Neural Stochastic Differential Equations on Compact State Spaces: Theory, Methods, and Application to Suicide Risk Modeling

Malinda Lu[1,*] **Yue-Jane Liu**[1,*] **Matthew K. Nock**[2] **Yaniv Yacoby**[1,†]

[1]Wellesley College
[2]Harvard University
[*]Equal contribution
[†]Correspondence to: yy109@wellesley.edu

Abstract

Ecological Momentary Assessment (EMA) studies enable the collection of high-frequency self-reports of suicidal thoughts and behaviors (STBs) via smartphones. Latent stochastic differential equations (SDEs) are a promising model class for EMA data, as it is irregularly sampled, noisy, and partially observed. But SDE-based models suffer from two key limitations. (a) These models often violate domain constraints, undermining scientific validity and clinical trust of the model. (b) Training is numerically unstable without ad hoc fixes (e.g. oversimplified dynamics) that are ill-suited for high-stakes applications. Here, we develop a novel class of expressive SDEs whose solutions are provably confined to a prescribed compact polyhedral state space, matching the domains of EMA data. In this work, (1) we show why chain-rule based constructions of SDEs on compact domains fail, theoretically and empirically; (2) we derive constraints on drift and diffusion for general and stationary SDEs so their solutions remain in the desired state space; and (3), we introduce a parameterization that maps arbitrary (neural or expert-given) dynamics into constraint-satisfying SDEs. On several real EMA datasets, including a large suicide-risk study, our parameterization improves forecasts and optimization dynamics over standard latent neural SDE baselines. These contributions pave the way for principled, trustworthy continuous-time models of suicide risk and other clinical time series and extend applications of SDE-based methods (e.g. diffusion models) to domains with hard state constraints.

1 Introduction and Related Work

Suicide is a leading cause of death worldwide (World Health Organization, 2025), yet our ability to understand and predict suicidal thoughts and behaviors (STBs) remains limited (Nock and Wang, 2026). A key difficulty is that transitions from suicidal ideation to action can occur rapidly (Millner et al., 2017), leaving narrow windows for observation. Ecological Momentary Assessment (EMA) studies (Shiffman et al., 2008) address this using smartphones to collect high-frequency self-reports of affects, emotions, and STBs in daily life. However, EMA time series are irregularly sampled, noisy, and partially observed, with dynamics influenced by unobserved external factors. These characteristics violate the assumptions of many conventional approaches, which rely on regular sampling, complete observations or ad hoc treatment of missingness.

Latent stochastic differential equations (SDEs) (e.g. Archambeau et al. (2007); Li et al. (2020)) are an expressive probabilistic framework that naturally accommodates irregular sampling, explicitly models latent psychological dynamics, and separates process noise from observation noise. SDEs take the form, $dz_t = h(t, z_t) \cdot dt + g(t, z_t) \cdot dB_t$, wherein the change in state, dz_t, is modeled as a sum of deterministic and stochastic components, the "drift," h, and "diffusion," g, respectively. By parameterizing the dynamics (drift and diffusion) with neural networks (NNs), "neural SDEs" can capture volatile, nonlinear trajectories, making them one of the few viable models for EMA data. However, these models suffer from two important limitations.

Limitation 1: Many applications of SDE-based models require SDE solutions to satisfy domain-specific constraints. Without additional structure, SDEs with expressive (e.g. NN-based) dynamics can encode inappropriate inductive biases for a given domain, producing invalid forecasts that call into question the validity of the entire model. In EMA data, most survey questions use a 0–10 Likert scale, so SDE solutions must also lie in a compact state space. But instead, these SDEs may yield nonsensical forecasts, like a "12 out of 10" intensity of suicidal ideation. To illustrate the importance of addressing such model misspecification, imagine a weather model that forecasts 70% chance of 70°F, 20% chance of 80°F, and 10% chance of 5,000°F. Once the model assigns significant probability mass to an impossible event, how correct are the 70% and 20% forecasts? And how should the impossible 10% be reallocated? For suicide prevention, analogous misspecifications are not merely inconvenient; they are dangerous. A model that allocates probability mass to impossible psychological states may (a) learn incorrect dynamics, leading to false scientific conclusions, and/or (b) mislead downstream clinical decision-making, either missing opportunities to prevent suicide attempts or triggering unnecessary alarms that erode clinicians' trust.

Limitation 2: SDE-based models are unstable during training. Numerical instabilities arise both when solving the SDE *and* when backpropagating through the solver (e.g. Zhang et al. (2024)). When SDEs are incorporated into deep probabilistic models, these numerical instabilities exacerbate existing training difficulties. In practice, the adoption of SDE-based models often hinges on simplified dynamics (e.g. Ansari et al. (2023); Oh et al. (2024)) and training tricks (like KL-annealing, e.g. Li et al. (2020)). For suicide prevention, however, these workarounds are problematic: (a) oversimplified dynamics may fail to capture the complex mechanisms that drive rapid shifts in STBs, and (b) dependence on training tricks undermines reproducibility, hinders clinical deployment, and reduces trust and uptake among domain experts. Finally, although recent work explores promising simulation-free inference for latent SDEs (e.g. Zhang et al. (2024); Bartosh et al. (2025)), approximate inference methods introduce their own, hard-to-characterize inductive biases (e.g. Yacoby et al. (2020); Foong et al. (2020); Yacoby et al. (2022); Coker et al. (2022)); thus, the empirical behavior and trade-offs implicit in simulation-free methods are not yet well understood.

Insight: Both limitations are addressed by enforcing that SDE solutions stay within the data domain. Doing so not only addresses model mismatch—it also improves the stability of model fit. For example, recent work on diffusion models for image data observed that parameterizing SDEs on compact state spaces can improve their performance (Saharia et al., 2022; Lou and Ermon, 2023; Fishman et al., 2023a; Christopher et al., 2024). We hypothesize that, in early training, unconstrained SDE trajectories often exit the compact domain of the data and require many gradient steps to return, increasing chances of landing in poor local optima. Later in training, even small perturbations to the dynamics can push trajectories outside the data region. By constraining dynamics to respect the natural compact data domain, we induce a bias toward admissible models, improving both training dynamics and generalization. In this work, we focus on compact polyhedra, which represent a large class of commonly-used spaces, including rectangular spaces and simplexes, useful for a variety of temporal natural phenomena (e.g. Cresson et al. (2016)) as well as models for image data (e.g. diffusion models (Lou and Ermon, 2023)).

Shortcomings of existing SDE-based models on compact state spaces. Of these recent works, SDEs with reflected dynamics are promising because they apply to *any* SDE-based model. Re-

flected SDEs (RSDEs) on a compact space K augment the original SDE with a third term (Pilipenko, 2014): $dz_t = h(t, z_t) \cdot dt + g(t, z_t) \cdot dB_t + r(z_t) \cdot dC_t$, where $r(z_t)$ is a reflecting vector field that pushes z_t back into the interior of K, $\text{int}(K)$, when z_t is on the boundary of the space, ∂K, and C_t is a non-decreasing process that increases only when $z_t \in \partial K$ (i.e. C_t satisfies $\int_0^t \mathbb{I}(z_s \notin \partial K) \cdot dC_s = 0$). Thus, RSDEs behave like SDEs on the interior of the space, but are reflected inwards at the boundary. Despite their rich theory, RSDEs have two shortcomings. First, the instantaneous equal-and-opposite push towards the interior, represented by r, may not faithfully describe many phenomena in physics, biology, engineering, and medicine (e.g. d'Onofrio (2013); Rohanizadegan et al. (2020)), including the dynamics of STBs. Second, RSDEs currently lack efficient, high-order solvers (e.g. Ding and Zhang (2008); Fishman et al. (2023b)). These challenges become barriers for applications in suicide research, where model interpretation is important.

Complementing work on RSDEs, recent work has drawn on stochastic viability theory (Aubin, 1991) to parameterize SDEs on compact state spaces (without reflections) via careful construction of the dynamics (e.g. Cai and Lin (1996); Cresson et al. (2012); d'Onofrio (2013); Cresson et al. (2016); Cresson and Sonner (2018); Rohanizadegan et al. (2020)). While promising, the dynamics proposed in these works are tailored to specific phenomena and do not readily generalize. Here, we generalize these ideas to obtain arbitrarily flexible SDE dynamics on any compact polyhedral state space. We summarize the position of this work relative to these prior approaches in Appendix A.

Contributions. In this paper, we propose a novel class of expressive SDEs on compact polyhedral spaces using insights from stochastic viability theory. **(1)** We explain why chain-rule based approaches struggle theoretically and empirically (Section 2). **(2)** We prove constraints on the drift/diffusion that ensure general and stationary SDEs have an inductive bias for compact state spaces (Section 3). **(3)** We propose a parameterization that provably satisfies these constraints, allowing us to transform any dynamics—whether NN-based or expert-specified—into SDEs whose solutions remain in a prescribed compact polyhedral state space (Section 4), enabling us to capture a different class of natural phenomena than RSDEs. Finally, **(4)** on several real EMA datasets, we empirically demonstrate that enforcing viability filters out psychologically inadmissible models, thereby steering learning toward dynamics closer to the data-generating process, resulting in improved forecasts and optimization dynamics relative to baselines (Section 5). We include a demo of our method at: `https://github.com/mogu-lab/wsp-demo`.

Broader Impact. This work impacts two communities: (i) ML researchers developing SDE-based models whose latent states live on constrained domains, and (ii) clinical psychology theorists for whom it is crucial that the dynamics respect clinical knowledge. For ML researchers, our method provides a general mechanism for imposing hard state constraints on expressive SDE dynamics, and can be integrated into a broad range of SDE-based frameworks (e.g. diffusion models (Song et al., 2021) and infinitely deep models (Xu et al., 2022a)), while remaining compatible with standard inference methods for latent SDEs (Archambeau et al., 2007; Kidger et al., 2021; Issa et al., 2023; Zhang et al., 2025), including simulation-free approaches (Bartosh et al., 2025). For clinical psychology theorists, our viability results have implications for ongoing efforts to formalize suicide and other mental health challenges as dynamical systems (e.g. Millner et al. (2020); Wang et al. (2023); Robinaugh et al. (2024)). Ultimately, this work paves way for hybrid models (Schweidtmann et al., 2024) of suicide, jointly guided by domain expertise *and* by data, to advance our understanding of suicide and improve our ability to identify individuals at imminent risk in time for intervention.

Notation. Consider the following Itô SDE:

$$dz_t = h(t, z_t) \cdot dt + (\text{diag} \circ g)(t, z_t) \cdot dB_t. \tag{1}$$

Here, $t \geq 0$ is time, $z_t \in K$ is the SDE's solution, which lies in a compact space, $K \subset \mathbb{R}^{D_z}$. Next, $h : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}$ and $g : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}_{\geq 0}$ are the drift and diffusion, respectively. We overload $\text{diag}(\cdot)$ to embed (or extract) a vector into (or from) the diagonal of a matrix, and define I_D as a D-dimensional identity matrix. Finally, we use e_d for the dth standard basis vector, ∇ for the Jacobian, $\langle \cdot, \cdot \rangle$ for inner products, and z_t^d, h^d, and g^d for the dth dimension of z_t, h and g, respectively.

2 Challenges with Chain-Rule Based Methods

Our goal is to find an expressive parameterization of h and g so that the SDE in Eq. 1 is *viable*:

Definition 1 (Milian (1995)). *A stochastic process z_t is* viable *in K if, for every initial condition $z_0 \in K$, $\mathbb{P}(z_t \in K, \forall t \in [0, \infty)) = 1$.*

Transforming SDE solutions on $\mathbb{R}^{D_z}$ to solutions on K. The simplest way to ensure z_t lies on a compact space K is to derive a closed-form SDE for $f(z_t)$, where $f : \mathbb{R}^{D_z} \to K$. This is achieved with Itô's lemma for Itô SDEs and with the standard chain-rule for Stratonovich SDEs. While simple, however, this approach does not work theoretically or empirically for three reasons.

Challenge 1: Theory. There does not exist a diffeomorphism, f, from a non-compact set, $\mathbb{R}^{D_z}$, to a compact set, K. In practice, we often ignore this and map $\mathbb{R}^{D_z}$ to the *interior* of K, as in classification models that use sigmoid/softmax to map Euclidean outputs to a unit cube/simplex. However, this can cause pathologies: for linearly separable classes, the logistic regression maximum likelihood estimation drives parameters to infinity (Santner and Duffy, 1986), undermining interpretability. Analogous issues can arise for SDEs parameterizing time-varying Bernoulli probabilities.

Challenge 2: Numerical Stability. If we are willing to overlook Challenge 1, we find ourselves with numerically unstable dynamics. To see this, consider a 1D SDE $y_t \in \mathbb{R}$ with drift $\tilde{h}$ and diffusion $\tilde{g}$, and suppose we construct $z_t \in (0, 1)$ via the sigmoid transform $z_t = f(y_t)$. We begin with Itô's lemma and plug in $y_t = f^{-1}(z_t) = \text{sigmoid}^{-1}(z_t)$:

$$\begin{aligned} dz_t &= \left[\tilde{h}(t, y_t) \cdot \frac{\partial f(y_t)}{\partial y_t} + \frac{1}{2} \cdot \tilde{g}(t, y_t)^2 \cdot \frac{\partial^2 f(y_t)}{\partial y_t^2}\right] \cdot dt + \left[\tilde{g}(t, y_t) \cdot \frac{\partial f(y_t)}{\partial y_t}\right] \cdot dB_t, \\ &= (z_t - z_t^2) \cdot \left[\tilde{h}(t, f^{-1}(z_t)) + \tilde{g}(t, f^{-1}(z_t))^2 \cdot \left(\tfrac{1}{2} - z_t\right)\right] \cdot dt + (z_t - z_t^2) \cdot \tilde{g}(t, f^{-1}(z_t)) \cdot dB_t. \end{aligned} \tag{2}$$

Here, f^{-1} is unbounded, and can induce unbounded, numerically unstable dynamics. Moreover, existence and uniqueness proofs for SDEs typically require linearly bounded dynamics (e.g. Oksendal (2013, Theorem 5.2.1)).

Challenge 3: Inductive Bias. If we are willing to overlook Challenge 1, we can overcome Challenge 2 as follows. We observe that arbitrarily expressive $\tilde{h}$ and $\tilde{g}$ can "undo" f^{-1} by internally composing it with f. Thus, we can define $h(t, z_t) = \tilde{h}(t, f^{-1}(z_t))$ and $g(t, z_t) = \tilde{g}(t, f^{-1}(z_t))$ and parameterize h and g directly, e.g. via NNs:

$$dz_t = (z_t - z_t^2) \cdot \left[h(t, z_t) + g(t, z_t)^2 \cdot \left(\tfrac{1}{2} - z_t\right)\right] \cdot dt + (z_t - z_t^2) \cdot g(t, z_t) \cdot dB_t. \tag{3}$$

This way, so long as h and g are bounded, this SDE has bounded dynamics, overcoming Challenge 2. But, this surfaces yet another challenge: the inductive bias of this SDE is appropriate for few phenomena. Because the sigmoid maps $\mathbb{R}$ to $(0, 1)$, large excursions of y_t toward $\pm\infty$ correspond to vanishingly small changes in z_t near 0 or 1. In the induced SDE, this appears as the multiplicative factor, $(z_t - z_t^2)$, on both the drift and diffusion, which tends to 0 at the boundary, causing trajectories to slow down and effectively "stick" there. We verify this behavior empirically in Section 5. This behavior is undesirable when modeling suicidal ideation, which can oscillate rapidly during short durations (e.g. Coppersmith et al. (2023)).

3 Constraints on Dynamics for SDEs on Compact Polyhedra

Motivated by the challenges from Section 2, we prove constraints on the drift/diffusion to ensure that both stationary and general SDEs have an inductive bias for compact polyhedral state spaces. To begin, we define polyhedral subspaces of Euclidean space:

Definition 2. *Let $u, v \in \mathbb{R}^{D_z}$ and $\mathcal{H}(u, v) = \{z \in \mathbb{R}^{D_z} : \langle z - u, v\rangle \geq 0\}$ denote a closed half-space. $K \subset \mathbb{R}^{D_z}$ is a polyhedron if it is a finite intersection of closed half-spaces: $K = \bigcap_{s \in \{1, \cdots, S\}} \mathcal{H}(u_s, v_s)$.*

General SDEs. In Theorem 3, Milian (1995) shows that, under standard linear-growth and Lipschitz assumptions, an Itô SDE is viable in a polyhedron, K, if and only if, on each half-space boundary, the drift and diffusion have no outward-pointing components: thus, stochastic motion cannot push it across the boundary, and any normal motion at the boundary is deterministic and non-outward. Compared to the sigmoid-based chain-rule construction for the unit cube (Section 2), these are less restrictive requirements, since they do not force the drift to be 0 at the boundary. While Theorem 3 also holds for non-compact polyhedra, we focus on compact polyhedra from here on. In Appendix B, we extend this result to Stratonovich SDEs on compact polyhedra.

Theorem 3 (Milian (1995)). ***Suppose*** *that the drift and diffusion, $h(t, z_t)$ and $g(t, z_t)$, of an Itô SDE, defined for $t \geq 0$ and $z_t \in \mathbb{R}^{D_z}$, satisfy three conditions:* ***(i)*** *For each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|h(t, z_t)\|^2 + \|g(t, z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.* ***(ii)*** *For all $T > 0$, $z_t, z'_t \in K$, and $t \in [0, T]$, $\|h(t, z_t) - h(t, z'_t)\| + \|g(t, z_t) - g(t, z'_t)\| \leq C_T \cdot \|z_t - z'_t\|$.* ***(iii)*** *For each $z_t \in K$, $h(t, z_t)$ and $g(t, z_t)$ are continuous.* ***Then*** *z_t is viable in K* ***if and only if****: for all $s \in \{1, \ldots, S\}$ and $z_t \in K$ such that $\langle z_t - u_s, v_s \rangle = 0$, we have* ***(a)*** *$\langle h(t, z_t), v_s \rangle \geq 0$ and* ***(b)*** *$\langle g(t, z_t) \odot e_d, v_s \rangle = 0$ for $t \geq 0$ and $d \in \{1, \ldots, D_z\}$.*

Stationary SDEs. In many applications, it is important to model stationary dynamics (e.g. Tank et al. (2015)), particularly over short time horizons (e.g. Tonekaboni et al. (2021); Weatherhead et al. (2022)). Similarly, in EMA studies of STBs, patients are typically tracked for relatively short periods—often from one to several weeks (Ammerman and Law, 2022)—and their survey responses are often highly noisy due to unobserved external influences. Assuming stationarity for the duration of the study may reduce model complexity and provide a more appropriate inductive bias. We therefore extend Theorem 3 to stationary SDEs on r-polyhedra, which are, informally speaking, compact polyhedra with non-zero "volume" (avoiding polyhedra for which every point is on ∂K):

Definition 4. *A set $K \subset \mathbb{R}^{D_z}$ is an r-polyhedron if it is a compact polyhedron that is also regular: for every $z \in \partial K$ and every $\epsilon > 0$, there exists $z' \in \mathcal{B}(z, \epsilon)$ in a ball of radius ϵ centered at z that lies in the interior of K.*

We do this by selecting a diffusion, g, that satisfies (i)-(iii) and (b) from Theorem 3, deriving a closed-form equation for a drift h as a function of g that ensures stationarity, and proving that h also satisfies all conditions from Theorem 3 (see proof in Appendix C).

Theorem 5. *Let K be an r-polyhedron, and let $h(z_t)$ and $g(z_t)$ be the drift and diffusion, respectively, of an autonomous Itô SDE, defined for $t \geq 0$ and $z_t \in \mathbb{R}^{D_z}$.* ***Suppose*** *that for all $T > 0$, $z_t, z'_t \in K$, and $t \in [0, T]$, there exists $C_T > 0$ such that:* ***(i)*** *$\|g(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.* ***(ii)*** *$\|g(z_t) - g(z'_t)\| \leq C_T \cdot \|z_t - z'_t\|$ and $\|\mathrm{diag}(\nabla_{z_t} g(z_t)) - \mathrm{diag}(\nabla_{z'_t} g(z'_t))\| \leq C_T \cdot \|z_t - z'_t\|$.* ***(iii)*** *The unnormalized time-marginal, $\tilde{p}(z_t)$, has support on $\mathrm{int}(K)$, and $g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t)$ admits a continuous extension to K such that $\left\|g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t) - g(z'_t) \odot \nabla_{z'_t} \log \tilde{p}(z'_t)\right\| \leq C_T \cdot \|z_t - z'_t\|$.* ***Suppose further that:*** *for all $z_t \in \mathrm{int}(K)$,* ***(iv)*** *$g^d(z_t) > 0$ for all $d \in \{1, \ldots, D_z\}$, and* ***(v)*** *for each $s \in \{1, \ldots, S\}$, there exists $L_s < \infty$ such that $\left\|n_s \odot g(z)\right\| \leq L_s \cdot d_s(z)$, where $n_s = v_s / \|v_s\|$ and $d_s(z) = \langle z - u_s, n_s \rangle$.* ***Then*** *for any $z_0 \sim p(z_t)$, the solution z_t is viable in K, with time-marginal $p(z_t)$,* ***provided that:*** ***(a)*** *$h(z_t) = \frac{1}{2}\mathrm{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) + \frac{1}{2} g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t)$, and* ***(b)*** *For all $s \in \{1, \ldots, S\}$ and $z_t \in K$ such that $\langle z_t - u_s, v_s \rangle = 0$, $\langle g(z_t) \odot e_d, v_s \rangle = 0$ for all $d \in \{1, \ldots, D_z\}$.*

The parameterization of $h(z_t)$ in Theorem 5 is easily implemented by obtaining the score function, $\nabla_{z_t} \log \tilde{p}(z_t)$, via auto-differentiation of $\log \tilde{p}(z_t)$, parameterized by an unconstrained NN. Also, we note that assumptions (i)-(v) are easily satisfiable—see discussion in Appendix E.

4 Parameterization of Expressive SDEs on r-Polyhedra

We propose a parameterization that maps arbitrary (neural or expert-given) dynamics into constraint-satisfying dynamics from Theorems 3 and 5.

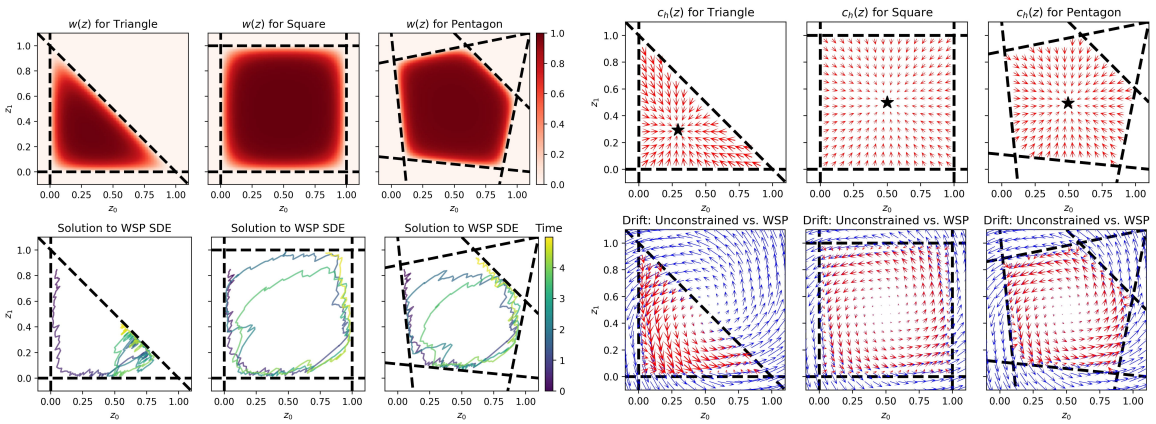

Figure 1. Intuition Underlying WSP. Top left: $w(z)$ from Eq. 4, approaching 0 at the boundaries and 1 in the interior. Top right: $c_h(z)$ from Eq. 5, pointing towards the Chebyshev center ★. Bottom left: solutions to WSP SDE, successfully remaining in K. Bottom right: an "expert-given" drift $\tilde{h}$ vs. WSP drift h (Eq. 5) matching in the interior of K, but differing near the boundary. Details in Appendix G.1.

Weighted Sums Parameterization (WSP). We observe that we can satisfy both constraints on the drift and diffusion using the same mechanism: $\text{WSP}(f, c, t, z) = w(z) \cdot f(t, z) + (1 - w(z)) \cdot c(z)$, using a different choice of $c(z)$ for each. Here, $f(\cdot)$ is the original, unconstrained dynamics, given by domain experts or by some flexible function class, like NNs; $c(\cdot)$ is a simple function that satisfies the constraints; finally, $w(z) \in [0, 1]$ weighs the sum of $f(\cdot)$ and $c(\cdot)$, approaching 0 at ∂K to favor $c(\cdot)$, and approaching 1 at the interior of K to favor $f(\cdot)$. Of many possible choices, we define:

$$w(z) = \tanh\left(\beta \cdot \prod_s \left[\frac{e^{-d_s(z)}}{\sum_{s'} e^{-d_{s'}(z)}} \cdot \tanh\left(\alpha \cdot d_s(z)\right)\right]\right), \qquad d_s(z) = \frac{\langle z - u_s, v_s \rangle}{\|v_s\|}. \tag{4}$$

Here, $d_s(z) \geq 0$ is the shortest distance from z to the boundary of $\mathcal{H}(u_s, v_s)$, and $\alpha, \beta > 0$ determine the transition rate between $f(z)$ and $c(z)$. They can be learned jointly with the model parameters. The intuition behind Eq. 4 is that $w(z)$ should approach 0 as z approaches the *closest* of the S boundaries. As such, we take a product of distances from z to each boundary, weighted by a softmin; this, in a sense, "selects the closest distance." By taking a product of these weighted distances, we obtain a function that is 0 at all boundaries and positive elsewhere, using tanh to ensure $w(z) \in [0, 1]$. Fig. 1 (top-left) visualizes $w(z)$ for different polyhedra.

