Beyond What If: Advancing Counterfactual Text Generation with Structural Causal Modeling

Ziao Wang , Xiaofeng Zhang* , Hongwei Du
School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
ziaowang@stu.hit.edu.cn, {zhangxiaofeng, hwdu}@hit.edu.cn

Abstract

Exploring the realms of counterfactuals, this paper introduces a versatile approach in text generation using structural causal models (SCM), broadening the scope beyond traditional singular causal studies to encompass complex, multi-layered relationships. To comprehensively explore these intricate, multi-layered causal relationships in text generation, we introduce a generalized approach based on the structural causal model (SCM), adept at handling complex causal interactions in a spectrum ranging from everyday stories to financial reports.Specifically, our method begins by disentangling each component of the text into pairs of latent variables, representing elements that remain unchanged and those subject to variation. Subsequently, counterfactual interventions are applied to these latent variables, facilitating the generation of outcomes that are influenced by complex causal dynamics. Extensive experiments have been conducted on both a public story generation dataset and a specially constructed dataset in the financial domain. The experimental results demonstrate that our approach achieves state-of-the-art performance across a range of automatic and human evaluation criteria, underscoring its effectiveness and versatility in diverse text generation contexts.

1 Introduction

Causal inference has always been a hot research topic in the field of Natural Language Processing (NLP), achieving significant results [Luo *et al.*, 2016; Feder *et al.*, 2022; Hu and Li, 2021]. In recent years, the study of text generation using counterfactuals has gradually gained traction [Qin *et al.*, 2019; Hao *et al.*, 2021]. This research avenue, anchored in conditional text generation, delves into the exploration of generated text in counterfactual worlds. Here, the text outcomes are examined in light of altered conditions, contrasting them with those under the original, actual-world conditions. One illustrating example of counterfactual story generation is given in Figure 1. Each text segment comprises

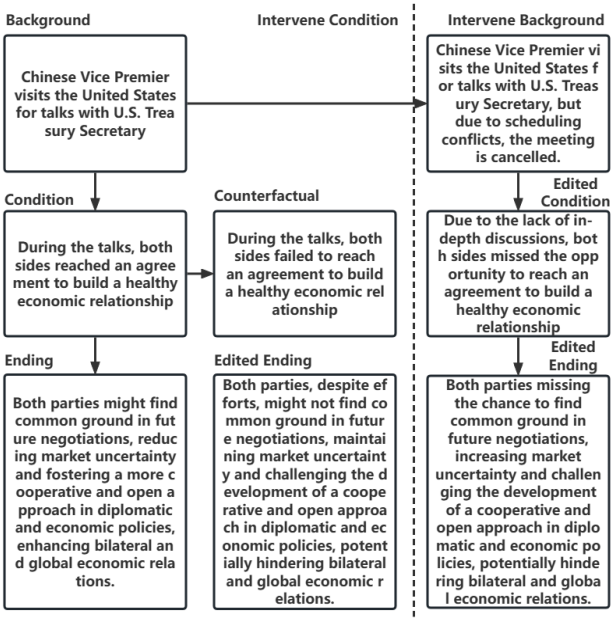


Figure 1: An illustrating example of the counterfactual story generation task.

three parts: background (marked as B), condition (marked as C), and outcome (marked as E). From the original text on the left, we learn that the Vice Premier is visiting the United States (B), with both parties seeking to prevent economic decoupling (C), leading to a meeting outcome where both sides welcome the establishment of a China-US economic working group, indicating a positive result (E). In this context, the counterfactual condition or background is the opposite of the original scenario. For example, the counterfactual condition could change from "reached an agreement" to "fail to reach an agreement", leading to "hindering bilateral and global economic relations", an outcome that completely contradicts the original.

Recent advancements in counterfactual text generation, moving beyond mere data augmentation, have led to innovative developments. For example, Qin et al. [Qin *et al.*, 2019] introduced an open-source dataset that has become essential

*Corresponding author. Email: zhangxiaofeng@hit.edu.cn

for evaluating model performance in this field. Their approach reframed the issue as a sequence-to-sequence generation problem, where background (B), altered condition (counterfactual C), and outcome (E) are concatenated as inputs for the GPT-2 model, generating counterfactual endings. Further, [Qin *et al.*, 2020] explored unsupervised generation of story endings, incorporating both past and future contexts for controlled generation. However, these methods primarily focus on generating endings under given counterfactual conditions, often overlooking the causal impact of the background on the conditions and consequently, the ending.

