Causal Hidden Markov Model for Time Series Disease Forecasting
Supplementary Material

A. Proof of Theorem 4.2

We post the Theorem 4.2 from the main text here for completeness.

Theorem A.1 (Identifiability)**.** *We assume that f_x, f_y, f_A are bijective. Denote $g_y^t(s_t, v_t) := \mathbb{E}(y_T | s_t, v_t, \boldsymbol{B}_{j\leq t})$. Under the following conditions:*

1. *$\{T^t_{\boldsymbol{o},i,j}\}$ are differentiable and non-zero almost everywhere for any $\boldsymbol{o} \in \{\boldsymbol{s}, \boldsymbol{v}, \boldsymbol{z}\}$ and $t \leq T$.*
2. *For every t, there exists at least $m := d * k + 1$ with $d := \max(d_s, d_v, d_z)$ and $k := \max(k_s, k_v, k_z)$ values of $\boldsymbol{B}_{t=0}$, i.e., $\boldsymbol{B}_{1,t=0}, ..., \boldsymbol{B}_{m,t=0}$ such that the $[\boldsymbol{\Gamma}^t_{\boldsymbol{o}}(\boldsymbol{B}_{2,t=0}) - \boldsymbol{\Gamma}^t_{\boldsymbol{o}}(\boldsymbol{B}_{1,t=0}), ..., \boldsymbol{\Gamma}^t_{\boldsymbol{o}}(\boldsymbol{B}_{m,t=0}) - \boldsymbol{\Gamma}^t_{\boldsymbol{o}}(\boldsymbol{B}_{1,t=0})]$ have full column rank and $\boldsymbol{o} \in \{\boldsymbol{s}, \boldsymbol{v}, \boldsymbol{z}\}$,*

we have that if θ and $\tilde{\theta}$ give rise to the same observational distribution, i.e., $p_\theta(\boldsymbol{x}_t, y_T, \boldsymbol{v}_t) = p_{\tilde{\theta}}(\boldsymbol{x}_t, y_T, \boldsymbol{v}_t)$ for any $\boldsymbol{x}_t, y_T, \boldsymbol{v}_t$ and $t < T$, then there exists invertible matrices $\{M^t_{\boldsymbol{o}}\}_{\boldsymbol{o}\in\{\boldsymbol{s},\boldsymbol{v},\boldsymbol{z}\}}$ and vectors $\{b^t_{\boldsymbol{o}}\}_{\boldsymbol{o}\in\{\boldsymbol{s},\boldsymbol{v},\boldsymbol{z}\}}$ such that:

Disentangle :

$$T^t_{\boldsymbol{s}}([f_x^{-1}]_{\mathcal{S}}(x_t)) = M^t_{\boldsymbol{s}}\tilde{T}^t_{\boldsymbol{s}}([\tilde{f}_x^{-1}]_{\mathcal{S}}(x_t)) + b^t_{\boldsymbol{s}}, \quad (1)$$

$$T^t_{\boldsymbol{v}}([f_x^{-1}]_{\mathcal{V}}(x_t)) = M^t_{\boldsymbol{v}}\tilde{T}^t_{\boldsymbol{v}}([\tilde{f}_x^{-1}]_{\mathcal{V}}(x_t)) + b^t_{\boldsymbol{v}}, \quad (2)$$

$$T^t_{\boldsymbol{z}}([f_x^{-1}]_{\mathcal{Z}}(x_t)) = M^t_{\boldsymbol{z}}\tilde{T}^t_{\boldsymbol{z}}([\tilde{f}_x^{-1}]_{\mathcal{Z}}(x_t)) + b^t_{\boldsymbol{z}}, \quad (3)$$

Prediction :

$$g^t_y([f_x^{-1}]_{\mathcal{S},\mathcal{V}}(x_t)) = \tilde{g}^t_y([\tilde{f}_x^{-1}]_{\mathcal{S},\mathcal{V}}(x_t)). \quad (4)$$

