Supplementary Material for "Model Specification Relaxation for Probabilistic Latent Variable Models: An Infinite-Horizon Optimal Control Approach"

NOMENCLATURE

$\boldsymbol{H}$	Hamiltonian function
Δ	first variation
δ	Dirac delta function
γ	a positive number
Λ	eigenvalue of kernel function
λ	adjoint state
$\langle\cdot,\cdot\rangle$	inner product
$\mathbb{D}_{\mathrm{KL}}$	Kullback-Leiber divergence
$\mathbb{E}$	expectation operator
$\mathbb{F}$	normalized density function family
$\mathbb{R}$	real number domain
$\mathbb{S}$	kernelized Stein discrepancy
$\mathcal{C}$	positive constant
$\mathcal{F}(z)$	flux function
$\mathcal{H}$	reproducing kernel Hilbert space
$\mathcal{K}(z)$	control gain
$\mathcal{N}$	normal distribution
$\mathcal{P}(z)$	prior distribution of latent variable z
$\mathcal{Q}(z)$	PDF of approximation distribution
$\mathcal{Q}_t(z)$	PDF of approximation distribution along time t
$\mathcal{Q}_\infty(z)$	approximation distribution at infinite time
$\mathcal{R}$	a real number
D	dimension
d	full derivative
$\mathrm{d}S(z)$	surface element
i.i.d.	independent and identically distributed
L	bootstrap sample size
M	instance number
s.t.	subjected to
Trace	matrix trace
T	end time
$\mathscr{A}$	matrix norm of the perturbation direction
$\mathscr{B}$	the matrix norm of the gradient of the perturbation direction
$\mathscr{C}_c^\infty(\mathbb{R}^{\mathrm{D}})$	the set of infinitely differentiable functions with compact support on $\mathbb{R}^{\mathrm{D}}$
$\mathscr{H}_0$	the light-tailedness constant
$\mathscr{Q}(z)$	empirical measure
$\mathscr{Q}_t(z)$	weak solution to the PDE
$\mathscr{R}$	the radius of the ball/neighborhood
$\mathscr{S}$	the functional measuring the discrepancy between the weak and strong solutions
μ	mean value
∇	gradient operator
ν	degree of freedom for Student's-t distribution
∂	partial derivative
$\phi(z)$	perturbation direction
$\psi(z)$	ansatz of perturbation direction
LV	latent variable
obs	observation
Σ	covariance
σ^2	variance
R^2	goodness-of-fit
θ	model parameter set
$\top$	matrix inversion
ε	discretization step
$\vec{n}(z)$	normal vector
$\widehat{\mathcal{Q}}_t(z)$	state estimation of $\mathcal{Q}_t(z)$
$\widehat{\psi}_i$	feature importance weight
$\widehat{x}$	predicted observational data
Ξ	orthonormal basis
ξ	learning rate
$C\left([0,\infty],\mathbb{R}^{\mathrm{D_{LV}}}\right)$	infinite-horizon path space
$C\left([0,\mathrm{T}],\mathbb{R}^{\mathrm{D_{LV}}}\right)$	path space
f	smooth test function
h	bandwidth
$K(z,z')$	kernel function
L	the smoothness constant of the negative log-likelihood function
St	Student's-t distribution
t	time index
$u(z)$	control policy
x	observational data
z	latent variable
CA	carbon-dioxide absorber
DC	debutanizer column
ELBO	evidence lower bound
EM	expectation maximization
InfO	Infinite-horizon Optimal control
MAE	mean-absolute error
MAPE	mean-absolute-percentage error
MoG	mixture of Gaussian
MoR	mixture of ring
ODE	ordinary differential equations
PDE	partial differential equation
PDF	probability density function
RBF	radius basis function
RKHS	reproducing kernel Hilbert space
RMSE	root-mean-square error
TM	two moon
WGS	water-gas-shift unit

S.I. Continuity Equation and Its Weak Solution

The notion of *weak solution* is central in connecting particle-based methods with PDEs represented by continuity equation:

$$\begin{cases} \dfrac{\partial \mathcal{Q}_t(z)}{\partial t} = -\nabla \cdot [\phi(z)\mathcal{Q}_t(z)] \\ \mathcal{Q}_t(z)|_{t=0} = \mathcal{Q}_0(z) \end{cases}. \tag{S1}$$

A classical solution to Equation (S1) requires $\mathcal{Q}_t(z)$ to be differentiable in both t and z, which may not hold when $\mathcal{Q}_t(z)$ is merely an empirical measure constructed from a finite set of particles $\{z_{i,t}\}_{i=1}^{\mathrm{M}}$. In contrast, a weak solution only requires that, for any test function within the the set $\mathscr{C}_c^\infty(\mathbb{R}^{\mathrm{D}})$ of all smooth functions with compact support on $\mathbb{R}^{\mathrm{D}}$:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\int f(z)\,\mathcal{Q}_t(z)\,\mathrm{d}z = \int \mathcal{Q}_t(z)\,\phi^\top(z)\cdot\nabla_z f(z)\,\mathrm{d}z, \\ &\forall f \in \mathscr{C}_c^\infty(\mathbb{R}^{\mathrm{D}}) \end{aligned} \tag{S2}$$

which is obtained by the following equation:

$$\begin{aligned} &\frac{\partial \mathcal{Q}_t(z)}{\partial t} = -\nabla\cdot[\phi(z)\mathcal{Q}_t(z)], \\ \overset{\text{(i)}}{\Rightarrow}&\frac{\mathrm{d}\mathcal{Q}_t(z)}{\mathrm{d}t} = -\mathcal{Q}_t(z)\nabla_z\cdot\phi(z), \\ \Rightarrow&\int f(z)\frac{\mathrm{d}\mathcal{Q}_t(z)}{\mathrm{d}t}\mathrm{d}z = -\int f(z)\mathcal{Q}_t(z)\nabla_z\cdot\phi(z)\mathrm{d}z \\ \overset{\text{(ii)}}{\Rightarrow}&\int f(z)\frac{\mathrm{d}\mathcal{Q}_t(z)}{\mathrm{d}t}\mathrm{d}z = \int \mathcal{Q}_t(z)\phi^\top(z)\nabla_z f(z)\mathrm{d}z \\ \Rightarrow&\frac{\mathrm{d}}{\mathrm{d}t}\int f(z)\,\mathcal{Q}_t(z)\,\mathrm{d}z = \int \mathcal{Q}_t(z)\,\phi^\top(z)\cdot\nabla_z f(z)\,\mathrm{d}z, \end{aligned} \tag{S3}$$

