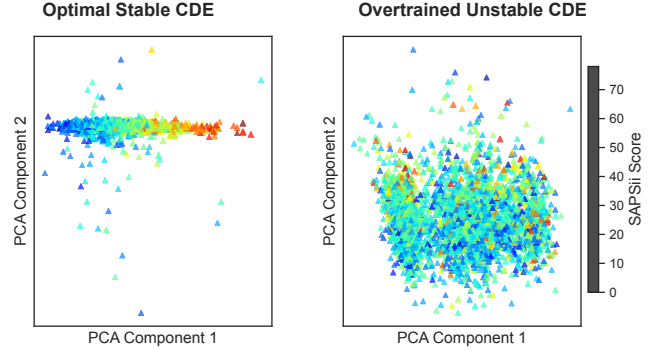

Figure 8: Heatmap of SAPS-II score distribution at *optimal and stable* and *overtrained and unstable* CDE representations on validation dataset

the same for all three acuity scores, yet these could be chosen independently of one another and yield a loss function :

$$\mathcal{L}_{\text{total}}(o_t, \hat{o}_t) = \mathcal{L}_{\text{MSE}}(o_t, \hat{o}_t) \\ - (\lambda_1 \cdot \rho_{\text{SOFA}}(\hat{s}_t) + \lambda_2 \cdot \rho_{\text{OASIS}}(\hat{s}_t) + \lambda_3 \cdot \rho_{\text{SAPSII}}(\hat{s}_t)) \tag{11}$$

B.1   Acuity score heatmaps

In section 4.2, we have compared the SOFA score heatmap of *optimal and stable* and *overtrained and unstable* CDE representations. Here we provide the heatmaps of other acuity scores, including OASIS and SAPS-II. The heatmaps are generated by projecting the latent features of validation trajectories into a lower dimensional space using PCA, and coloring the points by their respective acuity scores. The hyperparameters used for generating these heatmaps are the same as those used in Section 4.1, hidden size $= 64$, learning rate $= 2 \times 10^{-4}$, $\lambda = 1$, $\epsilon_1 = 0.1$, $p = 30$, $\epsilon_2 = 0.02$, $\rho_{\text{threshold}} = 0.7$.

As shown in Figure 7 and Figure 8, the *optimal and stable* CDE representation shows a clear clustering of patients with similar OASIS and SAPS-II scores, while the *overtrained and unstable* CDE representation exhibits a more scattered and random distribution, indicating a loss of clinical alignment. This further supports our finding that early stopping is crucial

Feature 1–8	Feature 9–16	Feature 17–24	Feature 25–33
Glasgow Coma Scale	Potassium	White Blood Cells	PaO_2/FiO_2
Heart Rate	Sodium	Platelets	Bicarbonate (HCO_3)
Systolic BP	Chloride	PTT	SpO_2
Diastolic BP	Glucose	PT	BUN
Mean BP	INR	Arterial pH	Creatinine
Respiratory Rate	Magnesium	Lactate	SGOT
Body Temp (°C)	Calcium	PaO_2	SGPT
FiO_2	Hemoglobin	$PaCO_2$	Bilirubin
			Base Excess

Table 5: List of 33 time-varying continuous physiological features used for state representation training.

Action Number	0	1	2	3	4
Vasopressors	0.00	$(0.00, 0.08]$	$(0.08, 0.22]$	$(0.22, 0.45]$	> 0.45
IV fluids	0.00	$(0.00, 50.00]$	$(50.00, 180.00]$	$(180.00, 530.00]$	> 530.00

Table 6: Discretized dosage bins for vasopressors and intravenous (IV) fluids used to define the 25-action reinforcement learning space

for maintaining acuity score alignment in CDE representations.

B.2 Correlation between MSE loss and Acuity Correlation loss

In Section 4.2, we have mentioned that the MSE loss and acuity correlation loss are consistent during CDE autoencoder training. Here we provide the mutual information plots between MSE loss and acuity correlation loss, which further supports our finding that the MSE loss and acuity correlation loss are capturing the same underlying latent quality. The mutual information is calculated using the empirical distribution of the losses over training epochs. As shown in Figure 9, the scatter plot of MSE loss and acuity correlation loss shows a strong linear relationship, with a high Pearson correlation coefficient with significant p-value; The high mutual information of 1.5296 ± 0.1847 nats indicates that the MSE loss and acuity correlation loss are capturing the same underlying latent quality, which further verifies our finding that early stopping is crucial for maintaining acuity score alignment in CDE representations.