Proof. To avoid the ambiguity, we denote the final step as $\bar{T}$. Denote $\theta := \{f_x, f_y, f_A, f_h, \mathbf{T}^t_h, \mathbf{\Gamma}^t_h\}$, where $f_h := \{f_s, f_v, f_z\}$, $\mathbf{T}^t_h := \{\mathbf{T}^t_s, \mathbf{T}^t_v, \mathbf{T}^t_z\}$ and $\mathbf{\Gamma}^t_h := \{\mathbf{\Gamma}^t_s, \mathbf{\Gamma}^t_v, \mathbf{\Gamma}^t_z\}$. For θ and $\tilde{\theta}$ that give rise to the same observational distribution, we have:

$$p_\theta(x_t|B_{<t}) = p_{\tilde{\theta}}(x_t|B_{<t})$$

$$\Longrightarrow \int p_{f_x}(x_t|h_t)p_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h}(h_t|B_{<t})dh_t$$

$$= \int p_{\tilde{f}_x}(x_t|h_t)p_{\tilde{\mathbf{T}}^t_h,\tilde{\mathbf{\Gamma}}^t_h}(h_t|B_{<t})dh_t$$

$$\Longrightarrow p_{\varepsilon_x}(x_t - \bar{x}_t)p_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h}(f_x^{-1}(\bar{x}_t)|B_{<t})|J_{f_x^{-1}}|d\bar{x}_t$$

$$= p_{\varepsilon_x}(x_t - \bar{x}_t)p_{\tilde{\mathbf{T}}^t_h,\tilde{\mathbf{\Gamma}}^t_h}(\tilde{f}_x^{-1}(\bar{x}_t)|B_{<t})|J_{\tilde{f}_x^{-1}}|d\bar{x}_t$$

$$\Longrightarrow (\tilde{p}_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h,f_x} * p_{\varepsilon_x})(x_t|B_{<t}) = (\tilde{p}_{\tilde{\mathbf{T}}^t_h,\tilde{\mathbf{\Gamma}}^t_h,\tilde{f}_x} * p_{\varepsilon_x})(x_t|B_{<t})$$

$$\Longrightarrow F[\tilde{p}_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h,f_x}](\omega)\varphi_{\varepsilon_x}(\omega) = F[\tilde{p}_{\tilde{\mathbf{T}}^t_h,\tilde{\mathbf{\Gamma}}^t_h,\tilde{f}_x}](\omega)\varphi_{\varepsilon_x}(\omega)$$

$$\Longrightarrow F[\tilde{p}_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h,f_x}](\omega) = F[\tilde{p}_{\tilde{\mathbf{T}}^t_h,\tilde{\mathbf{\Gamma}}^t_h,\tilde{f}_x}](\omega)$$

$$\Longrightarrow \tilde{p}_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h,f_x}(x_t|B_{<t}) = \tilde{p}_{\tilde{\mathbf{T}}^t_h,\tilde{\mathbf{\Gamma}}^t_h,\tilde{f}_x}(x_t|B_{<t}), \quad (5)$$

where the "J","F" stands for Jacobian matrix and Fourier Transformation; the $\phi(\omega)$ denotes the characteristic function of ε_x; and the $\tilde{p}_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h,f_x}(x_t)$ is denoted as $p_{\mathbf{T}^t_h,\mathbf{\Gamma}^t_h}(f_x^{-1}(x_t)|B_{<t})|J_{f_x^{-1}}|$. Follow the same derivation, we can similarly obtain that

$$\tilde{p}_{\mathbf{T}^t_v,\mathbf{\Gamma}^t_v,f_A}(A_t|B_{<t}) = \tilde{p}_{\tilde{\mathbf{T}}^t_v,\tilde{\mathbf{\Gamma}}^t_v,\tilde{f}_A}(A_t|B_{<t}). \quad (6)$$

For the label Y_T, we have

$$\tilde{p}_{\mathbf{T}^{\bar{T}}_{v,s},\mathbf{\Gamma}^{\bar{T}}_{v,s},f_y}(y_{\bar{T}}|B_{<\bar{T}}) = \tilde{p}_{\tilde{\mathbf{T}}^{\bar{T}}_{v,s},\tilde{\mathbf{\Gamma}}^{\bar{T}}_{v,s},f_y}(y_{\bar{T}}|B_{<\bar{T}})$$

$$\Longrightarrow p_{f_y}(f_y^{-1}(y_{\bar{T}})|B_{<\bar{T}})|J_{f_y^{-1}}| = p_{\tilde{f}_y}(\tilde{f}_y^{-1}(y_{\bar{T}})|B_{<\bar{T}})|J_{\tilde{f}_y^{-1}}| \quad (7)$$