Addressing this gap, our work aims to master this intricate task by allowing modifications in both background and conditions for hierarchical dependent causal text generation. For instance, altering the original background to "due to scheduling conflicts, the meeting is cancelled", as depicted in the right side of Figure 1, necessitates a change in the counterfactual condition to "missed the opportunity to reach an agreement", leading to a minimally edited ending like "increasing market uncertainty". To tackle this challenge, we propose a novel approach based on structural causal models (SCM) [Pearl *et al.*, 2016]. This involves adapting SCMs to our context and integrating them with a pre-trained model to generate both counterfactual C and E. A unique disentanglement component is designed to unravel the hierarchical causal relationships among B, C, and E, ensuring the separate influence of each variable in the underlying causal model. The disentangled variables in the latent space are then strategically combined to form the decoder's input, producing the counterfactual endings.

The major contributions of this paper are summarized below:

- We evolve the concept of counterfactual text generation into a hierarchical dependent causal text generation problem. This is achieved by enabling counterfactual modifications in both background and condition elements, and by developing an end-to-end approach based on structural causal models (SCM).
- A novel disentanglement component is introduced, which effectively maps multiple hierarchical dependent variables into a latent space. We also adapt and retrain a BART model, integrating it with our SCM component, specifically tailored for our task.
- We conduct comprehensive experiments not only on a public counterfactual story generation dataset but also on a custom-constructed dataset in the financial domain. Our experimental results affirm the superiority of our approach over existing methods, as evidenced by various automatic and human evaluation criteria.

2 Related Work

The task of counterfactual text generation, a subset of conditional text generation, has been explored in various contexts including context-based [Voita *et al.*, 2018], personalized [Lu *et al.*, 2017], and topic-based text generation [Yang *et al.*, 2021]. Despite the advancements, most approaches overlook causal reasoning, crucial for varying conditions [Roese, 1997; Schölkopf *et al.*, 2021; Fern and Pope, 2021; Wang *et al.*, 2019; Qin *et al.*, 2019; Hao *et al.*, 2021; Chen *et al.*, 2022; Zellers *et al.*, 2019]. Incorporating Structural Causal Models (SCM) [Pearl *et al.*, 2016], researchers have enhanced text generation with causal reasoning abilities. Examples include augmenting data for neural machine translation [Liu *et al.*, 2021], controllable text generation [Hu and Li, 2021], and counterfactual text optimization [Fern and Pope, 2021]. Various counterfactual generation methods have been proposed, employing standard sequence-to-sequence models and GPT-2 [Qin *et al.*, 2019; Hao *et al.*, 2021; Chen *et al.*, 2022], but they heavily rely on the generation capabilities of pre-trained models, which can misidentify counterfactual content [Zellers *et al.*, 2019]. Disentangled representation learning, crucial for interpretable and independent data variation, has been applied across sequential data analysis [Yamada *et al.*, 2019], weakly-supervised learning [Zhu *et al.*, 2023], high-fidelity synthesis [Lee *et al.*, 2020], and various NLP tasks [Vasilakes *et al.*, 2022; Dang-Nhu, 2021; Wang *et al.*, 2019]. Our work aligns with studies like [Ren *et al.*, 2022; Khrulkov *et al.*, 2021] that focus on modeling variations and extracting disentangled representations, respectively. Our work, therefore, lies at the convergence of conditional text generation, disentangled representation learning, and structural causal models. It pushes the boundaries of these fields to address the complex challenge of hierarchical dependent causal text generation.

3 Preliminaries and Task Setup

We consider a textual segment $T = \{B, C, E\}$, where B is the background context, C is a specific condition, and E is the consequent outcome. This structure, suitable for a wide range of domains, allows for the exploration of complex, multi-layered causal relationships. In counterfactual text generation, we aim to modify B or C with counterfactual content, leading to a new outcome E' that should be contextually coherent yet minimally edited. This process unveils the intricate causal layers within the text. The task is bifurcated into two primary sub-tasks:

Task 1. Intervening Background. Given T and a counterfactual background B', the task involves generating a condition C' and an outcome E' that logically follow the new background, creating $T' = \{B', C', E'\}$. This task emphasizes the exploration of causal layers from background to condition and outcome. The generation of E' encompasses multi-layered causal reasoning:

$$P(E') = P(B'|B)P(C'|B')P(E'|B', C')$$

Task 2. Intervening Condition. Given T and a counterfactual condition C', this task focuses on generating a new outcome E' consistent with C'. The updated segment is $T' = \{B, C', E'\}$, reflecting a different causal layer intervention. The process is formulated to capture the cascading effects of changing conditions on the outcome:

$$P(E') = P(C'|C)P(E'|B, C')$$

These tasks underscore our method's capacity to navigate and manipulate the multi-layered causal relationships inherent in text, showcasing its applicability in generating contextually rich and causally coherent counterfactual scenarios across diverse textual domains.

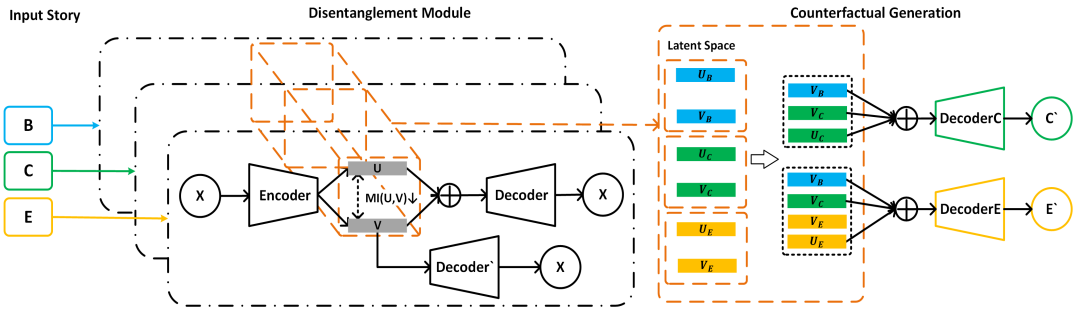


Figure 2: The architecture of the proposed approach.

4 The Proposed Approach

Our proposed approach is designed to generate counterfactual text by intervening at different layers of the background or condition. The approach comprises two primary components: (1) Disentanglement Component: This component disentangles the background, condition, and outcome variables into two types of variables in the latent space: U and V. Here, U symbolizes the unchanged content, aligning with the concept of minimal editing, while V represents the variable content, corresponding to the causal variables. This disentanglement facilitates the exploration and manipulation of multi-layered causal relationships within the text. (2) Counterfactual Text Generation Component: This component employs the Structural Causal Model (SCM) to enhance causal reasoning capabilities. It is seamlessly integrated with a pre-trained model to generate counterfactual text that adheres to the altered background or condition. This integration enables the generation of text that is not only contextually coherent but also reflects the intricate causal dynamics within the text.

The architecture of our approach, which caters to the generation of contextually rich and causally coherent counterfactual scenarios across various textual domains, is depicted in Figure 2.

4.1 Disentanglement Component

We design this component which respectively disentangles the input variables, i.e., B, C, E into each pair of varying variables and unchanged variables. Then the generation is conducted in the latent space with the combination of these latent variables (details will be given in the next subsection). The disentanglement processes are illustrated as follows.

The disentanglement component is built based on the encoder-decoder architecture, as shown in Figure 2. Let's take B as the input X, and the component is to disentangle X into variable U and V residing in the latent space. Let $q_\phi(U,V|X)$ and $P_\omega(X|V)$ respectively denote the encoder and decoder, variable V is then fed into a decoder D_V to reconstruct X,

$$\begin{aligned} L_{rV\to X} &= -E_{q_\phi(U,V|X)}[logP_\omega(X|V)] \\ &+ KL(q_\phi(U,V|X)||p(V))]. \end{aligned} \tag{1}$$

And another decoder D_{UV}, denoted as $P_\epsilon(X|U,V)$, takes variable U and V as the input to reconstruct X, the reconstruction loss of D_{UV} is written as,

$$\begin{aligned} L_{rUV\to X} &= -E_{q_\phi(U,V|X)}[logP_\epsilon(X|U,V)] \\ &+ KL(q_\phi(U,V|X)||p(U,V))]. \end{aligned} \tag{2}$$