WSP-based SDEs. Given any unconstrained dynamics, $\tilde{h} : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}$ and $\tilde{g} : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}_{\geq 0}$, WSP returns new dynamics, h and g, that satisfy (a)-(b) from Theorem 3:

$$\begin{aligned} h(t, z_t) &= \text{WSP}(\tilde{h}, c_h, t, z_t), \qquad & c_h(z_t) &= \gamma \cdot \frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}, \\ g(t, z_t) &= \text{WSP}(\tilde{g}, c_g, t, z_t), \qquad & c_g(z_t) &= 0, \end{aligned} \tag{5}$$

where z^* is the Chebyshev center of K, easily computed via linear programming once per polyhedron (Boyd, 2004). Because when $z_t \in \partial K$, $\langle z^* - z_t, v_s \rangle = \langle z^* - u_s, v_s \rangle - \langle z_t - u_s, v_s \rangle = \langle z^* - u_s, v_s \rangle > 0$, $c_h(z_t)$ points to the interior of K with magnitude controlled by $\gamma, \epsilon > 0$, learned jointly with the other model parameters. We choose the Chebyshev center due to its generality and its large margin to the boundary; in future work, we are excited to investigate alternative choices for $c_h(z_t)$ that incorporate domain expertise. Fig. 1 visualizes $c_h(z)$ and h for different polyhedra, showing WSP SDEs remain viable in K. To instantiate WSP for neural SDEs, we can simply set $\tilde{h}, \tilde{g}$ to NNs with linear and softplus activations, respectively. Next, we prove that, under the same mild conditions on $\tilde{h}$ and $\tilde{g}$ (e.g. Lipschitz continuity, etc.), WSP dynamics satisfy the conditions of Theorems 3 and 5

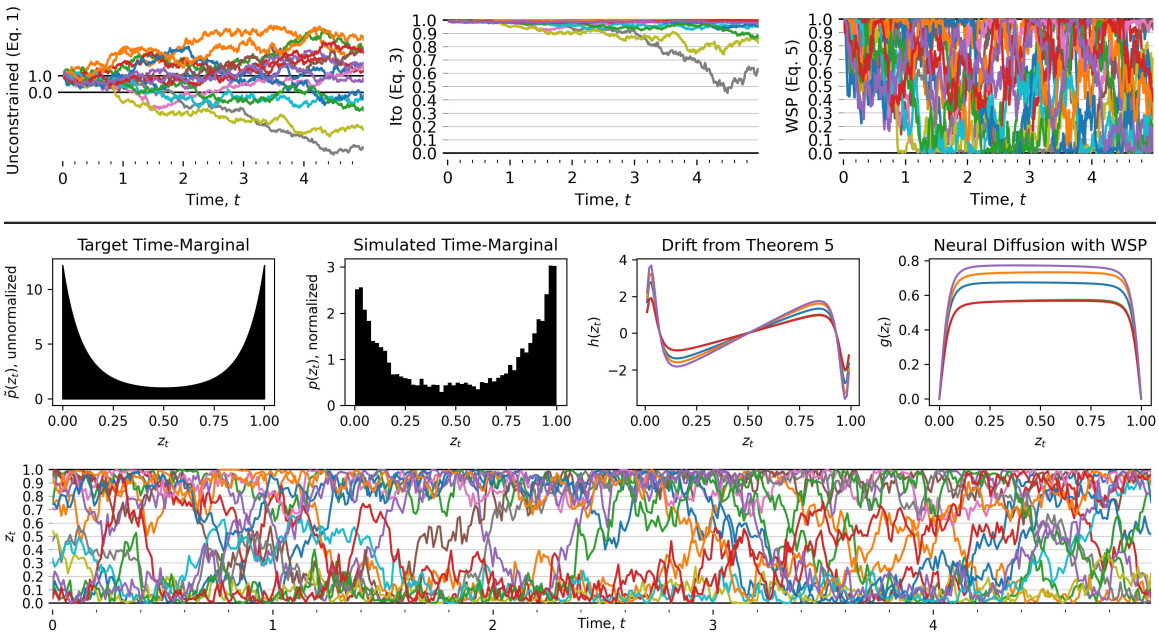

Figure 2. Top: Inductive bias of WSP vs. baselines. Left: unconstrained SDE (Eq. 1) with NN quickly leaves $K = [0, 1]$. Middle: SDE transformed via sigmoid (Eqs. 2 and 3) sticks to the boundary. Right: SDE with WSP (Eq. 5) successfully remains in K. **Bottom: Stationary WSP matches target time-marginal while remaining viable.** Additional results in Appendix I.2.

(proof in Appendix D). One limitation of our proposed $w(z)$ is that α and β need to grow quickly with the number of boundaries S. Below, there is a simple fix for the unit cube.

Theorem 6. *Let K be an r-polyhedron.* ***Suppose*** *that $\tilde{h}$ and $\tilde{g}$, defined above, satisfy conditions* ***(i)-(iii)*** *of Theorem 3.* ***Then*** *the drift and diffusion h and g, defined in Eq. 5, satisfy* ***(i)-(iii)*** *and* ***(a)-(b)*** *from Theorem 3.* ***Suppose further*** *that $\tilde{g}$ additionally satisfies conditions* ***(i)***, ***(ii)***, ***(iv)*** *of Theorem 5;* ***then*** *g, defined in Eq. 5, satisfies* ***(i)-(ii)***, ***(iv)-(v)*** *and* ***(b)*** *from Theorem 5.*

Theorem 6 does not cover condition (iii) from Theorem 5, which concerns the joint behavior of g *and* the score function, $\nabla_{z_t} \log \tilde{p}(z_t)$. One simple way to satisfy (iii) is by ensuring $\nabla_{z_t} \log \tilde{p}(z_t)$ is Lipschitz, enabling WSP to easily satisfy both Theorems 3 and 5.

WSP on the Unit Cube and Simplex. For a unit cube, $K = [0, 1]^{D_z}$, the constraints decouple across dimensions, so we apply WSP element-wise, reducing the number of elements in the product of Eq. 4 from $2D_z$ to 2. For a simplex, we first define $K \subset \mathbb{R}^{D_z-1}$ as the projection of a simplex onto $D_z - 1$ dimensions (e.g. the triangle in the top-left of Fig. 1 is the projection of a 3D simplex). We do this by setting $u, v \in \mathbb{R}^{D_z-1}$ from Definition 2 as follows. For $1 \leq s < D_z$, u_s is the zero vector and $v_s = e_s$. For $s = D_z$, u_s and v_s are vectors filled with $(D_z - 1)^{-1}$ and $-(D_z - 1)^{-1/2}$, respectively. Using WSP, we then define an SDE whose state, z_t, evolves in this projected space, K. The resulting process has $D_z - 1$ dimensions, each nonnegative with a sum ≤ 1. Finally, we recover the full D_z-dimensional simplex by adding the last component, $z_t^{D_z} = 1 - \sum_{s=1}^{D_z-1} z_t^s$, to z_t so that all components are nonnegative and sum to 1.

5 Experiments and Results

We present experiments that stress-test WSP against baselines, comparing inductive biases, faithfulness to expert knowledge, and effects on optimization dynamics.

5.1 Synthetic Data Experiments

WSP maps expert-given dynamics to viable dynamics on arbitrary polyhedra. We *stress-test* WSP by transforming the given dynamics, in which z_t *spirals out* of a target region, to remain viable in three arbitrary polyhedra: triangle, square, pentagon. The bottom-right of Fig. 1 compares the given and the WSP-transformed dynamics in blue and red, respectively, showing that they match on the interior of the space while differing near the boundary. The bottom-left of Fig. 1 then shows that sample trajectories from the WSP-transformed dynamics remain within the polyhedra while still spiraling according to the original dynamics. Details in Appendix G.1.

WSP produces volatile trajectories, qualitatively similar to EMA data, while baselines do not. Due to Limitation 2 (Section 1), initialization plays a crucial role in the success of expressive SDE-based models. As such, to empirically compare the inductive bias of WSP (Eq. 5) against baselines (Eqs. 1–3), we solve SDEs given by NNs h and g with randomly sampled weights. We define the viable region, $K = [0, 1]$, and specifically choose $z_0 = 0.99$ near the boundary to *stress-test* the chain-rule based SDEs in Eqs. 2 and 3 to show that, close to the boundary, they struggle to return to the interior of K (setup in Appendix G.2). While simple, *these experiments show WSP boasts a stark improvement over baselines in its ability to capture volatile data*. Fig. 2 (top) shows WSP ensures SDE samples are viable in K (faithful to expert knowledge), while freely moving across K (appropriate inductive bias for volatile data). Notably, WSP trajectories qualitatively resemble the dynamics of suicidal ideation in EMA data. In contrast, unconstrained SDEs quickly leave K, and SDEs based on Itô's lemma stick to the boundary. In Fig. 2 (with additional results in Appendices I.1 and I.2), we show these behaviors persist under the Karhunen-Loève approximation of Brownian motion (from Appendix F), and in the stationary construction from Theorem 5.

5.2 Real EMA Data Experiments

Latent SDEs. We model patient n's dth EMA response with an ordinal likelihood conditioned on their latent psychological state, $z_t^n \in [0, 1]^{D_z}$, evolving via an autonomous Stratonovich SDE:

$$z_{t_0}^n \sim \mathcal{N}_{[0,1]}(\mu_0, \sigma_0^2 \cdot I_{D_z}), \;\; dz_t^n = h(z_t^n; \theta) \cdot dt + \mathrm{diag}(g(z_t^n; \theta)) \circ dB_t, \;\; x_{t_m}^{n,d} \mid z_{t_m}^{n,d} \sim \mathrm{Cat}\left(\lambda(z_{t_m}^{n,d}; \sigma_\epsilon)\right),$$

where $\mathcal{N}_{[0,1]}$ is a truncated normal and $\lambda(\cdot; \sigma_\epsilon)$ is an ordinal link function with fixed cutpoints in $[0, 1]$. Since traditional variational methods for latent SDEs remain strong competitors to simulation-free methods, we choose the traditional method, but approximate Brownian motion with a truncated Karhunen-Loève expansion, as done by Ghosh et al. (2022), allowing us to replace the SDE solver with a stable, high-order, adaptive ODE solver. Details in Appendix F.

Viable Solvers. Although WSP is viable in continuous time, standard numerical solvers can step outside K due to discretization. Developing solvers with viability guarantees is an important future direction. In our real data experiments, we therefore include a pragmatic safeguard: we clip the state into K before evaluating drift/diffusion to improve numerical stability.

Baselines. We evaluate a set of ablations on 4 real EMA datasets (Appendix H) to isolate the effect of (i) imposing viability through WSP and (ii) applying the clipping safeguard used for numerical stability during training (since viable solvers are an open problem). Specifically, we compare latent SDE models with the dynamics below. Let $(\tilde{h}, \tilde{g})$ be NNs with linear and softplus output activations, respectively, and let $\mathrm{clip}(z) = \min(\max(z, 0), 1)$ denote an element-wise clipping onto the target unit cube. **Vanilla**: $(h, g) = (\tilde{h}, \tilde{g})$. **Vanilla+Clip**: $(h, g) = (\tilde{h} \circ \mathrm{clip}, \tilde{g} \circ \mathrm{clip})$. **WSP**: applies Eq. 5 to $(\tilde{h}, \tilde{g})$, yielding viable dynamics $(h_{\mathrm{WSP}}, g_{\mathrm{WSP}})$. **WSP+Clip**: $(h, g) = (h_{\mathrm{WSP}} \circ \mathrm{clip}, g_{\mathrm{WSP}} \circ \mathrm{clip})$.

Evaluation. We split each patient's time horizon at the median time step: observations after this point form the "forecast set" to stress-test forecasting, while 20% of time steps before it are held out

Table 1. WSP-based latent neural SDEs satisfy constraints; they improve training dynamics and forecasts on *all* EMA datasets. We report log posterior predictive on the interpolation and forecasting sets for models trained on a large-scale EMA suicide-risk study ("SMART") and several EMA mental health datasets (GLOBEM DS2-4)—details in Appendix H. We also report three constraint-satisfaction metrics, summarized below. ↑/↓ indicate whether higher/lower values are better. Details in Appendix G.3.

Data	Method	Log Predictive		Constraint Satisfaction		
		Interpolation ↑	Forecast ↑	DRVP ↑	DIVP ↑	DIDV ↓
SMART	WSP+Clip	**-34.17**	**-187.64**	**1.00**	**1.00**	**0.00**
	WSP	**-34.00**	**-189.23**	**1.00**	**1.00**	**0.00**
	Vanilla+Clip	-44.17	-218.20	**1.00**	0.00	3.62
	Vanilla	-46.50	-226.50	**1.00**	0.00	3.98
DS2	WSP+Clip	**-40.34**	**-103.73**	**1.00**	**1.00**	**0.00**
	WSP	**-46.97**	-124.19	**1.00**	**1.00**	**0.00**
	Vanilla+Clip	-108.19	-182.68	0.89	0.00	6.20
	Vanilla	-139.63	-423.54	0.55	0.00	9.23
DS3	WSP+Clip	**-34.37**	**-95.07**	**1.00**	**1.00**	**0.00**
	WSP	**-34.07**	**-96.12**	**1.00**	**1.00**	**0.00**
	Vanilla+Clip	-110.18	-268.36	0.96	0.00	6.78
	Vanilla	-129.48	-558.02	0.59	0.00	4.15
DS4	WSP+Clip	**-41.67**	**-93.73**	**1.00**	**1.00**	**0.00**
	WSP	**-41.75**	**-93.88**	**1.00**	**1.00**	**0.00**
	Vanilla+Clip	-114.86	-120.47	0.78	0.02	6.86
	Vanilla	-147.43	-246.90	0.58	0.00	8.49

Metrics: DRVP and DIVP measure the fraction of the boundary that satisfies Theorem 3's constraints on the drift and diffusion, respectively. **DIDV** measures the average difference between the diffusion value at the boundary and 0 (where 0 distance indicates a viable diffusion).

at random as the "interpolation set" (evaluating fit within the training window) to assess whether optimization got stuck in poor local optima. We evaluate the log posterior predictive on these two heldout sets, as well as constraint satisfaction metrics on the learned dynamics (Appendix G.3).

Isolating inductive bias under widening forecasts. Because STBs are influenced by external factors, they are intrinsically uncertain; thus, latent SDE posteriors quickly revert to the prior with growing forecast horizons. Naively comparing forecast distributions would therefore mostly reflect whether a model assigns mass outside the measurement domain, rather than whether its *dynamics* better capture the underlying process. We hypothesize that enforcing viability *filters out psychologically inadmissible dynamics, steering learning toward models closer to the ground truth.* We therefore fix the variational variances to force more confident forecasts, making differences in inductive bias (rather than uncertainty inflation) observable. Details in Appendix G.3.

WSP-based latent neural SDEs satisfy constraints, leading to better training dynamics and forecasts on real EMA data over Vanilla neural SDEs. Table 1 shows that: (1) WSP-based dynamics avoid assigning probability mass to impossible outcomes: the constraint-satisfaction metrics are perfect for WSP, by construction, and are far from perfect for the baselines. (2) These constraints guide optimization toward better optima: WSP-based models achieve higher interpolation log predictive performance. Since WSP constrains the dynamics, Vanilla could theoretically match it; thus, the gap suggests Vanilla training often converges to poorer optima. (3) Altogether, the WSP-based model exhibits a higher log-predictive for forecasting, indicating its inductive bias is more appropriate for mental health EMA data. Fig. 3 confirms this qualitatively: in comparison to WSP, the Vanilla dynamics escape the valid $[0, 1]$ space, fit the observed data poorly, and make worse forecasts. Finally, we note that although WSP substantially improves forecasts, it still cannot reliably forecast a patient's true state for the *entirety* of the second half of the study just from the first half—EMA data is too stochastic for *any* model to be this accurate. *Instead, these results suggest that embedding additional clinical knowledge in latent SDEs could yield similarly large gains.*

Ablation: Clipping yields modest gains for Vanilla and little to no gain for WSP. As shown in Table 1, Vanilla+Clip improves over Vanilla on interpolation and forecasting, yet remains substantially worse than WSP, while WSP+Clip is comparable to WSP. This confirms that WSP's gains stem from viability-constrained dynamics, not clipping.

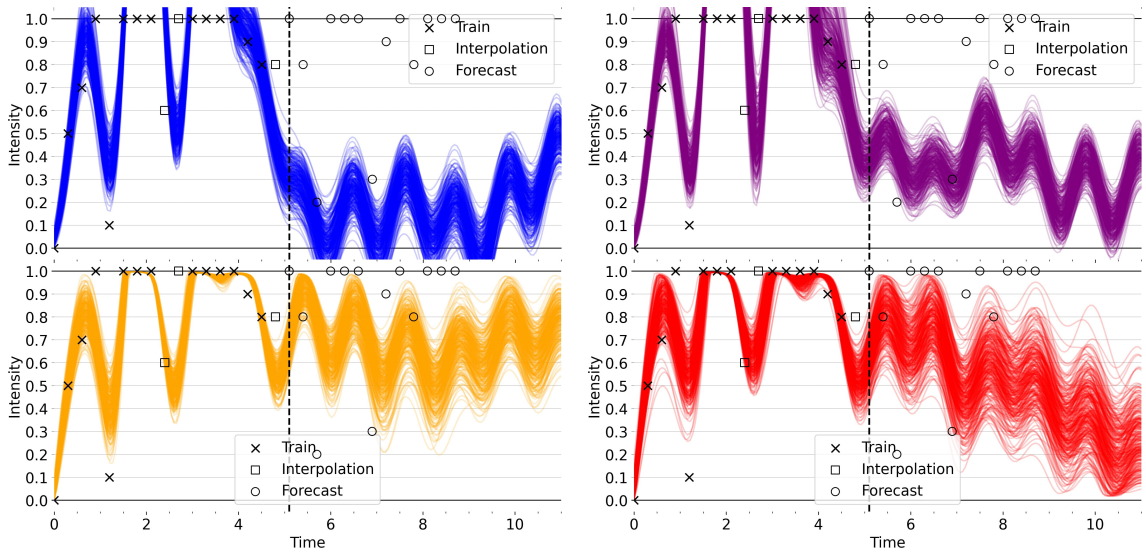

Figure 3. WSP-based latent neural SDE exhibits better inductive bias than Vanilla neural SDE baseline on the "SMART" data. Each plot visualizes posterior samples for a patient, corresponding to EMA item "urge to die," along with data 0-10 Likert-scale EMA items. Top-left: **Vanilla**. Top-right: **Vanilla+Clip**. Bottom-left: **WSP**. Bottom-right: **WSP+Clip**. Additional results in Appendix I.3.

6 Discussion and Future Work

In this work, we introduce a general method for transforming arbitrary—neural or expert-specified—dynamics into SDEs on arbitrary compact polyhedra. Empirically, our method substantially improves optimization and forecasts on EMA data for suicide risk and mental health. These results are a first step toward aligning expressive probabilistic models with clinical knowledge, but trustworthy models for scientific insight and intervention require further work, which we outline next.

Hybrid Modeling to Empirically Test Psychological Theories of Suicide. WSP naturally supports hybrid models that blend mechanistic, expert-specified structure with data-informed neural components, providing a vehicle for empirically testing psychological theories of suicide (Millner et al., 2020). Using WSP, domain experts can encode theoretical constraints into the drift and diffusion—such as directional influences among symptoms—while NNs learn the functional form of these influences. In future work, we plan to (a) instantiate specific theories as families of constrained SDEs, and (b) empirically assess these theories by comparing their forecasting performance.

Encoding Relationships Between EMA Items via Linear Inequality Constraints. Although our main application focuses on constraints for 0–10 Likert scales, the same framework accommodates a richer class of linear inequality constraints to encode relationships among symptoms. For example, if we believe suicidal ideation is driven by an intolerable internal state, we can impose that the "amount" of suicidal ideation is always lower than the "amount" of internal distress. In future work, we plan to systematically test such clinically motivated constraints.

Integration with Other Popular SDE-based Models. Our results may benefit other types of SDE-based models, like diffusion models (Song et al., 2021) and infinitely deep Bayesian NNs (Xu et al., 2022a), where the data often live in bounded or simplex-like domains (e.g. pixel values or probabilities), but the underlying SDEs are typically defined on the entire Euclidean space. In future work, we hope to evaluate whether our constraint-satisfying dynamics translate into improved optimization and performance for other SDE-based models.

Acknowledgements

The authors are grateful for funding from NIMH (U01MH116928), the Fuss Family Research Fund, the Chet and Will Griswold Suicide Prevention Fund, and a gift from an anonymous donor. The authors are also grateful to Wellesley College for supporting YL in the summer of 2025.

References

Brooke A Ammerman and Keyne C Law. Using intensive time sampling methods to capture daily suicidal ideation: a systematic review. *Journal of Affective Disorders*, 299:108–117, 2022.

Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International conference on machine learning*, pages 291–301. PMLR, 2019.

Abdul Fatir Ansari, Alvin Heng, Andre Lim, and Harold Soh. Neural continuous-discrete state space models for irregularly-sampled time series. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 926–951. PMLR, 23–29 Jul 2023.

Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor. Variational inference for diffusion processes. *Advances in neural information processing systems*, 20, 2007.

Jean-Pierre Aubin. *Viability Theory*. Modern Birkhäuser Classics. Birkhauser Boston, Secaucus, NJ, January 1991.

Jonathan T Barron. Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*, 2017.

Grigory Bartosh, Dmitry Vetrov, and Christian A. Naesseth. SDE matching: Scalable and simulation-free training of latent stochastic differential equations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0Hd1lh52Fi.

Denny Borsboom, Jonas MB Haslbeck, and Donald J Robinaugh. Systems-based approaches to mental disorders are the only game in town. *World Psychiatry*, 21(3):420, 2022.

Stephen Boyd. Convex optimization. *Cambridge UP*, 2004.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. Jax: composable transformations of python+ numpy programs. 2018.

Marie-France Bru. Wishart processes. *Journal of Theoretical Probability*, 4(4):725–751, 1991.

GQ Cai and YK Lin. Generation of non-gaussian stationary stochastic processes. *Physical Review E*, 54(1):299, 1996.

Jacob K Christopher, Stephen Baek, and Nando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4):385–396, 1983.

Beau Coker, Wessel P. Bruinsma, David R. Burt, Weiwei Pan, and Finale Doshi-Velez. Wide mean-field bayesian neural networks ignore the data. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5276–5333. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/coker22a.html.

Daniel DL Coppersmith, Oisín Ryan, Rebecca G Fortgang, Alexander J Millner, Evan M Kleiman, and Matthew K Nock. Mapping the timescale of suicidal thinking. *Proceedings of the National Academy of Sciences*, 120(17):e2215434120, 2023.

Jacky Cresson and Stefanie Sonner. A note on a derivation method for sde models: Applications in biology and viability criteria. *Stochastic Analysis and Applications*, 36(2):224–239, 2018.

Jacky Cresson, Bénédicte Puig, and Stefanie Sonner. Validating stochastic models: invariance criteria for systems of stochastic differential equations and the selection of a stochastic hodgkin-huxley type model. *arXiv preprint arXiv:1209.4520*, 2012.

Jacky Cresson, Bénédicte Puig, and Stefanie Sonner. Stochastic models in biology and the invariance problem. *Discrete & Continuous Dynamical Systems-Series B*, 21(7), 2016.

Deng Ding and Ying Ying Zhang. Numerical solutions for reflected stochastic differential equations, 2008.

Alberto d'Onofrio. *Bounded noises in physics, biology, and engineering*. Springer, 2013.

John R Dormand and Peter J Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Behnam Esmayli. Prove that the product of two lipschitz functions is locally lipschitz. Mathematics Stack Exchange, 2017. URL https://math.stackexchange.com/q/2171798. URL:https://math.stackexchange.com/q/2171798 (version: 2021-10-12).

Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael John Hutchinson. Diffusion models for constrained domains. *Transactions on Machine Learning Research*, 2023a. ISSN 2835-8856.

Nic Fishman, Leo Klarner, Emile Mathieu, Michael Hutchinson, and Valentin De Bortoli. Metropolis sampling for constrained diffusion models. *Advances in Neural Information Processing Systems*, 36:62296–62331, 2023b.

Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33: 15897–15908, 2020.

Sanmitra Ghosh, Paul J Birrell, and Daniela De Angelis. Differentiable bayesian inference of sde parameters using a pathwise series expansion of brownian motion. In *International Conference on Artificial Intelligence and Statistics*, pages 10982–10998. PMLR, 2022.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Ary L Goldberger, Luis AN 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.

Prashnna Gyawali, Sandesh Ghimire, and Linwei Wang. Enhancing mixup-based semi-supervised learningwith explicit lipschitz regularization. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1046–1051. IEEE, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations (ICLR'22)*, 2022.

Zacharia Issa, Blanka Horvath, Maud Lemercier, and Cristopher Salvi. Non-adversarial training of neural sdes with signature kernel scores. *Advances in Neural Information Processing Systems*, 36: 11102–11126, 2023.

Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

Patrick Kidger, James Foster, Xuechen Li, and Terry J Lyons. Neural sdes as infinite-dimensional gans. In *International conference on machine learning*, pages 5453–5463. PMLR, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

Kurt Kroenke, Robert L Spitzer, Janet B W Williams, and Bernd Löwe. An ultra-brief screening scale for anxiety and depression: the phq-4. *Psychosomatics*, 50(6):613–612, 2009.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pages 3870–3882. PMLR, 2020.