where $\mathbf{T}^{\bar{T}}_{v,s} := \{\mathbf{T}^{\bar{T}}_s, \mathbf{T}^{\bar{T}}_v\}$ and $\mathbf{\Gamma}^{\bar{T}}_{v,s} := \{\mathbf{\Gamma}^{\bar{T}}_s, \mathbf{\Gamma}^{\bar{T}}_v\}$. The Eq. (7) transforms the equivalence of the observational space to the equivalence of the $\mathcal{S} \times \mathcal{V}$ at time $\bar{T} - 1$. Further, since

$$\int p_{f_y}(f_y^{-1}(y_{\bar{T}})|h^{-z}_{\bar{T}-1}, B_{\bar{T}-1})p_{f_{s,v}}(h^{-z}_{\bar{T}-1}|B_{\leq\bar{T}-1})dh^{-z}_{\bar{T}-1}|J_{f_y^{-1}}|$$

$$= \int p_{\tilde{f}_y}(\tilde{f}_y^{-1}(y_{\bar{T}})|h^{-z}_{\bar{T}-1}, B_{\bar{T}-1})p_{\tilde{f}_{s,v}}(h^{-z}_{\bar{T}-1}|B_{\leq\bar{T}-1})dh^{-z}_{\bar{T}-1}|J_{\tilde{f}_y^{-1}}|,$$

where $h^{-z} := \{s, v\}$ and $f_{s,v}$ is such that $f_{s,v}(s_t, v_t, B_{t-1}, \varepsilon_s, \varepsilon_v) := [[f_s(s_t, B_{t-1}, \varepsilon_s)]^\top, [f_v(v_t, B_{t-1}, \varepsilon_v)]^\top]^\top$, it can then be similarly derived the transformation from time $\bar{T} - 1$ to time $\bar{T} - 2$:

$$p_{f_y}((f_{s,v}^{-1} \circ f_y^{-1})(y_{\bar{T}})_{\mathcal{S}\times\mathcal{V}}|B_{<\bar{T}-1})|J_{f_y^{-1}}||J_{f_{s,v}^{-1}}(\hat{h}^{-z,i}_{\bar{T}})|$$

$$=p_{\tilde{f}_y}((\tilde{f}_{s,v}^{-1} \circ \tilde{f}_y^{-1})(y_{\bar{T}})_{\mathcal{S}\times\mathcal{V}}|B_{<\bar{T}-1})|J_{\tilde{f}_y^{-1}}||J_{\tilde{f}_{s,v}^{-1}}(\hat{\tilde{h}}^{-z,i}_{\bar{T}})|, \quad (8)$$

where $\hat{h}^{-z,i}_t := ([f_{s,v}^{-1}]^{T-t+1} \circ f_y^{-1})(y_{\bar{T}})_{\mathcal{S}\times\mathcal{V}}$ with $[f]^k := \underbrace{f \circ f \circ ...f}_{k}$. Iteratively applying such a transformation, we would have that:

$$p_{f_y}([([f_{s,v}^{-1}]^{\bar{T}-t+1} \circ f_y^{-1})(y_{\bar{T}})]_{\mathcal{S}\times\mathcal{V}}|B_{<\bar{T}-1})|J_{f_y^{-1}}|=$$
$$p_{\tilde{f}_y}([([\tilde{f}_{s,v}^{-1}]^{\bar{T}-t+1} \circ \tilde{f}_y^{-1})(y_{\bar{T}})]_{\mathcal{S}\times\mathcal{V}}|B_{<\bar{T}-1})$$
$$* \, |J_{\tilde{f}_y^{-1}}| \frac{\Pi_{j=\bar{T}}^{t+1} |J_{\tilde{f}_{s,v}^{-1}}(\hat{\tilde{h}}_j^{-z,i})|}{\Pi_{j=\bar{T}}^{t+1} |J_{f_{s,v}^{-1}}(\hat{h}_j^{-z,i})|}, \tag{9}$$

which transforms to the latent space at time t at which the x, A are observed. Note that the v_t that generates the x_t, A_t is the same value, we have that

