Technical Appendix of Distribution-conditioned Adversarial Variational Autoencoder for Valid Instrumental Variable Generation

Anonymous submission

Theoretical Proof

Proof of Evidence Lower Bound

Proof. We proof the evidence lower bound of our VIV framework as below.

$$
\begin{aligned}
& log p(X,T,Y) \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log p(X,T,Y) dZdUdCdA \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{p(X,T,Y,Z,U,C,A)}{p(Z,U,C,A|X,T,Y)} dZdUdCdA \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{p(X,T,Y,Z,U,C,A)}{q(Z,U,C,A|X,T,Y)} \\
& \frac{q(Z,U,C,A|X,T,Y)}{p(Z,U,C,A|X,T,Y)} dZdUdCdA \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{p(X,T,Y,Z,U,C,A)}{q(Z,U,C,A|X,T,Y)} dZdUdCdA \\
& + \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{q(Z,U,C,A|X,T,Y)}{p(Z,U,C,A|X,T,Y)} dZdUdCdA \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{p(X,T,Y,Z,U,C,A)}{q(Z,U,C,A|X,T,Y)} dZdUdCdA \\
& + KL(q(Z,U,C,A|X,T,Y)||p(Z,U,C,A|X,T,Y))
\end{aligned} \tag{1}
$$

Based on the non-negativity of KL-divergence, we have:

$$
\begin{aligned}
& log p(X,T,Y) \\
\geq & \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{p(X,T,Y,Z,U,C,A)}{q(Z,U,C,A|X,T,Y)} dZdUdCdA \\
\equiv & ELBO_{VIV} \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{p(X,T,Y|Z,U,C,A)p(Z,U,C,A)}{q(Z,U,C,A|X,T,Y)} dZdUdCdA \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log p(X,T,Y|Z,U,C,A) dZdUdCdA \\
& + \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log \frac{p(Z,U,C,A)}{q(Z,U,C,A|X,T,Y)} dZdUdCdA \\
=& \iiiint_{Z,U,C,A} q(Z,U,C,A|X,T,Y) \\
& log p(X,T,Y|Z,U,C,A) dZdUdCdA \\
& - KL(q(Z,U,C,A|X,T,Y)||p(Z,U,C,A))
\end{aligned} \tag{2}
$$

Based on the local Markov assumption (Pearl et al. 2000), we have:

$$
\begin{aligned}
& p(X,T,Y|Z,U,C,A) \\
=& p(X|Z,U,C,A)p(T|Z,U,C,A)p(Y|T,Z,U,C,A)
\end{aligned} \tag{3}
$$

For generative process depicted in Figure 1, based on the d-separation Pearl et al. (2000), $X \perp\!\!\!\perp (Z,U)$, $T \perp\!\!\!\perp A$, $Y \perp\!\!\!\perp Z \mid T,U,C,A$. Then we have:

$$
\begin{aligned}
& p(X,T,Y|Z,U,C,A) \\
=& p(X|C,A)p(T|Z,U,C)p(Y|T,U,C,A)
\end{aligned} \tag{4}
$$

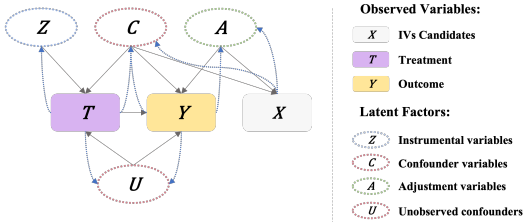

Figure 1: Bayesian network structure corresponds to the causal graph. Black arrows with solid lines denote the generative process, and blue arrows with dashed lines indicate the inference process. We study the more general scenario in real life, while IV candidates X is a view of C and A without Z.

Due to the exogeneity of the latent factors Z, U, C, A, we can factorize $q(Z, U, C, A|X, T, Y)$ as follows:

$$\begin{aligned} &q(Z, U, C, A|X, T, Y) \\ =&q(Z|X, T, Y)q(U|X, T, Y)q(C|X, T, Y)q(A|X, T, Y) \end{aligned} \tag{5}$$

Similar with Eq (4), for inference process depicted in Figure 1, we have $Z \perp\!\!\!\perp X$, $Z \perp\!\!\!\perp Y \mid T$, $U \perp\!\!\!\perp X$, $A \perp\!\!\!\perp T$, then:

$$\begin{aligned} &q(Z, U, C, A|X, T, Y) \\ =&q(Z|T)q(U|T, Y)q(C|X, T, Y)q(A|X, Y) \end{aligned} \tag{6}$$