where '(i)' is based on the relationship of the Lagrangian derivative, and '(ii)' is based on the integration by parts:

$$\begin{aligned} &\int \mathcal{Q}_t(z)\phi^\top(z)\nabla_z f(z)\mathrm{d}z + \int \mathcal{Q}_t(z)f(z)\nabla_z\cdot\phi(z)\mathrm{d}z = 0, \\ &\forall f \in \mathscr{C}_c^\infty(\mathbb{R}^{\mathrm{D}}). \end{aligned} \tag{S4}$$

This definition holds even if $\mathcal{Q}_t(z)$ is a measure, such as $\mathscr{Q}_t(z) = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\delta(z - z_{i,t})$, created from the following differential equations:

$$\begin{cases} \dfrac{\mathrm{d}z_{i,t}}{\mathrm{d}t} = \phi(z_{i,t}) \\ z_{i,0} \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_0(z) \end{cases}. \tag{S5}$$

As the number of particles $\mathrm{M}\to\infty$, the empirical measure $\mathscr{Q}_t(z)$ converges weakly to the solution $\mathcal{Q}_t(z)$ of the corresponding PDE. To demonstrate this property, we can define the functional $\mathscr{S}$ for $\mathscr{Q}_t(z)$ and $\mathcal{Q}_t(z)$ as follows:

$$\mathscr{S}_{\mathscr{Q}_t(z)} \coloneqq \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} f(z_{i,t}) = \int f(z)\mathscr{Q}_t(z)\mathrm{d}z, \tag{S6a}$$

$$\mathscr{S}_{\mathcal{Q}_t(z)} \coloneqq \int f(z)\mathcal{Q}_t(z)\mathrm{d}z. \tag{S6b}$$

Demonstrating the empirical measure $\mathscr{Q}_t(z)$ converges weakly to the solution $\mathcal{Q}_t(z)$ of the corresponding PDE is to prove that as the $\mathrm{M}\to\infty$, the following relationship holds:

$$\mathscr{S}_{\mathscr{Q}_t(z)} \xrightarrow[\mathrm{M}\to\infty]{} \mathscr{S}_{\mathcal{Q}_t(z)}, \quad \forall f(z) \in \mathscr{C}_c^\infty(\mathbb{R}^{\mathrm{D}}). \tag{S7}$$

To this end, we can take the derivative of $\mathscr{S}_{\mathscr{Q}_t(z)}$ with-respect-to time t as follows:

$$\begin{aligned} &\frac{\mathrm{d}\mathscr{S}_{\mathscr{Q}_t(z)}}{\mathrm{d}t} \\ =&\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}[\nabla f(z_{i,t})]^\top\frac{\mathrm{d}z_{i,t}}{\mathrm{d}t} \\ =&\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}[\nabla f(z_{i,t})]^\top\phi(z_{i,t}) \\ =&\int[\nabla_z f(z)]^\top\phi(z)\mathscr{Q}_t(z)\mathrm{d}z. \end{aligned} \tag{S8}$$

Similarly, the derivative of $\mathscr{S}_{\mathcal{Q}_t(z)}$ with-respect-to time t can be given as follows:

$$\begin{aligned} &\frac{\mathrm{d}\mathscr{S}_{\mathcal{Q}_t(z)}}{\mathrm{d}t} \\ =&\frac{\mathrm{d}}{\mathrm{d}t}\int f(z)\mathcal{Q}_t(z)\mathrm{d}z \\ =&\underbrace{\int\frac{\partial f(z)}{\partial t}\mathcal{Q}_t(z)\mathrm{d}z}_{=0} + \int f(z)\frac{\partial\mathcal{Q}_t(z)}{\partial t}\mathrm{d}z \\ =&\int f(z)\frac{\partial\mathcal{Q}_t(z)}{\partial t}\mathrm{d}z. \end{aligned} \tag{S9}$$

According to Equation (S1), we have:

$$\begin{aligned} &\int f(z)\frac{\partial\mathcal{Q}_t(z)}{\partial t}\mathrm{d}z \\ =&\int f(z)\{-\nabla_z\cdot[\phi(z)\mathcal{Q}_t(z)]\}\mathrm{d}z \\ =&\int\mathcal{Q}_t(z)\phi^\top(z)\cdot\nabla_z f(z)\mathrm{d}z \\ =&\int[\nabla_z f(z)]^\top\phi(z)\mathcal{Q}_t(z)\mathrm{d}z. \end{aligned} \tag{S10}$$

Under mild assumption, the difference between $\frac{\mathrm{d}\mathscr{S}_{\mathcal{Q}_t(z)}}{\mathrm{d}t}$ and $\frac{\mathrm{d}\mathscr{S}_{\mathscr{Q}_t(z)}}{\mathrm{d}t}$ satisfies the following inequality:

$$\left|\frac{\mathrm{d}\mathscr{S}_{\mathcal{Q}_t(z)}}{\mathrm{d}t} - \frac{\mathrm{d}\mathscr{S}_{\mathscr{Q}_t(z)}}{\mathrm{d}t}\right| \le \mathscr{L}|\mathscr{S}_{\mathcal{Q}_t(z)} - \mathscr{S}_{\mathscr{Q}_t(z)}|, \tag{S11}$$

which can be further reformulated as follows according to Gronwall's lemma [SR1]:

$$|\mathscr{S}_{\mathcal{Q}_t(z)} - \mathscr{S}_{\mathscr{Q}_t(z)}| \le |\mathscr{S}_{\mathcal{Q}_0(z)} - \mathscr{S}_{\mathscr{Q}_0(z)}|e^{\mathscr{L}t}. \tag{S12}$$

