A Latent Confounders and Proxy

It is infeasible to estimate causal effects without any observed confounding information [12, 2]. With unobserved confounders, a common approach is to introduce additional proxy variables [11, 1, 15]. For example, the socio-economic status of a patient is a confounder to the medication and outcome, which is unobserved and hard to measure directly. But we can use known auxiliary information such as the zip code and job type as a proxy to assess the confounder indirectly [8]. Note that directly treating the proxy variables as ordinary confounders can induce bias [13, 4, 3]. Instead, recent work [8, 16, 9, 10, 14] uses latent variable approaches. For example, Louizos et al. [8] use VAEs to infer latent confounders from proxy information. Our work develops a latent variable causal framework for controllable text generation. Due to the rich and abstract information in text, we introduce necessary training objectives to avoid training collapse and encourage confounder balance.

B Experimental Details

Model configuration The VAE backbone of our model largely follows the architecture of the VAE model in [6], except that our inference network (encoder) q_ϕ adopts the pretrained GPT-2 model (same as the decoder $p_\theta(\boldsymbol{x}|a, \boldsymbol{z})$) instead of BERT, in order to make sure both encoder q_ϕ and decoder $p_\theta(\boldsymbol{x}|a, \boldsymbol{z})$ have the same tokenization. We implement a to be either an all-zero or all-one vector of dimension 50, and set the dimension of $\boldsymbol{z}$ to 718, so that concatenating a and $\boldsymbol{z}$ leads to a vector of dimension 768, same as in [6]. We implement $p_\theta(\boldsymbol{c}|\boldsymbol{z})$ as a single-layer MLP and $p_\theta(a|\boldsymbol{z})$ as a two-layer MLP with the intermediate dimension the same as the input dimension (768).

More details of BIOS dataset For occupation which is the confounding factor, we subsample and merge the occupations into two groups, i.e., *{nurse, dietitian, paralegal, model, yoga teacher}*, and *{rapper, DJ, surgeon, software engineer, composer}*, which results in a correlation strength of 94% between occupation and gender.

Training confounding label classifier with data re-weighting For the `Conditional LM (full)` baseline for attribute-conditional generation, we first train a confounding label classifier with the limited confounding (proxy) labels in the dataset. Due to the strong correlation between the confounding factor and attribute (e.g., with a correlation strength of 90%), only a small fraction of (e.g., 10%) instances have opposite confounding label and attribute label. We thus train the classifier with data reweighting to reduce the bias. Specifically, we associate a weight of 0.9 to those instances with opposite confounding and attribute labels, and a weight of 0.1 to other instances with the same confounding and attribute labels. We tried other weights and obtained lower or similar classifier accuracy.

Human evaluation of text attribute transfer Following previous work [e.g., 7], we conduct comparison-based human evaluation for the output of different generation models. Specifically, for each test instance, we present the outputs of two comparison models to the human rater, and ask the human rater to rank which of the two outputs are better in terms of the goal of the task (i.e., accurately rewriting the text to possess desired attribute and meanwhile preserving all other characteristics of the original sentence). The human rater can also choose "no preference" if the two outputs are equally good or bad. We asked three human annotators (who are graduate students and proficient English speakers) to do the rating [7]. There were no potential participant risks. We evaluate on 50 test cases for each pair of comparison models. Table 1 shows the results, which are consistent with the observations from automatic evaluation.

	Ours better (%)	No preference (%)	Ours worse (%)
Hu et al. [5]	**62**	24	14
Ablation: Ours w/o cf-z/c	**54**	22	24

Table 1: Human evaluation of text attribute transfer on biased YELP. For example, the outputs of OURS are considered to be better than those of Hu et al. [5] on 62% test instances.

Classifiers used in training and evaluation We summarize the different classifiers used in training and evaluation in the experiment to serve as a reference and avoid confusion.

- For training:
  - *attribute classifier* f (Eq.4) is used to train our causal model. The classifier is pretrained with the biased (attribute, text) training corpus.
  - *confounding label classifier* is used in the baseline `Conditional LM (full)` for attribute-conditional generation. The classifier is trained with the available confounding (proxy) labels with data re-weighting, as discussed above.
- For evaluation:
  - *evaluation attribute classifier* is used to evaluate the generation accuracy of desired attribute. As an evaluation metric, the classifier is obtained by training on additional *unbiased* (attribute, text) data.
  - *evaluation confounding label classifier* is used to evaluate the correlation of attribute and confounding (proxy) labels in the generation. Similarly, the classifier is obtained by training on additional large unbiased data.