Combing Eq (2), (4) and (6), we have:

$$\begin{aligned} &logp_\theta(X, T, Y) \\ &\geq \mathbb{E}_{q_\phi(C|X,T,Y)q_\phi(A|X,Y)}[logp_\theta(X|C, A)] \\ &+ \mathbb{E}_{q_\phi(Z|T)q_\phi(U|T,Y)q_\phi(C|X,T,Y)}[logp_\theta(T|Z, C, U)] \\ &+ \mathbb{E}_{p(T)q_\phi(A|X,Y)q_\phi(U|T,Y)q_\phi(C|X,T,Y)} \\ &[logp_\theta(Y|A, C, U, T)] + KL(q_\phi(Z|T)||p(Z)) \\ &+ KL(q_\phi(U|T, Y)||p(U)) + KL(q_\phi(A|X, Y)||p(A)) \\ &+ KL(q_\phi(C|X, T, Y)||p(C)) \\ &\equiv ELBO_{VIV} \end{aligned} \tag{7}$$

□

Mathematical Foundation of Adversarial Learning

Our designed adversarial training helps to reduce the **Total Relation** Watanabe (1960) between different hidden factors. We provide the related proof as follows.

Proof. Based of the defination of Total Relation Loss Watanabe (1960), we have:

$$\begin{aligned} \mathcal{L}_{TC} =&KL(q(Z, U, C, A)||q(Z)q(U)q(C)q(A)) \\ =&\mathbb{E}_{q(Z,U,C,A)}[log\frac{q(Z, U, C, A)}{q(Z)q(U)q(C)q(A)}] \end{aligned} \tag{8}$$

Suppose in the adversarial training process, there are N data points sampled from $q(Z)q(U)q(C)q(A)$ with label Y=1 and N data points sampled from $q(Z, U, C, A)$ with label Y=0 in each batch. Then, we rewrite $q(Z, U, C, A)$ as $p(Z, U, C, A|Y = 0)$ and $q(Z)q(U)q(C)q(A)$ as $p(Z, U, C, A|Y = 1)$. We have:

$$\begin{aligned} \mathcal{L}_{TC} =&\mathbb{E}_{q(Z,U,C,A)}[log\frac{p(Z, U, C, A|Y = 0)}{p(Z, U, C, A|Y = 1)}] \\ =&\mathbb{E}_{q(Z,U,C,A)}[log\frac{p(Y = 0|Z, U, C, A)}{p(Y = 0)} \\ &\frac{p(Y = 1)}{p(Y = 1|Z, U, C, A)}] \end{aligned} \tag{9}$$

As the amounts for real samples and permuted samples are same for each batch, we have $p(Y = 1) = p(Y = 0)$, then:

$$\begin{aligned} \mathcal{L}_{TC} =&\mathbb{E}_{q(Z,U,C,A)}[log\frac{p(Z, U, C, A|Y = 0)}{p(Z, U, C, A|Y = 1)}] \\ =&\mathbb{E}_{q(Z,U,C,A)}[log\frac{p(Y = 0|Z, U, C, A)}{p(Y = 1|Z, U, C, A)}] \\ =&\mathbb{E}_{q(Z,U,C,A)}[log\frac{p(Y = 0|Z, U, C, A)}{1 - p(Y = 0|Z, U, C, A)}] \\ \approx&\mathbb{E}_{q(Z,U,C,A)}[log\frac{1 - D_\psi(Z, U, C, A)}{D_\psi(Z, U, C, A)}] \end{aligned} \tag{10}$$

Eq (10) is consistent with the objective function of adversarial training. □

Pseudo Codes

Algorithm 1 summarizes our proposed method.

Algorithm 1: Distribution-Conditioned Adversarial Variational Autoencoder for Instruments Generation

Input: Observed IVs Candidates X, Outcome Y, Treatment T, Batch Size B, Iteration n, Latent Factor Dimension m, Inference Network q_ϕ, Generative Network p_θ, Discriminator D_ψ.
Output: Generated Instruments Z
Initialize the parameters of q_ϕ, p_θ and D_ψ.
while objective is not converged **do**
  for $i = 1$ **to** n **do**
    Sample X^{bs}, T^{bs}, Y^{bs} from X, T, Y, respectively
    Sample Z^{bs}, U^{bs}, C^{bs}, A^{bs} from $q_\phi(Z, U, C, A|X^{bs}, T^{bs}, Y^{bs})$, respectively
    Update the parameters of q_ϕ, p_θ by minimizing $\mathcal{L}_{VIV}$
    for L in $[Z^{bs}, U^{bs}, C^{bs}, A^{bs}]$ do **do**
      $\pi \leftarrow$ random permutation on $\{1, ..., m\}$
      $(L^{(j)}_{perm})^m_j \leftarrow (L^{\pi(j)}_{perm})^m_j$
    end for
    Update the parameters of D_ψ by maxmizing $\mathcal{O}_{GAN}$
  end for
end while