B.3 RL policy comparison on CDE representations with / without acuity alignment

In this section we compare the downstream RL performance on CDE representations with and without acuity alignment. We use the CDE representations trained with ($\lambda = 1$) and without ($\lambda = 0$) acuity correlation loss, respectively, and train a dBCQ RL policy on each representation. Hyperparameters in Table 7 are tuned respectively to get an *optimal and stable* checkpoint for both settings.

The results are shown in Figure 10, where we can see that the dBCQ RL policy trained on CDE representations with acuity alignment achieves higher WIS returns 0.9195 than the
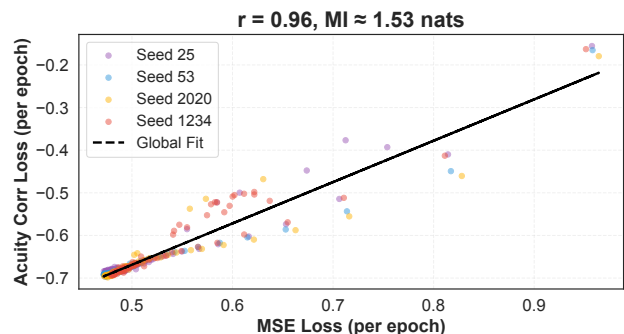

Figure 9: Scatter plot of MSE loss ($\mathcal{L}_{\text{MSE}}$) and Acuity Correlation loss ($\mathcal{L}_{\text{corr}}$) during CDE autoencoder training, across random seeds. The x-axis represents the MSE loss, and the y-axis represents the acuity correlation loss. Each scatter point represents the an epoch during CDE autoencoder training. The blackline represents the linear regression fit to the data. The Pearson correlation coefficient is 0.9578 ± 0.0127 with a significant p-value $< 10^{-45}$; The mutual information is 1.5296 ± 0.1847 (nats).

one trained on CDE representations without acuity alignment 0.6381. This further supports our finding that acuity correlation

C Early Stopping Criteria

Table 7 summarizes the best hyperparameters for the early stopping criteria in Section 3.3. The hyperparameters are tuned through grid search, and the best performing configuration is selected based on the downstream RL policy performance measured by WIS return.

Criterion	Metric	Hyperparameter	Best Hyperparameter
Low Validation Loss	Final validation loss should reach a minimal value during training	ϵ_1	0.1
Plateau	Number of consecutive epochs with small relative loss change	p ϵ_2	30 0.02
Acuity Correlation	Mean Pearson correlation between learned features and acuity scores	$\rho_{\text{threshold}}$	0.7

Table 7: Criteria for selecting optimal stopping epoch in CDE autoencoder training with hyperparameters to be tuned.
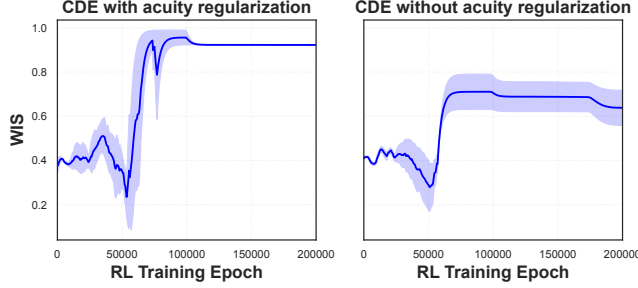

Figure 10: Comparison of WIS returns of dBCQ RL policies trained on CDE representations with and without acuity alignment, the left plot denotes the CDE autoencoder trained with $\lambda = 1$ on its best hyperparameters, the right plot denotes the CDE autoencoder trained with $\lambda = 0$ on its best hyperparameters. The x-axis represents the RL training epochs, and the y-axis represents the WIS returns.

D Stabilization methods

In this section, we provide a detailed description of the stabilization methods used in Section 4.3 to stabilize CDE autoencoder training. We evaluate three methods: implicit Adams solver, gradient clipping, and stiffness regularization. Each method is designed to improve the stability of the CDE autoencoder training process. Following the early stopping criteria, we evaluate the effectiveness of these methods by measuring the RL policy performance on the CDE representations, and also show the acuity alignment of each stabilization method on validation dataset in Figure 11.