C Generated Samples

CONDITIONAL LM (FULL)
$a = 0$ **(sentiment negative)**
`this was the worst experience i 've ever had at a glazier .`
`i even asked him if they could play on the tv channel .`
`this was pretty fun the first time i went . "`
`waited in line once but almost never reached the floor .`
`if you are ever up in chandler , tony will stop by .`
$a = 1$ **(sentiment positive)**
`very good and long wait time .`
`we loved our favorite harrah 's night ! "`
`i would love to try this restaurant again when they open . "`
`this place is great .`
`everything you will find in this restaurant !`
OURS
$a = 0$ **(sentiment negative)**
`no , it 's obvious that they were overcooked .`
`the seats were poorly done and basically sucked up .`
`it was n't enough to ask us if it was okay .`
`very disappointed with my food order yesterday .`
`i declined to replace it tho they were bad .`
$a = 1$ **(sentiment positive)**
`great for a relaxed evening out .`
`i 'm beyond impressed with the passion fruit and unbeatable service .`
`it 's a true pleasure to meet andrew .`
`jacksville became my go-to spot for dessert .`
`thank you for the technique , i am quite impressed .`

Table 2: Attribute-conditional generation trained on YELP dataset. CONDITIONAL LM (FULL) tends to generate non-restaurant reviews conditioning on $a = 0$, and restaurant reviews conditioning on $a = 1$.

OURS	
$a = 0$ **(sentiment negative)** $\rightarrow a = 1$ **(positive)**	
original:	`pick-up was just ok , but vehicle was filthy and had trash in it .`
output:	`pick-up was pretty good , but atmosphere was just incredible and comfortable .`
original:	`so that was nice but they served some sweet concoctions that made me sick .`
output:	`so good that they served some sweet and flavorful cocktails that made me super happy .`
original:	`similar to some of the other reviewers , the poutine was just that .`
output	`similar to some of the other reviewers , the poutine was just perfect .`
$a = 1$ **(sentiment positive)** $\rightarrow a = 0$ **(negative)**	
original:	`the santa fe salad is awesome .`
output:	`the santa fe salad is mediocre .`
original:	`the employees were super helpful , friendly and attentive .`
output:	`the employees were super rude , incompetent and unhelpful .`
original:	`i love their eggs benedict and pancakes both are amazing !`
output:	`i hate their eggs benedict and pancakes both are horrible .`

Table 3: Text attribute transfer on the biased YELP dataset.

References

[1] J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[2] A. D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3478–3486. PMLR, 2019.

[3] W. A. Fuller. *Measurement error models*, volume 305. John Wiley & Sons, 2009.

[4] Z. Griliches and J. A. Hausman. Errors in variables in panel data. *Journal of econometrics*, 31 (1):93–118, 1986.

[5] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. Xing. Toward controlled generation of text. In *International Conference on Machine Learning (ICML)*, 2017.

[6] C. Li, X. Gao, Y. Li, X. Li, B. Peng, Y. Zhang, and J. Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. *arXiv preprint arXiv:2004.04092*, 2020.

[7] S. Lin, W. Wang, Z. Yang, X. Liang, F. F. Xu, E. Xing, and Z. Hu. Record-to-text generation with style imitation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1589–1598, 2020.

[8] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems (NeurIPS)*, pages 6446–6456, 2017.

[9] D. Lu, C. Tao, J. Chen, F. Li, F. Guo, and L. Carin. Reconsidering generative objectives for counterfactual reasoning. In *NeurIPS*, 2020.

[10] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358, 2019.

[11] M. R. Montgomery, M. Gragnolati, K. A. Burke, and E. Paredes. Measuring living standards with proxy variables. *Demography*, 37(2):155–174, 2000.

[12] J. Pearl. *Causality*. Cambridge university press, 2009.

[13] J. Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

[14] R. Ranganath and A. Perotte. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*, 2018.

[15] J. H. Stock, M. W. Watson, et al. *Introduction to econometrics*, volume 3. Pearson New York, 2012.

[16] D. Tran and D. M. Blei. Implicit causal models for genome-wide association studies. In *International Conference on Learning Representations*, 2018.