Additional Experiments

Experiments on Demands-0.9 Dataset

For Demands dataset, we increase the value of ρ from 0.5 to 0.9 to amplify the confounding bias stemming from un-

In-Sample	Poly2SLS	NN2SLS	KernelIV	DualIV	DeepIV	OneSIV	DFIV	DeepGMM	AGMM
NoneIV	>1000	1.25(0.09)	0.59(0.06)	Nan	0.21(0.03)	0.32(0.04)	0.60(0.10)	1.27(0.06)	1.24(0.13)
UAS	>1000	1.21(0.13)	0.60(0.07)	6.29(2.84)	0.21(0.01)	**0.29(0.04)**	0.54(0.13)	1.27(0.05)	0.98(0.24)
WAS	>1000	1.18(0.10)	0.61(0.07)	4.56(1.29)	0.21(0.02)	0.30(0.03)	0.63(0.14)	1.28(0.07)	1.18(0.14)
ModeIV	>1000	1.21(0.13)	0.60(0.07)	6.29(2.84)	**0.20(0.03)**	0.29(0.04)	0.54(0.13)	1.27(0.05)	0.98(0.24)
AutoIV	6.46(4.02)	1.10(0.10)	0.60(0.07)	**3.46(1.79)**	0.21(0.02)	0.29(0.05)	0.70(0.14)	1.27(0.06)	1.19(0.11)
GIV-EM	2.91(1.78)	1.23(0.35)	0.61(0.06)	7.00(2.62)	0.21(0.03)	0.41(0.07)	0.47(0.19)	1.11(0.08)	1.06(0.10)
VIV	**0.72(0.08)**	**0.62(0.06)**	**0.55(0.07)**	**2.68(0.82)**	0.25(0.02)	0.42(0.02)	**0.33(0.03)**	**0.71(0.04)**	**0.57(0.05)**
TrueIV	**0.18(0.02)**	**0.10(0.01)**	**0.20(0.04)**	4.66(1.62)	**0.07(0.01)**	**0.17(0.02)**	**0.08(0.01)**	**0.25(0.02)**	**0.15(0.02)**
Out-Sample	Poly2SLS	NN2SLS	KernelIV	DualIV	DeepIV	OneSIV	DFIV	DeepGMM	AGMM
NoneIV	>1000	1.24(0.08)	0.59(0.06)	Nan	0.21(0.03)	0.32(0.04)	0.60(0.10)	1.26(0.06)	1.23(0.13)
UAS	>1000	1.21(0.13)	0.60(0.07)	6.29(2.83)	0.21(0.01)	**0.29(0.04)**	0.53(0.14)	1.26(0.05)	0.98(0.23)
WAS	>1000	1.18(0.10)	0.60(0.06)	4.62(1.29)	0.21(0.02)	0.30(0.03)	0.63(0.14)	1.27(0.06)	1.18(0.14)
ModeIV	>1000	1.21(0.13)	0.60(0.07)	6.29(2.83)	**0.20(0.03)**	0.29(0.04)	0.53(0.14)	1.26(0.05)	0.98(0.23)
AutoIV	6.58(4.22)	1.10(0.09)	0.59(0.06)	**3.44(1.71)**	0.21(0.02)	0.29(0.05)	0.69(0.14)	1.26(0.05)	1.19(0.11)
GIV-EM	2.89(1.77)	1.24(0.35)	0.61(0.06)	7.10(2.56)	0.21(0.03)	0.41(0.07)	0.47(0.20)	1.10(0.08)	1.06(0.12)
VIV	**0.71(0.08)**	**0.62(0.07)**	**0.55(0.07)**	**2.71(0.84)**	0.25(0.02)	0.42(0.03)	**0.33(0.03)**	**0.71(0.04)**	**0.56(0.05)**
TrueIV	**0.18(0.02)**	**0.10(0.01)**	**0.20(0.04)**	4.59(1.52)	**0.07(0.01)**	**0.17(0.02)**	**0.08(0.02)**	**0.24(0.02)**	**0.15(0.01)**

Table 1: Performance comparison of MSE of the counterfactual prediction on do(T) outcomes between VIV and the SOTA baselines on the Demands-0.9 datasets. Bold indicates the method with the best and second-best performance.