By minimizing Eq. 1, most of existing content of E could be preserved. If further minimize Eq. 2, U will be minimized as V is simultaneously trained to maximally generate X. Thus, the variance of U is under a well control. To guarantee the independence between U and V, we adopt mutual information to measure the statistical dependence between two variables, its general form could be written as

$$I(U,V) = H(U) - H(U|V)$$

where H is the Shannon entropy. However, this equation is intractable and thus a lower bound on $I(U,V)$ is desired to estimate. By considering the KL divergence as its distance, we can derive its supremum according to [Belghazi *et al.*, 2018], given as

$$\begin{aligned} I(U,V) &= D_{KL}(P(U,V)\|P(U)\otimes P(V)) \\ &\geq \sup_{\theta\in\Theta} E_{P_{UV}}[T_\theta] - \log(E_{P_U\otimes P_V}[e^{T_\theta}]), \end{aligned}$$

where $T_\theta : \mathcal{U}\times\mathcal{V}\to R$ is a transformation function parameterized by a neural network. The mutual information loss is then estimated as

$$L_{U\leftrightarrow V} = E_{P_{UV}}[T_\theta] - \log(E_{P_U\otimes P_V}[e^{T_\theta}]).$$

Therefore, the overall loss of this disentanglement component could be written as

$$L_{dis} = \alpha(L_{rV\to X} + L_{rUV\to X}) + (1-\alpha)L_{U\leftrightarrow V}, \tag{3}$$

where α is the coefficient. Note that this component will be executed three times respectively for B, C and E as the input.

4.2 Counterfactual Text Generation Component

To generate the counterfactual text, a structural causal model (SCM) is adapted in this section. First, we briefly review the conventional SCM as follows. It consists of two sets of variables: Endogenous variable $En = \{En_1, En_2, ..., En_i\}$ and Exogenous variable $Ex = \{Ex_1, Ex_2, ..., Ex_i\}$, and a

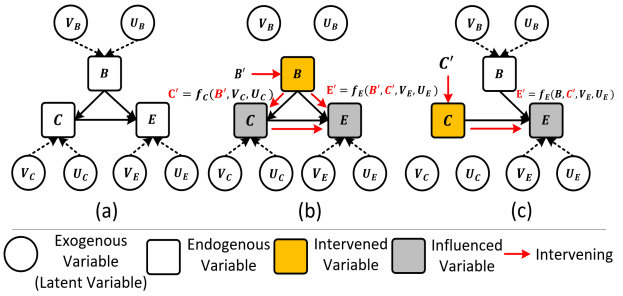


Figure 3: The causal inference process of our proposed counterfactual text generation component. (a) The original SCM. (b) Intervening the background B (using B') which causes the change of C, and thus E is affected by B' and C', written as $f_E(do(b = B'), C', V_E, U_E)$. (c) Intervening the condition C to C', apparently B will not affect C and thus E is only affected by $f_E(B, do(C = C'), V_E, U_E)$.

set of functions $f = \{f_X : W_X \rightarrow X | X \in En\}, W_X \subseteq (En \cup Ex) - \{X\}$ that assign values to endogenous variable in En, such that

$$En_i = f_{En_i}(Ex_i, En_j), \quad j \neq i \tag{4}$$

The exogenous variables are the direct cause of the endogenous variables [Pearl *et al.*, 2016]. To perform interventions on the endogenous variables E_n, the corresponding counterfactual changes could be described as,

$$En_i' = f_{En_i}(do(Ex_i = x_i'), En_j), \quad j \neq i \tag{5}$$

where *do-operator* assigns an intervention value to the variable. As $f_{En_i'}$ is not an observed value, according to [Pearl *et al.*, 2016], it is assumed that Based on the fundamental law of counterfactuals [Pearl *et al.*, 2016] that future eventualities does not alter the past, we have,

$$f_{En_i} = f_{En_i'}. \tag{6}$$

This formulation facilitates our counterfactual reasoning process as we do not need to learn new counterfactual functions for every interventions.

Our proposed causal graph is depicted in Figure 3(a), the corresponding causal relations among these variables are formulated as

$$\begin{aligned} En &= \{B, C, E\} \\ Ex &= \{U_B, V_B, U_C, V_C, U_E, V_E\} \\ f_c : C &= f_c(B, V_C, U_C) \qquad (7) \\ f_e : E &= f_e(B, C, V_E, U_E). \qquad (8) \end{aligned}$$

The causal inference process of our proposed two sub tasks are illustrated as follows.