D.1 Implicit Adams solver

When training the Neural CDE latent state $h(t)$ evolving under observation o_t as in Equation 1, implicit Adams methods advance $h(t)$ using future values of the target function in the integration step. An s step Adams-Moulton update rule from step t_n to t_{n+1} is defined as follows [Brouwer and Krishnan, 2023]:

$$h(n + 1) = h(n) + \Delta t \sum_{j=0}^{s} \beta_j \cdot f_\theta(h(n + 1 - j), o_{t_{n+1-j}}) \quad (12)$$

where $\Delta t = t_{n+1} - t_n$ and β_j are fixed coefficients. Note from Equation 12 that both sides include the term $f_\theta(h(n + 1 - j), o_{t_{n+1-j}})$, so a nonlinear equation needs to be solved

at each step. Here we list orders 0, 1, 2 and 4 for Adams-Moulton:

$$h(n) = h(n - 1) + \Delta t \cdot f_\theta(h(n), o_{t_n}); \quad (13)$$

$$h(n + 1) = h(n) + \Delta t \cdot \left(\frac{1}{2} f_\theta(h(n + 1), o_{t_{n+1}}) + \frac{1}{2} f_\theta(h(n), o_{t_n}) \right); \quad (14)$$

$$h(n + 2) = h(n + 1) + \Delta t \cdot \left(\frac{5}{12} f_\theta(h(n + 2), o_{t_{n+2}}) + \frac{8}{12} f_\theta(h(n + 1), o_{t_{n+1}}) - \frac{1}{12} f_\theta(h(n), o_{t_n}) \right); \quad (15)$$

$$h(n + 4) = h(n + 3) + \Delta t \cdot \left(\frac{251}{270} f_\theta(h(n + 4), o_{t_{n+4}}) + \frac{646}{720} f_\theta(h(n + 3), o_{t_{n+3}}) - \frac{264}{720} f_\theta(h(n + 2), o_{t_{n+2}}) + \frac{106}{720} f_\theta(h(n + 1), o_{t_{n+1}}) - \frac{19}{720} f_\theta(h(n), o_{t_n}) \right) \quad (16)$$

In our experiments in Section 4.3, we used the implicit Adams-Moulton method of order 4, which is a 4th-order implicit method. We do a grid search for *step size* $\Delta t \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$, and the best performance is achieved with $\Delta t = \frac{1}{8}$, resulting in a mean WIS return of 0.9206 on the downstream RL policy.

As shown in Figure 6 and Table 4, the implicit solver not only makes the curvature of the loss function smoother, makes the plateau longer and more stable, but also achieves the best performance in downstream RL policy. This indicates that implicit solver method balances the trade-off between smoothness and extracting sharp transits in the latent space (e.g, sudden deteriorations in septic patients), which is crucial for ICU time-series data.

D.2 Gradient clipping

Norm-based gradient clipping is a technique that limits the magnitude of the gradient vector during backpropagation. Given a parameter vector θ with gradient $g = \nabla_\theta L$ at a training step, the Euclidean norm $\|g\|_2$ is computed, then a threshold $\tau > 0$ is chosen, and the gradient is rescaled if its norm exceeds τ. Formally, the clipped gradient g_{clip} is defined by:

$$g_{\text{clip}} = \begin{cases} g, & \text{if } \|g\|_2 \leq \tau \\ \tau \frac{g}{\|g\|_2} & \text{if } \|g\|_2 > \tau \end{cases} \quad (17)$$

Mathematically, gradient clipping can be seen as a form of regularization on the optimization trajectory. It does not alter the loss function or objective explicitly, but it modifies the optimization dynamics to avoid excessively large parameter jumps.

When training Neural CDE model, as introduced in Equation 1, one backpropagates through the numerical solver to compute the gradients of the loss function with respect to the model parameters. If f_θ has large Lipschitz constant or the integration time is long, the backward gradients can suffer from gradient explosion. Previous works [Nguyen *et al.*, 2022; Coelho *et al.*, 2024] explicitly cite gradient clipping as an effective method for controlling gradient explosion and oscillatory in continuous-time models.