Hsueh-Ti Derek Liu, Francis Williams, Alec Jacobson, Sanja Fidler, and Or Litany. Learning smooth neural functions via lipschitz regularization. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–13, 2022.

Aaron Lou and Stefano Ermon. Reflected diffusion models. In *International Conference on Machine Learning*, pages 22675–22701. PMLR, 2023.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

Eberhard Mayerhofer, Oliver Pfaffel, and Robert Stelzer. On strong solutions for positive definite jump diffusions. *Stochastic processes and their applications*, 121(9):2072–2086, 2011.

Henry P McKean. *Stochastic integrals*, volume 353. American Mathematical Society, 2024.

Anna Milian. Stochastic viability and a comparison theorem. In *Colloquium Mathematicum*, volume 68, pages 297–316. Polska Akademia Nauk. Instytut Matematyczny PAN, 1995.

Alexander J Millner, Michael D Lee, and Matthew K Nock. Describing and measuring the pathway to suicide attempts: A preliminary study. *Suicide and Life-Threatening Behavior*, 47(3):353–369, 2017.

Alexander J Millner, Donald J Robinaugh, and Matthew K Nock. Advancing the understanding of suicide: The need for formal theory and rigorous descriptive research. *Trends in cognitive sciences*, 24(9):704–716, 2020.

G. N. Mil'shtejn. Approximate integration of stochastic differential equations. *Theory of Probability & Its Applications*, 19(3):557–562, 1975. doi: 10.1137/1119062. URL https://doi.org/10.1137/1119062.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Matthew K Nock and Shirley B Wang. Understanding, predicting, and preventing suicide: Recent advances using digital and computational methods. *Current Directions in Psychological Science*, page 09637214251414021, 2026.

Matthew K. Nock, Evan M. Kleiman, Kate H. Bentley, Rebecca G. Fortgang, Alexander J. Millner, Kelly L. Zuromski, Adam Bear, Alexis Christie, Merryn Daniel, Dylan DeMarco, Lia Follet, Flynn Kelly, Hope Neveux, Onyinye Obi-Obasi, Jordyn R. Ricard, Narise Ramlal, Tida Tambedou, Yaniv Yacoby, Suzanne A. Bird, Ralph Buonapane, Abigail Donovan, Patrick Mair, J. P. Onnela, Rosalind Picard, and Jordan W. Smoller. Using smartphone surveys to predict next-week suicide attempts. *Journal of Psychopathology and Clinical Science*, 2026.

YongKyung Oh, Dongyoung Lim, and Sungil Kim. Stable neural stochastic differential equations in analyzing irregular time series data. In *The Twelfth International Conference on Learning Representations*, 2024.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Chris Ormandy. Linking sampling and stochastic differential equations, 2019. URL https://chrisorm.github.io/SDE-S.html.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

Andrey Pilipenko. *An introduction to stochastic differential equations with reflection*, volume 1. Universitätsverlag Potsdam, 2014.

Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Springer Science & Business Media, 2013.

Donald J Robinaugh, Ria HA Hoekstra, Emma R Toner, and Denny Borsboom. The network approach to psychopathology: A review of the literature 2008–2018 and an agenda for future research. *Psychological medicine*, 50(3):353–366, 2020.

Donald J Robinaugh, Jonas Haslbeck, Lourens J Waldorp, Jolanda J Kossakowski, Eiko I Fried, Alexander J Millner, Richard J McNally, Oisín Ryan, Jill de Ron, Han LJ van der Maas, et al. Advancing the network theory of mental disorders: A computational model of panic disorder. *Psychological review*, 131(6):1482, 2024.

Yousef Rohanizadegan, Stefanie Sonner, and Hermann J Eberl. Discrete attachment to a cellulolytic biofilm modeled by an itô stochastic differential equation. 2020.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Thomas J Santner and Diane E Duffy. A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3):755–758, 1986.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Artur M Schweidtmann, Dongda Zhang, and Moritz Von Stosch. A review and perspective on hybrid modeling methodologies. *Digital Chemical Engineering*, 10:100136, 2024.

Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Alex Tank, Nicholas J. Foti, and Emily B. Fox. Bayesian structure learning for stationary time series. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 872–881, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

Shirley Wang, Donald Robinaugh, Alexander Millner, Rebecca Fortgang, Sharina Hamm, Coby Barrow, and Matthew Nock. Mathematical and computational modeling of suicidal thinking as a complex dynamical system. *PsyArXiv*, 2023. doi: 10.31234/osf.io/b29cs. URL https://doi.org/10.31234/osf.io/b29cs.

David Watson, Lee Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54 (6):1063–1070, 1988.

Addison Weatherhead, Robert Greer, Michael-Alice Moga, Mjaye Mazwi, Danny Eytan, Anna Goldenberg, and Sana Tonekaboni. Learning unsupervised representations for icu timeseries. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 152–168. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/weatherhead22a.html.

Eugene Wong and Moshe Zakai. On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564, 1965.

World Health Organization. Suicide. https://www.who.int/news-room/fact-sheets/detail/suicide, 2025. Accessed March 18, 2026.

Winnie Xu, Ricky TQ Chen, Xuechen Li, and David Duvenaud. Infinitely deep bayesian neural networks with stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pages 721–738. PMLR, 2022a.

Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, et al. Globem dataset: multi-year datasets for longitudinal human behavior modeling generalization. *Advances in neural information processing systems*, 35: 24655–24692, 2022b.

Yaniv Yacoby, Weiwei Pan, and Finale Doshi-Velez. Failure Modes of Variational Autoencoders and Their Effects on Downstream Tasks. *arXiv e-prints*, art. arXiv:2007.07124, July 2020. doi: 10.48550/arXiv.2007.07124.

Yaniv Yacoby, Weiwei Pan, and Finale Doshi-Velez. Mitigating the effects of non-identifiability on inference for bayesian neural networks with latent variables. *Journal of Machine Learning Research*, 23(244):1–54, 2022.

Jianxin Zhang, Josh Viktorov, Doosan Jung, and Emily Pitler. Efficient training of neural stochastic differential equations by matching finite dimensional distributions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xi Nicole Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 37:107198–107224, 2024.

Appendix

Table of Contents

A Summary of Related Work **18**

B Extending Theorem 3 to Stratonovich SDEs on Compact Polyhedra **18**

B.1 Proof of the forward direction 18

B.2 Proof of the reverse direction 20

C Proof of Theorem 5 **20**

C.1 Derivation of a drift sufficient for stationary dynamics 20

C.2 Proof that the sufficient drift can vanish at the boundary 21

C.3 Proof that the boundary is unreachable from the interior 22

C.4 Proof that the sufficient drift satisfies conditions from Theorem 3 25

D Proof of Theorem 6 **27**

D.1 Proof that WSP satisfies Theorem 3 27

D.2 Proof that WSP satisfies Theorem 5 30

E Discussion of Assumptions **32**

F Latent SDEs with Pathwise Expansions for Modeling Suicide Risk **32**

G Experimental Setups **34**

G.1 Experimental Setup for Fig. 1 34

G.2 Experimental Setup for Figs. 2, 4 and 5 35

G.3 Experimental Setup for Table 1 and Figs. 3 and 6–18 35

H Descriptions of Real Data **37**

I Results **39**

I.1 Inductive Bias with Smooth, Pathwise Expansion of Brownian Motion 39

I.2 Stationary SDEs on r-Polyhedra 39

I.3 Visualizations of Latent SDE Posterior Samples from the “SMART” Data 41

I.4 Visualizations of Latent SDE Prior Samples from the “SMART” Data 54

A Summary of Related Work

Approach	Smoothness of Drift/Diffusion	Any r-Polyhedron	Expressive
RSDE (e.g. Pilipenko (2014))	Non-smooth dynamics activated at boundary	✓	Expressive in interior; fixed at boundary
Chain Rule (Section 2)	Smooth on interior; degenerate at boundary	To the best of our knowledge, limited to spaces explored in prior work	✓
Handcrafted Viability Constraints (e.g. Cai and Lin (1996); Cresson et al. (2012); d'Onofrio (2013); Cresson et al. (2016); Cresson and Sonner (2018); Rohanizadegan et al. (2020))	✓	Limited to spaces explored in prior work	Limited to dynamics explored in prior work
This Work (WSP)	✓	✓	✓

B Extending Theorem 3 to Stratonovich SDEs on Compact Polyhedra

Corollary 7. ***Suppose*** *that the drift and diffusion, $h(t, z_t)$ and $g(t, z_t)$, of a Stratonovich SDE, defined for $t \geq 0$ and $z_t \in \mathbb{R}^{D_z}$, satisfy conditions (i)-(iii) from Theorem 3. Suppose further that for all $T > 0$, $z_t, z_t' \in K$, and $t \in [0, T]$, $\|\text{diag}(\nabla_{z_t} g(t, z_t)) - \text{diag}(\nabla_{z_t'} g(t, z_t'))\| \leq C_T \cdot \|z_t - z_t'\|$.* ***Then*** *$z_t$ is viable in compact polyhedron K if and only if (a)-(b) from Theorem 3 hold.*

B.1 Proof of the forward direction

Proof. Given the Stratonovich interpretation of the SDE in Eq. 1,

$$dz_t = h(t, z_t) \cdot dt + (\text{diag} \circ g)(t, z_t) \circ dB_t, \tag{6}$$

we can write the equivalent Itô SDE as follows:

$$dz_t = \hat{h}(t, z_t) \cdot dt + (\text{diag} \circ g)(t, z_t) \cdot dB_t, \tag{7}$$

where,

$$\hat{h}(t, z_t) = h(t, z_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t). \tag{8}$$

Since in Eq. 7, the diffusion is unchanged, we only need to show that when h satisfies (i)-(iii) and (a), so does $\hat{h}$.

Proof that $\hat{h}$ satisfies (i) from Theorem 3. We prove that for each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|\hat{h}(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$. We do this as follows:

$$\|\hat{h}(z_t)\| = \left\| h(t, z_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) \right\| \tag{9}$$

$$\leq \|h(t, z_t)\| + \frac{1}{2} \cdot \|\text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t)\| \tag{10}$$

$$\leq \|h(t, z_t)\| + \frac{1}{2} \cdot \|\text{diag}(\nabla_{z_t} g(t, z_t))\| \cdot \|g(t, z_t)\| \tag{11}$$

$$\leq \underbrace{\sqrt{C'_T \cdot (1 + \|z_t\|^2)}}_{\text{bounded via condition (i)}} + \underbrace{\frac{1}{2} \cdot \|\text{diag}(\nabla_{z_t} g(t, z_t))\|}_{\text{bounded by const. via condition (ii)}} \cdot \underbrace{\sqrt{C'_T \cdot (1 + \|z_t\|^2)}}_{\text{bounded via condition (i)}} \tag{12}$$

The above line can be written in the form $(1 + B) \cdot \sqrt{C'_T \cdot (1 + \|z_t\|^2)}$, which, when squared, gives us an inequality of the form, $\|\hat{h}(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

Proof that $\hat{h}$ satisfies (ii) from Theorem 3. Here, we prove that for all $T > 0$, $z_t, z'_t \in K$, and $t \in [0, T]$, $\|\hat{h}(z_t) - \hat{h}(z'_t)\| \leq C_T \cdot \|z_t - z'_t\|$:

$$\|\hat{h}(z_t) - \hat{h}(z'_t)\| = \left\| h(t, z_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) - h(t, z'_t) - \frac{1}{2} \cdot \text{diag}(\nabla_{z'_t} g(t, z'_t)) \odot g(t, z'_t) \right\| \tag{13}$$

$$= \left\| h(t, z_t) - h(t, z'_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) - \frac{1}{2} \cdot \text{diag}(\nabla_{z'_t} g(t, z'_t)) \odot g(t, z'_t) \right\| \tag{14}$$

$$\leq \|h(t, z_t) - h(t, z'_t)\| + \left\| \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) - \frac{1}{2} \cdot \text{diag}(\nabla_{z'_t} g(t, z'_t)) \odot g(t, z'_t) \right\| \tag{15}$$

Using the trick by Esmayli (2017), we have

$$\begin{aligned} \|\hat{h}(z_t) - \hat{h}(z'_t)\| \leq & \underbrace{\|h(t, z_t) - h(t, z'_t)\|}_{\leq C'_T \cdot \|z_t - z'_t\|} \\ & + \frac{1}{2} \cdot \underbrace{\left\| \text{diag}(\nabla_{z_t} g(t, z_t)) - \text{diag}(\nabla_{z'_t} g(t, z'_t)) \right\|}_{\leq C'_T \cdot \|z_t - z'_t\|} \cdot \|g(t, z_t)\| \\ & + \frac{1}{2} \cdot \left\| \text{diag}(\nabla_{z'_t} g(t, z'_t)) \right\| \cdot \underbrace{\|g(t, z_t) - g(t, z'_t)\|}_{\leq C'_T \cdot \|z_t - z'_t\|} \end{aligned} \tag{16}$$

Finally, since both $g(t, z_t)$ and $\text{diag}(\nabla_{z'_t} g(t, z'_t))$ are Lipschitz on a bounded domain, they can be bounded by a constant.

Proof that $\hat{h}$ satisfies (iii) from Theorem 3. Since $\hat{h}$ is comprised of addition and scaling operations on continuous functions, it is also continuous.

Proof that $\hat{h}$ satisfies (a) from Theorem 3. When $\langle z_t - u_s, v_s \rangle = 0$, we show (a) holds for $\hat{h}$ as follows:

$$\langle \hat{h}(t, z_t), v_s \rangle = \underbrace{\langle h(t, z_t), v_s \rangle}_{\geq 0} + \frac{1}{2} \cdot \sum_{d=1}^{D_z} \underbrace{\frac{\partial g^d(t, z_t)}{\partial z_t^d}}_{\text{bounded}} \cdot \underbrace{g^d(t, z_t) \cdot v_s^d}_{=0} \geq 0. \tag{17}$$

The first term is non-negative. The second term is 0 since $\langle g(t, z_t) \odot e_d, v_s \rangle = 0$ when condition (b) holds for g, and when 0 is multiplied by the partial (bounded thanks to condition (ii) for g), we get 0. Thus, $\langle \hat{h}(t, z_t), v_s \rangle \geq 0$.

□

B.2 Proof of the reverse direction

Proof. Suppose z_t is viable in K as a solution to the Stratonovich SDE. Since the Stratonovich SDE and its equivalent Itô form in Eq. 7 share the same sample paths, z_t is equally viable as a solution to the Itô SDE with drift $\hat{h}$ and diffusion g. From the forward-direction proof, $\hat{h}$ satisfies conditions (i)–(iii) of Theorem 3; and g satisfies conditions (i)–(iii) by assumption. Applying the *only-if* direction of Theorem 3 to the viable Itô SDE with dynamics $(\hat{h}, g)$ yields, for all $s \in \{1, \ldots, S\}$ and $z_t \in K$ such that $\langle z_t - u_s, v_s \rangle = 0$:

$$\langle \hat{h}(t, z_t), v_s \rangle \geq 0, \qquad \langle g(t, z_t) \odot e_d, v_s \rangle = 0 \quad \text{for all } d \in \{1, \ldots, D_z\}. \tag{18}$$

The second of these is exactly condition (b) from Theorem 3 for the Stratonovich SDE. It remains to recover condition (a) for h. At any boundary point where $\langle z_t - u_s, v_s \rangle = 0$, condition (b) gives $g^d(t, z_t) \cdot v_s^d = \langle g(t, z_t) \odot e_d, v_s \rangle = 0$ for every $d \in \{1, \ldots, D_z\}$. Therefore,

$$\langle \hat{h}(t, z_t), v_s \rangle = \left\langle h(t, z_t) + \frac{1}{2} \cdot \operatorname{diag}\left(\nabla_{z_t} g(t, z_t)\right) \odot g(t, z_t), v_s \right\rangle \tag{19}$$

$$= \langle h(t, z_t), v_s \rangle + \frac{1}{2} \cdot \sum_{d=1}^{D_z} \underbrace{\frac{\partial g^d(t, z_t)}{\partial z_t^d}}_{\text{finite by (ii)}} \cdot \underbrace{g^d(t, z_t) \cdot v_s^d}_{=0 \text{ by (b)}} \tag{20}$$

$$= \langle h(t, z_t), v_s \rangle. \tag{21}$$

Since $\langle \hat{h}(t, z_t), v_s \rangle \geq 0$, we conclude $\langle h(t, z_t), v_s \rangle \geq 0$, which is condition (a) from Theorem 3 for h.

□

C Proof of Theorem 5

To prove Theorem 5, we will first derive a sufficient form for the drift, h, as a function of the diffusion, g, so that it induces stationary dynamics with target (unnormalized) time-marginal $\tilde{p}(z_t)$ (Appendix C.1). Then, we show that, although the drift and diffusion vanish at the boundary (Appendix C.2), ∂K is unreachable from the interior, so the boundary does not form an absorbing state (Appendix C.3). Finally, we will show that the drift satisfies all conditions from Theorem 3, implying that it is viable in K (Appendix C.4).

C.1 Derivation of a drift sufficient for stationary dynamics

We find h by drawing inspiration from the derivation of Cai and Lin (1996), which sets the Fokker–Planck–Kolmogorov (FPK) equation to 0 to obtain stationarity and then solves for the dynamics. In contrast, instead of solving for the diffusion (which typically requires computing intractable integrals), we solve for the drift. This derivation parallels classical Langevin dynamics and stochastic gradient MCMC (e.g. Ma et al. (2015); Ormandy (2019)).

We begin with a general autonomous Itô SDE of the form,

$$dz_t = h(z_t) \cdot dt + g(z_t) \cdot dB_t, \tag{22}$$

where $g(z_t) \in \mathbb{R}^{D_z \times D_z}$ is a full matrix. Let $G(z_t) = g(z_t)\Sigma g(z_t)^\intercal$, where Σ is the covariance of the Brownian motion. Stationarity means $\frac{\partial}{\partial t}p(t, z_t) = 0$. Plugging in the FPK equation gives:

$$0 = -\sum_{d=1}^{D_z} \frac{\partial}{\partial z_t^d}\Big(h^d(z_t)p(z_t)\Big) + \frac{1}{2}\sum_{d=1}^{D_z}\sum_{d'=1}^{D_z} \frac{\partial^2}{\partial z_t^d \partial z_t^{d'}}\Big(G^{d,d'}(z_t)p(z_t)\Big) \tag{23}$$

$$= \sum_{d=1}^{D_z} \frac{\partial}{\partial z_t^d}\left(-h^d(z_t)p(z_t) + \frac{1}{2}\sum_{d'=1}^{D_z} \frac{\partial}{\partial z_t^{d'}}\Big(G^{d,d'}(z_t)p(z_t)\Big)\right). \tag{24}$$

For this theorem, we only consider identity-covariance Brownian motion and diagonal diffusion: $\Sigma = I$ and $g(z_t) \in \mathbb{R}^{D_z}_{\geq 0}$ interpreted as $\mathrm{diag}(g(z_t))$. Then G is diagonal with entries $g^d(z_t)^2$, and the stationary FPK equation simplifies to

$$0 = \sum_{d=1}^{D_z} \frac{\partial}{\partial z_t^d}\left(-h^d(z_t)p(z_t) + \frac{1}{2}\frac{\partial}{\partial z_t^d}\Big(g^d(z_t)^2 p(z_t)\Big)\right). \tag{25}$$

One sufficient (strong) way to satisfy this is to enforce the element-wise identities: for every $d \in \{1, \ldots, D_z\}$,

$$h^d(z_t)p(z_t) = \frac{1}{2}\frac{\partial}{\partial z_t^d}\Big(g^d(z_t)^2 p(z_t)\Big). \tag{26}$$

Solving for h^d and using $p(z_t) = \tilde{p}(z_t)/A$ yields:

$$h^d(z_t) = \frac{1}{2p(z_t)}\frac{\partial}{\partial z_t^d}\Big(g^d(z_t)^2 p(z_t)\Big) \tag{27}$$

$$= \frac{1}{2\tilde{p}(z_t)}\frac{\partial}{\partial z_t^d}\Big(g^d(z_t)^2 \tilde{p}(z_t)\Big) \tag{28}$$

$$= \frac{1}{2}\frac{\partial}{\partial z_t^d}\big[g^d(z_t)^2\big] + \frac{1}{2}g^d(z_t)^2 \frac{\partial}{\partial z_t^d}\log \tilde{p}(z_t). \tag{29}$$

Equivalently,

$$h(z_t) = \frac{1}{2}\mathrm{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) + \frac{1}{2}g(z_t)^2 \odot \underbrace{\nabla_{z_t}\log \tilde{p}(z_t)}_{\text{score function}}, \tag{30}$$

giving us condition (a) from Theorem 5.

C.2 Proof that the sufficient drift can vanish at the boundary

We now show that the stationary construction from Eq. 30 can be *degenerate* at boundary points if (i) the diffusion vanishes on ∂K as a convenient way to enforce viability and (ii) the score remains bounded on the boundary. In that case, $dz_t = 0$, indicating a potential absorbing state that would prevent the intended stationary density on $\mathrm{int}(K)$ from being realized. In Appendix C.3, we then prove that under the assumptions from Theorem 5, the boundary is unreachable from the interior of K, fixing the problem.

Lemma 8. *Let $z_t^\partial \in \partial K$ be any boundary point. Consider the stationary drift construction from Eq. 30, based on a corresponding diffusion, $g(z_t)$, that satisfies all conditions from Theorem 3, in part, using the following, convenient properties:*

(A1) $g^d(z_t) \geq 0$ for all $z_t \in \mathbb{R}^{D_z}$ and all $d \in \{1, \ldots, D_z\}$;

(A2) $g(z_t^\partial) = 0$;

(A3) $\|\nabla_{z_t} \log \tilde{p}(z_t)\|_2 < \infty$.

Then $h(z_t^\partial) = 0$. *In particular, at* z_t^∂ *both drift and diffusion vanish; hence if the SDE reaches* z_t^∂, *it remains on the boundary.*

Proof. By (A1), $g^d(z_t)^2 \geq 0$ and by (A2), $g^d(z_t^\partial) = 0$, z_t^∂ is the global minimizer of $g^d(z_t)^2$; therefore,

$$\frac{\partial}{\partial z_t^d}\left[g^d(z_t)^2\right]\Big|_{z_t = z_t^\partial} = 0 \qquad \forall d \in \{1, \ldots, D_z\}, \tag{31}$$

and

$$\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\Big|_{z_t = z_t^\partial} = 0. \tag{32}$$

Next, by (A2) and (A3),

$$g(z_t^\partial)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t^\partial) = 0. \tag{33}$$

Substituting these into Eq. 30 gives $h(z_t^\partial) = 0$. Finally, since $g(z_t^\partial) = 0$ as well, the SDE has $dz_t = 0$ at z_t^∂, so z_t^∂ is absorbing. □

C.3 Proof that the boundary is unreachable from the interior

We now prove that, under the assumptions from Theorem 5, the boundary is unreachable from the interior, ensuring the induced target marginal remains valid. Our proof draws on "McKean's Argument" (e.g. Section 2.9, Problem 7 from McKean (2024), Proposition 1 from Bru (1991), or Proposition 4.3 by Mayerhofer et al. (2011)). Specifically, we first derive a 1D SDE characterizing the distance from z_t to boundary s, and apply Itô's lemma to the logarithm of the distance to the boundary. By leveraging the linear bounds on the drift and diffusion, we show that the deterministic components of this log-distance remain finite. We then apply Doob's L^2 martingale convergence theorem to the remaining stochastic integral to show that it also converges to a finite limit almost surely. Since reaching the boundary in finite time would require the log-distance to diverge to $-\infty$, this finite limit yields a contradiction, thus proving the boundary is never reached.

Lemma 9. *Assume the conditions of Theorem 5, and let* $z_0 \in \operatorname{int}(K)$. *Then*

$$\mathbb{P}\left(\exists t \in [0, \infty) : \; z_t \in \partial K\right) = 0. \tag{34}$$

Equivalently, the boundary is unreachable from the interior.

Proof. Define the per-boundary hitting times and the first exit time from K:

$$\tau_s = \inf\{t \geq 0 : \; d_s(z_t) = 0\}, \qquad \tau = \inf\{t \geq 0 : \; z_t \in \partial K\} = \min_s \tau_s. \tag{35}$$

Since $z_0 \in \operatorname{int}(K)$ and z_t is continuous (the SDE has Lipschitz, linearly bounded dynamics by (i)–(ii), so a unique strong solution exists), we have $z_t \in \operatorname{int}(K)$ for all $t \in [0, \tau)$. On the event $\{\tau < \infty\}$, continuity of z_t and each $d_s(z_t)$ ensures there exists $s^* \in \{1, \ldots, S\}$ such that $d_{s^*}(z_\tau) = 0$ and $\tau_{s^*} = \tau$; fix such an s^*. We then define the distance-to-boundary process for boundary s^*:

$$y_t = d_{s^*}(z_t) = \langle z_t - u_{s^*}, n_{s^*} \rangle, \qquad n_{s^*} = \frac{v_{s^*}}{\|v_{s^*}\|}. \tag{36}$$

Since $z_0 \in \operatorname{int}(K)$, we have $y_0 > 0$, and $y_\tau = 0$ on $\{\tau < \infty\}$. All bounds below that are stated for $z_t \in K$ apply on $[0, \tau)$, since $z_t \in \operatorname{int}(K) \subset K$ throughout this interval.

