ICML Rebuttal

1 Response to Reviewer zymX

We thank the reviewer for pointing out inconsistencies in our earlier proofs. We have thoroughly worked through our theorems and proofs to improve their rigor based on reviewers' comments, and we invite the reviewer to examine our revised proofs in the **Response to Reviewer GWvX.** We further provide clarifications to several major mis-understandings below.

- **Identifiability of c**: The identifiability of the generative model and $\mathbf{c}$ is jointly defined in both Definition 1 and Definition 2. More specifically , the $\sim_A$-identifiablity of the generative models is equivalent to an invertible relation $\mathbf{T}(\mathbf{c}) = \mathbf{A}\mathbf{T}(\hat{\mathbf{c}}) + \mathbf{B}$ as defined in Definition 2. We have corrected typos in Definition 2 and the proof for Theorem 1 (see details in Overall Response) to address this confusion.

- **Transition from Theorem 1 to Theorem 2:** We have followed the suggestions from Reviewer GWvX to use the $\sim_A$-identifiablity results in Theorem 1 to derive the relation between $p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{c})$ and $p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{c}})$ for Theorem 2. Neither the proof for Theorem 1 or Theorem 2 requires a discrete assumption for $\mathbf{c}$. The key to our Theorem 2 (and a major differentiating aspect from Song et al. (2024)) is the $\sim_A$-identifiablity results established in Theorem 1, which is established for a continuous random variable $\mathbf{c}$ that follows an exponential family of distributions.

- **Definition & the continuous nature of c**: Perhaps there was a misunderstanding regarding the continuous nature of c as a random variable *vs.* its *continuous variation* over time. Our $\mathbf{c}$ is a continuous random variable that conditions the generation of $\mathbf{x}_{0:\tau}$. Beyond this, our generative models or theorems do not model or require a temporal structure for $\mathbf{c}$. This is different from existing works that considers hidden Markov models to describe and temporal transition of latent nonstationary states. Our model is more similar to existing works (such as iVAE) that leverage observed labels to condition locally stationary segments, with the difference that our conditioning variable is latent (which is also our main contribute to prior arts).

- **Validation of the identifiability of c**: Throughout the experiments, we presented identifiaiblity results on $\mathbf{c}$, using MCC if it is neural and

MSE if it is physics-based (see Figs 2-3).

We have thoroughly revised our proofs and Theorems as shown in the **Response to Reviewer GWvX**. We apologize for the typos and inconsistencies in our earlier proofs that may have caused these mis-understandings, and we welcome further discussions on the revised version.

2 Response to Reviewer pAEd

We thank the reviewer for the appreciation of our work. We have worked through our theorems and proofs to improve their rigor based on other reviewers' comments, and we invite the reviewer to examine our revised proofs in the **Response to Reviewer GWvX.**

3 Response to Reviewer AqQJ

We thank the reviewer for the appreciation of our work. We have worked through our theorems and proofs to improve their rigor based on other reviewers' comments, and we invite the reviewer to examine our revised proofs in the **Response to Reviewer GWvX.**

- **Whether λ is fine-tuned: $\boldsymbol{\Theta}$** in Equation (6) in our original submission includes the latent dynamic function f and the mixing function from $\mathbf{z}_t$ to $\mathbf{x}_t$. λ_ζ parameterizes the latent distribution of $\mathbf{c}$; instead of being fine-tuned, it is extracted from our feed-forward meta-model $q_\zeta(\mathbf{c}|\mathcal{D}^s)$ directly from context data $\mathcal{D}^s$.
- **Computational/memory cost of meta-models**: Because we utilize a feedforward meta-model as explained above, the computation and memory is efficient; it will increase compared to a non-meta base model, but the increase should not be significant. We regrettably did not have time to add this comparison during the rebuttal, but will be happy to add these results in the final version.
- The first sentence of the second Introduction paragraph is intended to point out that, without identifiability established, many different solutions of a generative model may fit the same observations as well. We will revise this sentence along the other citations and figure suggestions by the reviewer.

4 Response to Reviewer GWvX

We are sincerely thankful to the reviewer for taking the time to provide the detailed suggestions on how we can improve our proof of theorems. It greatly helped us to navigate through the various approaches of proofs that exist in the

literature. We have thoroughly revised and corrected the proofs for our theorems. Before presenting the detailed proofs, we summarize the major revisions made:

- **Clarifications on Theorem 1**: We would like to clarify that our *non-stationary state* $\mathbf{c}$ is a continuous random variable whose conditional distribution $p(\mathbf{c}|u)$ follows an exponential family distribution. With Theorem 1, we prove that **our generative model is $\sim_A$-identifiable with an invertible relation $\mathbf{T}(\mathbf{c}) = \mathbf{AT}(\hat{\mathbf{c}}) + \mathbf{B}$ that is jointly defined in Definition 1 and Definition 2**. Our original Definition 2 and proofs did have many typos that may have confused the reviewers, which we corrected below.