In section 3.4, we apply gradient clipping with a grid search on the threshold $\tau \in \{0.5, 1.0, 1.5\}$, and as a result, when $\tau = 1.0$, the CDE autoencoder achieves the best performance in downstream RL policy. With a mean WIS return of 0.6547, it is lower than the baseline CDE autoencoder without stabilization, which achieves a mean WIS return of 0.9195. An explanation for this could be that in irregular ICU time-series, patient trajectories are sparse and short, and gradient clipping uniformly dampens high-magnitude updates, however, some large gradients might be clinically meaningful. Also, in ICU trajectories, there are sudden deteriorations in septic patients, which causes large updates to θ, while gradient clipping will suppress that update, leading to a less informative representation.

D.3 Stiffness Regularization

This regularization technique leverages stiffness indicators to shape the training dynamics of Neural CDEs. Recall from Section 2.2 that the encoder is defined as :

$$\partial h(t) = f_\theta(h(t))\partial o_t \tag{18}$$

Assume that we solve it over time interval $[t_0, t_1]$, and reconstruct the downstream target $\hat{o}_{t_1} = \phi(h(t_1))$. Let $\mathcal{L}_{\text{total}}$ in Equation 5 be the reconstruction loss, $\{t_j\}_{j=1}^N$ be the solver time steps. Specifically, for each time step t_j, the stiffness score $\mathcal{S}_j$ is apprixmated via the real parts of the eigenvalues of the local Jacobian:

$$\mathcal{S}_j = \max \left\{ |\Re(\lambda_i)| : \lambda_i \in \text{eig}\left(\frac{\partial f_\theta(h(t_j))}{\partial h(t)}\right) \right\} \tag{19}$$

where

- $\frac{\partial f_\theta(h(t_j))}{\partial h(t)} \in \mathbb{R}^{d \times d}$ denotes the Jacobian of the vector field w.r.t. the hidden state at time t_j.

- $\text{eig}(\cdot)$ denotes the set of eigenvalues.

- $\Re(\lambda_i)$ denotes the real part of λ_i.

Then the total loss becomes:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{total}} + \lambda_{\text{reg}} \cdot \underbrace{\sum_{j=1}^N S_j}_{\text{Stiffness Regularization}} \tag{20}$$

This objective encourages the encoder's neural vector field f_θ to not only fit the data but also generate dynamics that are less stiff, leading to smoother trajectories in the latent space. The hyperparameter λ_{reg} controls the strength of this regularization, which balances the tradeoff between fitting the data and minimizing the solver complexity. As mentioned in Table 1, we do a grid search on $\lambda_{\text{reg}} \in \{0.005, 0.01, 0.015\}$, and find that $\lambda_{\text{reg}} = 0.01$ achieves the best performance in downstream RL policy, with a mean WIS return of 0.9189. Figure 6 shows that the stiffness regularization stabilizes the CDE training and lead to a longer, smoother plateau, which is beneficial for picking the optimal stopping epoch. And its downstream RL policy reaches a comparable optimal WIS return.

In summary, stiffness regularzation is a promising method for stabilizing CDE autoencoder training for learning MIMIC-III data representation, as it encourages the model to learn smoother dynamics, which is crucial for capturing the underlying patterns in irregular ICU time-series data. It also helps to extend the plateau of the training loss, making it easier to find the optimal stopping epoch.

E Limitation and Future Work

When evaluating the state representations by RL policy, the bias-variance tradeoff in Weighted Importance Sampling (WIS) is tricky. Particulary, for sepsis treatment, the variance of WIS is highly sensitive to the behaviour policy threshold τ_{BC}. As τ_{BC} decreases, the variance of WIS gets higher. Like the prior works, our work does not discuss potential pitfalls of WIS variance with respect to τ_{BC} and how to mitigate it. Future work should analyze the impact of τ_{BC}, and explore strategies such as doubly robust estimators to reduce WIS variance.