Causal Inference by Intervening Background. The causal inference process is plotted in Figure 3(b). As shown in the red arrow, after intervening B to B' by performing *do-operator* as $do(B = B')$, the condition C is counterfactually changed to C' via a causal function f_c, and thus the story endings E' is causally dependent on B' and C'. The corresponding causal functions are directly given as

$$\begin{aligned} f_c : C' &= f_c(do(B = B'), V_C, U_C). \qquad (9) \\ f_e : E' &= f_e(do(B = B'), C', V_E, U_E). \qquad (10) \end{aligned}$$

Causal Inference by Intervening Condition. The causal inference process by intervening condition C is plotted using red arrow in Figure 3(c). It is obvious that the intervened C' is independent of B, and thus E' relies on C' and B, estimated as

$$f_e : E' = f_e(B, do(C = C'), V_E, U_E). \tag{11}$$

Learning Causal Functions. We approximate the causal functions f_c and f_e using two DNNs, two decoders are employed, respectively denoted as D_C and D_E.

Our counterfactual text generation task is essentially to answer the question like "Given the whole story, what is value of E if B had been a counterfactual B' or C had been a counterfactual C'?". The steps to answer the question is given as follows.

- Abduction: Training the network depicted in Figure 2 to acquire the latent representations of the variables as well as the causal functions f approximated by the decoder.
- Action: Intervening B or C to generate the counterfactual embedding $V_{B'}$ or $V_{C'}$ via the disentanglement component.
- Prediction: Using the learnt causal function f and exogenous variables shown in Eq. 10 and 11 to infer the counterfactual content.

Model Training. During the model training phase, counterfactual content $B^{'}$ and $C^{'}$ are treated as known variables since the dataset has provided $C^{'}$ and $B^{'}$ is acquired through applying method from [Wu *et al.*, 2021] where they designed certain control codes like "negation" or "delete" for the counterfactual generation and used a fill-in-the-blank structure to specify the position where the perturbation occurs. And thus the $V_{B'}$ is obtained by inputting $B^{'}$ to the disentanglement module.

Decoder D_C is trained to approximate the causal function f_c defined in Eq. 7, calculated as

$$P(C|B, V_C, U_C) = \prod_t^T P(c_t|V_B, V_C, U_C, c_{<t}). \tag{12}$$

Decoder D_E is trained to approximate the causal function f_e defined in Eq. 8, calculated as

$$P(E|B, C, V_E, U_E) = \prod_t^T P(e_t|V_B, V_C, V_E, U_E, e_{<t}). \tag{13}$$

Without loss of generality, the cross-entropy loss is adopted as the model loss, formulated as

$$L_{cf} = -log(P(C|B, V_C, U_C) + P(E|B, C, V_E, U_E)).$$

5 Experiments

In this section, we present a comprehensive overview of our experimental approach. We begin by introducing the datasets used in our experiments, including the newly constructed counterfactual financial text generation dataset. Following this, we outline the evaluation criteria and the models used for comparison purposes. Extensive experiments are performed to answer following research questions:

- RQ 1: Whether the proposed approach is superior to the state-of-the-art counterfactual story generation approaches or not?
- RQ 2: Whether the proposed component, e.g., disentanglement component, works or not? (ablation study).
- RQ 3: What is the quality of the generated counterfactual story (case study)?

5.1 Dataset Construction

The dataset was constructed to validate our model's versatility across various domains; detailed construction details are provided in the appendix A due to page limit.

5.2 Experimental Datasets and Parameter Settings

We use publicly available counterfactual story generation dataset [Qin *et al.*, 2019] and our constructed counterfactual financial text generation dataset to valid our proposed method. The detail of the story dataset and the parameter settings are provided in the appendix B due to page limit.

5.3 Evaluation Criteria

We adopt the automatic and human evaluation criteria to evaluate the model performance of the proposed approach as well as the compared models described as follows.