A 1D SDE for y_t on $[0, \tau)$. Because $d_{s^*}(\cdot)$ is affine, $\nabla_{z_t} d_{s^*}(z) = n_{s^*}$ and $\nabla^2_{z_t} d_{s^*}(z) = 0$. Applying Itô's lemma yields, for $t < \tau$,

$$dy_t = \langle n_{s^*}, h(z_t)\rangle dt + (n_{s^*} \odot g(z_t))^\intercal \cdot dB_t. \tag{37}$$

We define the scalar drift and diffusion magnitudes,

$$\mu_{s^*}(z_t) = \langle n_{s^*}, h(z_t)\rangle, \qquad \sigma_{s^*}(z_t) = \|n_{s^*} \odot g(z_t)\|_2. \tag{38}$$

By assumption (iv) in Theorem 5, $g(z) > 0$ for all $z \in \text{int}(K)$, hence $\sigma_{s^*}(z_t) > 0$ for all $t < \tau$. Next, we define the 1D process,

$$\widetilde{B}_t = \int_0^t \frac{(n_{s^*} \odot g(z_u))^\intercal}{\sigma_{s^*}(z_u)} \cdot dB_u, \qquad t < \tau, \tag{39}$$

where dB_u denotes the D_z-dimensional Brownian increment. By Lévy's characterization of Brownian motion, $\widetilde{B}_t$ is a standard 1D Brownian motion on $[0, \tau)$, because $\widetilde{B}_0 = 0$ and because its quadratic variation up to t equals t:

$$\int_0^t \left\| \frac{(n_{s^*} \odot g(z_u))}{\sigma_{s^*}(z_u)} \right\|_2^2 \cdot du = \int_0^t 1 \cdot du = t, \qquad t < \tau. \tag{40}$$

Therefore, for $t < \tau$, we can write Eq. 37 as:

$$dy_t = \mu_{s^*}(z_t) \cdot dt + \sigma_{s^*}(z_t) \cdot d\widetilde{B}_t. \tag{41}$$

Bounding $\sigma_{s^*}(z)$ and $\mu_{s^*}(z)$ linearly in $d_{s^*}(z)$. By assumption (v), there exists $L_{s^*} < \infty$ such that for all $z \in K$,

$$\sigma_{s^*}(z) = \|n_{s^*} \odot g(z)\|_2 \leq L_{s^*} d_{s^*}(z). \tag{42}$$

Next, we bound $\mu_{s^*}(z_t) = \langle n_{s^*}, h(z_t)\rangle$ linearly in $d_{s^*}(z_t)$. Using condition (a) from Theorem 5,

$$\mu_{s^*}(z_t) = \left\langle n_{s^*}, \frac{1}{2}\text{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\right\rangle + \left\langle n_{s^*}, \frac{1}{2} g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t)\right\rangle. \tag{43}$$

For diagonal diffusion, the dth component of $\text{diag}(\nabla_{z_t}[g(z_t)^2])$ is $\frac{\partial}{\partial z_t^d}(g^d(z_t)^2) = 2g^d(z_t)\frac{\partial}{\partial z_t^d} g^d(z_t)$; hence,

$$\left\langle n_{s^*}, \frac{1}{2}\text{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\right\rangle = \sum_{d=1}^{D_z} n^d_{s^*} g^d(z_t) \frac{\partial}{\partial z_t^d} g^d(z_t) \tag{44}$$

$$= \langle n_{s^*} \odot g(z_t), \text{diag}(\nabla_{z_t} g(z_t))\rangle. \tag{45}$$

Similarly,

$$\left\langle n_{s^*}, \frac{1}{2} g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t)\right\rangle = \frac{1}{2}\sum_{d=1}^{D_z} n^d_{s^*} g^d(z_t)^2 (\nabla_{z_t} \log \tilde{p}(z_t))^d \tag{46}$$

$$= \frac{1}{2}\langle n_{s^*} \odot g(z_t), g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t)\rangle. \tag{47}$$

This gives us:

$$\mu_{s^*}(z_t) = \langle n_{s^*} \odot g(z_t), \text{diag}(\nabla_{z_t} g(z_t))\rangle + \frac{1}{2}\langle n_{s^*} \odot g(z_t), g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t)\rangle \tag{48}$$

Therefore, by the Cauchy-Schwarz inequality,

$$\|\mu_{s^*}(z_t)\|_2 \leq \|n_{s^*} \odot g(z_t)\|_2 \cdot \|\text{diag}(\nabla_{z_t} g(z_t))\|_2 + \frac{1}{2}\|n_{s^*} \odot g(z_t)\|_2 \cdot \|g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t)\|_2 \tag{49}$$

Now, we define the finite constants (since K is compact and the relevant quantities are continuous by (i)–(iii)):

$$M_{\nabla g} = \sup_{z_t \in K} \|\mathrm{diag}(\nabla_{z_t} g(z_t))\|_2, \quad M_{gs} = \sup_{z \in K} \|g(z) \odot \nabla \log \tilde{p}(z)\|_2. \tag{50}$$

Using assumption (v), $\|n_{s^*} \odot g(z_t)\|_2 \leq L_{s^*} d_{s^*}(z_t)$, and plugging into Eq. 49 yields, for all $z_t \in K$,

$$\|\mu_{s^*}(z_t)\|_2 \leq L_{s^*} d_{s^*}(z_t) M_{\nabla g} + \frac{1}{2} L_{s^*} d_{s^*}(z_t) M_{gs} = \underbrace{L_{s^*} \cdot \left(M_{\nabla g} + \frac{1}{2} M_{gs} \right)}_{A_{s^*}} \cdot d_{s^*}(z_t). \tag{51}$$

Thus, both $\sigma_{s^*}(z)$ and $\mu_{s^*}(z)$ are linearly bounded in $d_{s^*}(z)$.

Un-reachability of $y_t = 0$ in finite time. For $t < \tau$, we have $y_t = d_{s^*}(z_t) > 0$, so $\log y_t$ is well-defined. Applying Itô's lemma to $f(y_t) = \log y_t$ in Eq. 41 gives,

$$d \log y_t = \left(\frac{\mu_{s^*}(z_t)}{y_t} - \frac{1}{2} \left(\frac{\sigma_{s^*}(z_t)}{y_t} \right)^2 \right) dt + \frac{\sigma_{s^*}(z_t)}{y_t} d\widetilde{B}_t. \tag{52}$$

By the linear bounds in Eqs. 42 and 51 and since $y_t = d_{s^*}(z_t) > 0$ for $t < \tau$, the ratios are uniformly bounded:

$$\left| \frac{\mu_{s^*}(z_t)}{y_t} \right| \leq A_{s^*}, \qquad 0 \leq \frac{\sigma_{s^*}(z_t)}{y_t} \leq L_{s^*}. \tag{53}$$

Integrating Eq. 52 from 0 to $t < \tau$ yields:

$$\log y_t = \underbrace{\log y_0}_{\text{finite}} + \int_0^t \underbrace{\left(\frac{\mu_{s^*}(z_u)}{y_u} - \frac{1}{2} \left(\frac{\sigma_{s^*}(z_u)}{y_u} \right)^2 \right)}_{\text{bounded by } A_{s^*} + \frac{1}{2} L_{s^*}^2} du + \underbrace{\int_0^t \frac{\sigma_{s^*}(z_u)}{y_u} d\widetilde{B}_u}_{M_t}. \tag{54}$$

The term inside the deterministic integral is bounded by $A_{s^*} + \frac{1}{2} L_{s^*}^2$, so the deterministic integral is finite for any finite $t < \tau$. The stochastic term, M_t, is a continuous local martingale on $[0, \tau)$ (as is any Itô integral with a locally square-integrable integrand). We next show that $\lim_{t \uparrow \tau} M_t$ exists and is finite a.s. in the event $\{\tau < \infty\}$.

The difficulty is that M_t is only defined on $[0, \tau)$, since the integrand $\frac{\sigma_{s^*}(z_t)}{y_t}$ involves $1/y_t$, which may blow up as $t \uparrow \tau$. In particular, although M_t is a local martingale, we cannot directly apply martingale convergence to it. Instead, for each fixed $T < \infty$, we introduce the stopped process M_t^T, to which we can apply the martingale convergence theorem and connect it back to our original M_t. We define the stopped process,

$$M_t^T = \int_0^{t \wedge \tau \wedge T} \frac{\sigma_{s^*}(z_u)}{y_u} d\widetilde{B}_u, \tag{55}$$

which agrees with M_t for all $t < \tau \wedge T$, but is frozen at its value at time $\tau \wedge T$ thereafter. The key advantage is that stopping at the deterministic time T bounds the integration horizon; combined with the bound from Eq. 53, the integrand of M_t^T is bounded, so M_t^T is a well-defined Itô integral with continuous sample paths. Because $M_0^T = 0$ and the quadratic variation of M_t^T, $\langle M_t^T \rangle_\infty$, satisfies,

$$\mathbb{E}\left[\langle M_t^T \rangle_\infty \right] = \mathbb{E}\left[\lim_{t \to \infty} \int_0^{t \wedge \tau \wedge T} \left(\frac{\sigma_{s^*}(z_u)}{y_u} \right)^2 du \right] = \mathbb{E}\left[\int_0^{\tau \wedge T} \left(\frac{\sigma_{s^*}(z_u)}{y_u} \right)^2 du \right] \leq L_{s^*}^2 \cdot T < \infty, \tag{56}$$

both conditions of Revuz and Yor (2013, Proposition 1.23) are satisfied, so M_t^T is a (true) square-integrable martingale. We can therefore apply Doob's L^2 martingale convergence theorem, which says, $M_\infty^T := \lim_{t\to\infty} M_t^T$ exists and is finite a.s. Relating M_t^T back to M_t, in the event $\{\tau \le T\}$,

$$\lim_{t\uparrow\tau} M_t = \lim_{t\uparrow\tau} M_t^T = M_\tau^T = M_\infty^T, \tag{57}$$

which is finite a.s. And since T is arbitrary, $\lim_{t\uparrow\tau} M_t$ is finite a.s. in the event $\{\tau < \infty\}$.

Finally, we derive a contradiction to show the impossibility of the event $\{\tau < \infty\}$. In this event, since all three terms in Eq. 54 are a.s. finite, $\lim_{t\uparrow\tau} \log y_t$ is finite, and therefore,

$$\lim_{t\uparrow\tau} y_t = \exp\left(\lim_{t\uparrow\tau} \log y_t\right) > 0. \tag{58}$$

However, by continuity of z_t and d_{s^*}, and since $\tau_{s^*} = \tau$ means z_τ lies exactly on boundary s^*,

$$\lim_{t\uparrow\tau} y_t = y_\tau = d_{s^*}(z_\tau) = 0. \tag{59}$$

These two conclusions are contradictory, so the event $\{\tau < \infty\}$ must have probability zero. Therefore $\mathbb{P}(\tau < \infty) = 0$, and so $\mathbb{P}\left(\exists t \in [0,\infty) : z_t \in \partial K\right) = 0$.

□

C.4 Proof that the sufficient drift satisfies conditions from Theorem 3

Since Appendix C.3 shows the solution to the SDE will never reach the boundary, we already know it is viable. Still, for completeness, we show here that the drift for the stationary dynamics satisfies conditions from Theorem 3, allowing us to apply Theorem 3 as well.

Proof that h satisfies (i) from Theorem 3. Here, we prove that for each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0,T]$, $\|h(z_t)\|^2 \le C_T \cdot (1 + \|z_t\|^2)$. We do this as follows:

$$\begin{aligned}
\|h(z_t)\| &= \frac{1}{2} \cdot \left\|\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) + g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t)\right\| && (60)\\
&\le \frac{1}{2} \cdot \left\|\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\right\| + \frac{1}{2}\left\|g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t)\right\| && (61)\\
&\le \frac{1}{2} \cdot \underbrace{\left\|\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\right\|}_{①} + \frac{1}{2}\underbrace{\|g(z_t)\|}_{M_g} \cdot \underbrace{\|g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t)\|}_{M_{gs}} && (62)
\end{aligned}$$

where $M_g = \sup_{z_t \in K} \|g(z_t)\|_2$ by continuity of $g(z_t)$ on a compact K. Next, we bound the gradient of g using condition (ii):

$$\begin{aligned}
① &= \left\|\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\right\| && (63)\\
&= \left\|\lim_{\epsilon\to 0} \begin{bmatrix} \frac{g^1(z_t+\epsilon\cdot e_1)^2 - g^1(z_t)^2}{\epsilon} \\ \vdots \\ \frac{g^{D_z}(z_t+\epsilon\cdot e_{D_z})^2 - g^{D_z}(z_t)^2}{\epsilon} \end{bmatrix}\right\| && (64)\\
&= \lim_{\epsilon\to 0}\left\|\begin{bmatrix} \frac{g^1(z_t+\epsilon\cdot e_1)^2 - g^1(z_t)^2}{\epsilon} \\ \vdots \\ \frac{g^{D_z}(z_t+\epsilon\cdot e_{D_z})^2 - g^{D_z}(z_t)^2}{\epsilon} \end{bmatrix}\right\| \quad \text{(by continuity of the norm)} && (65)\\
&\le \sum_{d=1}^{D_z} \lim_{\epsilon\to 0}\left\|\frac{g^d(z_t+\epsilon\cdot e_d)^2 - g^d(z_t)^2}{\epsilon}\right\| \quad \text{(since the } \ell_2\text{-norm is upper bounded by the } \ell_1\text{-norm)} && (66)
\end{aligned}$$

$$= \sum_{d=1}^{D_z} \lim_{\epsilon \to 0} \frac{1}{\epsilon} \cdot \left\| g^d(z_t + \epsilon \cdot e_d)^2 - g^d(z_t)^2 \right\| \tag{67}$$

$$\leq \sum_{d=1}^{D_z} \lim_{\epsilon \to 0} \frac{1}{\epsilon} \cdot \left\| g(z_t + \epsilon \cdot e_d)^2 - g(z_t)^2 \right\| \tag{68}$$

$$= \sum_{d=1}^{D_z} \lim_{\epsilon \to 0} \frac{1}{\epsilon} \cdot \left\| (g(z_t + \epsilon \cdot e_d) - g(z_t)) \odot g(z_t + \epsilon \cdot e_d) - g(z_t) \odot (g(z_t) - g(z_t + \epsilon \cdot e_d)) \right\| \tag{69}$$

$$\leq \sum_{d=1}^{D_z} \lim_{\epsilon \to 0} \frac{1}{\epsilon} \cdot \left\| (g(z_t + \epsilon \cdot e_d) - g(z_t)) \odot g(z_t + \epsilon \cdot e_d) \right\| + \frac{1}{\epsilon} \cdot \left\| g(z_t) \odot (g(z_t) - g(z_t + \epsilon \cdot e_d)) \right\| \tag{70}$$

$$\leq \sum_{d=1}^{D_z} \lim_{\epsilon \to 0} \frac{1}{\epsilon} \cdot \left\| g(z_t + \epsilon \cdot e_d) - g(z_t) \right\| \cdot \left\| g(z_t + \epsilon \cdot e_d) \right\| + \frac{1}{\epsilon} \cdot \left\| g(z_t) \right\| \cdot \left\| g(z_t) - g(z_t + \epsilon \cdot e_d) \right\| \tag{71}$$

$$\leq \sum_{d=1}^{D_z} \lim_{\epsilon \to 0} \frac{M_g}{\epsilon} \cdot \left\| g(z_t + \epsilon \cdot e_d) - g(z_t) \right\| + \frac{M_g}{\epsilon} \cdot \left\| g(z_t) - g(z_t + \epsilon \cdot e_d) \right\| \tag{72}$$

$$\leq \sum_{d=1}^{D_z} \lim_{\epsilon \to 0} \frac{2 \cdot M_g}{\epsilon} \cdot C_T \cdot \left\| \epsilon \cdot e_d \right\| \tag{73}$$

$$= 2 \cdot D_z \cdot M_g \cdot C_T \cdot \left\| e_d \right\| \tag{74}$$

$$= 2 \cdot D_z \cdot M_g \cdot C_T \tag{75}$$

where, in Eq. 69, we use the trick from Esmayli (2017). Putting all of this together, we have that $\|h(z_t)\|$ is bounded by some constant, for which we can always find a new constant C_T to further bound it: $\|h(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

Proof that h satisfies (ii) from Theorem 3. Here, we prove that for all $T > 0$, $z_t, z_t' \in K$, and $t \in [0, T]$, $\|h(z_t) - h(z_t')\| \leq C_T \cdot \|z_t - z_t'\|$.

We begin as follows:

$$\begin{aligned} \|h(z_t) - h(z_t')\| \leq \frac{1}{2} \cdot & \underbrace{\left\| \text{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) - \text{diag}\left(\nabla_{z_t'}[g(z_t')^2]\right) \right\|}_{②} \\ & + \frac{1}{2} \cdot \underbrace{\left\| g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) - g(z_t')^2 \odot \nabla_{z_t'} \log \tilde{p}(z_t') \right\|}_{③}. \end{aligned} \tag{76}$$

Using the trick by Esmayli (2017) again, we bound ② as follows:

$$② = \left\| \text{diag}\left(\nabla_{z_t'}[g(z_t')^2]\right) - \text{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) \right\| \tag{77}$$

$$= 2 \cdot \left\| g(z_t') \odot \text{diag}\left(\nabla_{z_t'} g(z_t')\right) - g(z_t) \odot \text{diag}\left(\nabla_{z_t} g(z_t)\right) \right\| \tag{78}$$

$$= 2 \cdot \left\| (g(z_t') - g(z_t)) \odot \text{diag}\left(\nabla_{z_t'} g(z_t')\right) - g(z_t) \odot \left(\text{diag}\left(\nabla_{z_t} g(z_t)\right) - \text{diag}\left(\nabla_{z_t'} g(z_t')\right)\right) \right\| \tag{79}$$

$$\leq 2 \cdot \left\| (g(z_t') - g(z_t)) \odot \text{diag}\left(\nabla_{z_t'} g(z_t')\right) \right\| + 2 \cdot \left\| g(z_t) \odot \left(\text{diag}\left(\nabla_{z_t} g(z_t)\right) - \text{diag}\left(\nabla_{z_t'} g(z_t')\right)\right) \right\| \tag{80}$$

$$\leq 2 \cdot \left\| g(z_t') - g(z_t) \right\| \cdot \left\| \text{diag}\left(\nabla_{z_t'} g(z_t')\right) \right\| + 2 \cdot \left\| g(z_t) \right\| \cdot \left\| \text{diag}\left(\nabla_{z_t} g(z_t)\right) - \text{diag}\left(\nabla_{z_t'} g(z_t')\right) \right\| \tag{81}$$

$$\leq 2 \cdot C_T \cdot \left\| z_t - z_t' \right\| \cdot \underbrace{\left\| \text{diag}\left(\nabla_{z_t'} g(z_t')\right) \right\|}_{\text{bounded by const. via (ii)}} + 2 \cdot M_g \cdot \underbrace{\left\| \text{diag}\left(\nabla_{z_t} g(z_t)\right) - \text{diag}\left(\nabla_{z_t'} g(z_t')\right) \right\|}_{\leq C_T \cdot \|z_t - z_t'\| \text{ via (ii)}} \tag{82}$$

We similarly bound ③ as follows:

$$③ = \left\| g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) - g(z_t')^2 \odot \nabla_{z_t'} \log \tilde{p}(z_t') \right\| \tag{83}$$

$$= \left\| g(z_t) \odot \left(g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t) \right) - g(z_t') \odot \left(g(z_t') \odot \nabla_{z_t'} \log \tilde{p}(z_t') \right) \right\| . \tag{84}$$

Using the trick by Esmayli (2017),

$$③ \leq \underbrace{\|g(z_t) - g(z_t')\|}_{\leq C_T \cdot \|z_t - z_t'\| \text{ via (ii)}} \cdot \underbrace{\|g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t)\|}_{\leq M_{gs}} \tag{85}$$

$$+ \underbrace{\|g(z_t')\|}_{\leq M_g} \cdot \underbrace{\left\| g(z_t) \odot \nabla_{z_t} \log \tilde{p}(z_t) - g(z_t') \odot \nabla_{z_t'} \log \tilde{p}(z_t') \right\|}_{\leq C_T \cdot \|z_t - z_t'\| \text{ via (iii)}} \tag{86}$$

$$\leq (M_{gs} + M_g) \cdot C_T \cdot \|z_t - z_t'\|, \tag{87}$$

Proof that h satisfies (iii) from Theorem 3. Here, we prove that for each $z_t \in K$, $h(z_t)$, defined for $t \geq 0$, is continuous.

Since continuous functions are closed under all operations used to define $h(z_t)$, and since $h(z_t)$ is defined in terms of other continuous functions, it is also continuous.

Proof that h satisfies (a) from Theorem 3. Here, we prove that, for all $s \in \{1, \ldots, S\}$ and $z_t \in K$ such that $\langle z_t - u_s, v_s \rangle = 0$, we have $\langle h(z_t), v_s \rangle \geq 0$.

$$\langle h(z_t), v_s \rangle = \langle g(z_t) \odot \operatorname{diag}\left(\nabla_{z_t} g(z_t)\right), v_s \rangle + \frac{1}{2} \cdot \langle g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t), v_s \rangle \tag{88}$$

$$= v_s^\intercal \cdot \left(g(z_t) \odot \operatorname{diag}\left(\nabla_{z_t} g(z_t)\right)\right) + \frac{1}{2} \cdot v_s^\intercal \cdot \left(g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) \right) \tag{89}$$

$$= v_s^\intercal \cdot \left(\sum_{d=1}^{D_z} (g(z_t) \odot e_d) \cdot \nabla_{z_t^d} g^d(z_t) \right) + \frac{1}{2} \cdot v_s^\intercal \cdot \left(g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) \right) \tag{90}$$

$$= \left(\sum_{d=1}^{D_z} v_s^\intercal \cdot (g(z_t) \odot e_d) \cdot \nabla_{z_t^d} g^d(z_t) \right) + \frac{1}{2} \cdot v_s^\intercal \cdot \left(g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) \right) \tag{91}$$

$$= \frac{1}{2} \cdot v_s^\intercal \cdot \left(g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) \right) \quad (\text{since } \langle g(z_t) \odot e_d, v_s \rangle = 0) \tag{92}$$

$$= \frac{1}{2} \cdot v_s^\intercal \cdot \left(\sum_{d=1}^{D_z} (g(z_t) \odot e_d) \cdot g^d(z_t) \cdot \nabla_{z_t^d} \log \tilde{p}(z_t) \right) \tag{93}$$

$$= \frac{1}{2} \cdot \left(\sum_{d=1}^{D_z} v_s^\intercal \cdot (g(z_t) \odot e_d) \cdot g^d(z_t) \cdot \nabla_{z_t^d} \log \tilde{p}(z_t) \right) \tag{94}$$

$$= 0 \tag{95}$$

D Proof of Theorem 6

D.1 Proof that WSP satisfies Theorem 3

Proof. To prove Theorem 6, we will show that $h(t, z_t)$ and $g(t, z_t)$, defined in Eq. 5, satisfy (i)-(iii) and (a)-(b) in Theorem 3.

Proof that WSP satisfies (i) from Theorem 3. Here, we prove that for each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|h(t, z_t)\|^2 + \|g(t, z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

First, we show that $w(z_t)$, defined in Eq. 4, lies in $[0, 1]$. Since $\alpha > 0$, and for any $z_t \in K$, $d_s(z_t) \geq 0$ (since distances are non-negative), we have,

$$0 \leq \tanh\left(\alpha \cdot d_s(z_t)\right) \leq 1. \tag{96}$$

Next, since $\beta > 0$, we know that:

$$0 \leq \beta \cdot \prod_s \underbrace{\frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}}}_{\in[0,1]} \cdot \underbrace{\tanh\left(\alpha \cdot d_s(z_t)\right)}_{\in[0,1]} \tag{97}$$

This then gives us,

$$0 \leq \underbrace{\tanh\left(\beta \cdot \prod_s \frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}} \cdot \tanh\left(\alpha \cdot d_s(z_t)\right)\right)}_{\geq 0} \leq 1, \tag{98}$$

thereby showing that $w(z_t) \in [0, 1]$. Using this, we go on to show that h and g satisfy condition (i) from Theorem 3.

$$\begin{aligned}
\|h(t, z_t)\| &= \|\mathrm{WSP}(\tilde{h}, c_h, t, z_t)\| && (99)\\
&= \|w(z_t) \cdot \tilde{h}(t, z_t) + (1 - w(z_t)) \cdot c_h(z_t)\| && (100)\\
&\leq \|w(z_t) \cdot \tilde{h}(t, z_t)\| + \|(1 - w(z_t)) \cdot c_h(z_t)\| && (101)\\
&\leq \|1 \cdot \tilde{h}(t, z_t)\| + \|(1 - 0) \cdot c_h(z_t)\| && (102)\\
&= \|\tilde{h}(t, z_t)\| + \|c_h(z_t)\| && (103)\\
&= \|\tilde{h}(t, z_t)\| + \left\|\gamma \cdot \frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}\right\| && (104)\\
&= \|\tilde{h}(t, z_t)\| + \gamma \cdot \underbrace{\left\|\frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}\right\|}_{<1} && (105)\\
&\leq \|\tilde{h}(t, z_t)\| + \gamma && (106)\\
&\leq \sqrt{C'_T \cdot (1 + \|z_t\|^2)} + \gamma && (107)
\end{aligned}$$

Thus,

$$\begin{aligned}
\|h(t, z_t)\|^2 &\leq \left(\sqrt{C'_T \cdot (1 + \|z_t\|^2)} + \gamma\right)^2 && (108)\\
&= C'_T \cdot (1 + \|z_t\|^2) + \gamma^2 + 2 \cdot \gamma \cdot \sqrt{C'_T \cdot (1 + \|z_t\|^2)} && (109)\\
&\leq C_T \cdot (1 + \|z_t\|^2) && (110)
\end{aligned}$$

for some $C_T > 0$.