- **Corrections on Theorem 2**: We made major corrections to **more explicitly connect Theorem 1 and Theorem 2**. More specifically, following Reviewer GWvX's suggestions, we introduced a Lemma 2 to establish the equivalence between $p(\mathbf{x}_{0:t}|\mathbf{c})$ and $p(\mathbf{x}_{0:t}|\hat{\mathbf{c}})$ based on $\sim_A$-identifiablity established in Theorem 1, from which we derive the relation between $p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{c})$ and $p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{c}})$ and use that to establish the conditions for the identifiaiblity of $\mathbf{z}_t$.

- **Stochasticity of latent dynamic functions:** Our latent dynamic function is stochastic as its parameters are generated from $\mathbf{c}$ which is random variable. Our $\mathbf{z}_0$ is also a random variable. We do not assume additional sources of stochasticity, *i.e.*, when conditioned on $\mathbf{c}$ and $\mathbf{z}_0$, $p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{c})$ is a Delta function. We have revised our proof process for Theorem 2 to reflect this.

- **Video data:** This is an excellent suggestion. In the current datasets considered, the data in the three simulated physics systems (pendulum, cart-pole, double pendulum) and the real pendulum data are in the form of sequence of images, *i.e.,* video data. In the future, we are very interested in exploring more video data from more complex and real-world systems.

- **Scalability in terms of structure learning:** Indeed. Our models work well for the current image data considered. We expect that trade-off between the structure learning and reconstruction/generation may be more affected by the complexity of the data to be modeled, especially those with complex "texture" features beyond the dynamics features.

We will address the rest of the writing and reference suggestions in the final manuscript. We believe that the rigor of our theorems have been substantially improved thanks to the detailed comments from the reviewer. We are eager for further discussion on the revised version presented.

Time-Series Generative Model

Invertible Mixing Function: $\mathbf{z}_t \in \mathcal{Z} \in \mathbb{R}^n \to \mathbf{x}_t \in \mathcal{X} \in \mathbb{R}^m$:

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$$

Dynamic Function:

$$\frac{dz_{kt}}{dt} = f_{kt}(\mathbf{z}_t, \mathbf{c}) \Leftrightarrow \frac{d\mathbf{z}_t}{dt} = \mathbf{f}(\mathbf{z}_t, \mathbf{c}), \quad \mathbf{c} \in \mathcal{C} \in \mathbb{R}^d \tag{1}$$

$$\mathbf{x}_t = \mathbf{g}\left(\mathbf{z}_0 + \int_{\tau=0}^{t} \mathbf{f}(\mathbf{z}_t, \mathbf{c})ds\right) = F_{\mathbf{g},\mathbf{f},t}(\mathbf{z}_0, \mathbf{c}), \quad t = 0, 1, \ldots, \tau \tag{2}$$

$$\mathbf{x}_{0:\tau} = [F_{\mathbf{g},\mathbf{f},0}(\mathbf{z}_0, \mathbf{c}), \ldots, F_{\mathbf{g},\mathbf{f},\tau}(\mathbf{z}_0, \mathbf{c})] = \mathbf{F}_{\mathbf{g},\mathbf{f}}(\mathbf{z}_0, \mathbf{c}) \tag{3}$$

Suppose $\mathbf{F}_{\mathbf{g},\mathbf{f}}(\mathbf{z}_0, \mathbf{c})$ is invertible function: $(\mathbf{z}_0, \mathbf{c}) = \mathbf{F}^{-1}_{\mathbf{g},\mathbf{f}}(\mathbf{x}_{0:\tau})$. We further denote $\mathbf{c} = \mathbf{F}^{-1,\mathbf{c}}_{\mathbf{g},\mathbf{f}}(\mathbf{x}_{0:\tau})$ and $\mathbf{z}_0 = \mathbf{F}^{-1,\mathbf{z}_0}_{\mathbf{g},\mathbf{f}}(\mathbf{x}_{0:\tau})$.