$$\tilde{p}_{\mathbf{T}_h^t, \mathbf{\Gamma}_h^t, l_{x,A}}(x_t, A_t|B_{<t}) = \tilde{p}_{\tilde{\mathbf{T}}_h^t, \tilde{\mathbf{\Gamma}}_h^t, \tilde{l}_{x,A}}(x_t, A_t|B_{<t}), \tag{10}$$

where $l_{x,A}([x_t, A_t]) := [[f_x^{-1}(x_t)]_{\mathcal{S}\times\mathcal{Z}}]^\top, [f_A^{-1}(A_t)]^\top]^\top$. Taking logarithmic on both sides of Eq. (6), we have:

$$\log|J_{f_A}(A_t)| + \sum_{i=1}^{d_v}\left(\log C_i(f_{A,i}^{-1}(A_t)) - \log Q_i(B_{<t})\right)$$
$$+ \sum_{i=1}^{d_v}\left(\sum_{j=1}^{k_v} T_{i,j}^v(f_{A,i}^{-1}(A_t))\Gamma_{i,j}^A(B_{<t})\right)$$
$$= \log|J_{\tilde{f}_A}(A_t)| + \sum_{i=1}^{d_v}\left(\log \tilde{C}_i(\tilde{f}_{A,i}^{-1}(A_t)) - \log \tilde{Q}_i(B_{<t})\right)$$
$$+ \sum_{i=1}^{d_v}\left(\sum_{j=1}^{k_v} \tilde{T}_{i,j}^v(\tilde{f}_{A,i}^{-1}(A_t))\tilde{\Gamma}_{i,j}^v(B_{<t})\right). \tag{11}$$

According to the assumption (2), we have

$$\langle \mathbf{T}^v(f_A^{-1}(A_t)), \overline{\mathbf{\Gamma}}^v(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^v(\tilde{f}_A^{-1}(A_t)), \overline{\tilde{\mathbf{\Gamma}}}^v(B_{k,<t})\rangle + \tilde{b}_v, \tag{12}$$

for all $k \in [m]$ for some $\tilde{b}_v$, where $\bar{\mathbf{\Gamma}}(B) = \mathbf{\Gamma}(B) - \mathbf{\Gamma}(B_{0,<t})$. Similarly, from Eq. (10), we have that

$$\langle \mathbf{T}^v(f_A^{-1}(A_t)), \overline{\mathbf{\Gamma}}^v(B_{k,<t})\rangle +$$
$$\langle \mathbf{T}^{s,z}([f_x^{-1}(x_t)]_{\mathcal{S}\times\mathcal{Z}}), \overline{\mathbf{\Gamma}}^{s,z}(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^v(\tilde{f}_A^{-1}(A_t)), \overline{\tilde{\mathbf{\Gamma}}}^v(B_{k,<t})\rangle +$$
$$\tilde{\mathbf{T}}^{s,z}([\tilde{f}_x^{-1}(x_t)]_{\mathcal{S}\times\mathcal{Z}}), \overline{\tilde{\mathbf{\Gamma}}}^{s,z}(B_{k,<t})\rangle + \tilde{b}_v + \tilde{b}_{s,z}, \tag{13}$$

for some $\tilde{b}_{s,z}$. Then we have:

$$\langle \mathbf{T}^{s,z}([f_x^{-1}](x_t)]_{\mathcal{S}\times\mathcal{Z}}), \overline{\mathbf{\Gamma}}^{s,z}(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^{s,z}([\tilde{f}_x^{-1}(x_t)]_{\mathcal{S}\times\mathcal{Z}}), \overline{\tilde{\mathbf{\Gamma}}}^{s,z}(B_{k,<t})\rangle + \tilde{b}_{s,z}, \tag{14}$$

Similarly, the Eq. (5) can imply:

$$\langle \mathbf{T}^{s,z,v}(f_x^{-1}(x_t)), \overline{\mathbf{\Gamma}}^{s,z,v}(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^{s,z,v}(\tilde{f}_x^{-1}(x_t)), \overline{\tilde{\mathbf{\Gamma}}}^{s,z,v}(B_{k,<t})\rangle + \tilde{b}_{s,z} + \tilde{b}_v, \tag{15}$$