Similarly for $\|g(t, z_t)\|^2$, we have

$$\begin{aligned}
\|g(t, z_t)\| &= \|\mathrm{WSP}(\tilde{g}, c_g, t, z_t)\| && (111)\\
&= \|w(z_t) \cdot \tilde{g}(t, z_t) + (1 - w(z_t)) \cdot c_g(z_t)\| && (112)\\
&\leq \|w(z_t) \cdot \tilde{g}(t, z_t)\| + \|(1 - w(z_t)) \cdot c_g(z_t)\| && (113)\\
&\leq \|1 \cdot \tilde{g}(t, z_t)\| + \|(1 - 0) \cdot c_g(z_t)\| && (114)\\
&= \|\tilde{g}(t, z_t)\| + \|c_g(z_t)\| && (115)\\
&= \|\tilde{g}(t, z_t)\| + \|0\| && (116)\\
&= \|\tilde{g}(t, z_t)\| && (117)\\
&\leq \sqrt{C_T \cdot (1 + \|z_t\|^2)} && (118)
\end{aligned}$$

Thus, $\|g(t, z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

Proof that WSP satisfies (ii) from Theorem 3. We now prove that for all $T > 0$, $z_t, z_t' \in K$, and $t \in [0, T]$, $\|h(t, z_t) - h(t, z_t')\| + \|g(t, z_t) - g(t, z_t')\| \leq C_T \cdot \|z_t - z_t'\|$.

We do this as follows:

$$\|h(t, z_t) - h(t, z_t')\| = \|\text{WSP}(\tilde{h}, c_h, t, z_t) - \text{WSP}(\tilde{h}, c_h, t, z_t')\| \tag{119}$$

$$= \left\| \Big(w(z_t) \cdot \tilde{h}(t, z_t) + (1 - w(z_t)) \cdot c_h(z_t) \Big) - \Big(w(z_t') \cdot \tilde{h}(t, z_t') + (1 - w(z_t')) \cdot c_h(z_t') \Big) \right\| \tag{120}$$

$$= \left\| \Big(w(z_t) \cdot \tilde{h}(t, z_t) - w(z_t') \cdot \tilde{h}(t, z_t') \Big) + ((1 - w(z_t)) \cdot c_h(z_t) - (1 - w(z_t')) \cdot c_h(z_t')) \right\| \tag{121}$$

$$\leq \|w(z_t) \cdot \tilde{h}(t, z_t) - w(z_t') \cdot \tilde{h}(t, z_t')\| + \|(1 - w(z_t)) \cdot c_h(z_t) - (1 - w(z_t')) \cdot c_h(z_t')\| \tag{122}$$

Using the trick by Esmayli (2017), we have:

$$\begin{aligned} \|h(t, z_t) - h(t, z_t')\| &\leq \|w(z_t) - w(z_t')\| \cdot \|\tilde{h}(t, z_t)\| + \|w(z_t')\| \cdot \|\tilde{h}(t, z_t) - \tilde{h}(t, z_t')\| \\ &\quad + \|(1 - w(z_t)) - (1 - w(z_t'))\| \cdot \|c(z_t)\| + \|1 - w(z_t')\| \cdot \|c(z_t) - c(z_t')\| \end{aligned} \tag{123}$$

$$\begin{aligned} &\leq \|w(z_t) - w(z_t')\| \cdot \|\tilde{h}(t, z_t)\| + 1 \cdot \|\tilde{h}(t, z_t) - \tilde{h}(t, z_t')\| \\ &\quad + \|w(z_t) - w(z_t')\| \cdot \|c(z_t)\| + 1 \cdot \|c(z_t) - c(z_t')\| \end{aligned} \tag{124}$$

$$= \|w(z_t) - w(z_t')\| \cdot \underbrace{(\|\tilde{h}(t, z_t)\| + \|c(z_t)\|)}_{\text{bounded by const.}} + \underbrace{\|\tilde{h}(t, z_t) - \tilde{h}(t, z_t')\|}_{< C_T' \cdot \|z_t - z_t'\|} + \|c(z_t) - c(z_t')\| \tag{125}$$

Since K is compact and $\tilde{h}(z_t)$ is continuous and linearly bounded (i.e. $\|\tilde{h}(z_t)\|^2 \leq C_T(1 + \|z_t\|^2)$), we know that $\|\tilde{h}(t, z_t)\|$ is bounded above by a constant. Similarly, $c(z_t)$ is continuous and bounded:

$$\|c(z_t)\|^2 = \gamma^2 \cdot \left\| \frac{z^* - z_t}{\|z^* - z_t\| + \epsilon} \right\|^2 \leq \gamma^2 \cdot \left\| \frac{z^* - z_t}{\epsilon} \right\|^2 = \frac{\gamma^2}{\epsilon^2} \cdot \|z^* - z_t\|^2, \tag{126}$$

so $\|c(z_t)\|$ is bounded above by a constant. This leaves us to show that $w(z_t)$ and $c(z_t)$ are Lipschitz. This is true since both functions are comprised of either composition of Lipschitz functions, or of multiplications of bounded Lipschitz functions, and both of these operations are closed under Lipschitz continuity.

Similarly for $\|g(t, z_t) - g(t, z_t')\|$, we have,

$$\|g(t, z_t) - g(t, z_t')\| = \|\text{WSP}(\tilde{g}, c_g, t, z_t) - \text{WSP}(\tilde{g}, c_g, t, z_t')\| \tag{127}$$

$$= \|(w(z_t) \cdot \tilde{g}(t, z_t) + (1 - w(z_t)) \cdot c_g(z_t)) - (w(z_t') \cdot \tilde{g}(t, z_t') + (1 - w(z_t')) \cdot c_g(z_t'))\| \tag{128}$$

$$= \|w(z_t) \cdot \tilde{g}(t, z_t) - w(z_t') \cdot \tilde{g}(t, z_t')\| \tag{129}$$

Since $w(z_t) \cdot \tilde{g}(t, z_t)$ is the product of a bounded Lipschitz function and a Lipschitz function, we know that $g(t, z_t)$ is also Lipschitz.

Proof that WSP satisfies (iii) from Theorem 3. Here, we prove that for each $z_t \in K$, $h(t, z_t)$ and $g(t, z_t)$ are continuous.

Since all functions involved are continuous and continuity is closed under addition, subtraction, multiplication and composition, $h(t, z_t)$ and $g(t, z_t)$, defined for $t \geq 0$, are continuous for each $z_t \in K$.

Proof that WSP satisfies (a) from Theorem 3. Here we prove that for all $s \in \{1, \ldots, S\}$ and $z_t \in K$ such that $\langle z_t - u_s, v_s \rangle = 0$, we have $\langle h(t, z_t), v_s \rangle \geq 0$.

First, when $\langle z_t - u_s, v_s \rangle = 0$,

$$d_s(z_t) = \frac{\langle z_t - u_s, v_s \rangle}{\|v_s\|} = \frac{0}{\|v_s\|} = 0. \tag{130}$$

This means that,

$$w(z_t) = \tanh\left(\beta \cdot \prod_s \frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}} \cdot \tanh\left(\alpha \cdot d_s(z_t)\right)\right) = 0. \tag{131}$$

Plugging this into $\langle h(t, z_t), v_s \rangle$, we get:

$$\begin{aligned}
\langle h(t, z_t), v_s \rangle &= \langle \mathrm{WSP}(\tilde{h}, c_h, t, z_t), v_s \rangle && (132)\\
&= \langle w(z_t) \cdot \tilde{h}(t, z_t) + (1 - w(z_t)) \cdot c_h(z_t), v_s \rangle && (133)\\
&= \langle 0 \cdot \tilde{h}(t, z_t) + (1 - 0) \cdot c_h(z_t), v_s \rangle && (134)\\
&= \langle c_h(z_t), v_s \rangle && (135)\\
&= \left\langle \gamma \cdot \frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}, v_s \right\rangle && (136)\\
&= \underbrace{\frac{\gamma}{\|z^* - z_t\| + \epsilon}}_{>0} \cdot \langle z^* - z_t, v_s \rangle && (137)
\end{aligned}$$

Because z^* is the Chebyshev center, it lies strictly in $\mathrm{int}(K)$, so $\langle z^* - u_s, v_s \rangle > 0$. Thus, when $z_t \in \partial K$, $\langle z^* - z_t, v_s \rangle = \langle z^* - u_s, v_s \rangle - \langle z_t - u_s, v_s \rangle = \langle z^* - u_s, v_s \rangle > 0$, completing the proof.

Proof that WSP satisfies (b) from Theorem 3. Here we prove that for all $s \in \{1, \ldots, S\}$ and $z_t \in K$ such that $\langle z_t - u_s, v_s \rangle = 0$, we have $\langle g(t, z_t) \odot e_d, v_s \rangle = 0$ for $t \geq 0$ and $d \in \{1, \ldots, D_z\}$. We do this as follows:

$$\begin{aligned}
\langle g(t, z_t) \odot e_d, v_s \rangle &= \langle \mathrm{WSP}(\tilde{g}, c_g, t, z_t) \odot e_d, v_s \rangle && (138)\\
&= \langle (w(z_t) \cdot \tilde{g}(t, z_t) + (1 - w(z_t)) \cdot c_g(z_t)) \odot e_d, v_s \rangle && (139)\\
&= \langle (0 \cdot \tilde{g}(t, z_t) + (1 - 0) \cdot c_g(z_t)) \odot e_d, v_s \rangle && (140)\\
&= \langle c_g(z_t) \odot e_d, v_s \rangle && (141)\\
&= \langle 0 \odot e_d, v_s \rangle && (142)\\
&= \langle 0, v_s \rangle && (143)\\
&= 0 && (144)
\end{aligned}$$

□

D.2 Proof that WSP satisfies Theorem 5

Proof. Since $c_g = 0$, we have $g(t, z_t) = w(z_t) \cdot \tilde{g}(t, z_t)$ throughout. Note that condition (iii) concerns the score function $\tilde{p}$ and is independent of g; it must be verified separately by the practitioner.

Proof that g satisfies (i) from Theorem 5. This follows immediately from the proof of (i) from Theorem 3 in Appendix D.1, which already establishes $\|g(t, z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

Proof that g satisfies (ii) from Theorem 5. The Lipschitz condition $\|g(t,z_t) - g(t,z_t')\| \leq C_T \cdot \|z_t - z_t'\|$ follows from the proof of (ii) from Theorem 3 in Appendix D.1. Now, we prove the Lipschitz condition on $\mathrm{diag}(\nabla_{z_t} g(t,z_t))$. Since $g(t,z_t) = w(z_t) \cdot \tilde{g}(t,z_t)$, the product rule gives:

$$\mathrm{diag}(\nabla_{z_t} g(t,z_t)) = \nabla_{z_t} w(z_t) \odot \tilde{g}(t,z_t) + w(z_t) \cdot \mathrm{diag}(\nabla_{z_t} \tilde{g}(t,z_t)). \tag{145}$$

Therefore,

$$\|\mathrm{diag}(\nabla_{z_t} g(t,z_t)) - \mathrm{diag}(\nabla_{z_t'} g(t,z_t'))\| \tag{146}$$

$$\leq \underbrace{\|\nabla_{z_t} w(z_t) \odot \tilde{g}(t,z_t) - \nabla_{z_t'} w(z_t') \odot \tilde{g}(t,z_t')\|}_{Ⓐ} + \underbrace{\|w(z_t) \cdot \mathrm{diag}(\nabla_{z_t} \tilde{g}(t,z_t)) - w(z_t') \cdot \mathrm{diag}(\nabla_{z_t'} \tilde{g}(t,z_t'))\|}_{Ⓑ}. \tag{147}$$

Using the trick by Esmayli (2017), we bound Ⓐ as follows:

$$Ⓐ \leq \underbrace{\|\nabla_{z_t} w(z_t) - \nabla_{z_t'} w(z_t')\|}_{\leq C_T \cdot \|z_t - z_t'\|} \cdot \underbrace{\|\tilde{g}(t,z_t)\|}_{\leq M} + \underbrace{\|\nabla_{z_t'} w(z_t')\|}_{\text{bounded by const.}} \cdot \underbrace{\|\tilde{g}(t,z_t) - \tilde{g}(t,z_t')\|}_{\leq C_T \cdot \|z_t - z_t'\| \text{ via (ii)}}, \tag{148}$$

where $\nabla_{z_t} w(z_t)$ is Lipschitz since w is a smooth composition of smooth functions (tanh, softmin, product), and $\|\nabla_{z_t} w(z_t)\|$ is bounded on the compact set K. Similarly, we bound Ⓑ as follows:

$$Ⓑ \leq \underbrace{\|w(z_t) - w(z_t')\|}_{\leq C_T \cdot \|z_t - z_t'\|} \cdot \underbrace{\|\mathrm{diag}(\nabla_{z_t} \tilde{g}(t,z_t))\|}_{\text{bounded by const. via (ii)}} + \underbrace{\|w(z_t')\|}_{\leq 1} \cdot \underbrace{\|\mathrm{diag}(\nabla_{z_t} \tilde{g}(t,z_t)) - \mathrm{diag}(\nabla_{z_t'} \tilde{g}(t,z_t'))\|}_{\leq C_T \cdot \|z_t - z_t'\| \text{ via (ii)}}. \tag{149}$$

Altogether, $\|\mathrm{diag}(\nabla_{z_t} g(t,z_t)) - \mathrm{diag}(\nabla_{z_t'} g(t,z_t'))\| \leq C_T \cdot \|z_t - z_t'\|$ for some $C_T > 0$.

Proof that g satisfies (iv) from Theorem 5. We must show $g^d(t,z_t) > 0$ for all $d \in \{1,\ldots,D_z\}$ and $z_t \in \mathrm{int}(K)$. Since $g(t,z_t) = w(z_t) \cdot \tilde{g}(t,z_t)$, it suffices to show $w(z_t) > 0$ and $\tilde{g}^d(t,z_t) > 0$ separately. The latter holds by condition (iv) of Theorem 5 for $\tilde{g}$. For the former, since $z_t \in \mathrm{int}(K)$, we have $d_s(z_t) > 0$ for all $s \in \{1,\ldots,S\}$, and therefore:

$$\frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}} > 0 \qquad \text{and} \qquad \tanh\left(\alpha \cdot d_s(z_t)\right) > 0, \tag{150}$$

so the product $\prod_s \frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}} \cdot \tanh(\alpha \cdot d_s(z_t)) > 0$, and hence $w(z_t) = \tanh(\beta \cdot \text{positive}) > 0$.

Proof that g satisfies (v) from Theorem 5. We must show that for each $s \in \{1,\ldots,S\}$, there exists $L_s < \infty$ such that $\|n_s \odot g(z_t)\|_2 \leq L_s \cdot d_s(z_t)$ for all $z_t \in K$. Let $M_{\tilde{g}} = \sup_{z_t \in K} \|\tilde{g}(t,z_t)\|_2$, which is finite since K is compact and $\tilde{g}$ is continuous. We first bound $w(z_t)$ linearly in $d_s(z_t)$:

$$w(z_t) = \tanh\left(\beta \cdot \prod_s \frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}} \cdot \tanh\left(\alpha \cdot d_s(z_t)\right)\right) \tag{151}$$

$$\leq \beta \cdot \prod_s \frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}} \cdot \tanh\left(\alpha \cdot d_s(z_t)\right) \qquad \text{since } \tanh(x) \leq x \text{ for } x \geq 0 \tag{152}$$

$$\tag{153}$$

Since all factors in the product above lie in $[0,1]$, the product can be bounded by any single factor:

$$w(z_t) \leq \beta \cdot \underbrace{\frac{e^{-d_s(z_t)}}{\sum_{s'} e^{-d_{s'}(z_t)}}}_{\leq 1} \cdot \tanh\left(\alpha \cdot d_s(z_t)\right) \tag{154}$$

$$\leq \beta \cdot \tanh\left(\alpha \cdot d_s(z_t)\right) \tag{155}$$

$$\leq \alpha\beta \cdot d_s(z_t) \qquad \text{since } \tanh(x) \leq x \text{ for } x \geq 0. \tag{156}$$

Therefore,

$$\|n_s \odot g(z_t)\|_2 \leq \|g(z_t)\|_2 = w(z_t) \cdot \|\tilde{g}(t, z_t)\|_2 \leq \alpha\beta \cdot d_s(z_t) \cdot M_{\tilde{g}} = L_s \cdot d_s(z_t), \tag{157}$$

where $L_s = \alpha\beta M_{\tilde{g}} < \infty$.

Proof that g satisfies (b) from Theorem 5. This is identical to the proof of (b) from Theorem 3 in Appendix D.1, since the condition is the same.

□

E Discussion of Assumptions

The assumptions in Theorems 3 and 5 are easily satisfied when h, g, and $\log \tilde{p}(z_t)$ are parameterized by NNs.

Lipschitz Continuity with Respect to Inputs. Lipschitz continuous functions are closed under composition, making a large class of NNs Lipschitz continuous by construction. Additionally, there exist many easy and empirically effective methods for explicitly obtaining Lipschitz continuity, for example via weight normalization (e.g. Miyato et al. (2018)), regularization (e.g. Liu et al. (2022)), and architecture design (e.g. Anil et al. (2019)). Altogether, this allows us to conveniently satisfy the Lipschitz continuity assumptions for h, g, and $\log \tilde{p}(z_t)$—(ii) in Theorem 3, and (ii) and (iii) in Theorem 5.

Next, Hurault et al. (2022) (Proposition 2) proved that if every function in a collection is differentiable with uniformly bounded, Lipschitz gradients with respect to the inputs, the gradients of their composition are themselves Lipschitz with respect to the inputs. This makes a large class of NNs satisfy the second half of (ii) from Theorem 5. This property is known as "Lipschitz smoothness," and can also be encouraged explicitly, for example via mixup regularization (Gyawali et al., 2020).

Linearly Bounded NNs. We parameterize all NNs here with a composition of continuous functions, thereby making them continuous. And since continuous functions on compact spaces are bounded, we easily satisfy (i) from Theorems 3 and 5.

Differentiability and Continuity of Partials of NNs. All NNs here use continuously differentiable activation functions, so they are continuously differentiable with continuous partials.

F Latent SDEs with Pathwise Expansions for Modeling Suicide Risk

Notation. We observe each patient $n \in [1, \ldots, N]$ at times $t_1, \ldots, t_M$. Note that the observation times and the number of observations differ by patient, but for notational simplicity, we will denote them as if each patient has observations at the same times. Let $x_t^{n,d} \in \{0, \ldots, 10\}$ denote patient n's response to Likert-scale question d at time t. Similarly, denote $z_t^{n,d}$ as the dth component of patient n's latent psychological state at time t. Let $x_t^n \in \{0, \ldots, 10\}^{D_x}$ denote patient n's response to all D_x survey questions, and let $z_t^n \in [0, 1]^{D_z}$ denote the patient's D_z-dimensional latent space, confined to the unit cube.

Generative Process. We assume patient n's data is generated via:

$$z_{t_0}^n \sim \mathcal{N}_{[0,1]}(\mu_0, \sigma_0^2 \cdot I_{D_z}), \qquad \text{(initial state, drawn from a } [0,1]\text{-truncated normal)} \tag{158}$$

$$z_{t_m}^n | z_{t_{m-1}}^n \sim p(\cdot | z_{t_{m-1}}^n) = z_{t_{m-1}}^n + \int_{t_{m-1}}^{t_m} \underbrace{h(z_t; \theta)}_{\text{prior drift}} \cdot dt + \underbrace{g(z_t; \theta)}_{\text{prior diffusion}} \circ dB_t, \tag{159}$$

$$x_{t_m}^{n,d} | z_{t_m}^{n,d} \sim \text{Cat}(\lambda(z_{t_m}^{n,d}; \sigma_\epsilon)), \qquad \text{(ordinal likelihood of survey data given latent state)} \tag{160}$$

wherein the dynamics, $h : K \to \mathbb{R}^{D_z}$ and $g : K \to \mathbb{R}^{D_z}_{\geq 0}$, are time-independent, and where,

$$\lambda(z_t^d; \sigma_\epsilon) = \Phi\left(\frac{b^{1:12} - z_t^d}{\sigma_\epsilon}\right) - \Phi\left(\frac{b^{0:11} - z_t^d}{\sigma_\epsilon}\right), \quad b = \begin{bmatrix} -\infty & 0.05 & 0.15 & 0.25 & \dots & 0.95 & \infty \end{bmatrix},$$

in which $\Phi(\cdot)$ is the CDF of a standard normal and b represents the boundaries for 0-10 range ordinal likelihood.

Latent Dimensionality. Note that in these experiments, we set the latent dimensionality equal to the data dimensionality, $D_z = D_x$. In many applications, one typically chooses $D_z < D_x$, treating the latent space as a compressed representation of the observations. However, when modeling psychological state, there is often no single low-dimensional "underlying condition" that generates the observed symptoms; instead, mental health challenges are characterized by the way in which the symptoms themselves interact, regulate, or reinforce one another—this is known as the "networks approach" in clinical psychology (Borsboom et al., 2022; Robinaugh et al., 2020). For this reason, each latent dimension directly corresponds to an observed dimension.

Eliminating SDE Solvers with a Pathwise Series Expansion. Fitting the Latent SDE above requires backpropagation through a slow, numerically unstable SDE solver. In contrast, ODE solvers are known to be more numerically stable, accurate, and well-behaved when used with adaptive step-sizes (e.g. Lou and Ermon (2023)). Thus, we replace Brownian motion with the Karhunen-Loève expansion (Särkkä and Solin, 2019):

$$d\widehat{B}_t|\xi = \sum_{r=1}^{R} \sqrt{\frac{2}{T}} \cos\left(\frac{(2 \cdot r - 1) \cdot \pi \cdot t}{2T}\right) \cdot \xi^r \cdot dt, \qquad \xi \sim \mathcal{N}(0, I_R), \tag{161}$$

where T is the end-time of the process (in this work, it's the end-time of the EMA study) and I_R is an R-dimensional identity matrix. This expansion consists of a sum of R randomly weighted ODEs that, as $R \to \infty$, converge to dB_t, leading the overall equation to converge to the Stratonovich SDE (Wong and Zakai, 1965). Using a finite R, we obtain an approximation of the above model that we can fit using ODE solvers. Following Ghosh et al. (2022), we replace Eq. 159 above with:

$$z_{t_m}^n | z_{t_{m-1}}^n, \xi^n = z_{t_{m-1}}^n + \int_{t_{m-1}}^{t_m} h(z_t; \theta) \cdot dt + g(z_t; \theta) \cdot d\widehat{B}_t, \qquad \xi^n \sim \mathcal{N}(0, I_R), \tag{162}$$

wherein z_t^n is now a *deterministic* function of a new latent variable, ξ^n.

Fitting Latent SDEs to Data. Our goal is to find parameters, $\Theta = \{\mu_0, \sigma_0, \theta, \sigma_\epsilon\}$, that maximize the log marginal likelihood (LML) of the observed data:

$$\Theta^* = \text{argmax}_\Theta \frac{1}{N} \sum_{n=1}^{N} \log p(X^n; \Theta) = \text{argmax}_\Theta \frac{1}{N} \sum_{n=1}^{N} \log \int_{z_{t_0}} \int_{\xi} p(X^n, \xi, z_{t_0}; \Theta) \cdot d\xi \cdot dz_{t_0},$$

where X^n represents all of patient n's training data. Since the above integrals are intractable, we compute a variational lower bound to the LML instead:

$$\log p(X^n; \Theta) \geq \mathbb{E}_{q(\xi, z_{t_0}|X^n;\Phi)} \left[\log \frac{p(X^n, \xi, z_{t_0}; \Theta)}{q(\xi, z_{t_0}|X^n; \Phi)}\right] = \text{ELBO}(X^n; \Theta, \Phi), \tag{163}$$

where

$$q(\xi, z_{t_0}|X^n; \Phi) = \mathcal{N}(\mu_\xi^n, \sigma_\xi^n \cdot \sigma_\xi^n \cdot I_R) \cdot \mathcal{N}_{[0,1]}(\mu_{z_0}^n, \sigma_{z_0}^n \cdot \sigma_{z_0}^n \cdot I_{D_z}), \tag{164}$$

is our variational family, and $\Phi = \{\mu_\xi^n, \sigma_\xi^n, \mu_{z_0}^n, \sigma_{z_0}^n\}_{n=1}^N$ is the set of all variational parameters. We maximize the ELBO with respect to the model and inference parameters, Θ and Φ, using stochastic gradient descent.

Approximate Predictive Log-Likelihood. Let Θ^* and Φ^* denote the model and inference parameters that maximize the ELBO as obtained by gradient descent. Then the predictive log-likelihood is,

$$\log p(x^n_{t^*}|X^n;\Theta^*) = \log \mathbb{E}_{p(\xi,z_{t_0}|X^n;\Theta^*)}\left[p(x^n_{t^*}|\xi,z_{t_0};\Theta^*)\right] \approx \log \mathbb{E}_{q(\xi,z_{t_0}|X^n;\Phi^*)}\left[p(x^n_{t^*}|\xi,z_{t_0};\Theta^*)\right], \tag{165}$$

obtained via Monte Carlo approximation.

Forecasting. We sample from the approximate posterior and solve the corresponding differential equation to obtain forecasts. That is, for each draw from $q(\xi, z_{t_0}|X^n;\Phi^*)$, we evaluate:

$$z^n_{t^*}|z_{t_0},\xi = z_{t_0} + \int_{t_0}^{t^*} h(z_t;\theta)\cdot dt + g(z_t;\theta)\cdot d\widehat{B}_t. \tag{166}$$

G Experimental Setups

Software. All experiments were conducted in Jax (Bradbury et al., 2018) with NumPyro (Phan et al., 2019), Diffrax (Kidger, 2021) and Chex.

