APPENDIX

A. Identifiability

Proposition 1. *If we can recover $P(Z, X, Y_0, Y_1)$, then we can recover the CATE under the causal model in Fig. 1a.*

Proof. We can first change the variables using the formula $Y := Y_0(1-t) + Y_1 t$ and $T := t$. Then, $P(Z, X, Y_0, Y_1)$ can readily be transformed to $P(Z, X, T, Y)$, which in turn can prove the identifiability according to Theorem 1 in [1]. □

B. Pseudo Code

Algorithm 1 CEMVAE

Input: Dataset $\mathcal{D}_{obs} = \{x_i, t_i, y_i\}_{i=1}^{n}$, coefficient α, sampling size k

Initialize: $q_{\phi_0}(Y|X,T)$, $q_{\phi_1}(Y|X,T)$, $q_\theta(Z|X,Y_0,Y_1)$, $p_\psi(X|Z)$, $p_{\epsilon_0}(Y_0|Z)$, and $p_{\epsilon_1}(Y_1|Z)$

Pretrain auxiliary models until convergence.
while $q_{\phi_0}(Y|X,T)$ and $q_{\phi_1}(Y|X,T)$ are not converged **do**
    $\phi_0, \phi_1 \leftarrow \underset{\phi_0,\phi_1}{\arg\max} \sum_{i=1}^{n} (1 - t_i) \log q_{\phi_0}(y_i|x_i, t_i) + t_i \log q_{\phi_1}(y_i|x_i, t_i)$.
end while
Train all models end-to-end
while $q_{\phi_0}(\cdot), q_{\phi_1}(\cdot), q_\theta(\cdot), p_\psi(\cdot), p_{\epsilon_0}(\cdot)$, and $p_{\epsilon_1}(\cdot)$ are not converged **do**
    sample k counterfactuals $y_{1-t_i,i}$ from $q_{\phi_{1-t_i}}(Y|X = x_i, T = 1 - t_i)$
    $\phi_0, \phi_1, \psi, \theta, \epsilon_0, \epsilon_1 \leftarrow \underset{\phi_0,\phi_1,\psi,\theta,\epsilon_0,\epsilon_1}{\arg\max} \sum_{i=1}^{n} \mathcal{L}_{CEMVAE}(x_i, y_{t_i})$
end while

Output: $q_{\phi_0}(Y|X,T)$, $q_{\phi_1}(Y|X,T)$, $q_\theta(Z|X,Y_0,Y_1)$, $p_\psi(X|Z)$, $p_{\epsilon_0}(Y_0|Z)$, and $p_{\epsilon_1}(Y_1|Z)$

C. Dataset Summarization

TABLE I: Dataset Statistics

	IHDP	eICU	Synthetic
Size	747	1824	1000
$P(T = 1)$	0.18	0.32	0.71
Dimension of discrete features	19	7	3
Dimension of continuous features	6	28	10
test set size	113	274	150
validation set size	190	100	100

D. Hyperparameters

The hyperparameters for grid search for our models and some key baseline models are given. Weight decay is short for "wdecay", the dimension of Z is shorted for "dz", learning rate is short for "lr", the number of hidden layers is short for "hidden", and the size of a hidden layer is short for "dl".

TABLE II: Hyperparameters for CEMVAE and CEMVAE-D.

Hyperparameters	eICU	Synthetic	IHDP
wdecay	$[10^{-4}]$	$[10^{-4}]$	$[10^{-4}]$
dz	[21,25,29]	[6,10,14]	[17, 22, 27]
lr	[0.001]	[0.001]	[0.001]
hidden	[2,3,4]	[2,3,4]	[3,4,5]
dl	[170,210,250]	[80,110,140]	[150, 200, 250]
α	[0.8,1.0,1.2,1.8]	[0.8,1.0,1.2,1.8]	[0.8,1.0,1.2,1.8]

TABLE III: Hyperparameters for CEVAE.

Hyperparameters	eICU	Synthetic	IHDP
wdecay	$[10^{-4}]$	$[10^{-4}]$	$[10^{-4}]$
dz	[17,22,27]	[6,10,14]	[17, 22, 27]
lr	[0.001]	[0.001]	[0.001]
hidden	[3,4,5]	[2,3,4]	[3,4,5]
dl	[150,200,250]	[80,110,140]	[150, 200, 250]

REFERENCES

[1] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, page 6449–6459, Red Hook, NY, USA, 2017. Curran Associates Inc.

TABLE IV: Hyperparameters for TEDVAE.

Hyperparameters	eICU	Synthetic	IHDP
wdecay	$[10^{-4}]$	$[10^{-4}]$	$[10^{-4}]$
dz	[17,22,27]	[5,7, 10]	[13,15,17,18,21]
dz_c	dz	dz	dz
dz_y	$int(0.5dz)$	$int(0.5dz)$	$int(0.5dz)$
dz_t	$int(0.5dz)$	$int(0.5dz)$	$int(0.5dz)$
lr	[0.001]	[0.001]	[0.001]
hidden	[3,4,5]	[2,3,4]	[3,4,5]
dl	[150,200,250]	[80,110,140]	[150, 200, 250]
α_t	[0.8,1.0,1.2,1.8]	[0.8,1.0,1.2,1.8]	[0.8,1.0,1.2,1.8]
α_y	α_t	α_t	α_t

TABLE V: Hyperparameters for X-MLP.

Hyperparameters	eICU	Synthetic	IHDP
lr	[0.001]	[0.001]	[0.001]
hidden	[3,4,5]	[2,3,4]	[3,4,5]
dl	[150, 180, 210]	[80,100,120]	[120, 150, 180]