Subtracting Eq. (14) from Eq. (15), we have

$$\langle \mathbf{T}^v([f_x^{-1}(x_t)]_{\mathcal{V}}), \overline{\mathbf{\Gamma}}^v(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^v([\tilde{f}_x^{-1}(x_t)]_{\mathcal{V}}), \overline{\tilde{\mathbf{\Gamma}}}^v(B_{k,<t})\rangle + \tilde{b}_v. \tag{16}$$

Denote $f_y^t := [f_{s,v}^{-1}]^{\bar{T}-t+1} \circ f_y^{-1}$. Similarly, according to Eq. (9), we have:

$$\langle \mathbf{T}^{s,v}([f_y^t(y_{\bar{T}})_{\mathcal{S}\times\mathcal{V}}), \overline{\mathbf{\Gamma}}^{s,v}(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^{s,v}(\tilde{f}_y^t(y_{\bar{T}})_{\mathcal{S}\times\mathcal{V}}), \overline{\tilde{\mathbf{\Gamma}}}^{s,v}(B_{k,<t})\rangle + \tilde{b}_{s,v}. \tag{17}$$

Besides, we have

$$\tilde{p}_{\mathbf{T}_h^t, \mathbf{\Gamma}_h^t, l_{x,y}}(x_t, y_T|B_{<t}) = \tilde{p}_{\tilde{\mathbf{T}}_h^t, \tilde{\mathbf{\Gamma}}_h^t, \tilde{l}_{x,y}}(x_t, y_T|B_{<t}), \tag{18}$$

where $l_{x,y}([x_t, y_T]) := [[f_x^{-1}(x_t)]_{\mathcal{Z}}^\top, [[f_y^t]^{-1}(y_T)]^\top]^\top$. Then we will have:

$$\langle \mathbf{T}^z([f_x^{-1}(x_t)]_{\mathcal{Z}}), \overline{\mathbf{\Gamma}}^z(B_{k,<t})\rangle$$
$$+ \langle \mathbf{T}^{s,v}([[f_y^t]^{-1}(y_T)]_{\mathcal{S}\times\mathcal{V}}), \overline{\mathbf{\Gamma}}^{s,v}(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^z([\tilde{f}_x^{-1}(x_t)]_{\mathcal{Z}}), \overline{\mathbf{\Gamma}}^z(B_{k,<t})\rangle$$
$$+ \langle \tilde{\mathbf{T}}^{s,v}([[\tilde{f}_y^t]^{-1}(y_T)]_{\mathcal{S}\times\mathcal{V}}), \overline{\tilde{\mathbf{\Gamma}}}^{s,v}(B_{k,<t})\rangle + \tilde{b}_z + \tilde{b}_{s,v}. \tag{19}$$

Subtracting Eq. (17) from Eq. (19), we have:

$$\langle \mathbf{T}^z([f_x^{-1}(x_t)]_{\mathcal{Z}}), \overline{\mathbf{\Gamma}}^z(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^z([\tilde{f}_x^{-1}(x_t)]_{\mathcal{Z}}), \overline{\tilde{\mathbf{\Gamma}}}^z(B_{k,<t})\rangle + \tilde{b}_z. \tag{20}$$

Subtracting Eq. (21), (16) from Eq. (15), we have:

$$\langle \mathbf{T}^s([f_x^{-1}(x_t)]_{\mathcal{S}}), \overline{\mathbf{\Gamma}}^s(B_{k,<t})\rangle$$
$$= \langle \tilde{\mathbf{T}}^s([\tilde{f}_x^{-1}(x_t)]_{\mathcal{S}}), \overline{\tilde{\mathbf{\Gamma}}}^s(B_{k,<t})\rangle + \tilde{b}_s. \tag{21}$$

Denote $M_o := \left(\overline{\mathbf{\Gamma}}^o \overline{\mathbf{\Gamma}}^{o,\top}\right)^{-1} \overline{\tilde{\mathbf{\Gamma}}}^{o,\top}$ in which $\overline{\mathbf{\Gamma}}^o := [\overline{\mathbf{\Gamma}}^o(B_{2,<t}), ..., \overline{\mathbf{\Gamma}}^o(B_{m,<t})]$, then applying the assumption (1) and the result from [1], we have that the M_o for $o \in \{s, z, v\}$ is invertible. Finally, since we have

$$\mathbb{E}_{f_x, f_y}(y_T|x_t, B_{\le t}) = \mathbb{E}_{\tilde{f}_x, \tilde{f}_y}(y_T|x_t, B_{\le t}), \tag{22}$$

then we have

$$\int g_y^t([f_x^{-1}(\bar{x}_t)]_{\mathcal{S}\times\mathcal{V}}) p_{\varepsilon_x}(x - \bar{x}) d\bar{x}_t$$
$$= \int \tilde{g}_y^t([\tilde{f}_x^{-1}(\bar{x}_t)]_{\mathcal{S}\times\mathcal{V}}) p_{\varepsilon_x}(x - \bar{x}) d\bar{x}_t \tag{23}$$