G.1 Experimental Setup for Fig. 1

Suppose we were given dynamics from a domain expert and we wanted to transform them via WSP to ensure they remained viable in any polyhedron.

Expert-Given Dynamics. Consider the dynamics,

$$\tilde{h}(t,z_t) = \begin{bmatrix} 2\cdot z^0_t + 5.25\cdot z^1_t - 3.625 \\ 5.25\cdot z^0_t - 2\cdot z^1_t - 1.625 \end{bmatrix}, \qquad \tilde{g}(t,z_t) = \begin{bmatrix} 0.3 \\ 0.3 \end{bmatrix}, \tag{167}$$

which spiral counterclockwise outward from the point $(0.5, 0.5)$.

Polyhedra. We instantiate Definition 2 as follows to obtain three polyhedra:

- Right-Angle Triangle:

$$\begin{aligned} u_1 &= \begin{bmatrix}0.0 & 0.0\end{bmatrix}, & u_2 &= \begin{bmatrix}0.0 & 0.0\end{bmatrix}, & u_3 &= \begin{bmatrix}0.5 & 0.5\end{bmatrix}, \\ v_1 &= \begin{bmatrix}1.0 & 0.0\end{bmatrix}, & v_2 &= \begin{bmatrix}0.0 & 1.0\end{bmatrix}, & v_3 &= \begin{bmatrix}-\sqrt{0.5} & -\sqrt{0.5}\end{bmatrix}. \end{aligned}$$

- Unit Square:

$$\begin{aligned} u_1 &= \begin{bmatrix}0.0 & 0.0\end{bmatrix}, & u_2 &= \begin{bmatrix}0.0 & 0.0\end{bmatrix}, & u_3 &= \begin{bmatrix}1.0 & 1.0\end{bmatrix}, & u_4 &= \begin{bmatrix}1.0 & 1.0\end{bmatrix}, \\ v_1 &= \begin{bmatrix}1.0 & 0.0\end{bmatrix}, & v_2 &= \begin{bmatrix}0.0 & 1.0\end{bmatrix}, & v_3 &= \begin{bmatrix}-1.0 & 0.0\end{bmatrix}, & v_4 &= \begin{bmatrix}0.0 & -1.0\end{bmatrix}. \end{aligned}$$

- Lopsided Pentagon:

$$\begin{aligned} &u_1 = \begin{bmatrix}0.1 & 0.1\end{bmatrix}, u_2 = \begin{bmatrix}0.1 & 0.1\end{bmatrix}, u_3 = \begin{bmatrix}1.1 & 1.1\end{bmatrix}, u_4 = \begin{bmatrix}1.1 & 1.1\end{bmatrix}, u_5 = \begin{bmatrix}0.8 & 0.8\end{bmatrix}, \\ &v_1 = \begin{bmatrix}1.0 & 0.1\end{bmatrix}, v_2 = \begin{bmatrix}0.1 & 1.0\end{bmatrix}, v_3 = \begin{bmatrix}-1.0 & 0.2\end{bmatrix}, v_4 = \begin{bmatrix}0.2 & -1.0\end{bmatrix}, v_5 = \begin{bmatrix}-\sqrt{0.5} & -\sqrt{0.5}\end{bmatrix}. \end{aligned}$$

Dynamics. We transform the dynamics from Eq. 167 via WSP in Eq. 5 to remain within each of the above polyhedra. We use the following hyper-parameters:

- Right-Angle Triangle: $\alpha = 5.0$, $\beta = 100.0$, $\gamma = 2.0$, $\epsilon = 0.1$.
- Unit Square: $\alpha = 5.0$, $\beta = 1000.0$, $\gamma = 2.0$, $\epsilon = 0.1$.
- Lopsided Pentagon: $\alpha = 10.0$, $\beta = 8000.0$, $\gamma = 2.0$, $\epsilon = 0.1$.

Differential Equation Solver. All SDE trajectories began at $z_0 = \begin{bmatrix}0.05 & 0.85\end{bmatrix}^\intercal$ and were solved for time interval $t \in [0.0, 5.0]$ via the Itô-Milstein solver (Mil'shtejn, 1975) with a step-size of 0.01.

G.2 Experimental Setup for Figs. 2, 4 and 5

Dynamics. Due to Limitation 2 (Section 1), initialization plays a crucial role in the success of expressive SDE-based models. This is because many datasets (e.g. images) lie in compact Euclidean subspaces. In early-stage training, SDE trajectories often leave the region, requiring a large number of gradient steps just to return to it (while not necessarily fitting the data well), causing optimization to get stuck in poor local optima. In late-stage training, small perturbations to the dynamics may, again, yield trajectories that lie outside the region. As such, to empirically compare the inductive bias of WSP (Eq. 5) against baselines (Eqs. 1–3), we solve SDEs given by NNs h and g with randomly sampled weights. We define the viable region, $K = [0, 1]$, to be an interval, and specifically choose to set $z_0 = 0.99$ near the boundary to stress-test the chain-rule based SDEs in Eqs. 2 and 3 to show that once close to the boundary, they will struggle to return to the interior of K. While simple, *these preliminary experiments already show WSP boasts a stark improvement in inductive bias in comparison to baselines.*

Architecture. In all experiments presented here, we used 3-layer NNs with 64 hidden units and CELU activation (Barron, 2017). We repeated these experiments with 2-layer and 4-layer NNs and observed the *exact same behavior*, so we have omitted them for brevity. We also repeated these experiments with other continuous activation functions—GeLU (Hendrycks and Gimpel, 2016), ELU (Clevert et al., 2015), SELU (Klambauer et al., 2017), and SiLU (Elfwing et al., 2018)—and we observed the *exact same type of behavior*, so we have omitted them for brevity.

Random Restarts. For each SDE in Eqs. 1–3 and 5, we randomly drew the weights using the Glorot normal initialization (Glorot and Bengio, 2010). In each plot, we repeated this initialization 5 times, drawing 3 samples for each initialization.

Differential Equation Solver. In Fig. 2, we used the Itô-Milstein SDE solver (Mil'shtejn, 1975). In Fig. 4, we used the Dormand-Prince 8/7 ODE solver (Dormand and Prince, 1980). In all experiments, we simulated the dynamics for $t \in [0, 5]$ with a step size of 0.001. We purposefully chose a small step size to ensure the faithfulness of the SDE solutions to the dynamics.

Pathwise Expansion. We used a truncation of $R = 40$ terms in the pathwise expansion (Eq. 161) for the experiment in Fig. 4. We repeated the experiments with $R = 20$, 100, and 200 and observed the *exact same behavior*, so we have omitted them for brevity.

G.3 Experimental Setup for Table 1 and Figs. 3 and 6–18

Dynamics and Architecture. In all experiments, we used 3-layer NNs with 64 hidden units and GELU activations for the unconstrained drift and diffusion, $\tilde{h}$ and $\tilde{g}$. For the diffusion, we additionally applied a softplus activation (Dugas et al., 2000) at the output to enforce positivity.

When using WSP to transform $\tilde{h}$ and $\tilde{g}$ so that the SDE solution remains viable, we additionally clipped the state z_t to ensure it remains viable before passing it into the dynamics. This was necessary because, although our dynamics are viable, standard ODE/SDE solvers are not. Developing solvers that are themselves viable is an important direction for future work.

Differential Equation Solver. Thanks to the pathwise expansion (described in Appendix F), we used the Dormand-Prince 8/7 ODE solver (Dormand and Prince, 1980) using an adaptive step size initialized at 0.01, with a minimum of 0.001, and with a tolerance of 0.001.

Warm Starts. Since optimizing the ELBO for neural latent SDEs is challenging, we propose to start optimization with the following warm starts:

1. **Drift.** We encourage the drift to be 0 to encourage the model to explore both positive and negative values of drift. We do this by minimizing the following loss function for 5000 gradient steps:

$$\text{argmin}_\theta \|\tilde{h}(z_t;\theta) - 0.0\|_2^2, \tag{168}$$

   wherein $\tilde{h}$ is the drift before applying WSP.

2. **Diffusion.** We initialized the diffusion term at 0.5 to encourage the model to learn a nonzero diffusion. Without this initialization, the model frequently collapsed to solutions with diffusion equal to 0. In that case, the noise variable ξ is unused in the generative process, so its posterior matches its prior, the KL term in the ELBO becomes zero, and optimization gets trapped in a local optimum where only the reconstruction term contributes. We do this by minimizing the following loss function for 5000 gradient steps:

$$\text{argmin}_\theta \|\tilde{g}(z_t;\theta) - 0.5\|_2^2, \tag{169}$$

   wherein $\tilde{g}$ is the diffusion before applying WSP.

Optimization. We used the Adam optimizer (Kingma and Ba, 2014) with a linear learning rate scheduler going from 1e-4 to 5e-5 and a total of 80,000 gradient steps. To be able to evaluate the inductive bias of the model, we also fixed the variational standard deviations for all N posteriors to a small value of $\sigma_\xi^n = 0.05$ throughout optimization, such that the model does not simply increase uncertainty to match the stochasticity in the data (see Section 5 for additional discussion).

Random Restarts. We repeated each experiment 5 times, each time randomly drawing the weights of the drift and diffusion NNs using a Glorot normal initializer (Glorot and Bengio, 2010), which adapts scaling to the arithmetic average of the numbers of inputs and outputs.

Train/Interpolation/Forecast Data Split. We split the study time horizon into interpolation and forecasting regimes. To create a challenging forecasting task, we defined the test set as all observations after the median time step (computed across all patients and time points) and evaluated models on predicting this held-out second half, thereby stress-testing their inductive bias. For interpolation, we randomly held out 20% of the time steps in the first half of the time horizon for each patient (10% of all time steps). This interpolation set lies within the training window and evaluates how well the learned dynamics fit observed data.

Model Selection. Across the random restarts, we selected the model with the highest posterior predictive on the interpolation set, treating it like a validation set.

Model Evaluation. We assessed the quality of learned models along several axes:

- **Inductive Bias.** We evaluated each model's posterior predictive on the test set.
- **Learning Dynamics.** We evaluated each model's posterior predictive on the interpolation set. Optimization for models that have worse interpolation set performance got stuck in poor local optima.
- **Constraint Satisfaction.** We evaluated how well each model satisfied the constraints proposed in Theorem 3. Since our real data experiments all require the SDE solution to lie in the unit cube, the constraints are: (a) $h^d(z) \geq 0$ if $z^d = 0$, and $h^d(z) \leq 0$ if $z^d = 1$, and (b) $g^d(z) = 0$ if $z^d \in \{0, 1\}$. We do this via the metrics proposed below, which we approximate via 100 Monte Carlo samples. In these metrics, $z \cdot (1 - e_d) + e_d$ is z with dimension d set to 1, and $z \cdot (1 - e_d)$ is z with dimension d set to 0.

– Drift Validity Proportion (DRVP):

$$\text{DRVP} = \frac{1}{2D_z} \cdot \sum_{d=1}^{D_z} \mathbb{E}_{z\sim\mathcal{U}[0,1]} \left[\mathbb{I}\left(h^d(z\cdot(1-e_d)+e_d) \leq 0\right) + \mathbb{I}\left(h^d(z\cdot(1-e_d)) \geq 0\right) \right]. \tag{170}$$

This computes the fraction of boundary for which the drift satisfies the constraints (higher is better).

– Diffusion Validity Proportion (DIVP):

$$\text{DIVP} = \frac{1}{2D_z} \cdot \sum_{d=1}^{D_z} \mathbb{E}_{z\sim\mathcal{U}[0,1]} \left[\mathbb{I}\left(g^d(z\cdot(1-e_d)+e_d) < \epsilon\right) + \mathbb{I}\left(g^d(z\cdot(1-e_d)) < \epsilon\right) \right], \tag{171}$$

for some small $\epsilon = 0.001$ to allow for very small positive values, since $g(z)$ uses a softplus activation. This computes the fraction of boundary for which the diffusion satisfies the constraints (higher is better).

– Diffusion Distance to Valid (DIDV):

$$\text{DIDV} = \frac{1}{2D_z} \cdot \sum_{d=1}^{D_z} \mathbb{E}_{z\sim\mathcal{U}[0,1]} \left[|g^d(z\cdot(1-e_d)+e_d)| + |g^d(z\cdot(1-e_d))| \right]. \tag{172}$$

This metric computes the average distance from the diffusion's value on the boundary to 0 (lower is better).

H Descriptions of Real Data

The PhysioNet GLOBEM Data. GLOBEM is a public, multi-year collection of datasets for longitudinal human behavior collected from 2018 to 2022 (Goldberger et al., 2000; Xu et al., 2022b). A total of 497 unique undergraduates in the United States were recruited via email to participate in the study. The collection is split into 4 datasets, one for each year, and each containing 155, 218, 137, and 195 participants, respectively.

In their study, participants downloaded a mobile app and wore a fitness tracker for 10 weeks. In this work, we specifically focused on the EMA data collected via mobile surveys pertaining directly to mental health, which consisted of 6 different psychometric instruments:

- Perceived Stress Scale (PSS-4) (Cohen et al., 1983): measures stress levels during the last month, ranging from 0 to 16, with higher values indicating more perceived stress. This was collected every Wednesday.
- Positive Affect and Negative Affect Schedule (PANAS) (Watson et al., 1988): captures positive and negative affects, respectively, each on a range from 0 to 20. Higher values indicate higher levels of positive/negative affect, respectively. This was collected every Wednesday and Sunday.
- Patient Health Questionnaire Mental Health (PHQ-4-MH) (Kroenke et al., 2009) ranges from 0 to 12, with higher values indicating a higher risk of mental health struggles. This was collected every Sunday.
- Patient Health Questionnaire Anxiety (PHQ-4-A) (Kroenke et al., 2009) ranges from 0 to 6, with higher values indicating a higher risk of anxiety. This was collected every Sunday.
- Patient Health Questionnaire Depression (PHQ-4-D) (Kroenke et al., 2009) ranges from 0 to 6, with higher values indicating a higher risk of depression. This was collected every Sunday.

Of the four years of data available, we chose to use "DS2" from 2019, "DS3" from 2020, and "DS4" from 2021 because these three datasets share the same EMA questions ("DS1" from 2018 did not use the same EMA questions as the other 3 years). To ensure that we had sufficient data from each patient used in training, we only included participants who had at least 10 observations recorded, omitting 8 participants from DS2, 6 participants from DS3, and 4 participants from DS4.

For instruments with response scales extending beyond 0–10, we observed a small proportion of values exceeding 10; specifically, across all patients and all time steps in the 3 datasets, we found 112 PHQ-4 mental health scores (0.1%), 400 PSS-4 scores (0.3%), 500 positive affect scores (0.3%), and 2,445 negative affect scores (1.5%) greater than 10. As such, we truncated all values greater than 10 to 10, such that a score of 10 represents the category "≥ 10."

Because participants entered the study on different calendar dates, we converted all dates to patient-specific, zero-anchored time indices, with time step 0 corresponding to each participant's own study start date. In our model, we measured time in days and then rescaled it by a factor of 1/5, compressing the overall time horizon to reduce the computational cost of solving the SDEs.

The "SMART" EMA Data. A total of 623 individuals reporting suicidal thoughts and/or recent suicidal behavior were enrolled from two Boston-area hospitals, who participated in the Sensing and Mobile Assessment in Real Time (SMART) study at Harvard University (Nock et al., 2026). This sample included 315 adults (18 years and older) recruited from a psychiatric emergency service and 308 adolescents (ages 12–19) recruited from a psychiatric inpatient unit.

Individuals were excluded if they did not own an iOS or Android smartphone, were unable to provide informed consent or assent, lacked fluency in spoken or written English, or exhibited substantial cognitive or behavioral impairment (e.g. florid psychosis, intellectual disability, dementia, acute intoxication, or marked agitation or violence).

After enrollment, participants provided written consent/assent, completed a baseline questionnaire, and installed the LifeData app on their smartphones. The app delivered brief self-report surveys. Participants received $10 for the baseline assessment and $1 for each ecological momentary assessment (EMA) survey they completed. All procedures were reviewed and approved by the institutional review boards of the participating institutions.

EMA surveys captured momentary suicidal thinking—specifically the urge to engage in suicidal behavior, suicidal intent, and perceived ability to resist suicidal urges—as well as 17 affective states (negative, hopeless, trapped, isolated, burdensome, angry, self-hate, agitated, worried, numb, fatigued, humiliated, desire to escape, desire to avoid, energetic, and positive) rated on a 0–10 Likert scale. In this work, we modeled four EMA survey items, all on a 0-10 Likert scale, with higher values indicating higher intensity: stress, negative affect, desire to escape, and urge to die. Surveys were administered six times per day over a three-month period. The first and last surveys of each day occurred at fixed times selected collaboratively with each participant, whereas the remaining surveys were delivered at random times between these anchors.

Participants also had the option to initiate additional surveys at any time, for instance to report a suicide attempt, non-suicidal self-injury, or other significant events. A dedicated risk-monitoring team reviewed incoming data in real time and intervened when participants reported elevated suicidal intent (further details available upon request).

Similar to the PhysioNet GLOBEM dataset, we also converted all dates to zero-anchored time indices. Since surveys were delivered at random times, we discretized time into 6 intervals per day such that responses within the same interval were considered to be administered in the same time bucket. If multiple responses from the same patient fell into the same bucket (84% of observations), we took the maximum to capture peak distress. This reduces the amount of noisy data while preserving critical psychological signals. The final time indices were rescaled by a factor of 1/20 to compress the time horizon and reduce the computational cost of solving the SDEs. Finally, after this preprocessing, to ensure that we had sufficient data from each patient used in training, we only

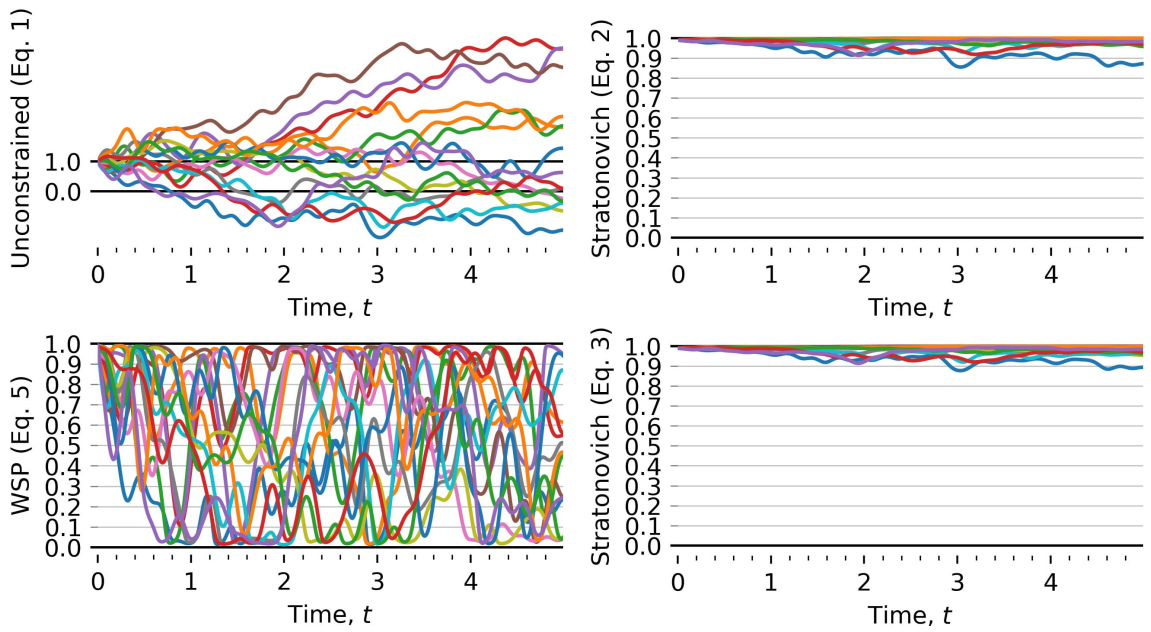

Figure 4. WSP exhibits better inductive bias than baselines given smooth, pathwise expansion of Brownian motion. Top left: Stratonovich-SDE with NN quickly leaves $K = [0, 1]$. Top & bottom right: Stratonovich-SDE transformed via sigmoid sticks to the boundary. Bottom left: Stratonovich-SDE with WSP successfully remains in K. Note: for Eqs. 2 and 3, we used the Stratonovich chain-rule instead of Itô's lemma.

included participants who had at least 20 observations recorded, resulting in a final cohort of 201 patients.

I Results

I.1 Inductive Bias with Smooth, Pathwise Expansion of Brownian Motion

Since SDE solvers are slow and unstable, prior work focused on finding mechanisms to use ODE solvers instead. ODE solvers are known to be more numerically stable, accurate, and well-behaved when used with adaptive step-sizes. Prior work has used several strategies to accomplish this. For example, prior work approximates the first two moments of the time-marginal using the Fokker-Planck-Kolmogorov (FPK) equation, which can then be solved using an ODE solver—this is known as the "Gaussian Assumed Approximation" (Särkkä and Solin, 2019). In diffusion models, prior work derived an FPK-based, fast, numerically stable process that samples from the same distribution as the SDE using an ODE solver—this is called the "probability flow ODE" (Song et al., 2021). Finally, prior work (e.g. Ghosh et al. (2022)) replaces Brownian motion with the Karhunen-Loève Expansion (Särkkä and Solin, 2019)—see Eq. 161. In Fig. 4, we empirically demonstrate that WSP exhibits the same inductive bias in comparison to baselines, even under this pathwise expansion.

I.2 Stationary SDEs on r-Polyhedra

In Fig. 5, we show that, for any neural diffusion with WSP (Eq. 5) with randomly generated weights, we can always construct a drift, given by Theorem 5, with dynamics that are viable in $K = [0, 1]$ and induce the target time-marginal. Moreover, like their general counterparts, these stationary dynamics overcome the shortcomings of the baseline dynamics in Eqs. 1–3.

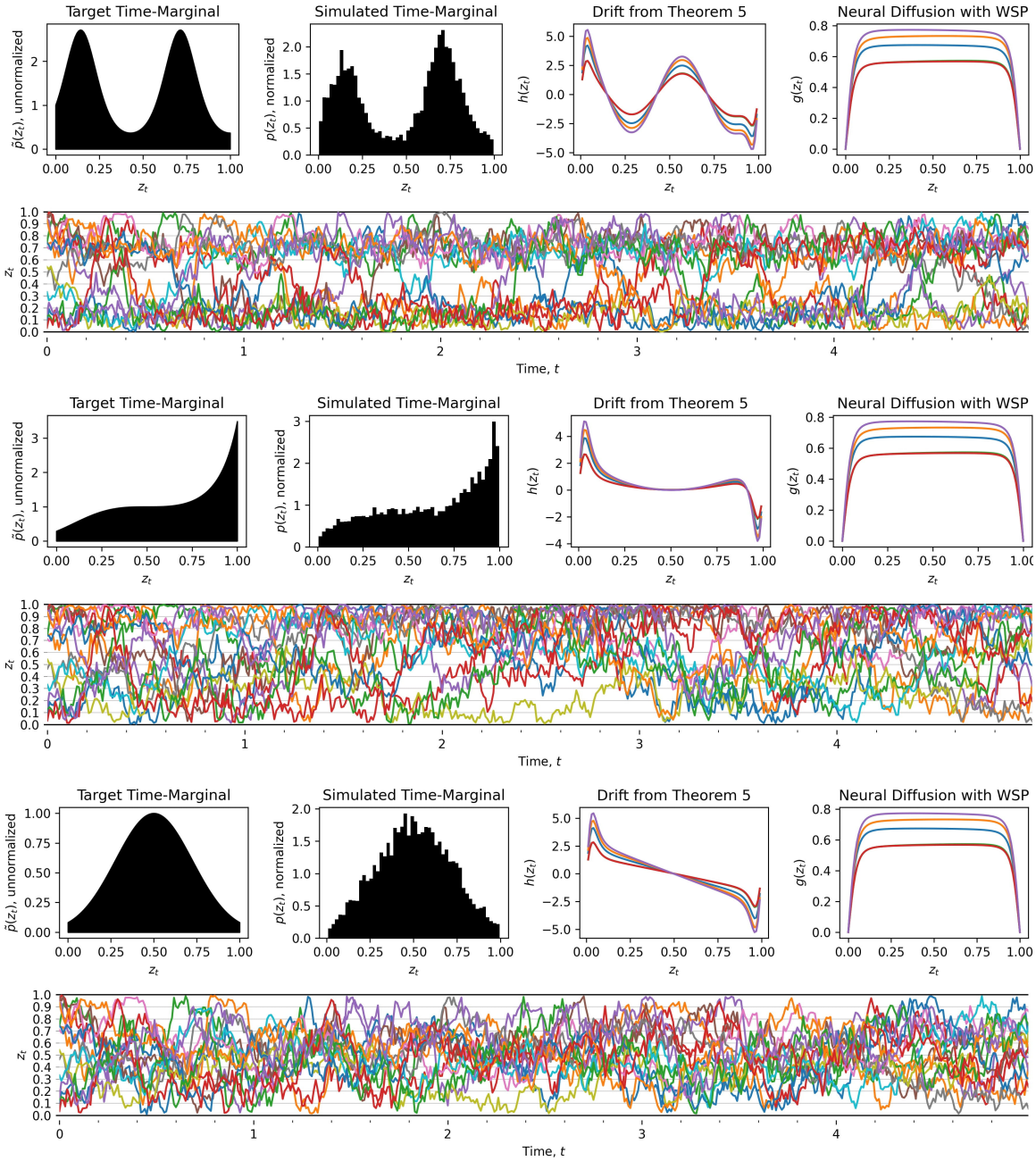

Figure 5. Stationary SDE exhibits better inductive bias than baselines. Given a target time-marginal, given any diffusion with WSP, we can always derive a corresponding drift via Theorem 5 that is viable in K and has the target stationary distribution. Like the general dynamics, these dynamics overcome the shortcomings of the baseline dynamics in Eqs. 1–3. Here, our diffusion is a NN with randomly initialized weights, with each color corresponding to a different seed. Note: the target time-marginal is *not* normalized.