Meanwhile, at the initial time $t = 0$, according to the celebrated "Weak Law of Large Number" [SR2], $\mathscr{Q}_0(z)$ and $\mathcal{Q}_0(z)$ satisfies the following relationship:

$$\lim_{\mathrm{M}\to\infty}|\mathscr{S}_{\mathcal{Q}_0(z)} - \mathscr{S}_{\mathscr{Q}_0(z)}| = 0. \tag{S13}$$

Thus, we can get the following relationship, which proves the weak convergence of the particle solution:

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathscr{S}_{\mathcal{Q}_t(z)} - \mathscr{S}_{\mathscr{Q}_t(z)}] \le \mathscr{L}|\mathscr{S}_{\mathcal{Q}_t(z)} - \mathscr{S}_{\mathscr{Q}_t(z)}|, \tag{S14}$$

with $|\mathscr{S}_{\mathcal{Q}_t(z)} - \mathscr{S}_{\mathscr{Q}_t(z)}| \to 0$ as $\mathrm{M}\to\infty$.

S.II. Detailed Proof of the Convergence

In this part, we proposed a detailed proof of the convergence theorem in our main content. At first, let us review the discretization approach we use in our main content. Specifically, in our main content, we discretize the weak solution given by Equation (S5) by the forward Euler's method [SR3] as follows:

$$z_{i,t+\varepsilon} = z_{i,t} + \varepsilon\phi(z_{i,t}), \text{ for } i = 1, \ldots, \mathrm{M}. \tag{S15}$$

On this basis, we propose the following theorem in the main content to demonstrate the convergence property of the E-step for InfO-EM algorithm as follows:

Theorem S1. *Suppose that $\|\nabla_z\phi(z)\| \leq \mathscr{B}$, $\|\phi(z)\| \leq \mathscr{A}$, and $\mathscr{H}_0$ is the constant that control the light-tailness of $\mathcal{P}(z|x)$. Let $\{\mathcal{Q}_t(z) \mid t = 1, 2, \ldots, \mathrm{T}\}$ denote the sequence of variational distributions generated by the E-step of the InfO-EM algorithm. Then, the sequence of KL divergences $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_t(z)\|\mathcal{P}(z|x)]$ converges to a finite value as $t \to \infty$ given that the step size ε satisfies the following inequality:*

$$0 < \varepsilon < \min(\frac{1}{\mathscr{B}}, \frac{\mathscr{H}_0}{\mathscr{A}}), \tag{S16}$$

Before proposing the proof, we should introduce the light-tailness property on the target distribution $\mathcal{P}(z|x)$ in order to ensure the validity of our Taylor expansion and to control higher-order discretization errors. Specifically, we say that $\mathcal{P}(z|x)$ is light-tailed if there exists a universal constant $\mathscr{H}_0 < \infty$ such that

$$\int \|\nabla_z \log \mathcal{P}(z|x)\| \, \mathcal{P}(z|x)\mathrm{d}z < \mathscr{H}_0. \tag{S17}$$

This condition requires that the expectation (under $\mathcal{P}(z|x)$) of the norm of the score function $\nabla_z \log \mathcal{P}(z|x)$ is finite. Intuitively, this ensures that $\mathcal{P}(z|x)$ decays sufficiently rapidly in the tails so that the gradients do not blow up at infinity. On this basis, the proof is articulated as follows:

Proof. When we obtain the weak solution by simulating Equation (S15), suppose at iteration t, our variational distribution is $\mathcal{Q}_t(z)$, and the target distribution is $\mathcal{P}(z|x)$. We perform a small transport transformation on z:

$$z_{t+\varepsilon} = \mathcal{T}(z) \coloneqq z + \varepsilon\phi(z). \tag{S18}$$

where ε is a small step size, and $\phi(z)$ is a smooth perturbation direction.

The pushforward distribution (after applying Equation (S18)) is denoted as $\mathcal{Q}_{t+\varepsilon}(z)$. The aim is to Taylor expand the evolution of

$$\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{t+\varepsilon}(z)\|\mathcal{P}(z|x)] = \int \mathcal{Q}_{t+\varepsilon}(z) \log \frac{\mathcal{Q}_{t+\varepsilon}(z)}{\mathcal{P}(z|x)} dz \tag{S19}$$

with respect to ε, around $\varepsilon = 0$. The new probability density, for small ε, can be given by Liouville's theorem:

$$\mathcal{Q}_{t+\varepsilon}(z) = \mathcal{Q}_t(\mathcal{T}^{-1}(z)) \cdot |\det \mathcal{J}_{\mathcal{T}^{-1}}(z)| \tag{S20}$$

where $\mathcal{J}_{\mathcal{T}^{-1}}(z)$ is the Jacobian matrix of the inverse map, and for small ε:

$$\mathcal{T}^{-1}(z) \approx z - \varepsilon\phi(z) \tag{S21}$$

so expanding to the first order in ε:

$$\begin{aligned} \mathcal{Q}_{t+\varepsilon}(z) &\approx \mathcal{Q}_t(z - \varepsilon\phi(z)) \left[1 - \varepsilon\nabla \cdot \phi(z)\right] \\ &\approx \mathcal{Q}_t(z) - \varepsilon\nabla_z \left[q_t(z)\phi(z)\right] \end{aligned} \tag{S22}$$

Define (with $\mathscr{F}(\varepsilon) \coloneqq \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{t+\varepsilon}(z)\|\mathcal{P}(z|x)]$):

$$\mathscr{F}(\varepsilon) \coloneqq \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{t+\varepsilon}(z)\|\mathcal{P}(z|x)] = \int \mathcal{Q}_{t+\varepsilon}(z) \log \frac{\mathcal{Q}_{t+\varepsilon}(z)}{\mathcal{P}(z|x)} \mathrm{d}z. \tag{S23}$$

Applying the Taylor's expansion at $\varepsilon = 0$, we get:

$$\mathscr{F}(\varepsilon) = \mathscr{F}(0) + \varepsilon\mathscr{F}'(0) + \mathcal{O}(\varepsilon^2). \tag{S24}$$