Conditional Prior:

$$p_\lambda(\mathbf{c}|\mathbf{u}) = \frac{Q(\mathbf{c})}{Z(u)} \exp(\langle \mathbf{T}(\mathbf{c}), \boldsymbol{\lambda}(\mathbf{u}) \rangle) \tag{4}$$

$$\mathbf{T}(\mathbf{c}) = (\mathbf{T}_1(c_1), \ldots, \mathbf{T}_n(c_d)) = (T_{1,1}(c_1), \ldots, T_{d,k}(c_d)) \in \mathbb{R}^{dk} \tag{5}$$

$$\boldsymbol{\lambda}(\mathbf{u}) = (\boldsymbol{\lambda}_1(\mathbf{u}), \ldots, \boldsymbol{\lambda}_n(\mathbf{u})) = (\lambda_{1,1}(\mathbf{u}), \ldots, \lambda_{d,k}(\mathbf{u})) \in \mathbb{R}^{dk} \tag{6}$$

Observation Joint Distribution:

$$p_\Theta(\mathbf{x}_{0:\tau}|\mathbf{u}) = \iint p_{\mathbf{g},\mathbf{f}}(\mathbf{x}_{0:\tau}|\mathbf{c}, \mathbf{z}_0) p_{\boldsymbol{\lambda}}(\mathbf{c}|\mathbf{u}) p(\mathbf{z}_0) d\mathbf{c} d\mathbf{z}_0, \quad \Theta = (\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda}) \tag{7}$$

Theorem 1

Definition 1: *Formally let $\mathbf{x}_{0:\tau}$ be an observed time series generated by the generative processes specified by Equation (7) with $\Theta = (\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda})$. A model $\hat{\Theta} = (\hat{\mathbf{g}}, \hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}})$ is observationally equivalent to Θ if $p_{\hat{\Theta}}(\mathbf{x}_{0:\tau})$ matches $p_\Theta(\mathbf{x}_{0:\tau})$ everywhere. Let $\sim$ be an equivalence relation on Θ, we say that Equation (7) is $\sim$ identifiable if*

$$p_\Theta(\mathbf{x}_{0:\tau}) = p_{\hat{\Theta}}(\mathbf{x}_{0:\tau}) \Rightarrow \Theta \sim \hat{\Theta}. \tag{8}$$

Definition 2: *Let $\sim$ be the equivalence relation on $\Theta = (\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda})$ defined as follows:*

$$(\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda}) \sim (\hat{\mathbf{g}}, \hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}}) \Leftrightarrow \tag{9}$$

$$\exists \mathbf{A}, \mathbf{B} \mid \mathbf{T}\left(\mathbf{F}^{-1,\mathbf{c}}_{\mathbf{g},\mathbf{f}}(\mathbf{x}_{0:\tau})\right) = \mathbf{A}\mathbf{T}\left(\mathbf{F}^{-1,\mathbf{c}}_{\hat{\mathbf{g}},\hat{\mathbf{f}}}(\mathbf{x}_{0:\tau})\right) + \mathbf{B}, \forall \mathbf{x}_{0:\tau} \in \mathcal{X}^{\tau+1} \tag{10}$$

where $\mathbf{A}$ is an $dk \times dk$ matrix and $\mathbf{B}$ is a dim-dk vector. If A is invertible, we denote this relation by $\sim_A$.

Theorem 1: *Assume that we observe data sampled from the above generative model with parameters $\Theta = (\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda})$. Assume the following holds:*

(i) *The function* $\mathbf{F}_{\mathbf{g},\mathbf{f}}(\mathbf{z}_0, \mathbf{c})$ *is injective.*

(ii) *The sufficient statistics* $T_{i,j}$ *are differentiable almost everywhere, and* $(T_{i,j})_{1 \leq j \leq k}$ *are linearly independent on any subset of* $\mathcal{X}$ *of measure greater than zero.*

(iii) *There exist* $dk + 1$ *distinct points* $\mathbf{u}^0, ..., \mathbf{u}^{dk}$ *such that the matrix:*

$$\mathbf{L} = (\boldsymbol{\lambda}(\mathbf{u}_1) - \boldsymbol{\lambda}(\mathbf{u}_0), ..., \boldsymbol{\lambda}(\mathbf{u}_{dk}) - \boldsymbol{\lambda}(\mathbf{u}_0)) \tag{11}$$

Of size $dk \times dk$ *is invertible.*

then the parameters $\Theta = (\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda})$ *are* $\sim_A$*-identifiable.*

Before proving Theorem 1, we first introduce definition 2 and Lemma 1.