Loss	In-Sample				Out-Sample			
	Poly2SLS	KernelIV	DeepIV	Average	Poly2SLS	KernelIV	DeepIV	Average
GAN	0.32(0.09)	0.36(0.24)	0.35(0.05)	0.34(0.02)	4.25(0.76)	4.80(1.64)	0.90(0.10)	3.32(2.11)
VAE	0.23(0.06)	0.34(0.15)	**0.33(0.03)**	0.30(0.06)	2.03(1.89)	0.72(0.18)	0.81(0.08)	1.19(0.73)
Total	**0.22(0.03)**	**0.26(0.12)**	0.33(0.04)	**0.27(0.06)**	**0.47(0.13)**	**0.62(0.18)**	**0.72(0.08)**	**0.60(0.13)**

Table 2: Ablation Study.

observed confounders. This allows us to evaluate the performance of instrumental variable generation methods in a more challenging scenario. As can be seen from Figure 1, compared to the experimental results on the Demands-0.5 dataset, all instrumental variable generation methods exhibit poorer performance on downstream counterfactual prediction tasks in the Demands-0.9 dataset. This indicates that unobserved confounding bias indeed poses significant challenges for counterfactual prediction tasks. Nevertheless, our algorithm continues to outperform the other IV-generation methods under this scenario.

Experiments on Twins Dataset in Out-Sample Setting

The experimental results on out-sample setting in Twins dataset are presented in Table 3, which are consistent with the findings in the main paper.

Counterfactual Prediction Visualization

We present the graphs depicting the estimated values of the effect function using the intervention T=do(t) on all SOTA IV-based counterfactual prediction backbones mentioned in the main paper. We arrange the outcomes based on the actual observed outcomes (Ground Truth, GT) on Demands-0.5 dataset for both in-sample and out-sample settings. The results shown in Figure 3 are consistent with the findings in the main paper.

Ablation Study

We perform the ablation experiments on Twins dataset to examine the contributions of the VAE network and the GAN network in total loss function on final inference performance. We select three backbones to display the related results as presented in Table 2. We have the following findings: 1) Only retaining the adversarial training module (GAN) in the framework will impair the counterfactual prediction performance of the model, which demonstrates that achieving a low Total Correlation (TC) is only valuable when we are able to preserve the information related to the underlying factors, as emphasized by Kim and Mnih (2018); 2) By exclusively utilizing variational autoencoder, we have managed to surpass the performance of all comparative methods. This un-

Out-Sample	Poly2SLS	NN2SLS	KernelIV	DeepIV	OneSIV	DeepGMM	AGMM	Average
NoneIV	>100	1.15(0.22)	1.08(0.22)	0.85(0.09)	0.78(0.06)	1.21(0.21)	1.12(0.16)	79.1(207)
UAS	3.46(2.80)	1.04(0.22)	1.03(0.19)	0.83(0.12)	0.70(0.12)	1.18(0.18)	0.92(0.12)	1.31(0.96)
WAS	5.52(6.95)	1.00(0.18)	1.02(0.20)	0.85(0.10)	0.64(0.10)	1.26(0.20)	0.89(0.15)	1.60(1.74)
ModeIV	3.46(2.80)	1.04(0.22)	1.03(0.19)	0.82(0.08)	0.70(0.12)	**1.18(0.18)**	0.92(0.12)	1.31(0.96)
AutoIV	19.1(38.0)	1.06(0.26)	1.05(0.19)	0.76(0.12)	**0.55(0.06)**	1.21(0.25)	1.04(0.25)	3.54(6.87)
GIV-EM	7.52(13.1)	**0.97(0.22)**	1.04(0.20)	0.81(0.16)	0.76(0.06)	1.21(0.23)	1.37(0.24)	1.95(2.46)
VIV	**0.47(0.13)**	1.33(0.21)	**0.62(0.18)**	**0.72(0.08)**	0.90(0.09)	1.65(0.33)	**0.68(0.09)**	**0.91(0.43)**
TrueIV	**0.52(0.38)**	**0.91(0.25)**	**0.53(0.15)**	**0.66(0.10)**	**0.59(0.05)**	**0.50(0.10)**	**0.12(0.02)**	**0.55(0.24)**

Table 3: Performance comparison of MSE of the counterfactual prediction on do(T) outcomes between VIV and the SOTA baselines on out-sample setting of the Twins dataset. Bold indicates the method with the best and second-best performance.