Applying the Fourier Transformation on both sides, we have

$$g_y^t([f_x^{-1}(\bar{x}_t)]_{\mathcal{S}\times\mathcal{V}}) = \tilde{g}_y^t([\tilde{f}_x^{-1}(\bar{x}_t)]_{\mathcal{S}\times\mathcal{V}}).$$

The proof is completed. □

B. Posterior Reparameterization

The reparameterization of p_ψ is given by:

$$p_\psi(\boldsymbol{h}_{<T}|\boldsymbol{u}_{<T}, y_T) = \frac{p_\psi(\boldsymbol{h}_{<T}, y_T|\boldsymbol{u}_{<T})}{p_\psi(y_T|\boldsymbol{u}_{<T})}$$
$$= \frac{p_\psi(\boldsymbol{h}_{<T}, \boldsymbol{u}_{<T}) p_\psi(y_T|\boldsymbol{s}_{T-1}, \boldsymbol{v}_{T-1})}{p_\psi(y_T|\boldsymbol{u}_{<T}) p_\psi(\boldsymbol{u}_{<T})}$$
$$= \frac{p_\psi(\boldsymbol{h}_{<T}|\boldsymbol{u}_{<T}) p_\psi(y_T|\boldsymbol{s}_{T-1}, \boldsymbol{v}_{T-1})}{p_\psi(y_T|\boldsymbol{u}_{<T})}, \tag{24}$$

where the $p_\psi(\boldsymbol{h}_{<T}|\boldsymbol{u}_{<T})$ can be further factorized using mean-field approach due to our Markov assumption, *i.e.*,

$$p_\psi(\boldsymbol{h}_{<T}|\boldsymbol{u}_{<T}) = \Pi_{t<T} p_\psi(\boldsymbol{h}_t|\boldsymbol{u}_t, \boldsymbol{h}_{t-1}). \tag{25}$$

Since the q_ϕ is expected to mimic the behavior of p_ψ (also p), it shares the same way of reparameterization with p_ψ.

Then the reformulation in Eq. (24) and mean-field factorization together imply that

$$q_\phi(\boldsymbol{h}_{<T}|\boldsymbol{u}_{<T}, y_T) = \frac{q_\phi(y_T|\boldsymbol{s}_{T-1}, \boldsymbol{v}_{T-1})}{q_\phi(y_T|\boldsymbol{u}_{<T})} * \Pi_{t<T} q_\phi(\boldsymbol{h}_t|\boldsymbol{u}_t, \boldsymbol{h}_{t-1}), \tag{26}$$

where $q_\phi(\boldsymbol{h}_t|\boldsymbol{u}_t, \boldsymbol{h}_{t-1}) \sim \mathcal{N}(\mu(\boldsymbol{h}_{t-1}, \boldsymbol{u}_t), \Sigma(\boldsymbol{h}_{t-1}, \boldsymbol{u}_t))$.

C. Personal Attributes and Clinical Measurements

We also collect corresponding personal attributes and clinical measurements for the 507 students in primary school. The personal attributes contains 16 attributes of each person, containing the Age, Gender, Father's height, Father's weight, Mother's height, Birth length, Height, Head circumference, Weight, Waist circumference, Pulse, Diastolic blood pressure, Systolic blood pressure, Number of parents with eyes, Close working hours, Outdoor time. The clinical measurements contains 15 clinical measurement indexes of visual acuity examination, containing Diopter, Long-distance accommodation response, Short-distance accommodation response, Oct parameter1, Oct parameter2, Oct parameter, Right eye naked vision, Peripheral refractive power, Right eye IOP, Axial length, Anterior chamber depth, Corneal thickness, Corneal curvature1, Corneal curvature2 and Corneal diameter of the corresponding person.