Definition 2: *We say that an exponential family distribution is strongly exponential if for any subset* $\mathcal{X}$ *of* $\mathbb{R}$ *the following is true:*

$$\left(\exists \boldsymbol{\theta} \in \mathbb{R}^k \mid \forall x \in \mathcal{X}, \langle \mathbf{T}(x), \boldsymbol{\theta} \rangle = const\right) \Rightarrow (l(\mathcal{X}) = 0 \text{ or } \boldsymbol{\theta} = \mathbf{0}) \tag{12}$$

where l *is the Lebesgue measure.*

Lemma 1: *Consider a strongly exponential distribution of size* $k \geq 2$ *with sufficient statistic* $\mathbf{T}(x) = (T_1(x), \ldots, T_k(x))$*. Further assume that* $\mathbf{T}$ *is differentiable almost everywhere. Then there exist* k *distinct values* x_1 *to* x_k *that* $(\mathbf{T}(x_1), \ldots, \mathbf{T}(x_k))$ *are linearly independent in* $\mathbb{R}^k$*.*

Proof of Theorem 1: Suppose we have two sets of parameters $\Theta = (\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda})$ and $\hat{\Theta} = (\hat{\mathbf{g}}, \hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}})$ such that:

$$p_\Theta(\mathbf{x}_{0:\tau}|\mathbf{u}) = p_{\hat{\Theta}}(\mathbf{x}_{0:\tau}|\mathbf{u}) \tag{13}$$

$$\int p_{\mathbf{g},\mathbf{f}}(\mathbf{x}_{0:\tau}|\mathbf{c}, \mathbf{z}_0) p_{\boldsymbol{\lambda}}(\mathbf{c}|\mathbf{u}) p(\mathbf{z}_0) d(\mathbf{c}, \mathbf{z}_0) = \int p_{\hat{\mathbf{g}},\hat{\mathbf{f}}}(\mathbf{x}_{0:\tau}|\mathbf{c}, \mathbf{z}_0) p_{\hat{\boldsymbol{\lambda}}}(\mathbf{c}|\mathbf{u}) p(\mathbf{z}_0) d(\mathbf{c}, \mathbf{z}_0) \tag{14}$$

To simplify without loss of generality, we rewrite the above equation as:

$$\int p(\mathbf{x}_{0:\tau}|\mathbf{c}, \mathbf{z}_0) p(\mathbf{c}|\mathbf{u}) p(\mathbf{z}_0) d(\mathbf{c}, \mathbf{z}_0) = \int p(\mathbf{x}_{0:\tau}|\hat{\mathbf{c}}, \hat{\mathbf{z}}_0) p(\hat{\mathbf{c}}|\mathbf{u}) p(\hat{\mathbf{z}}_0) d(\hat{\mathbf{c}}, \hat{\mathbf{z}}_0) \tag{15}$$

$$\int \delta(\mathbf{x}_{0:\tau} - \mathbf{F}_{\mathbf{g},\mathbf{f}}(\mathbf{c}, \mathbf{z}_0)) p(\mathbf{c}|\mathbf{u}) p(\mathbf{z}_0) d(\mathbf{c}, \mathbf{z}_0) = \int \delta(\mathbf{x}_{0:\tau} - \mathbf{F}_{\hat{\mathbf{g}},\hat{\mathbf{f}}}(\hat{\mathbf{c}}, \hat{\mathbf{z}}_0)) p(\hat{\mathbf{c}}|\mathbf{u}) p(\hat{\mathbf{z}}_0) d(\hat{\mathbf{c}}, \hat{\mathbf{z}}_0) \tag{16}$$

$$p(\mathbf{c}|\mathbf{u}) p(\mathbf{z}_0) |\mathbf{J}_{\mathbf{F}^{-1}_{\mathbf{g},\mathbf{f}}}(\mathbf{x}_{0:\tau})| = p(\hat{\mathbf{c}}|\mathbf{u}) p(\hat{\mathbf{z}}_0) |\mathbf{J}_{\mathbf{F}^{-1}_{\hat{\mathbf{g}},\hat{\mathbf{f}}}}(\mathbf{x}_{0:\tau})| \tag{17}$$

Equation (17) can also be derived from Equation (15) with a change of variables.

Now consider two different values of $\mathbf{u}$: $\mathbf{u}_i$, $\mathbf{u}_0$

$$\begin{aligned}&\log p(\mathbf{c}|\mathbf{u}_i)p(\mathbf{z}_0)|\det J_{\mathbf{F}_{\mathbf{g},\mathbf{f}}^{-1}}(\mathbf{x}_{0:\tau})| - \log p(\mathbf{c}|\mathbf{u}_0)p(\mathbf{z}_0)|\det J_{\mathbf{F}_{\mathbf{g},\mathbf{f}}^{-1}}(\mathbf{x}_{0:\tau})| \\ &= \log\frac{Z(\mathbf{u}_0)}{Z(\mathbf{u}_i)} + \langle \mathbf{T}(\mathbf{c}), \boldsymbol{\lambda}(\mathbf{u}_i) - \boldsymbol{\lambda}(\mathbf{u}_0)\rangle = \log\frac{Z(\mathbf{u}_0)}{Z(\mathbf{u}_i)} + \langle \mathbf{T}(\mathbf{c}), \bar{\boldsymbol{\lambda}}(\mathbf{u}_i)\rangle \end{aligned} \tag{18}$$