Clearly, $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_t(z)\|\mathcal{P}(z|x)]$. Now, we compute $\mathscr{F}'(0)$. Take the derivative inside the integral:

$$\begin{aligned} \mathscr{F}'(\varepsilon) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \int \mathcal{Q}_{t+\varepsilon}(z) \log \frac{\mathcal{Q}_{t+\varepsilon}(z)}{\mathcal{P}(z|x)} \mathrm{d}z \\ &= \int \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathcal{Q}_{t+\varepsilon}(z) \left(1 + \log \frac{\mathcal{Q}_{t+\varepsilon}(z)}{\mathcal{P}(z|x)}\right) \mathrm{d}z \end{aligned} \tag{S25}$$

At $\varepsilon = 0$, $\mathcal{Q}_{t+\varepsilon}(z) = \mathcal{Q}_t(z)$:

$$\mathscr{F}'(0) = \int \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathcal{Q}_{t+\varepsilon}(z)\Big|_{\varepsilon=0} [1 + \log \frac{\mathcal{Q}_t(z)}{\mathcal{P}(z|x)}] \mathrm{d}z \tag{S26}$$

Now, using the result from the calculus of variations:

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathcal{Q}_{t+\varepsilon}(z)\Big|_{\varepsilon=0} = -\nabla_z \cdot (\mathcal{Q}_t(z)\phi(z)) \tag{S27}$$

Thus,

$$\mathscr{F}'(0) = -\int \nabla_z \cdot (\mathcal{Q}_t(z)\phi(z)) \left[1 + \log \frac{\mathcal{Q}_t(z)}{\mathcal{P}(z|x)}\right] \mathrm{d}z. \tag{S28}$$

Now, use integration by parts:

$$\int -\nabla_z \cdot (q_t(z)\phi(z)) f(z) \mathrm{d}z = \int q_t(z)\phi(z)^\top \nabla_z f(z) \mathrm{d}z. \tag{S29}$$

Set $f(z) = 1 + \log \frac{q_t(z)}{p(z|x)}$. Its gradient is:

$$\nabla_z \psi(z) = \nabla_z \log q_t(z) - \nabla_z \log p(z|x) \tag{S30}$$

Hence,

$$\begin{aligned} \mathscr{F}'(0) &= \int \mathcal{Q}_t(z)\phi(z)^\top \left(\nabla_z \log \mathcal{Q}_t(z) - \nabla_z \log \mathcal{P}(z|x)\right) \mathrm{d}z \\ &= \mathbb{E}_{\mathcal{Q}_t(z)} \left[\phi(z)^\top (\nabla_z \log \mathcal{Q}_t(z) - \nabla_z \log \mathcal{P}(z|x))\right] \end{aligned} \tag{S31}$$

But with a negative sign because the original derivative is minus divergence:

$$\mathscr{F}'(0) = -\mathbb{E}_{\mathcal{Q}_t(z)} \left[(\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z))^\top \phi(z)\right]. \tag{S32}$$

Putting all together, we get:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}} \left(\mathcal{Q}_{t+\varepsilon}(z)\|p(z|x)\right) \\ =&\mathbb{D}_{\mathrm{KL}} \left(\mathcal{Q}_t(z)\|\mathcal{P}(z|x)\right) - \varepsilon \, \mathbb{E}_{\mathcal{Q}_t(z)} \big[(\nabla_z \log \mathcal{P}(z|x) \\ &- \nabla_z \log \mathcal{Q}_t(z))^\top \phi(z)\big] + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{S33}$$

Since $\|\nabla_z\phi(z)\| \leq \mathscr{B}$, there exists a positive constant $\mathcal{C}$ such that:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{t+\varepsilon}(z)\|\mathcal{P}(z|x)] \leq \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_t(z)\|\mathcal{P}(z|x)] \\ &- \varepsilon\mathbb{E}_{\mathcal{Q}_t(z)}\{[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z|x)]^\top \phi(z)\} + \mathcal{C}\varepsilon^2. \end{aligned} \tag{S34}$$

In addition, when no restrictions are imposed on the hypothesis space for $\psi(z)$, the maximum value of the optimization problem

$$\arg\max_{\psi(x)} \int \mathcal{Q}_t(z)\ [\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z)]^\top \times [\nabla_z \log \mathcal{P}(z|x) + \psi(z)]\ \mathrm{d}z. \quad \text{(S35)}$$

is attained when:

$$\psi(z) = -\nabla_z \log \mathcal{Q}_t(z). \quad \text{(S36)}$$

Specfically, since

$$\begin{aligned}
&\int \mathcal{Q}_t(z)\ [\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z)]^\top \\
&\qquad \times [\nabla_z \log \mathcal{P}(z|x) + \psi(z)]\ \mathrm{d}z \\
=&\int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{P}(z|x)]^\top\, [\nabla_z \log \mathcal{P}(z|x)]\, \mathrm{d}z \\
&+ \int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{P}(z|x)]^\top\, [\psi(z)]\, \mathrm{d}z \\
&- \int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{Q}_t(z)]^\top\, [\nabla_z \log \mathcal{P}(z|x)]\, \mathrm{d}z \\
&- \int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{Q}_t(z)]^\top\, [\psi(z)]\, \mathrm{d}z.
\end{aligned}$$

Notably, $\int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{P}(z|x)]^\top\, [\nabla_z \log \mathcal{P}(z|x)]\, \mathrm{d}z$ and $\int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{Q}_t(z)]^\top\, [\nabla_z \log \mathcal{P}(z|x)]\, \mathrm{d}z$ do not depend on $\psi(z)$, the optimization problem can be simplified as follows:

$$\begin{aligned}
\arg\max_{\psi(z)} &\int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{P}(z|x)]^\top\, [\psi(z)]\, \mathrm{d}z \\
&- \int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{Q}_t(z)]^\top\, [\psi(z)]\, \mathrm{d}z.
\end{aligned}$$