Automatic Evaluation. A good number of automatic evaluation criteria have been adopted in the experiments. The ROUGE [Lin, 2004] is adopted to evaluate the hit rate, the Word Mover's Similarity (WMS) [Kusner *et al.*, 2015] and the BERTScore [Zhang* *et al.*, 2020] (BERTS) are adopted to evaluate semantic similarities, and we also fine-tune a Bert model on the story dataset which is denoted as 'BERT-FT'. To evaluate the consistency of the generated ending, we adopt a factual consistency metric [Huang *et al.*, 2021] (FactScore) as well as a model-based semantic consistency metric, denoted as 'NSPScore'. To calculate the 'NSPScore' score, we also fine-tune a Bert model for the next sentence prediction (NSP) task using this story dataset.

Human Evaluation. For human evaluation criteria, we randomly select 100 generated samples from testing set and seek 2 groups of annotators where each group contains 3 independent annotators to evaluate the quality of the generated stories. Annotators of group A and group B have similar backgrounds to alleviate human bias. The selected text is anonymized and sent to each group for human evaluation. The average results of these two groups are reported as our final results. The human evaluation criteria is to measure the consistency and similarity of the generated content to the background, condition and original ending. To this end, the human annotators are required to answer following questions by giving scores from 1 to 3, and the higher the score the better the results. These questions are listed as follows.

- **BACKGROUND (BG)**: To what extent is the generated ending consistent with the background? Whether the generated ending conflicts with the background or not.
- **CONDITION (CF)**: Does the generated ending well reflect the counterfactual condition?
- **PLOT**: Is the generated ending similar to the plot of the original ending?

	ROUGE-L	BERTS	BERTS-FT	NSPSore	WMS	FactScore
GPT2-S+ZS	3.75	47.55	51.06	77.07	20.60	3.71
GPT2-M+ZS	5.43	48.14	51.62	80.97	21.25	4.68
GPT2-XL+ZS	6.19	46.74	50.58	85.20	21.97	5.65
BART-base+ZS	4.88	41.89	40.38	31.89	22.25	0.00
BART-large+ZS	4.86	42.28	41.63	31.08	22.21	0.00
T5-base+ZS	4.41	48.73	52.15	80.93	23.91	0.69
T5-large+ZS	4.37	48.05	53.16	85.68	22.36	3.75
GPT2-S+FT	4.02	52.79	56.21	94.06	25.92	6.71
GPT2-M+FT	5.78	53.31	56.64	95.26	25.94	6.66
BART-base+FT	4.90	46.58	50.06	67.02	25.44	0.76
BART-large+FT	4.01	46.29	49.62	66.75	25.43	0.92
T5-base+FT	4.41	48.73	52.15	70.93	26.91	0.69
T5-large+FT	5.93	51.84	56.04	77.65	26.12	2.03
GPT2-S+FT+CF	5.20	52.62	56.24	95.08	26.98	6.76
GPT2-M+FT+CF	6.23	52.69	56.33	95.49	27.05	6.76
BART-base+FT+CF	8.42	50.83	54.63	92.89	25.43	0.84
BART-large+FT+C	6.11	54.77	58.12	90.10	25.48	1.10
T5-base+FT+CF	6.41	49.73	52.15	90.93	26.91	0.69
T5-large+FT+CF	5.93	51.84	56.04	92.65	26.12	2.03
GPT2-S+RC+CF	10.96	62.48	66.15	96.81	33.96	7.82
GPT2-M+RC+CF	11.06	62.57	66.22	97.67	33.91	8.74
BART-base+RC+CF	8.35	61.42	65.09	96.11	32.41	6.63
BART-large+RC+C	8.28	61.06	64.81	93.94	32.42	7.17
T5-base+RC+CF	9.41	62.15	68.73	95.92	33.91	7.69
T5-large+RC+CF	11.37	63.16	69.05	**97.67**	32.36	7.75
DELOREAN	9.33	61.26	59.39	89.12	33.16	6.78
EDUCAT	9.68	64.13	65.11	92.09	34.58	6.82
Our-base	12.06	68.89	69.07	95.28	34.98	11.44
Our-large	**13.21**	**70.09**	**71.32**	97.33	**35.94**	**12.64**
GPT2-S+SUP	13.36	65.90	72.42	96.46	38.17	15.63
GPT2-M+SUP	13.95	66.25	72.83	96.24	38.14	15.67
BART-base+SUP	13.92	67.51	74.24	97.98	37.27	14.28
BART-large+SUP	14.32	67.58	74.35	98.13	37.26	14.98
T5-base+SUP	15.41	72.15	78.73	97.92	38.91	16.69
T5-large+SUP	16.37	73.16	78.05	98.67	37.36	17.75

Table 1: Automatic evaluation results for the financial dataset.