Considering the same for $p(\hat{\mathbf{c}}|\mathbf{u})p(\hat{\mathbf{z}}_0)|\mathbf{J}_{\mathbf{F}_{\hat{\mathbf{g}},\hat{\mathbf{f}}}}(\bar{\mathbf{x}}_{0:\tau})|$, we get:

$$\log\frac{Z(\mathbf{u}_0)}{Z(\mathbf{u}_i)} + \langle \mathbf{T}(\mathbf{c}), \bar{\boldsymbol{\lambda}}(\mathbf{u}_i)\rangle = \log\frac{\hat{Z}(\mathbf{u}_0)}{\hat{Z}(\mathbf{u}_i)} + \langle \mathbf{T}(\hat{\mathbf{c}}), \bar{\boldsymbol{\lambda}}(\mathbf{u}_i)\rangle \tag{19}$$

$$\langle \mathbf{T}(\mathbf{c}), \bar{\boldsymbol{\lambda}}(\mathbf{u}_i)\rangle = \langle \mathbf{T}(\hat{\mathbf{c}}), \bar{\boldsymbol{\lambda}}(\mathbf{u}_i)\rangle + \log\frac{\hat{Z}(\mathbf{u}_0)Z(\mathbf{u}_i)}{Z(\mathbf{u}_i)Z(\mathbf{u}_0)} = \langle \mathbf{T}(\hat{\mathbf{c}}), \bar{\boldsymbol{\lambda}}(\mathbf{u}_i)\rangle + b_i \tag{20}$$

Based on assumption (iii), for $dk+1$ distinct points $\mathbf{u}^0, ..., \mathbf{u}^{dk}$, we get

$$\mathbf{L}^T\mathbf{T}(\mathbf{c}) = \hat{\mathbf{L}}^T\mathbf{T}(\hat{\mathbf{c}}) + \mathbf{b} \tag{21}$$

$$\mathbf{T}(\mathbf{c}) = \mathbf{A}\mathbf{T}(\hat{\mathbf{c}}) + \bar{\mathbf{b}}, \quad \mathbf{A} = \mathbf{L}^{-T}\hat{\mathbf{L}} \text{ and } \bar{\mathbf{b}} = \mathbf{L}^{-T}\mathbf{b} \tag{22}$$

$$\mathbf{T}(\mathbf{c}) = \mathbf{A}\mathbf{T}(\hat{\mathbf{c}}) + \bar{\mathbf{b}} \tag{23}$$

Based on assumption (ii) and definition 2, we know $p(\mathbf{c}|\mathbf{u})$ is strongly exponential. Since $J_{\mathbf{T}}(\mathbf{c})$ exists and is an $dk \times d$ matrix of rank d, which implies the rank of $\mathbf{A}$ is d. We distinguish two cases:

- If $k = 1$, then this means that $\mathbf{A}$ is invertible (because $\mathbf{A}$ is $d \times d$).
- Based on Lemma 1, for each $i \in [1, \ldots, d]$, define $\mathbf{T}_i(c_i) = (T_{i,1}(c_i), \ldots, T_{i,k}(c_i))$, there exist k points $c_i^1, \ldots, c_i^k$ that $(\mathbf{T}_i(c_i^1), \ldots, \mathbf{T}_i(c_i^k))$ are linearly independent. Collect those points into k vectors $(\mathbf{c}^1, \ldots, \mathbf{c}^k)$, and concatenate the k Jacobians the matrix $\mathbf{Q} = (\mathbf{J}_{\mathbf{T}}(\mathbf{c}^1), \ldots, \mathbf{J}_{\mathbf{T}}(\mathbf{c}^k))$. Then the matrix $\mathbf{Q}$ is invertible (through a combination of Lemma 1 and the fact that each component of T univariate). Then define,

  $$\hat{\mathbf{Q}} = \left(\mathbf{J}_{\mathbf{T}\left(\mathbf{F}_{\mathbf{g},\mathbf{f}}^{-1,\mathbf{c}} \circ \mathbf{F}_{\hat{\mathbf{g}},\hat{\mathbf{f}}}\right)}(\mathbf{c}^1), \ldots, \mathbf{J}_{\mathbf{T}\left(\mathbf{F}_{\mathbf{g},\mathbf{f}}^{-1,\mathbf{c}} \circ \mathbf{F}_{\hat{\mathbf{g}},\hat{\mathbf{f}}}\right)}(\mathbf{c}^k)\right) \tag{24}$$

  we get $\mathbf{Q} = \mathbf{A}\hat{\mathbf{Q}}$. The invertibility of $\mathbf{Q}$ implies the invertibility of $\mathbf{A}$ and $\hat{\mathbf{Q}}$.