I.3 Visualizations of Latent SDE Posterior Samples from the "SMART" Data

Figs. 6–17 visualize the posteriors of latent neural SDEs of various patients in the "SMART" dataset across all baselines—**Vanilla**, **Vanilla+Clip**, WSP, and **WSP+Clip**. These results consistently show that:

1. WSP-based dynamics respect the specified clinical constraints and avoid assigning probability mass to impossible outcomes: the constraint-satisfaction metrics are perfect for WSP, by construction, and are far from perfect for Vanilla *and* Vanilla+Clip.
2. These constraints guide optimization toward better minima: the WSP-based model exhibits better interpolation performance, indicating optimization converged to better optima. In contrast, the Vanilla and Vanilla+Clip model often struggles to fit the interpolation set, even though it lies in the same time region as the training data.
3. Altogether, the WSP-based model exhibits a better inductive bias for forecasting.

We, again, note that although WSP substantially improves forecasts, it still cannot reliably infer a patient's true state in the second half of the study just from the first half—EMA data is too stochastic for *any* model to be this accurate. *Accordingly, we present examples where WSP forecasts are strikingly good (closely tracking the forecasting set or its overall trend) alongside examples where all dynamics miss the target entirely.* Taken together, these results show that WSP's inductive bias can markedly improve forecasting, and suggest that incorporating additional clinical knowledge could yield similarly large gains.

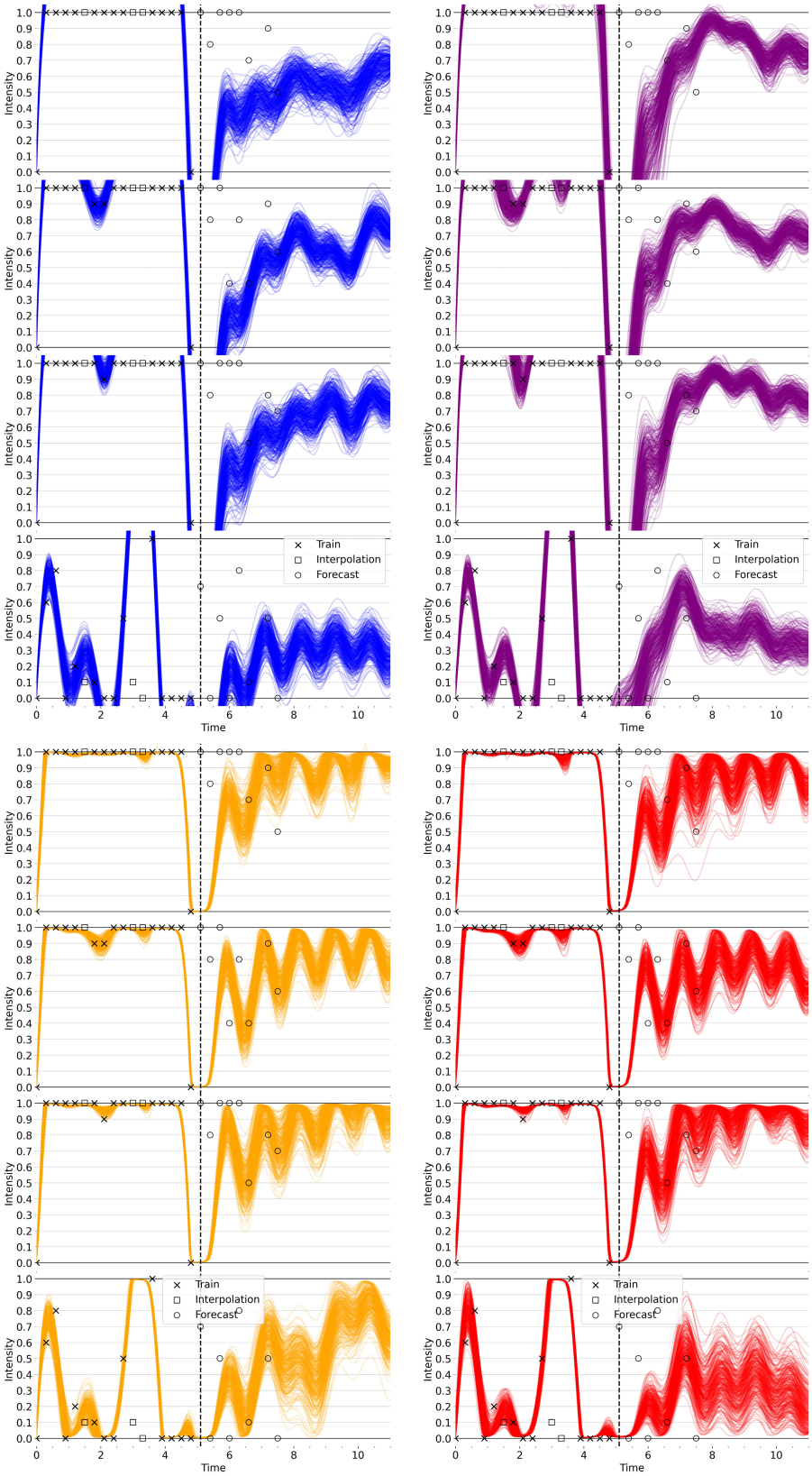

Figure 6. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

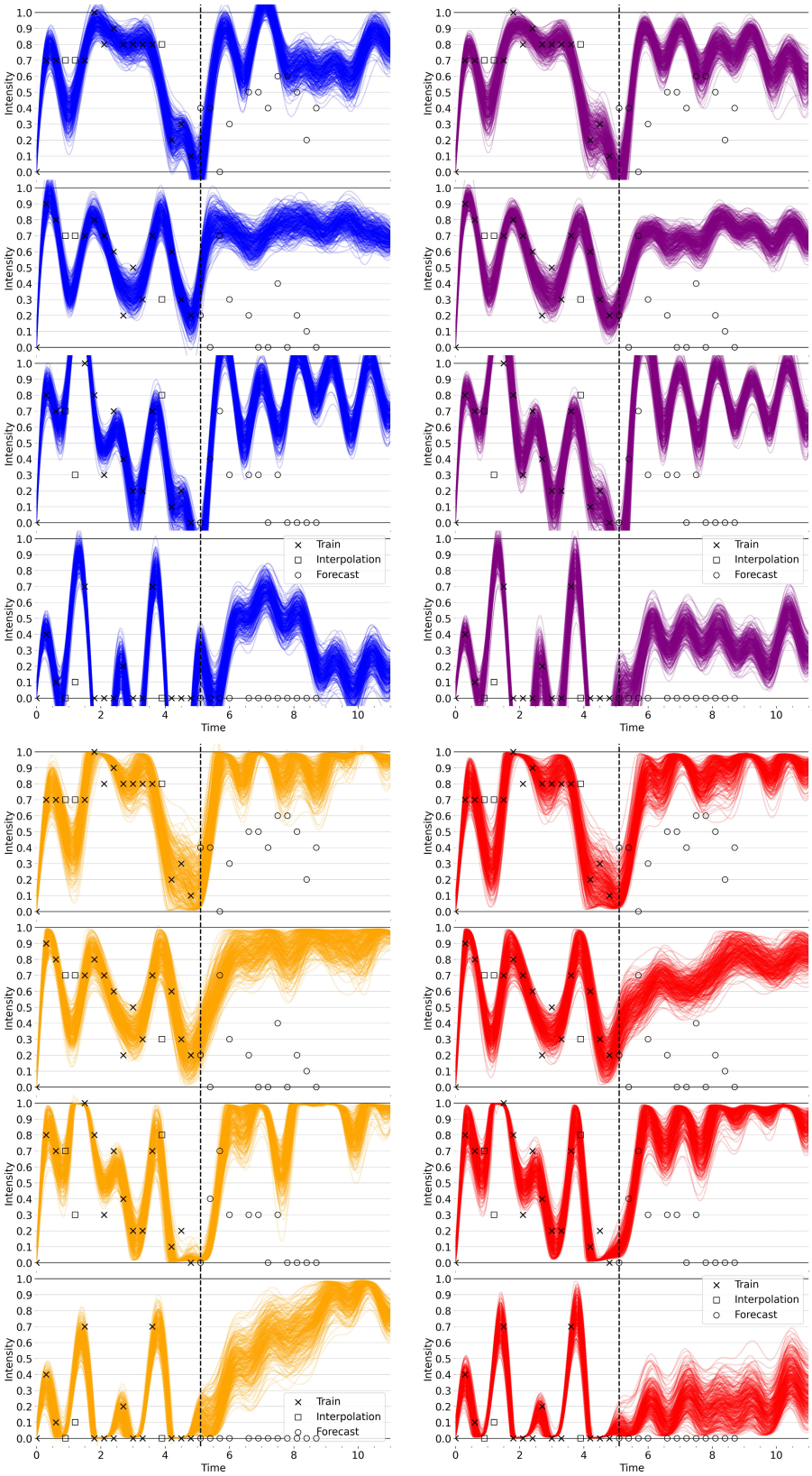

Figure 7. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

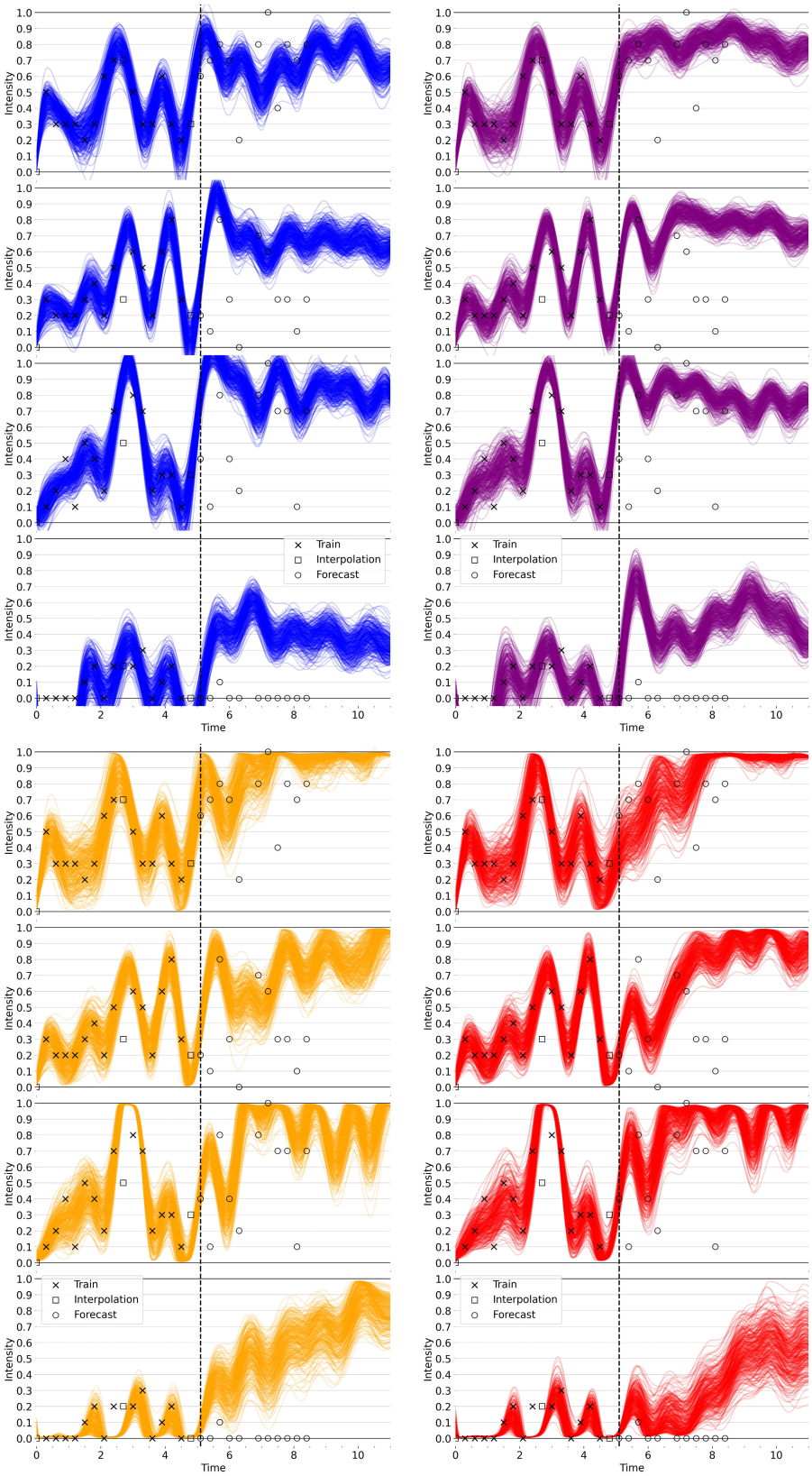

Figure 8. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

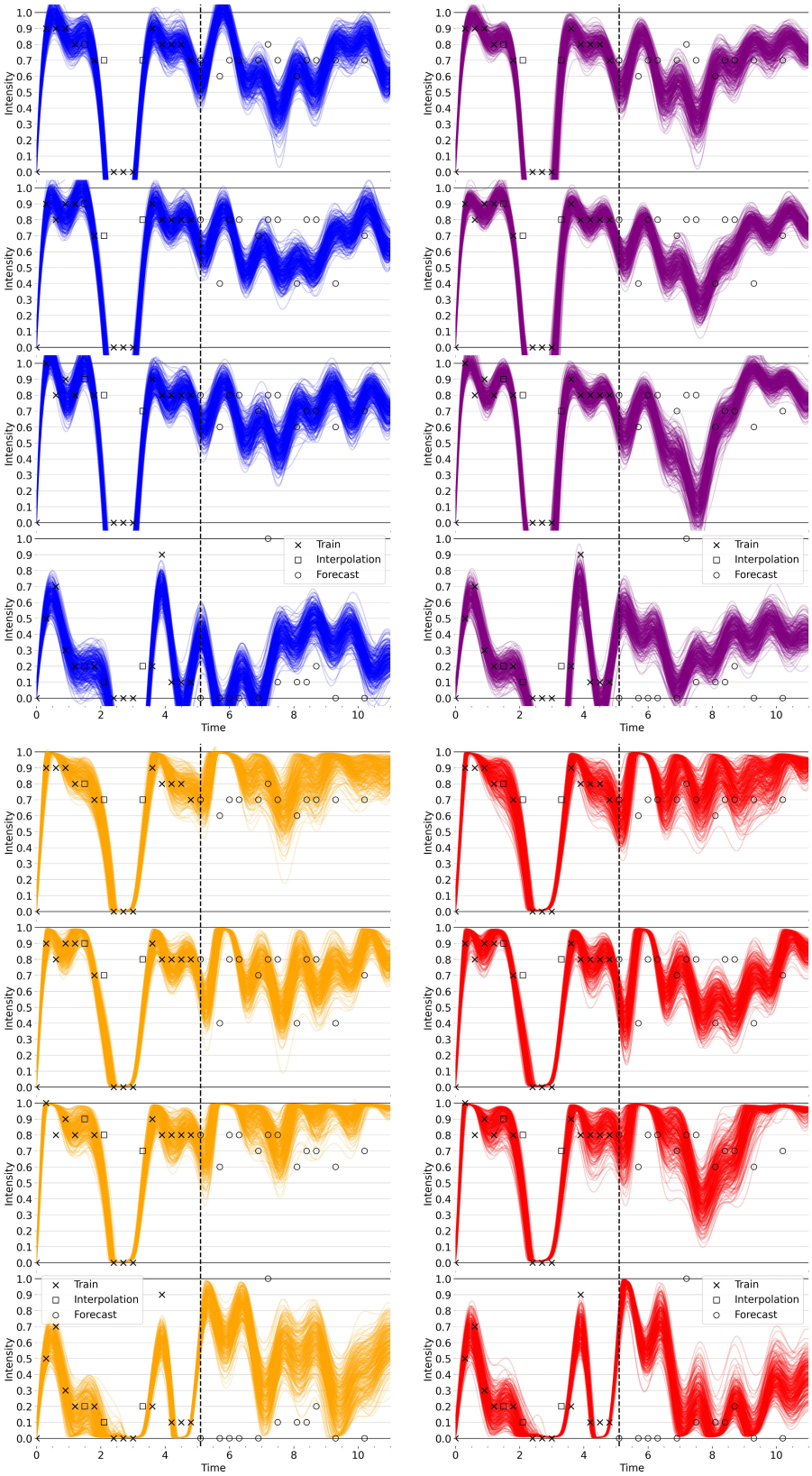

Figure 9. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

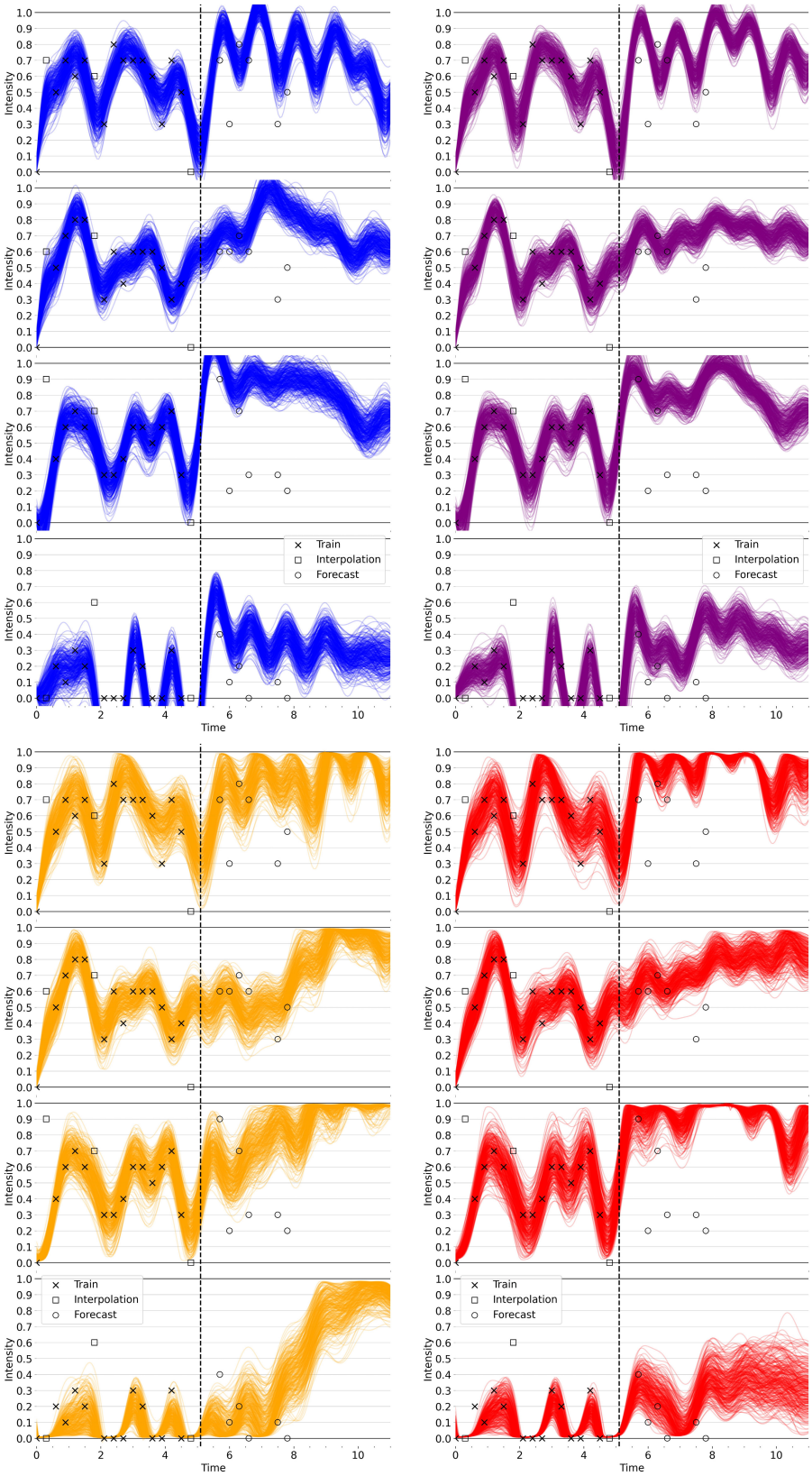

Figure 10. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

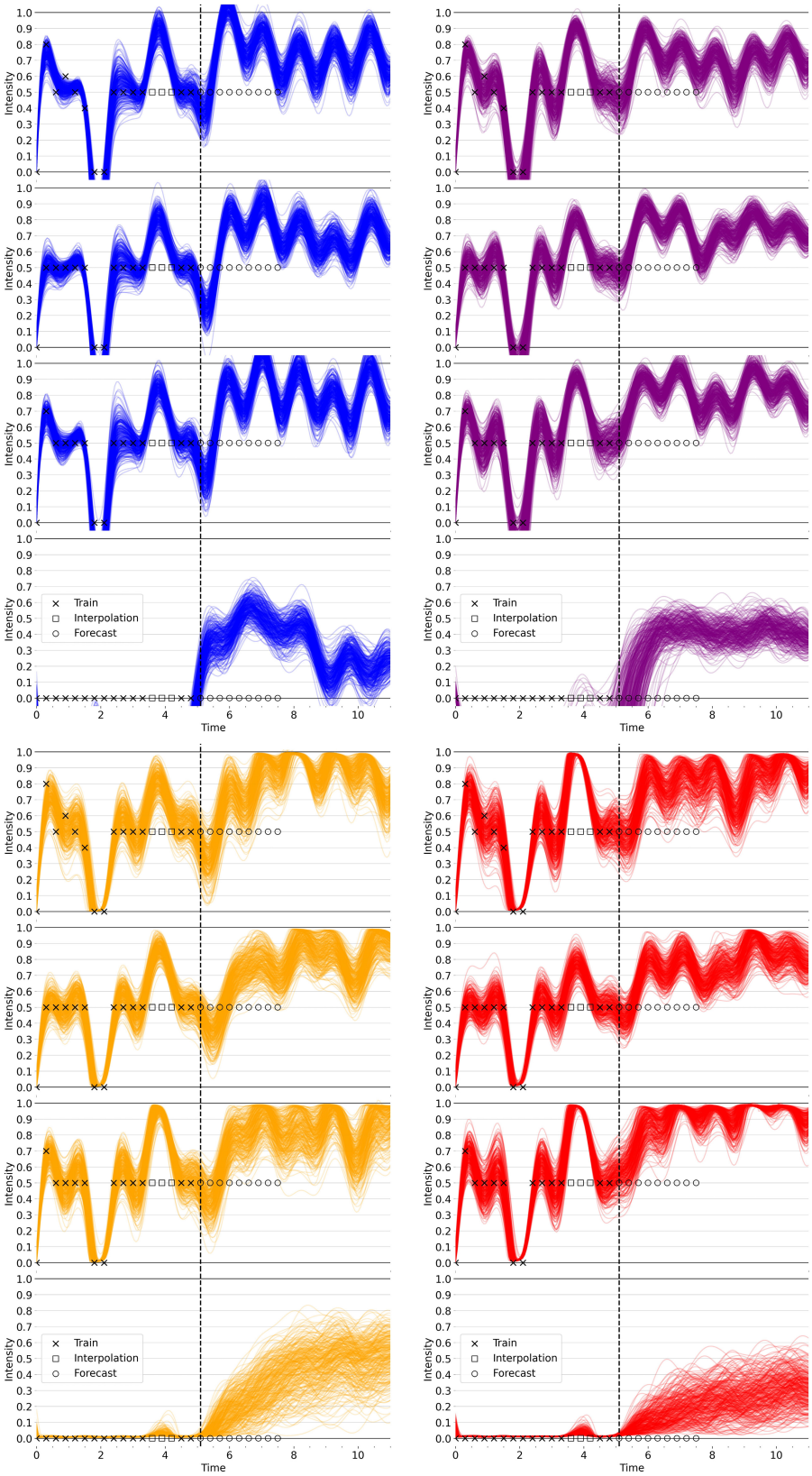

Figure 11. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. For this particular patient, all models make poor forecasts for the top three dimensions, but WSP does substantially improve forecasting for the fourth dimension (bottom row). See discussion in Appendix I.3.