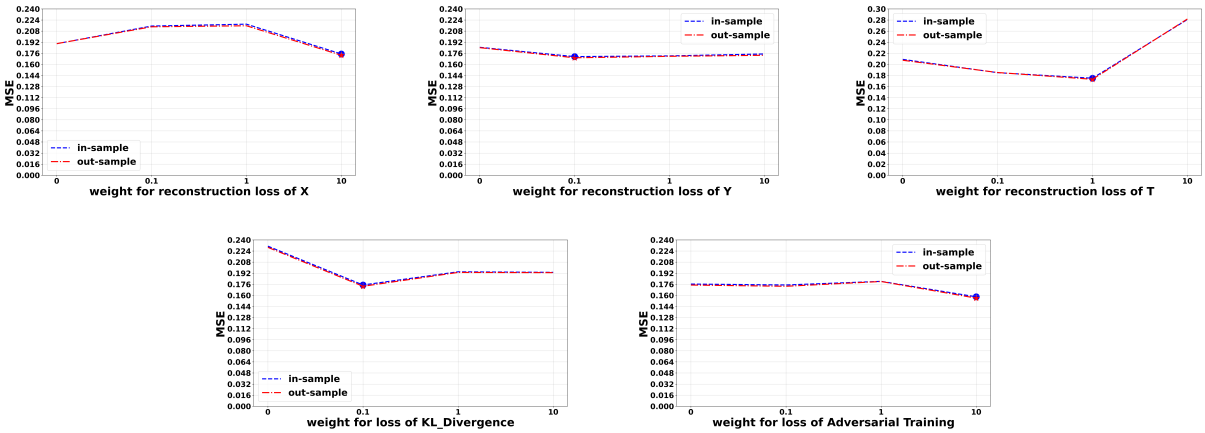

Figure 2: Hyper-parameters sensitivity analysis on Demands-0.5 dataset. The blue and red lines show the MSE results of these parameters in in-sample and out-sample settings, respectively. The blue circle and red star represent the best parameters for the setting.

derscores the soundness of our constructed causal strcture and the effectiveness of the variational network derived from the causal structure in generating valid instrumental variables; 3) Building upon variational autoencoder, the incorporation of GAN further enhances the performance of the instrumental variables we generate in downstream counterfactual prediction tasks. This illustrates how adversarial learning aids in the separation of the latent factors.

Hyper-Parameters Analysis

We select DFIV Xu et al. (2020) as the backbone method to showcase effects of hyper-parameter changes considering its relatively shorter run-time. Specifically, we change the weight for reconstruction loss of X, T and Y, KL-divergence loss, adversarial learning loss in the scope $\{0, 0.1, 1, 10\}$. When analysing a specific parameter, we keep other parameters in the same setting. The related experiments are conducted on the Demands-0.5 dataset.

As can be seen from Figure 2, for every crucial parameter, even when we subject them to substantial fluctuations within a broader range, the performance of VIV remains largely unaffected. Across each parameter configuration, VIV consistently outperforms all baselines whose results have been reported in the main paper, demonstrating the robustness of VIV.

Hardware Setting

In this work, we perform all experiments on a cluster with two 12-core Intel Xeon E5-2697 v2 CPUs and a total 768 GiB Memory RAM.

Optimal Hyper-Parameters

We report the optimal hyper-parameters of VIV for Demands-0.5, Demands-0.9 and Twins datasets in Table 4.

References

Kim, H.; and Mnih, A. 2018. Disentangling by Factorising. In *International Conference on Machine Learning*.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19: 2.