Theorem 2:

Theorem 2: *Suppose we have two sets of parameters and they are $\sim_A$-identifiable that, $(\mathbf{g}, \mathbf{f}, \boldsymbol{\lambda}) \sim_A (\hat{\mathbf{g}}, \hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}})$ and $\hat{\mathbf{g}}$ is an invertible function, that $\hat{\mathbf{z}}_t = \hat{\mathbf{g}}(\mathbf{x}_t)$. Let*

$\eta_{kt}(\mathbf{c}) = z_{k,t-1} + \int_{t-1}^{t} f_k(\mathbf{z}_s, \mathbf{c})ds$ and:

$$\mathbf{v}_{kt}(\mathbf{c}) \triangleq \left(\frac{\partial^2 \eta_{kt}(\mathbf{c})}{\partial z_{kt} \partial z_{1,t-1}}, \ldots, \frac{\partial^2 \eta_{kt}(\mathbf{c})}{\partial z_{kt} \partial z_{n,t-1}}, \frac{\partial \eta_{kt}(\mathbf{c})}{\partial z_{kt}} \right)^T \tag{25}$$

$$\dot{\mathbf{v}}_{kt}(\mathbf{c}) \triangleq \left(\frac{\partial^3 \eta_{kt}(\mathbf{c})}{\partial z_{kt}^2 \partial z_{1,t-1}}, \ldots, \frac{\partial^3 \eta_{kt}(\mathbf{c})}{\partial z_{kt}^2 \partial z_{n,t-1}}, \frac{\partial^2 \eta_{kt}(\mathbf{c})}{\partial z_{kt}^2} \right) \tag{26}$$

If for each value of $\mathbf{z}_t (t > 0)$, $\mathbf{v}_{1t}, \dot{\mathbf{v}}_{1t}, \ldots, \mathbf{v}_{nt}, \dot{\mathbf{v}}_{nt}$, as $2n$ function vectors are linearly independent, then $\hat{\mathbf{z}}_t$ must be an invertible, component-wise transformation of a permuted version of $\mathbf{z}_t$, that $\hat{\mathbf{g}}^{-1}(\mathbf{x}_t) = T \circ \pi \circ \mathbf{g}^{-1}(\mathbf{x}_t)$, where π is a permutation and T is a component-wise invertible transformation.

Before proving Theorem 2, we first introduce Lemma 2.

Lemma 2: *Consider* $p_\Theta(\mathbf{x}_{0:t}|\mathbf{c})$ *and* $p_{\hat{\Theta}}(\mathbf{x}_{0:t}|\hat{\mathbf{c}})$ *for* $t > 0$, *where* $\Theta \sim_A \hat{\Theta}$ *and* $\mathbf{T}(\mathbf{c}) = \mathbf{A}\mathbf{T}(\hat{\mathbf{c}}) + \bar{\mathbf{b}}$ *as discussed in Definition 2 and Theorem 1. Then* $p_\Theta(\mathbf{x}_{0:t}|\mathbf{c}) = p_{\hat{\Theta}}(\mathbf{x}_{0:t}|\hat{\mathbf{c}})$.

Since $\mathbf{A}$ is invertible and $\mathbf{T}$ is sufficient statistics for the exponential family distribution as defined in Equation (4), $\mathbf{c}$ and $\hat{\mathbf{c}}$ follows an invertible relation which we denote as $\hat{\mathbf{c}} = w(\mathbf{c})$, with which we have $p(\hat{\mathbf{c}}) = p(\mathbf{c})|\mathbf{J}_w(\mathbf{c})|^{-1}$. Then:

$$p(\mathbf{x}_{0:t}|\hat{\mathbf{c}}) = \frac{p(\mathbf{x}_{0:t}, \hat{\mathbf{c}})}{p(\hat{\mathbf{c}})} = \frac{p(\mathbf{x}_{0:t}, \mathbf{c})|\mathbf{J}_{\mathbf{w}}(\mathbf{c})|^{-1}}{p(\mathbf{c})|\mathbf{J}_{\mathbf{w}}(\mathbf{c})|^{-1}} = p(\mathbf{x}_{0:t}|\mathbf{c}) \tag{27}$$