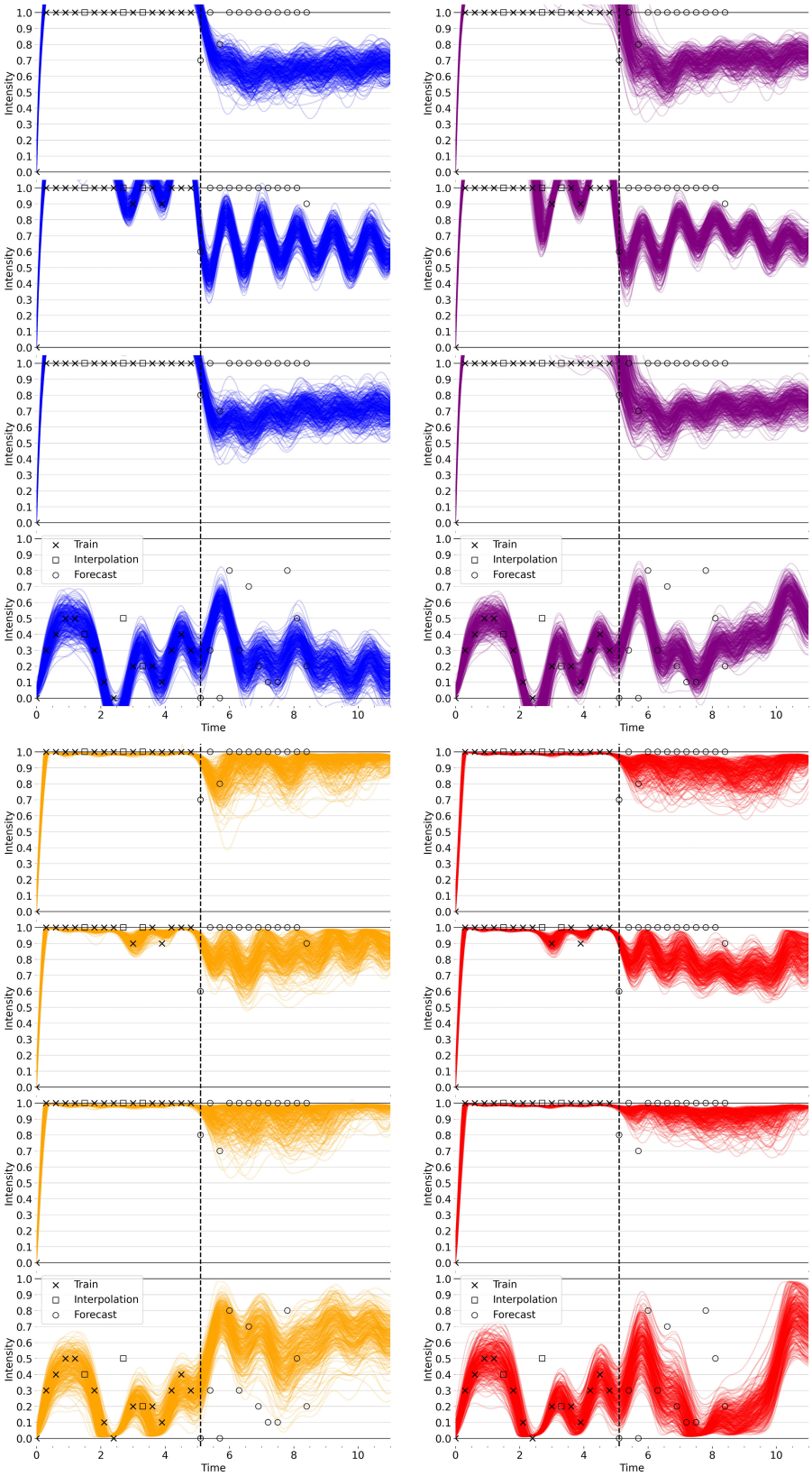

Figure 12. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

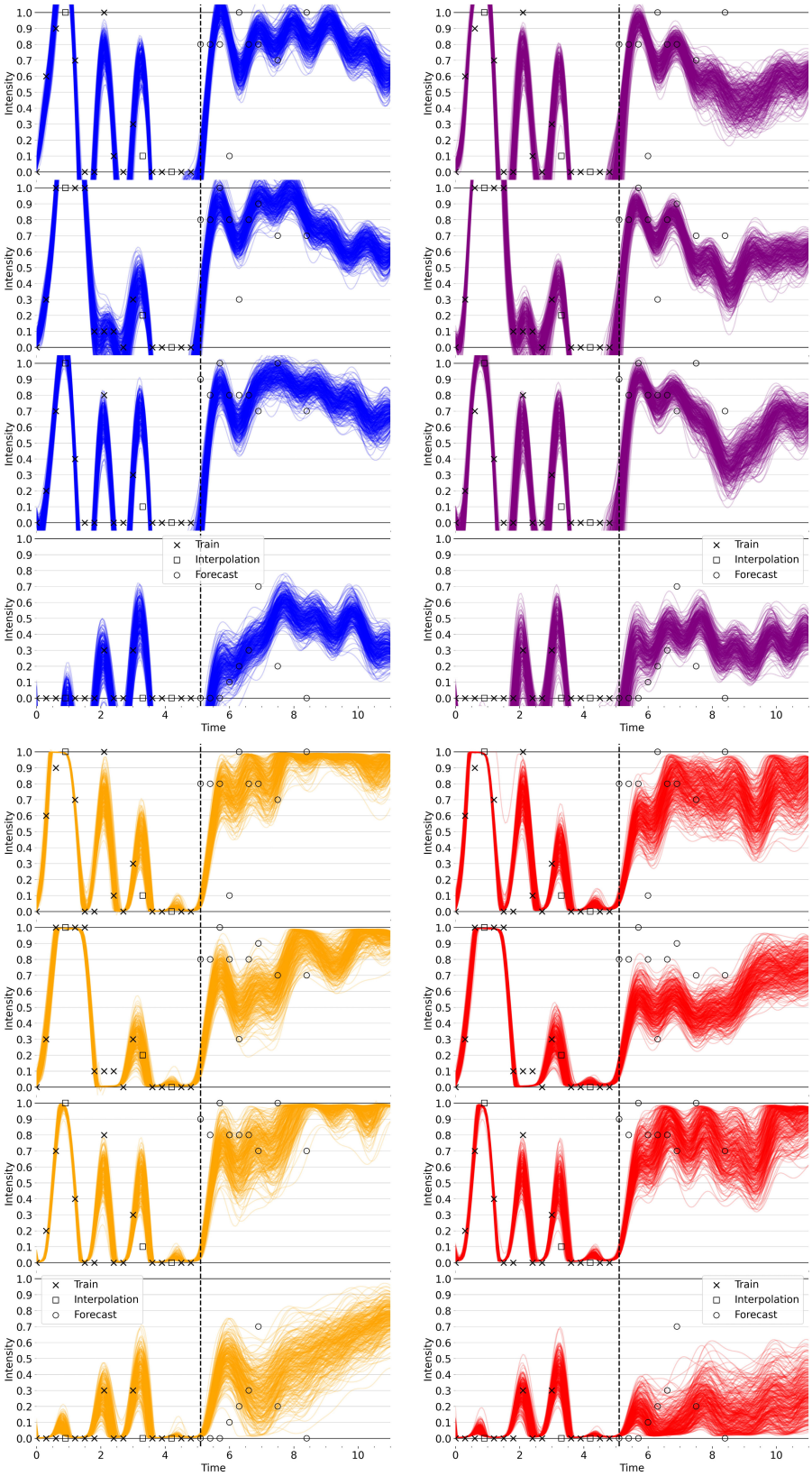

Figure 13. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

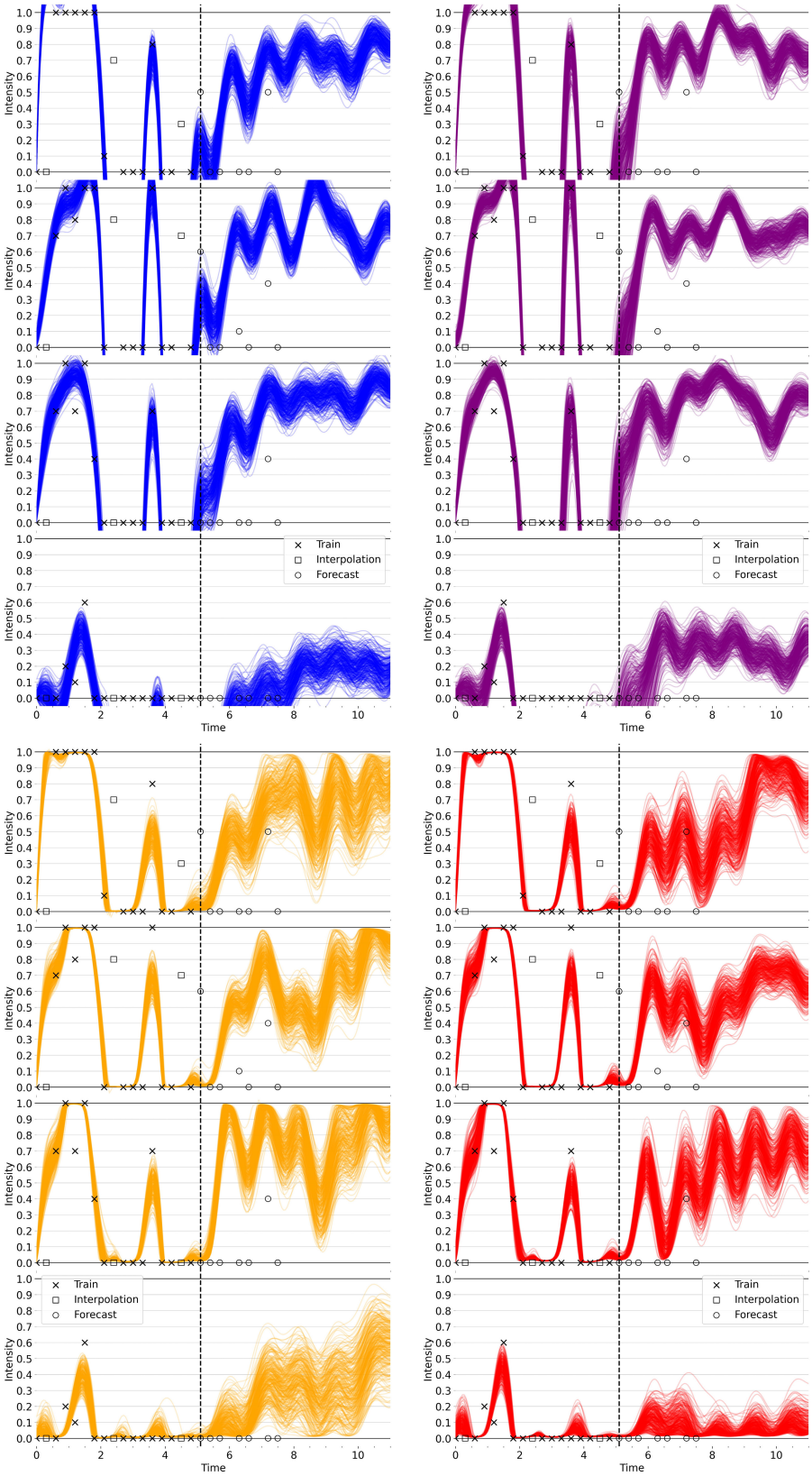

Figure 14. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

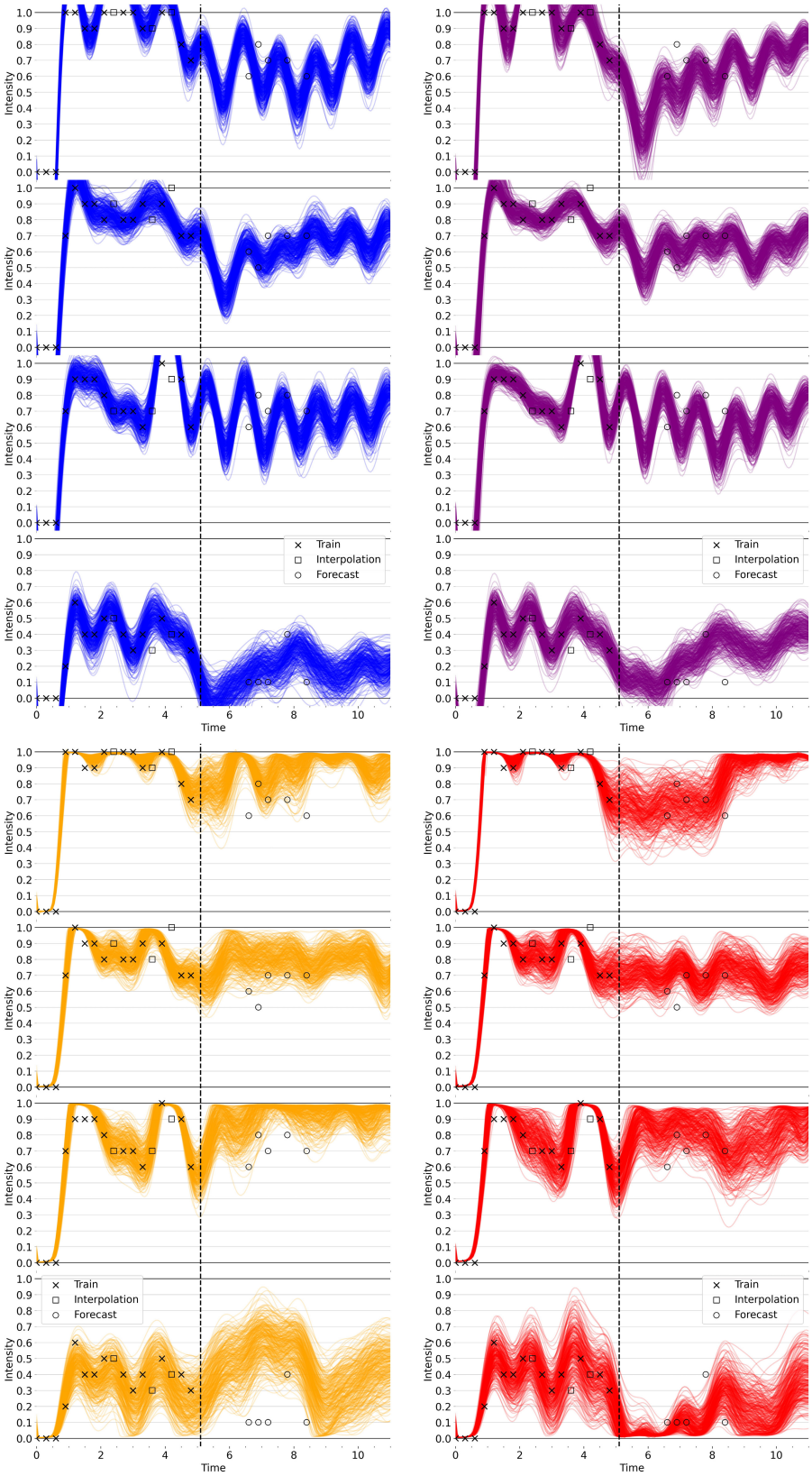

Figure 15. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

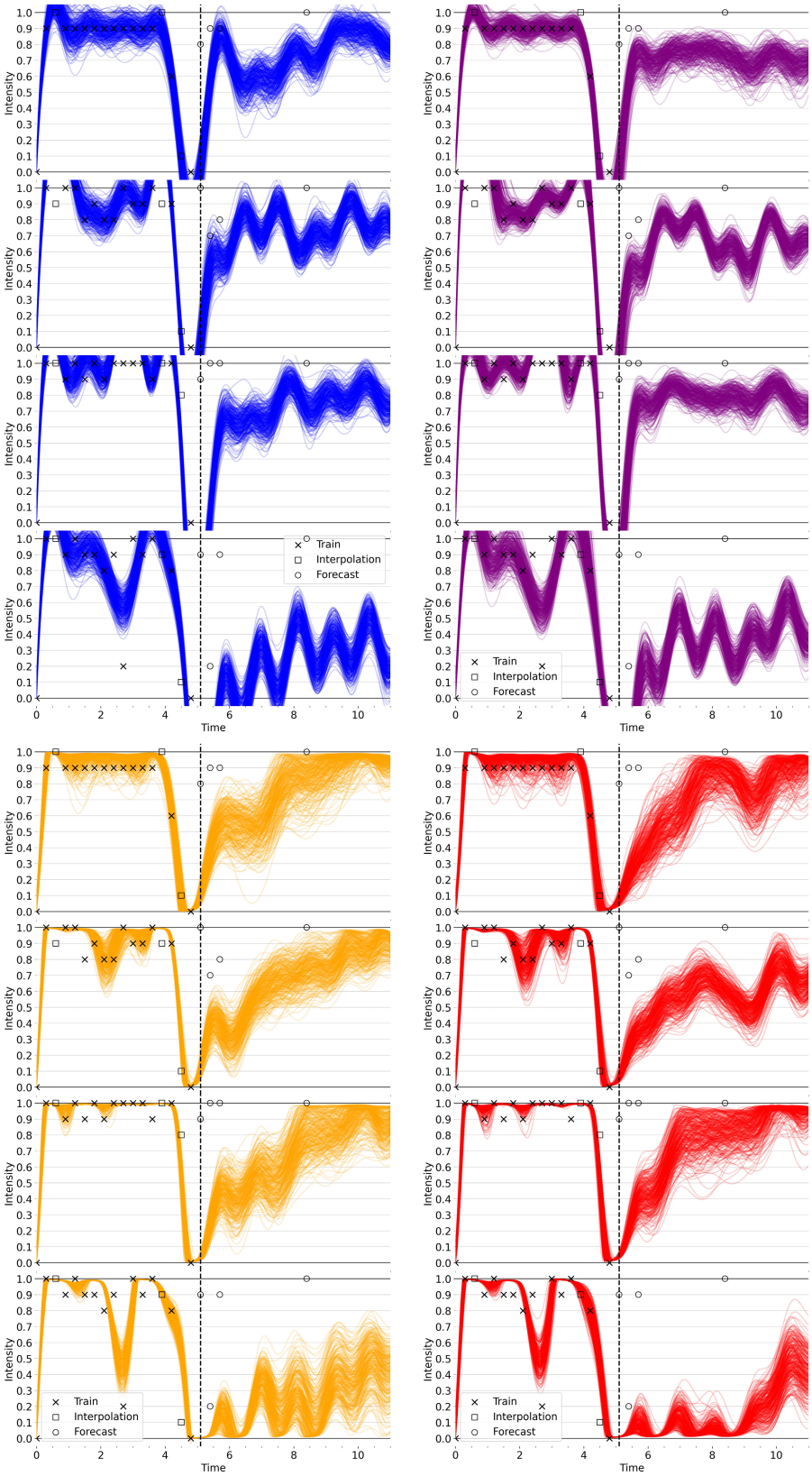

Figure 16. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

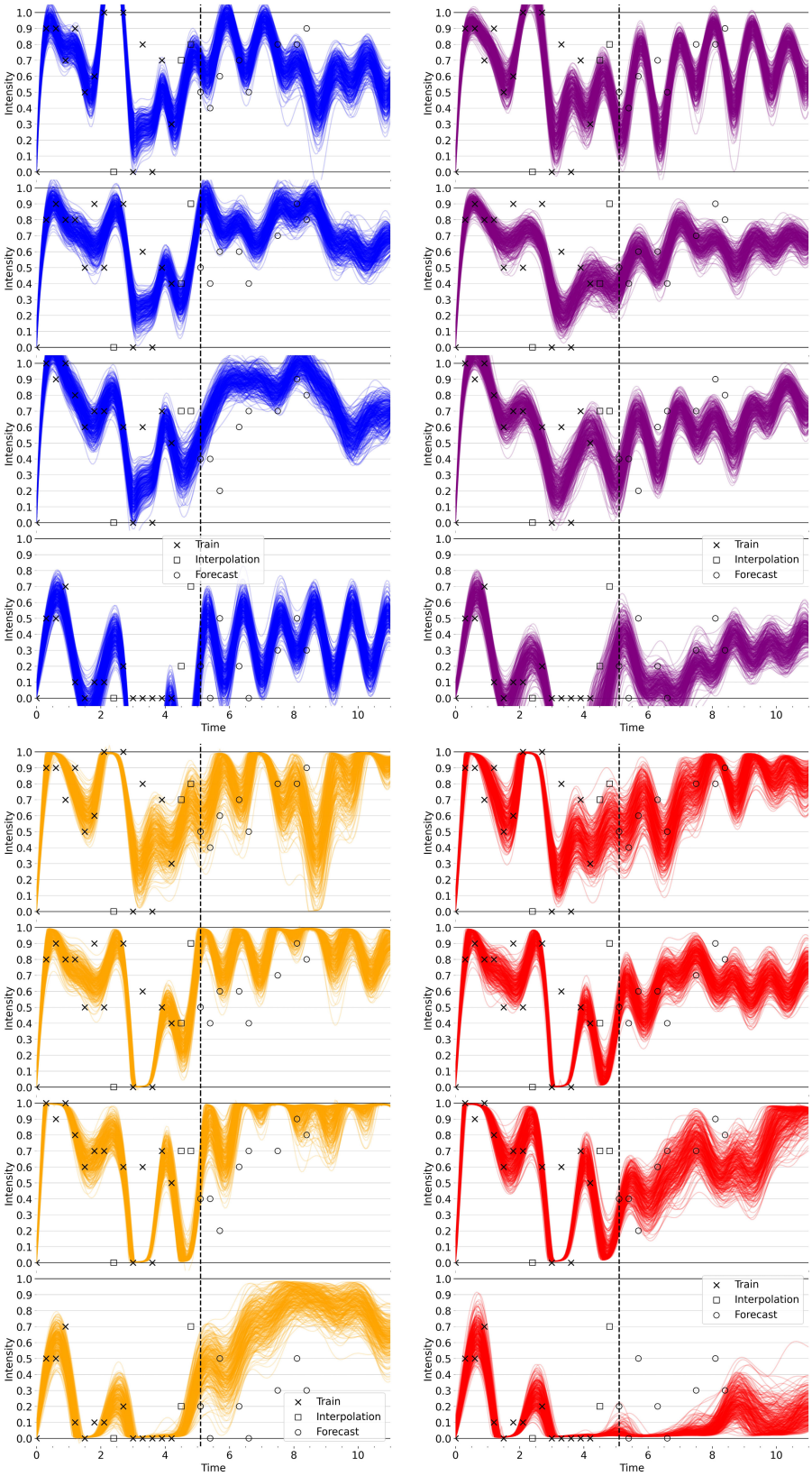

Figure 17. Qualitative Comparison of Posterior for Specific Patient from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. See discussion in Appendix I.3.

I.4 Visualizations of Latent SDE Prior Samples from the "SMART" Data

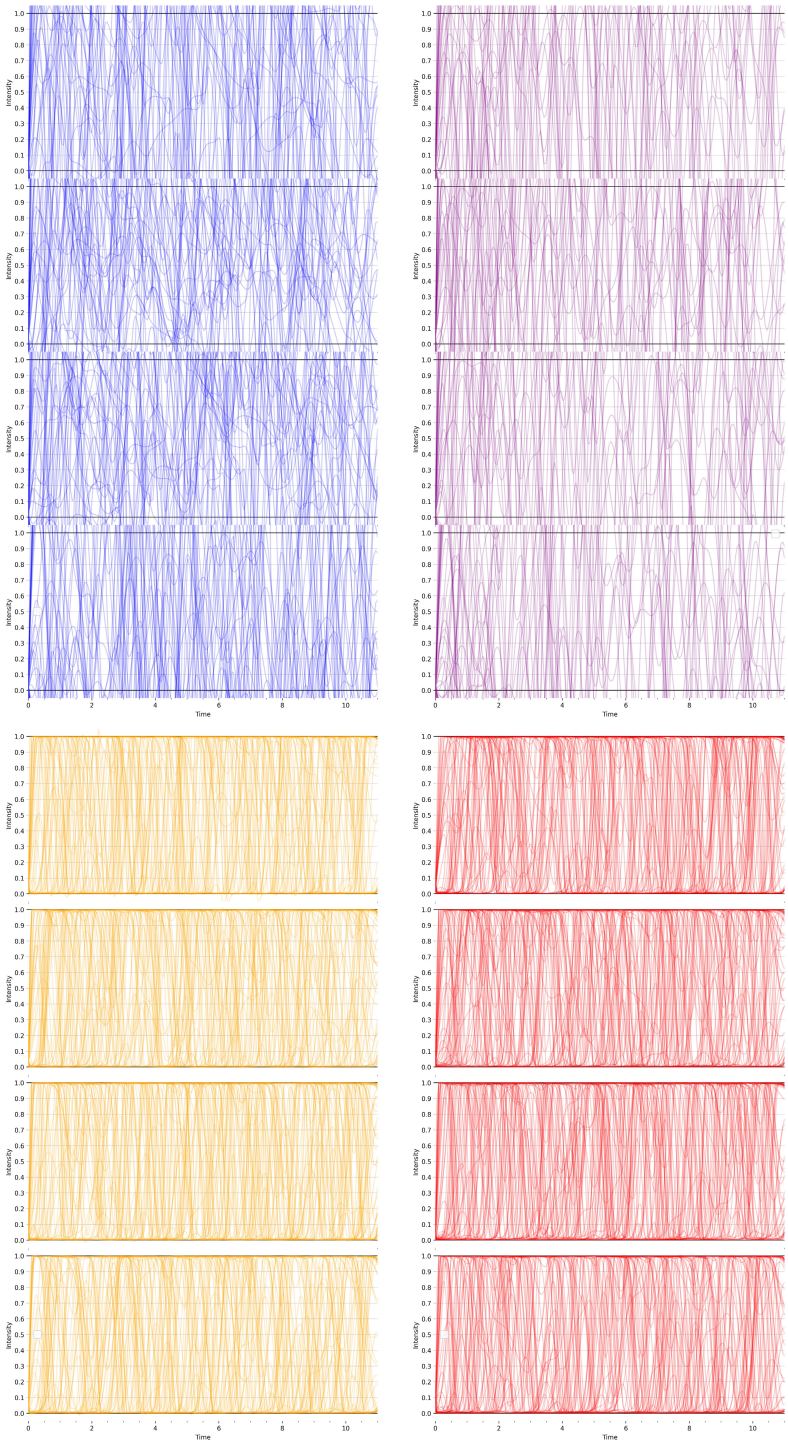

Figure 18. Qualitative Comparison of Prior Samples from Vanilla (top left), Vanilla+Clip (top right), WSP (bottom left), and WSP+Clip (bottom right). Each row represents a different EMA survey item, listed in Appendix H. WSP ensures prior samples remain viable whereas baselines do not.