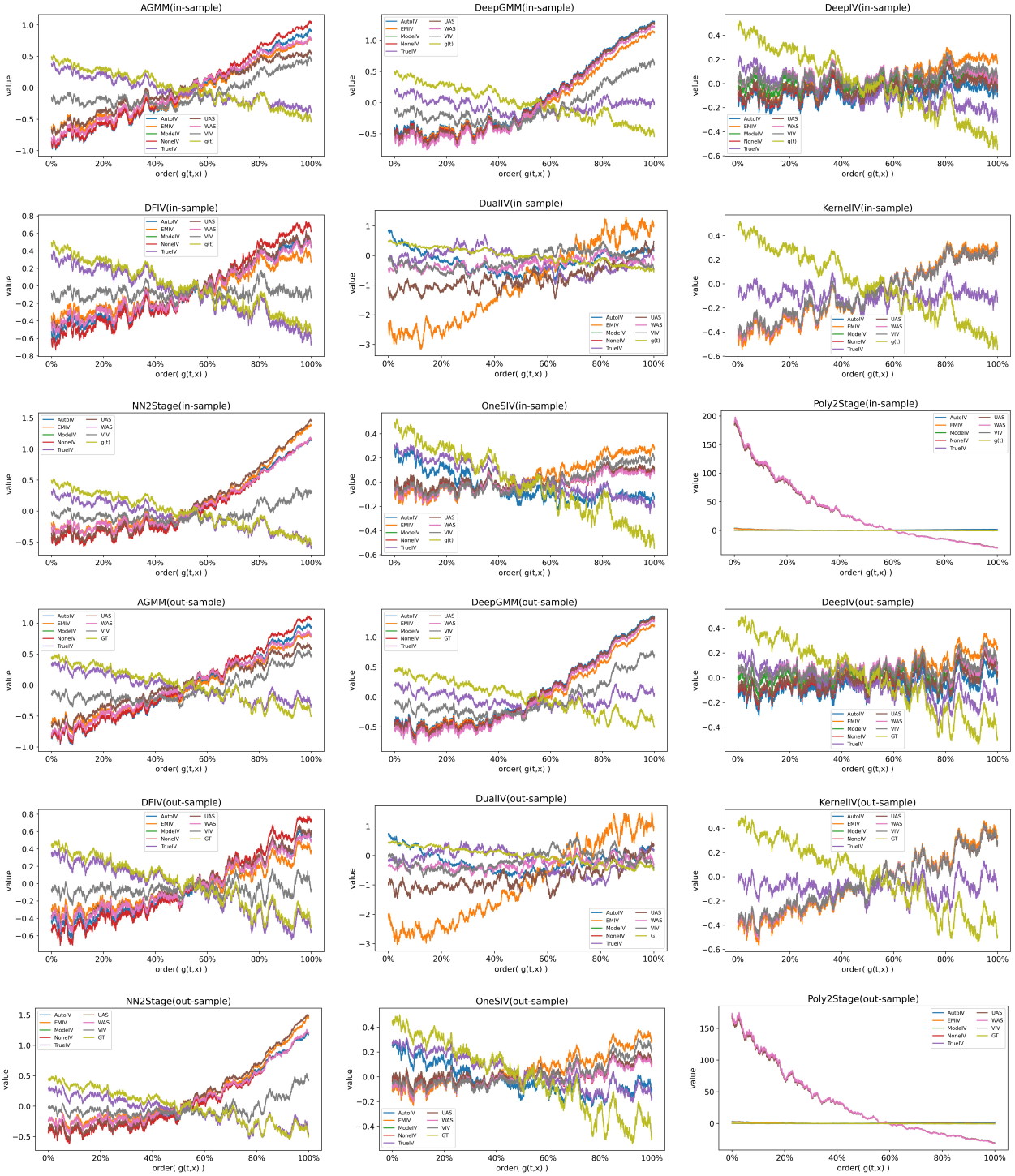

Figure 3: Counterfactual prediction curves with different IV-based generation methods. Compared to other IV-generation methods, the counterfactual prediction curve based on VIV closely approximates the ground truth curve and the counterfactual prediction curve based on TrueIV.

Watanabe, S. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1): 66–82.

Xu, L.; Chen, Y.; Srinivasan, S.; de Freitas, N.; Doucet, A.; and Gretton, A. 2020. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*.

Hyper-parameters	**Demands-0.5**	**Demands-0.9**	**Twins**
Learning rate for max-stage	1e-4	1e-4	1e-4
Learning rate for min-stage	1e-3	1e-3	1e-3
weight for reconstruction loss of X	10	10	1
weight for reconstruction loss of Y	10	10	0.1
weight for reconstruction loss of T	1	1	10
weight for KL-divergence loss	0.1	0.1	0.1
weight for adversarial training loss	0.1	0.1	10
Depth of inference network	2	2	2
Depth of generative network	2	2	2
Depth of discriminator	2	2	2
Dim of inference network	$(128)_2$	$(128)_2$	$(128)_2$
Dim of generative network	$(128)_2$	$(128)_2$	$(128)_2$
Dim of discriminator	$(128)_2$	$(128)_2$	$(128)_2$
Dim of latent factors	2	2	2
Epoch	2	2	10
Batch size	256	256	256

Table 4: Optimal hyper-parameters