In the future, we plan to extend our work to more stabilization methods including Gaussian noise injection [Alain and Bengio, 2014] and dropout in vector field [Le *et al.*, 2018]. To evaluate the state representations, other SOTA offline RL methods such as Fitted Q-Iteration, Conservative Offline Model Based Policy Optimization should be considered. There is still big potential to extend acuity alignment to other time-series medical tasks including septic shock prediction [Giannini *et al.*, 2019], ICU stroke recovery [Choo and Chang, 2022], and post-surgical complication prediction [Hassan *et al.*, 2023].
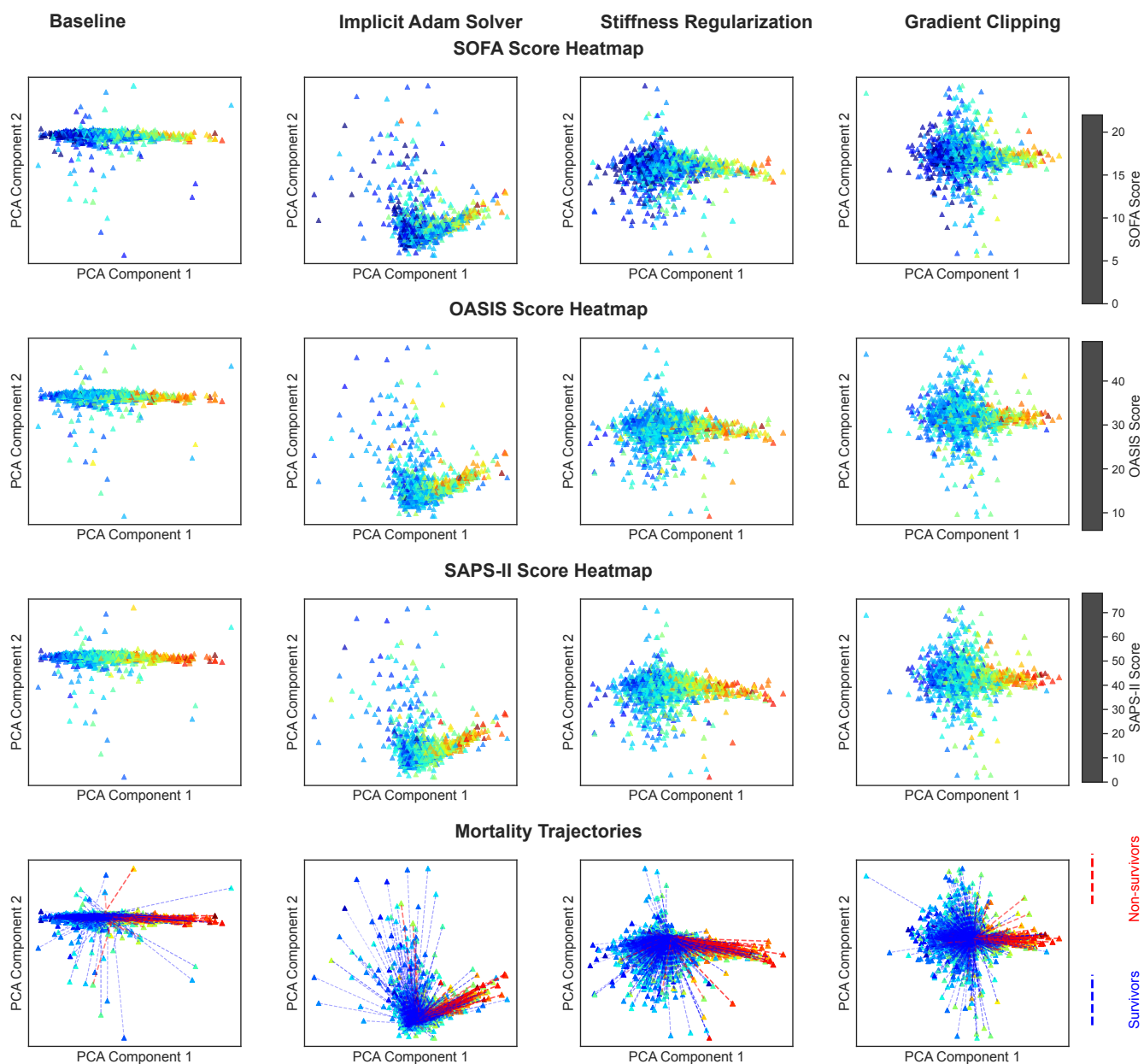
Figure 11: SOFA, OASIS, SAPS-II score distribution and mortality trajectories of CDE autoencoders for each stabilization method on validation dataset. It can be shown that following the early stopping criteria, all methods achieve a clear clustering of patients with similar acuity scores, and the non-survivor trajectories are well separated from survivor trajectories.