It should be pointed out that the expression to be maximized can be further reformulated as follows:

$$\arg\max_{\psi(z)} \int \mathcal{Q}_t(z)\, [\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z)]^\top\, [\psi(z)]\, \mathrm{d}z,$$

which is an inner product between the vector $\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z)$ and $\psi(z)$. Notably, we wish to maximize the functional with-respect-to $\psi(z)$, therefore we take the first variation of the objective with-respect-to $\psi(z)$ as follows:

$$\begin{aligned}
&\frac{\delta}{\delta \psi(z)} \int [\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z)] \psi(z) \mathcal{Q}_t(z) \mathrm{d}z \\
=&\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z) \\
=&0,
\end{aligned} \quad \text{(S37)}$$

which indicates that we can get the following results when Equation (S35) reaches to the maximum value:

$$\nabla_z \log \mathcal{Q}_t(z) = \nabla_z \log \mathcal{P}(z|x). \quad \text{(S38)}$$

Comparing Equation (S38) to Equation (S35), it can be observed that the optimal $\psi(z)$ that maximizes Equation (S35) is expressed as follows:

$$\psi(z) = -\nabla_z \log \mathcal{Q}_t(z).$$

As a result, we have the following inequality:

$$\begin{aligned}
&[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z|x)]^\top \phi(z) \\
\geq &\|\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z|x)\|^2 \geq 0.
\end{aligned} \quad \text{(S39)}$$

which indicates that the following inequality holds:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{t+\varepsilon}(z) \| \mathcal{P}(z|x)] \leq \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_t(z) \| \mathcal{P}(z|x)] \\
&- \varepsilon \mathbb{E}_{\mathcal{Q}_t(z)} \{ \|\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_t(z|x)\|^2 \} + \mathcal{C}\varepsilon^2.
\end{aligned} \quad \text{(S40)}$$

Notably, since $0 < \varepsilon < \min(\frac{1}{\mathscr{B}}, \frac{\mathscr{H}_0}{\mathscr{A}})$, $\mathcal{C}\varepsilon^2$ will not dominate the right-hand-side of Equation (S40). Consequently the iterative sequence for the of the E-step for InfO-EM algorithm will progressively reduce the KL divergence between $\mathcal{Q}_t(z)$ and $\mathcal{P}(z|x)$. In addition, the KL divergence is non-negative $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_t(z) \| \mathcal{P}(z|x)] \geq 0$, thus the iterative process of the E-step for InfO-EM algorithm converges. □

After that, we propose the detailed elaboration for the convergence of the M-step for the InfO-EM algorithm.

Theorem S2. *Suppose that the empirical negative log-likelihood $\mathscr{L}(\theta) \coloneqq -\frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} \log p_\theta(x_i|z_i)$, is L-smooth; that is, there exists $L > 0$ such that for any θ and θ', $\|\nabla_\theta \log p_\theta(x|z) - \nabla_{\theta'} \log p_{\theta'}(x|z)\| \leq L\|\theta' - \theta\|$. Furthermore, assume $-\log p_\theta(x|z)$ is lower bounded, i.e., there exists $\mathscr{R} > 0$ such that $-\log p_\theta(x|z) > \mathscr{R}$ for all θ. Under these assumptions, for gradient descent with a fixed step size $\xi \in (0, \frac{2}{L})$, we have: $\|\nabla_\theta \mathscr{L}(\theta^\tau)\| \to 0$ as $\tau \to \infty$.*

Before start proving the theorem, we should demonstrate the basic concept of the L-smooth. Suppose the empirical negative log-likelihood $\mathscr{L}(\theta) \coloneqq -\frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} \log p_\theta(x_i|z_i)$ is L-smooth; that is, there exists $L > 0$ such that for any θ, θ',

$$\|\nabla_\theta \log p_\theta(x|z) - \nabla_{\theta'} \log p_{\theta'}(x|z)\| \leq L\|\theta' - \theta\|. \quad \text{(S41)}$$

Proof. By the smoothness assumption, for any θ, θ',

$$\mathscr{L}(\theta') \leq \mathscr{L}(\theta) + \nabla_\theta \mathscr{L}(\theta)^\top (\theta' - \theta) + \frac{L}{2}\|\theta' - \theta\|^2, \quad \text{(S42)}$$

Let the update at iteration τ be

$$\theta^{\tau+1} = \theta^\tau - \xi \nabla_\theta \mathscr{L}(\theta^\tau), \quad \text{(S43)}$$

where $\xi \in (0, \frac{2}{L})$ is the (fixed) step size. Applying the L-smoothness upper bound with $\theta = \theta^\tau$ and $\theta' = \theta^{\tau+1}$,

$$\begin{aligned}
&\mathscr{L}(\theta^{\tau+1}) \\
\leq &\mathscr{L}(\theta^\tau) + \nabla_\theta \mathscr{L}(\theta^\tau)^\top (\theta^{\tau+1} - \theta^\tau) + \frac{L}{2}\|\theta^{\tau+1} - \theta^\tau\|^2 \\
= &\mathscr{L}(\theta^\tau) + \nabla_\theta \mathscr{L}(\theta^\tau)^\top (-\xi \nabla_\theta \mathscr{L}(\theta^\tau)) + \frac{L}{2}\xi^2 \|\nabla_\theta \mathscr{L}(\theta^\tau)\|^2 \\
= &\mathscr{L}(\theta^\tau) - \xi \|\nabla_\theta \mathscr{L}(\theta^\tau)\|^2 + \frac{L}{2}\xi^2 \|\nabla_\theta \mathscr{L}(\theta^\tau)\|^2 \\
= &\mathscr{L}(\theta^\tau) - (\xi - \frac{L\xi^2}{2}) \|\nabla_\theta \mathscr{L}(\theta^\tau)\|^2.
\end{aligned} \quad \text{(S44)}$$

Consequently, we can get the following result:

$$\mathscr{L}(\theta^{\tau+1}) \leq \mathscr{L}(\theta^\tau) - (\xi - \frac{L\xi^2}{2}) \|\nabla_\theta \mathscr{L}(\theta^\tau)\|^2. \quad \text{(S45)}$$