Proof of Theorem 2: From Lemma 2 and with a change of variable, we have

$$p(\hat{\mathbf{z}}_{0:t}|\hat{\mathbf{c}}) \prod_{i=1}^{t} |\mathbf{J}_{\mathbf{g}}(\hat{\mathbf{z}}_i)|^{-1} = p(\mathbf{z}_{0:t}|\mathbf{c}) \prod_{i=1}^{t} |\mathbf{J}_{\mathbf{g}}(\mathbf{z}_i)|^{-1} \tag{28}$$

where

$$p(\mathbf{z}_{0:t}|\mathbf{c}) = p(\mathbf{z}_0) \prod_{i=1}^{t} p(\mathbf{z}_i|\mathbf{z}_{i-1}, \mathbf{c}) \tag{29}$$

Because $p(\mathbf{x}_{0:t}|\hat{\mathbf{c}}) = p(\mathbf{x}_{0:t}|\mathbf{c})$ and $p(\mathbf{x}_{0:t-1}|\hat{\mathbf{c}}) = p(\mathbf{x}_{0:t-1}|\mathbf{c})$, we have

$$\frac{p(\mathbf{x}_{0:t}|\hat{\mathbf{c}})}{p(\mathbf{x}_{0:t-1}|\hat{\mathbf{c}})} = \frac{p(\mathbf{x}_{0:t}|\mathbf{c})}{p(\mathbf{x}_{0:t-1}|\mathbf{c})} \tag{30}$$

which gives:

$$p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{c}})|\mathbf{J}_{\hat{\mathbf{g}}}(\hat{\mathbf{z}}_t)|^{-1} = p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{c})|\mathbf{J}_{\mathbf{g}}(\mathbf{z}_t)|^{-1} \tag{31}$$

$$p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{c}}) = p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{c})|\mathbf{J}_{\mathbf{g}}(\mathbf{z}_t)|^{-1}|\mathbf{J}_{\hat{\mathbf{g}}}(\hat{\mathbf{z}}_t)| \tag{32}$$

Now Define $\mathbf{z}_t = \left(\mathbf{g}^{-1} \circ \hat{\mathbf{g}}\right)(\hat{\mathbf{z}}_t) = \mathbf{h}(\hat{\mathbf{z}}_t)$. Since both $\mathbf{g}$ and $\hat{\mathbf{g}}$ are invertible, $\mathbf{h}$ is invertible. Let $\mathbf{H}_t = \mathbf{J}_{\mathbf{h}}(\hat{\mathbf{z}}_t)$ be the Jacobian matrix of the transformation $\mathbf{h}(\hat{\mathbf{z}}_t)$ and denote by $\mathbf{H}_{kit}$ its (k,i)th entry. Equation (32) is equivalent to:

$$p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{c}}) = p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{c})|\mathbf{H}_t| \tag{33}$$

$$\delta\left(\hat{\mathbf{z}}_t - \left(\hat{\mathbf{z}}_{t-1} + \int_{t-1}^{t} \hat{\mathbf{f}}(\hat{\mathbf{z}}_s, \hat{\mathbf{c}})ds\right)\right) = |\mathbf{H}_t|\delta\left(\mathbf{z}_t - \left(\mathbf{z}_{t-1} + \int_{t-1}^{t} \mathbf{f}(\mathbf{z}_s, \mathbf{c})ds\right)\right) \tag{34}$$

$$\delta\left(h^{-1}(\mathbf{z}_t) - \left(\hat{\mathbf{z}}_{t-1} + \int_{t-1}^{t} \hat{\mathbf{f}}(\hat{\mathbf{z}}_s, \hat{\mathbf{c}})ds\right)\right) = |\mathbf{H}_t|\delta\left(\mathbf{z}_t - \left(\mathbf{z}_{t-1} + \int_{t-1}^{t} \mathbf{f}(\mathbf{z}_s, \mathbf{c})ds\right)\right) \tag{35}$$

$$\hat{\mathbf{z}}_{t-1} + \int_{t-1}^{t} \hat{\mathbf{f}}(\hat{\mathbf{z}}_s, \hat{\mathbf{c}})ds = h^{-1}(\mathbf{z}_t) = h^{-1}(\mathbf{z}_{t-1} + \int_{t-1}^{t} \mathbf{f}(\mathbf{z}_s, \mathbf{c})ds) \tag{36}$$