D. Implementation Details

We crop and resize the raw retinal images into 128 × 128 images. The clinical measurements and personal attributes are normalized by mean-variance normalization. The weights for classification loss, reconstruction loss, and the KL loss are 1, 1, 0.1. All results have been rerun five times for robustness. We use Adam as the optimizer. The learning rate is set to 0.000005 with 0.9 decay rate after 10,30,50,80 epochs. The batch size is set to 16. We use Xavier initialization for the model and train it in 300 epochs. For the classifier network at second stage, we use two layers fully connected layer with the dimension set to 512. When training the new classifier for $\boldsymbol{s} + \boldsymbol{v}$ and $\boldsymbol{z}$, we load our trained Causal-HMM model and fix its parameters. The learning rate is set to 0.001 with 0.9 decay rate after 10,30,50,80 epochs. The batch size is set to 16. We train the classifier model in 100 epochs.

E. Robustness to the Dimension of Hidden Variables

For robustness to the dimension of hidden variables, we repeat for 5 times for different dimensions of $\boldsymbol{s}, \boldsymbol{v}, \boldsymbol{z}$. Results are shown in Tab. 1. We can see that the different dimensions did not affect the performance too much, with mean ACC fluctuated within 1.61 and mean AUC fluctuated within 4.48. This results show that our method achieves robust generalization to the dimension of hidden variables.

F. More Visualization Results

Due to the space limitation of the main text, we put more visualization results here. The results are shown in Fig. 1.

References

[1] Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23th International Conference on Artificial Intelligence and Statistics (AISTATS-23)*, volume 108, Palermo, Italy, 2020. AISTATS Committee, PMLR. 2