Note that for $\xi \in (0, \frac{2}{L})$, we have $\xi - \frac{L\xi^2}{2} > 0$, so unless $\nabla_\theta \mathscr{L}(\theta^\tau) = 0$, the objective strictly decreases each step. In particular, summing both sides from $\tau = 0$ to $\mathcal{E} - 1$,

$$\mathscr{L}(\theta^{\mathcal{E}}) \leq \mathscr{L}(\theta^0) - \left(\xi - \frac{L\xi^2}{2}\right) \sum_{\tau=0}^{\mathcal{E}-1} \|\nabla \mathscr{L}(\theta^\tau)\|^2 \tag{S46}$$

Since $\mathscr{L}(\theta)$ is lower bounded, both sides are finite and

$$\sum_{\tau=0}^{\infty} \|\nabla_\theta \mathscr{L}(\theta^\tau)\|^2 < \infty, \tag{S47}$$

which means the sequence $\|\nabla_\theta \mathscr{L}(\theta^\tau)\|^2$ is summable. If $\|\nabla_\theta \mathscr{L}(\theta^\tau)\|$ does not go to zero, then for some $\mathscr{R} > 0$, $\|\nabla \mathscr{L}(\theta^\tau)\| > \mathscr{R}$ infinitely often; the sum would diverge, which is contradiction. Hence,

$$\lim_{\tau \to \infty} \|\nabla \mathscr{L}(\theta^\tau)\| = 0. \tag{S48}$$

That is, the sequence converges to a stationary point. □

S.III. Detailed Experimental Information

A. KSD-based Test of the Goodness-of-fit

In this subsection, we detail the procedure for conducting a goodness-of-fit test based on the KSD. First, we recall the definition of KSD:

$$\begin{aligned} \mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) :=& \mathbb{E}_{z,z' \sim \mathcal{Q}(z)} \left[\mathcal{V}_{\mathcal{P}(z|x)}(z, z')\right] \\ \approx& \frac{1}{\mathrm{M}^2} \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} \mathcal{V}_{\mathcal{P}(z|x)}\left(z_i, z_j\right), \end{aligned} \tag{S49a}$$

$$\begin{aligned} \mathcal{V}_{\mathcal{P}(z|x)}(z, z') :=& \left[\nabla_z \log \mathcal{P}(z|x)\right]^\top K(z, z') \left[\nabla_{z'} \log \mathcal{P}(z'|x)\right] \\ &+ \left[\nabla_z \log \mathcal{P}(z|x)\right]^\top \nabla_{z'} K(z, z') \\ &+ \left[\nabla_z K(z, z')\right]^\top \left[\nabla_{z'} \log \mathcal{P}(z'|x)\right] \\ &+ \mathrm{Trace}\left(\nabla_{z,z'} K(z, z')\right). \end{aligned} \tag{S49b}$$

Based on [SR4], we compute the bootstrap samples $\mathbb{S}^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right)$ repeatedly for L iterations (referred to as the bootstrap sample size) using the following equation:

$$\begin{aligned} &\mathbb{S}_l^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) \\ =& \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} (w_{i,l} - \frac{1}{\mathrm{M}})(w_{j,l} - \frac{1}{\mathrm{M}}) \mathcal{V}_{\mathcal{P}(z|x)}\left(z_i, z_j\right), \end{aligned} \tag{S50}$$

where $l \in \{1, 2, \cdots, \mathrm{L}\}$ is the index for the bootstrap iteration, and $(w_1, \cdots, w_{\mathrm{M}}) \sim \mathrm{Mult}(\mathrm{M}; \frac{1}{\mathrm{M}}, \cdots, \frac{1}{\mathrm{M}})$. Here, Mult denotes the multinomial distribution, and M is the sample size.

Next, we compute the proportion $\widehat{\alpha}$ of bootstrap samples satisfying the following condition:

$$\widehat{\alpha} = \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \mathbb{I}\left[\mathbb{S}_l^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) > \mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right)\right], \tag{S51}$$

where $\mathbb{I}$ is the indicator function defined as:

$$\begin{aligned} &\mathbb{I}\left[\mathbb{S}_l^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) > \mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right)\right] \\ :=& \begin{cases} 1 & \text{if } \mathbb{S}_l^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) > \mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right), \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{S52}$$

Finally, for a predefined confidence level α and the following hypotheses:

- H_0: The samples $\{z_i\}_{i=1}^{\mathrm{M}} \sim \mathcal{Q}(z)$ are drawn from $\mathcal{P}(z|x)$,
- H_1: The samples $\{z_i\}_{i=1}^{\mathrm{M}} \sim \mathcal{Q}(z)$ are not drawn from $\mathcal{P}(z|x)$,

The corresponding decision rule is as follows:

- Reject H_0 if $\widehat{\alpha} > \alpha$.
- Accept H_0 if $\widehat{\alpha} \leq \alpha$.

The entire procedure is summarized in Algorithm S1.

Algorithm S1 Bootstrap Goodness-of-fit Test based on KSD.

1: **Input:** Samples $\{z_i\}_{i=1}^{\mathrm{M}}$, score function $\nabla_z \log \mathcal{P}(z|x)$, significance level α, and bootstrap sample size L.
2: **Test:** H_0: $\{z_i\}_{i=1}^{\mathrm{M}}$ are drawn from $\mathcal{P}(z|x)$, versus, H_1: $\{z_i\}_{i=1}^{\mathrm{M}}$ is not drawn from $\mathcal{P}(z|x)$.
3: $\mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) \leftarrow$ Eq. (S49a)
4: $\widehat{\alpha} \leftarrow$ Eq. (S51)
5: **if** $\widehat{\alpha} > \alpha$ **then**
6: **Output:** H_1
7: **else**
8: **Output:** H_0
9: **end if**

B. Detailed Information for Inferential Sensor Dataset

1) Debutanizer Column: Figure S1 presents the flowsheet of the DC, a benchmark dataset [SR5]. The DC is required to maximize the pentane (C5) content in the overhead distillate and simultaneously minimize the bottom flow's butane (C4) content. To measure the butane concentration from the bottom flow in real-time to improve downstream product quality, seven process variables marked in the red zone in Fig. S1 are chosen for inferential sensor modeling. A total of 2,396 samples are collected from the process. The detailed descriptions of the process variables are given in Table SI. Based on [SR5], we extend the process variables into 13 dimensions as follows:

$$\begin{bmatrix} u_1(t), u_2(t), u_3(t), u_4(t), u_5(t), u_5(t-1), \\ u_5(t-2), u_5(t-3), (u_1(t) + u_2(t))/2, \\ x(t-1), x(t-2), x(t-3), x(t-4) \end{bmatrix}^\top. \tag{S53}$$