Define $\eta_{kt}(\mathbf{c}) = z_{k,t-1} + \int_{t-1}^{t} f_k(\mathbf{z}_s, \mathbf{c})ds$, we have:

$$\frac{\partial\left(\hat{\mathbf{z}}_{t-1} + \int_{t-1}^{t} \hat{\mathbf{f}}(\hat{\mathbf{z}}_s, \hat{\mathbf{c}})ds\right)}{\partial \hat{z}_{it}} = \mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{z}_t) \begin{bmatrix} \frac{\partial \eta_{1t}(\mathbf{c})}{\partial z_{1t}}\mathbf{H}_{1it} \\ \vdots \\ \frac{\partial \eta_{nt}(\mathbf{c})}{\partial z_{nt}}\mathbf{H}_{nit} \end{bmatrix} \tag{37}$$

$$\frac{\partial^2\left(\hat{\mathbf{z}}_{t-1} + \int_{t-1}^{t} \hat{\mathbf{f}}(\hat{\mathbf{z}}_s, \hat{\mathbf{c}})ds\right)}{\partial \hat{z}_{it}\partial \hat{z}_{jt}} = \frac{\partial \mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{z}_t)}{\partial \hat{z}_{jt}} \begin{bmatrix} \frac{\partial \eta_{1t}(\mathbf{c})}{\partial z_{1t}}\mathbf{H}_{1it} \\ \vdots \\ \frac{\partial \eta_{nt}(\mathbf{c})}{\partial z_{nt}}\mathbf{H}_{nit} \end{bmatrix} + \mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{z}_t) \begin{bmatrix} \frac{\partial^2 \eta_{1t}(\mathbf{c})}{\partial z_{1t}^2}\mathbf{H}_{1it}\mathbf{H}_{1jt} + \frac{\partial \eta_{1t}(\mathbf{c})}{\partial z_{1t}}\frac{\partial \mathbf{H}_{1it}}{\partial \hat{z}_{jt}} \\ \vdots \\ \frac{\partial^2 \eta_{nt}(\mathbf{c})}{\partial z_{nt}^2}\mathbf{H}_{nit}\mathbf{H}_{njt} + \frac{\partial \eta_{nt}(\mathbf{c})}{\partial z_{nt}}\frac{\partial \mathbf{H}_{nit}}{\partial \hat{z}_{jt}} \end{bmatrix} \tag{38}$$

$$\frac{\partial^2\left(\hat{\mathbf{z}}_{t-1} + \int_{t-1}^{t} \hat{\mathbf{f}}(\hat{\mathbf{z}}_s, \hat{\mathbf{c}})ds\right)}{\partial \hat{z}_{it}\partial \hat{z}_{jt}\partial z_{l,t-1}} = \frac{\partial \mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{z}_t)}{\partial \hat{z}_{jt}} \begin{bmatrix} \frac{\partial^2 \eta_{1t}(\mathbf{c})}{\partial z_{1t}\partial z_{l,t-1}}\mathbf{H}_{1it} \\ \vdots \\ \frac{\partial^2 \eta_{nt}(\mathbf{c})}{\partial z_{nt}\partial z_{l,t-1}}\mathbf{H}_{nit} \end{bmatrix} + \mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{z}_t) \begin{bmatrix} \frac{\partial^3 \eta_{1t}(\mathbf{c})}{\partial z_{1t}^2\partial z_{l,t-1}}\mathbf{H}_{1it}\mathbf{H}_{1jt} + \frac{\partial^2 \eta_{1t}(\mathbf{c})}{\partial z_{1t}\partial z_{l,t-1}}\frac{\partial \mathbf{H}_{1it}}{\partial \hat{z}_{jt}} \\ \vdots \\ \frac{\partial^2 \eta_{nt}(\mathbf{c})}{\partial z_{nt}^2\partial z_{l,t-1}}\mathbf{H}_{nit}\mathbf{H}_{njt} + \frac{\partial^2 \eta_{nt}(\mathbf{c})}{\partial z_{nt}\partial z_{l,t-1}}\frac{\partial \mathbf{H}_{nit}}{\partial \hat{z}_{jt}} \end{bmatrix} \tag{39}$$

Since for each value of $\mathbf{z}_t (t > 0)$, $\mathbf{v}_{1t}, \dot{\mathbf{v}}_{1t}, \ldots, \mathbf{v}_{nt}, \dot{\mathbf{v}}_{nt}$, as $2n$ function vectors are linearly independent, we get $\mathbf{H}_{kit}\mathbf{H}_{kjt} = 0$ or $i \neq j$, That is, in each row of $\mathbf{H}_t$ there is only one non-zero entry. Since $\mathbf{h}$ is invertible, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.