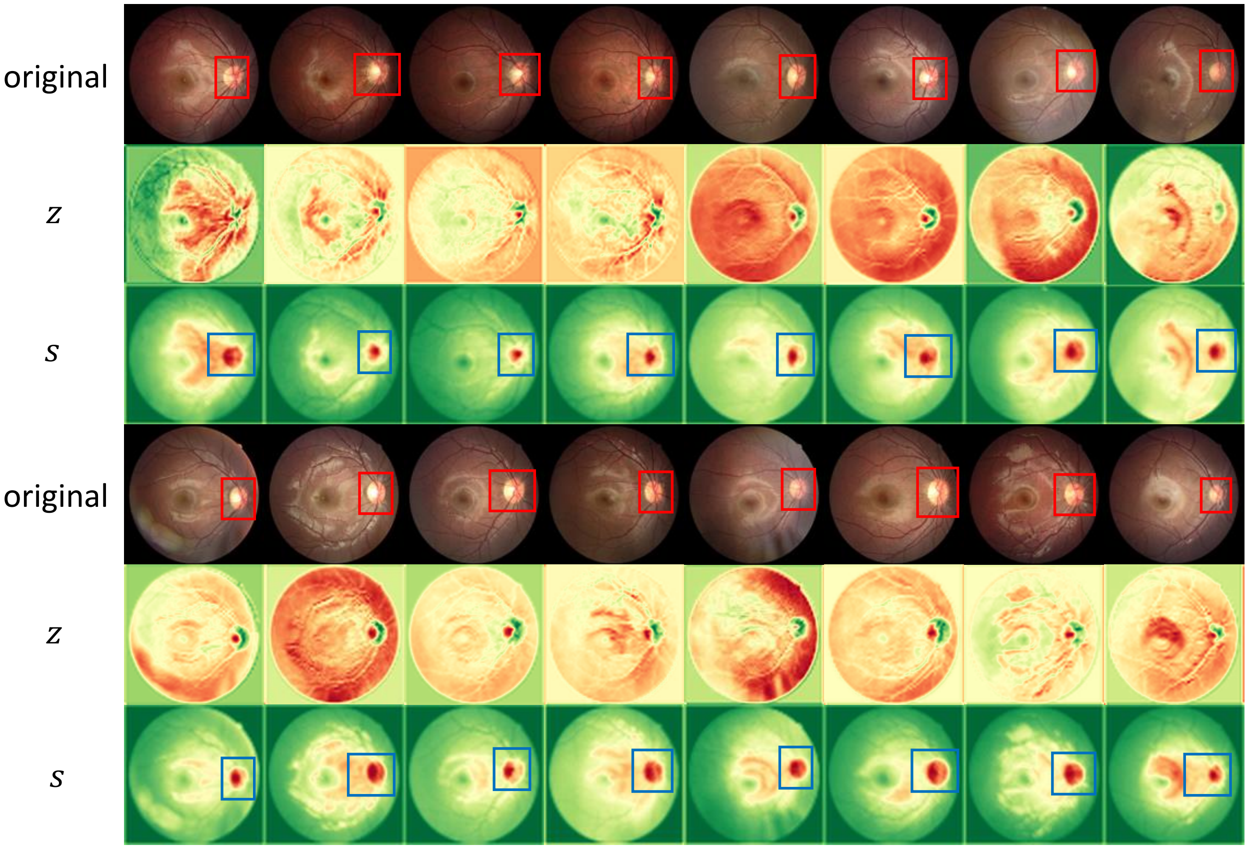

Figure 1. More visualization of learned feature maps by s and z on the test dataset. For each three line picture, the top row is the original images, with the disease areas marked by red rectangles; the middle row is the feature maps of z by Grad-CAM; and the bottom row is feature maps of s by Grad-CAM, with the found areas of disease are marked by blue rectangles. The red to green corresponds to high to low response of corresponding hidden variable. As shown, the high response areas of s are concentrated on the optic disc, while the response areas of z are distributed in other regions.

Methods	**z48-s64-v16**		**z192-s256-v64**		**z128-s96-v32**		**z96-s128-v32**	
Grades \ Metric	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
G1 to G5	74.39 ± 1.42	84.15 ± 0.95	76.63 ± 2.75	84.78 ± 1.11	72.34 ± 5.86	84.02 ± 1.16	77.19 ± 1.69	85.43 ± 1.76
G1 to G4	69.16 ± 2.38	79.33 ± 1.57	71.40 ± 4.10	76.13 ± 6.17	68.78 ± 1.69	81.18 ± 1.02	72.89 ± 2.64	78.99 ± 1.53
G1 to G3	64.11 ± 2.52	72.54 ± 3.21	62.43 ± 1.38	65.75 ± 4.98	65.04 ± 3.99	73.07 ±5.11	62.43 ± 2.03	68.24 ± 2.93
G1 to G2	62.80 ± 7.81	63.98 ± 4.47	56.45 ± 3.41	57.46 ± 2.59	62.24 ± 0.84	69.78 ± 4.28	65.42 ± 1.47	65.09 ± 2.29
G2 to G5	77.38 ± 4.20	85.98 ± 3.52	75.51 ± 5.85	84.36 ± 1.63	78.13 ± 2.52	86.59 ± 1.04	76.26 ± 2.44	86.71 ± 0.89
G2 to G4	73.83 ± 4.90	80.96 ± 1.09	73.64 ± 4.92	77.64 ± 1.53	71.59 ± 2.99	81.15 ± 1.34	71.22 ± 5.17	80.62 ± 1.36
G2 to G3	67.85 ± 1.42	76.75 ± 1.12	62.99 ± 0.84	67.45 ± 3.26	68.04 ± 1.67	76.50 ± 1.31	66.91 ± 2.69	75.07 ± 1.31
G3 to G5	79.07 ± 4.05	86.53 ± 5.65	78.69 ± 3.46	85.23 ± 1.43	77.19 ± 2.34	86.79 ± 1.39	77.01 ± 3.41	86.22 ± 1.34
G3 to G4	71.77 ± 5.50	81.16 ± 1.51	71.40 ± 2.99	79.65 ± 0.53	74.02 ± 2.59	83.74 ± 1.15	71.77 ± 2.59	82.22 ± 1.29
G4 to G5	78.13 ± 2.99	86.63 ± 6.38	73.96 ± 5.34	86.07 ± 0.85	78.50 ±2.38	86.51 ± 0.54	78.13 ± 3.21	86.92 ± 1.53
Mean	71.85 ± 3.72	79.80 ± 2.95	70.31 ± 3.50	76.45 ± 2.41	71.59 ± 2.69	80.93 ± 1.83	71.92 ± 2.73	79.55 ± 1.62

Table 1. Changing dimensions of the hidden variables. Results of ACC (accuracy) and AUC (Area Under the Curve) on the test dataset on 10 time series settings.