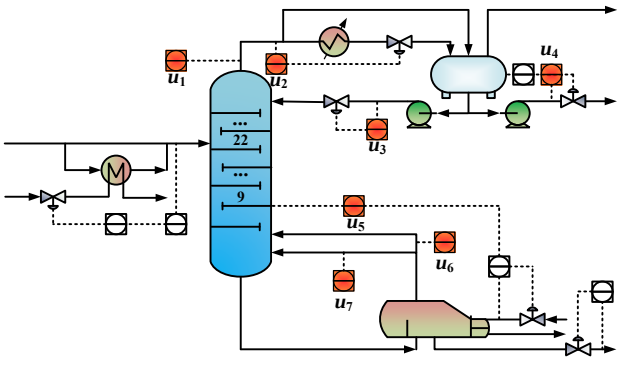

Fig. S1. The flowsheet of debutanizer column.

TABLE SI
THE PROCESS VARIABLES IN THE DC

Process variables	Unit	Description
u_1	°C	Top temperature
u_2	$kg \cdot cm^{-2}$	Top pressure
u_3	$m^3 \cdot h^{-1}$	Reflux flow rate
u_4	$m^3 \cdot h^{-1}$	Top distillate rate
u_5	°C	Temperature of the 9th tray
u_6	°C	Bottom temperature A
u_7	°C	Bottom temperature B

2) Carbon-Dioxide Absorber: Figure S2 presents the flowsheet of the carbon-dioxide (CO_2) absorber column (CA) [SR6]. The CA is a vital equipment in ammonia synthesis process to handle the CO_2 by-product in the hydrogen from upstream unit. The sodium hydroxide solvent is chosen to be absorption liquid and the corresponding chemical reaction can be given in (S54):

$$\begin{cases} CO_2 + 2KOH = K_2CO_3 + H_2O \\ K_2CO_3 + H_2O + CO_2 = 2KHCO_3 \end{cases}, \tag{S54}$$

where KOH, K_2CO_3, and $KHCO_3$ are potassium hydroxide, potassium carbonate, and potassium bicarbonate, respectively.

To enhance product quality in the downstream urea synthesis process, it is crucial to minimize the CO_2 concentration in the hydrogen stream, necessitating real-time monitoring of CO_2 levels in the outlet gas. However, traditional gas chromatography methods for measuring CO_2 concentration are hindered by significant time delays.

To address this measurement challenge and enhance the control quality of the CO_2 absorber, industrial data-driven models have been employed to estimate the CO_2 concentration in the absorber's outlet stream in real time. Despite a comprehensive understanding of the absorber's operations, constructing a rigorous process simulation is challenging due to the absence of thermodynamic electrolyte binary parameters, making a data-driven model the preferred approach for developing the industrial data-driven task.

For the purposes of quality monitoring and control, several hard sensors are installed within the plant to collect data on process variables, which serve as secondary variables for the industrial data-driven. Eleven process variables, identified within the red zone, have been selected for model construction. Table SII provides a detailed description of these variables. In total, 6,000 samples were collected from the process.

3) Water-Gas-Shift Unit: Figure S3 presents the flowsheet of the water-gas-shift (WGS) unit. This process is crucial for increasing the hydrogen content in the gas stream, which is essential for ammonia synthesis. The WGS process involves a series of fixed-bed reactors connected in series, where the following heterogeneous catalytic reaction takes place:

$$CO + H_2O \leftrightarrow CO_2 + H_2. \tag{S55}$$

This reaction, known as the water-gas shift reaction, converts carbon monoxide (CO) and water vapor (H_2O) into carbon dioxide (CO_2) and hydrogen (H_2). The reaction is exothermic and is typically facilitated by a catalyst, such as iron oxide or copper-based catalysts, to enhance the reaction rate and selectivity. To ensure that the technology requirement of the carbon-to-hydrogen ratio is met for efficient ammonia synthesis, it is crucial to monitor and control the carbon monoxide (CO) concentration in the gas stream. For this purpose, several hard sensors are installed in the WGS section to collect process variables in real time. These sensors provide vital data for process control and optimization.

Furthermore, thirteen process variables marked in the red zone in Figure S3 are selected for data-driven modeling. These process variables are chosen based on their relevance to the process dynamics and their impact on the CO concentration. The industrial data-driven model developed using these process variables enables the real-time estimation of CO concentration, providing a valuable tool for process control and optimization. The detailed descriptions of these process variables, including their measurement locations and process significance, are provided in Table SIII.

TABLE SII
PROCESS VARIABLES FOR CARBON-DIOXIDE ABSORBER.

Input Variables	Description
u_1	Pressure of inlet gas
u_2	Liquid level of buffer vessel
u_3	Temperature of inlet barren liquor
u_4	Flowrate of inlet lean solution
u_5	Flowrate of inlet semi-lean solution
u_6	Temperature of inlet gas
u_7	Pressure drop of absorber
u_8	Temperature of rich solution
u_9	Liquid level of absorber
u_{10}	Liquid level of the separator
u_{11}	Pressure of outlet gas
x	CO_2 concentration

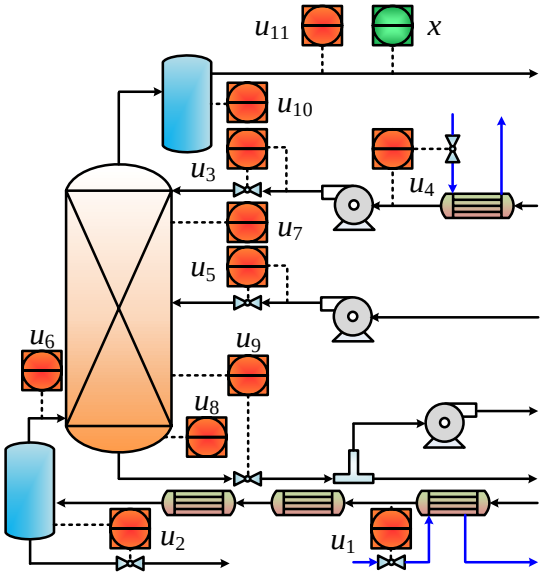

Fig. S2. The flowsheet of carbon-dioxide absorber.

C. Experiment Protocols

In this paper, we choose the following baseline models to demonstrate the efficacy of the proposed InfO-

TABLE SIII
THE PROCESS VARIABLES IN THE WGS UNIT

Process variables	Description
u_1	High temperature bed temperature 1
u_2	High temperature bed temperature 2
u_3	High temperature bed temperature 3
u_4	Outlet temperature of high temperature bed
u_5	Outlet temperature of cooling water
u_6	Split-gas temperature
u_7	Inlet temperature of low temperature bed
u_8	Low temperature bed temperature 1
u_9	Low temperature bed temperature 2
u_{10}	Low temperature bed temperature 3
u_{11}	Outlet temperature of low temperature bed
u_{12}	Outlet pressure of low temperature bed
u_{13}	Product gas pressure
x	Carbon monoxide concentration

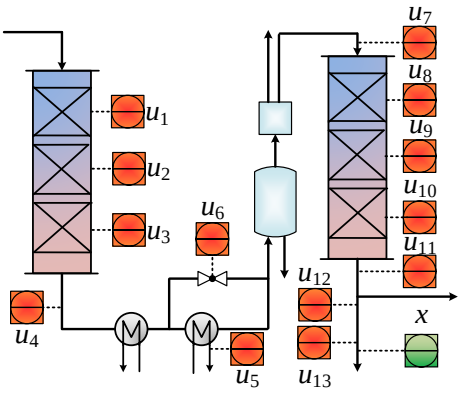

Fig. S3. The flowsheet of water-gas-shift unit.

PLVM: Deep Bayesian Probabilistic Slow Feature Analysis (DBPSFA) [SR7], Modified Unsupervised VAE-Supervised Deep VAE (MUDVAE-SDVAE) [SR8], Nonlinear Probabilistic Latent Variable Regression (NPLVR) [SR9], and Gaussian Mixture-Variational Autoencoder (GMVAE) [SR10].

All experiments are conducted on a workstation equipped with an Intel Xeon E5 processor, 8 Nvidia GTX 1080 GPUs, and 64 GB of RAM. To maintain consistency and fairness across evaluations, the Adam optimizer, as detailed by [SR11], is employed uniformly across all experimental runs. The model training and inference processes are carried out using Python 3.8 and PyTorch 1.13 [SR12].

For all datasets, the data is sorted in ascending order by timestamp. On this basis, the first 60% of the data is selected for training, the first 60% to 80% of the data is selected for validation, and the rest of the data is selected for testing.

1) Reasons for Baseline Models: To demonstrate the effectiveness of the proposed Info-PLVM, several PLVMs designed for industrial inferential sensor modeling—specifically, NPLVR, DBPSFA, and MUDVAE-SDVAE—are selected as baseline models. In addition, we noticed that the approximation distribution for NPLVR, DBPSFA, and MUDVAESDVAE is specified as a Gaussian distribution. To further investigate the influence of the specification function family type of approximation distribution, we add the GMVAE as a baseline model.

TABLE SIV
HYPER-PARAMETER SETTINGS

Model Name	$\mathcal{B}$	ξ	$\mathrm{D_{LV}}$
DBPSFA	32	0.005	5
MUDVAE-SDVAE	32	0.01	5
NPLVR	64	0.01	5
GMVAE	128	0.0001	5
InfO-PLVM	64	0.1	5

2) Hyperparameter Settings: For fairness, we use the multi-layer-perceptron to parameterize $p_\theta(x|z)$, and the hidden unit for the multi-layer-perceptron is set as $[10, 7, 5]$ based on reference [SR9]. On this basis, the particle number M, bandwidth h, discretization step-size ε, and simulation time T for InfO-PLVM are set as 20, 1.0, 0.01, and 200, respectively. Other parameters like learning rate ξ, batch size $\mathcal{B}$, and dimension of latent space $\mathrm{D_{LV}}$ for the baseline models and InfO-PLVM are listed in Table SIV.

REFERENCES

[SR1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

[SR2] R. Vershynin, "High-dimensional probability," 2009.

[SR3] J. C. Butcher, *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[SR4] Q. Liu, J. Lee, and M. Jordan, "A kernelized stein discrepancy for goodness-of-fit tests," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 276–284.

[SR5] L. Fortuna, S. Graziani, and M. G. Xibilia, "Soft sensors for product quality monitoring in debutanizer distillation columns," *Control Eng. Pract.*, vol. 13, no. 4, pp. 499–508, 2005.

[SR6] B. Shen, L. Yao, and Z. Ge, "Nonlinear probabilistic latent variable regression models for soft sensor application: From shallow to deep structure," *Control Eng. Pract.*, vol. 94, p. 104198, 2020.

[SR7] C. Jiang, Y. Lu, W. Zhong, B. Huang, W. Song, D. Tan, and F. Qian, "Deep bayesian slow feature extraction with application to industrial inferential modeling," *IEEE Trans. Ind. Inform.*, pp. 1–1, 2021.

[SR8] R. Xie, N. M. Jan, K. Hao, L. Chen, and B. Huang, "Supervised variational autoencoders for soft sensor modeling with missing data," *IEEE Trans. Ind. Inform.*, vol. 16, no. 4, pp. 2820–2828, 2019.

[SR9] B. Shen, L. Yao, and Z. Ge, "Nonlinear probabilistic latent variable regression models for soft sensor application: From shallow to deep structure," *Control Eng. Pract.*, vol. 94, p. 104198, 2020.

[SR10] F. Guo, B. Wei, and B. Huang, "A just-in-time modeling approach for multimode soft sensor based on gaussian mixture variational autoencoder," *Comput. Chem. Eng.*, vol. 146, p. 107230, 2021.

[SR11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–8.

[SR12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.