5.4 Baseline Models

To evaluate the effectiveness of the proposed approach, a number of baseline models [Radford *et al.*, 2019; Lewis *et al.*, 2020; Raffel *et al.*, 2020] and state-of-the-art models [Qin *et al.*, 2020; Chen *et al.*, 2022] are compared. The baseline models are pre-trained models under Zero-shot, Unsupervised and Supervised settings. The details of the comparison models are provided in the appendix C due to page limit.

5.5 Experimental Results

RQ 1: Performance Comparison

Automatic Evaluation Results. The corresponding automatic evaluation results of all approaches are reported in Table 1 and Table 2. From these tables, we have the following

	ROUGE-L	BERTS	BERTS-FT	NSPSore	WMS	FactScore
GPT2-S+ZS	11.90	2.50	9.93	40.66	50.53	0.09
GPT2-M+ZS	16.09	11.24	17.24	57.85	52.77	0.33
GPT2-XL+ZS	17.05	22.67	21.00	65.79	54.08	0.36
BART-base+ZS	0.33	0.00	0.00	0.08	46.20	0.00
BART-large+ZS	0.58	0.00	0.00	0.29	44.75	0.00
T5-base+ZS	16.53	16.44	13.70	48.57	52.76	0.42
T5-large+ZS	11.42	24.49	23.17	97.66	52.77	0.32
GPT2-S+FT	16.23	31.77	28.78	93.79	55.16	0.53
GPT2-M+FT	15.82	33.88	29.88	96.60	55.20	0.46
BART-base+FT	17.28	36.93	33.60	96.79	56.34	0.95
BART-large+FT	18.55	35.95	33.95	96.94	56.38	1.23
T5-base+FT	15.87	28.38	28.22	97.10	54.28	0.42
T5-large+FT	17.15	34.04	33.04	97.72	55.84	0.97
GPT2-S+FT+CF	16.06	31.86	28.67	92.53	55.18	0.45
GPT2-M+FT+CF	15.67	33.74	29.38	95.81	55.15	0.55
BART-base+FT+CF	17.35	37.06	33.91	96.34	56.24	1.13
BART-large+FT+CF	18.67	35.79	34.09	96.22	56.46	1.32
T5-base+FT+CF	13.87	27.13	27.11	96.94	53.22	0.20
T5-large+FT+CF	17.26	33.32	32.49	97.46	55.83	0.87
GPT2-S+RC+CF	40.92	70.54	76.80	90.95	72.38	20.99
GPT2-M+RC+CF	41.15	70.62	75.29	87.67	71.72	18.15
BART-base+RC+CF	42.07	78.79	82.66	94.98	75.71	20.39
BART-large+RC+CF	43.72	80.36	83.70	96.19	75.58	21.28
T5-base+RC+CF	41.89	82.01	82.73	96.28	75.68	22.83
T5-large+RC+CF	43.03	82.72	83.64	96.04	76.65	23.08
DELOREAN	24.78	67.33	66.30	97.98	59.80	8.30
EDUCAT	33.51	75.81	78.37	94.61	58.20	28.22
Our-base	43.58	80.04	85.63	95.82	77.07	29.51
Our-large	**44.11**	**83.44**	**86.49**	**98.38**	**77.50**	**29.98**
GPT2-S+SUP	43.24	82.99	84.70	96.21	76.63	40.78
GPT2-M+SUP	43.86	84.50	85.85	94.68	73.04	40.68
BART-base+SUP	43.83	84.46	86.15	96.61	77.57	43.85
BART-large+SUP	44.78	85.32	86.98	97.14	77.49	43.62
T5-base+SUP	43.99	86.67	86.59	96.46	77.90	44.49
T5-large+SUP	44.04	86.82	86.77	96.55	77.97	44.97

Table 2: Automatic evaluation results for the story dataset.

observations. First, the proposed model demonstrated outstanding performance, particularly in the aspect of factual accuracy as measured by the FactScore metric. This exceptional result on the FactScore is a testament to the model's proficiency in generating text that is not only coherent and contextually relevant but also factually consistent, a crucial aspect in counterfactual text generation.

Second, the significance of the reconstruction mode (RC) in the model's architecture was profoundly evident in the experimental outcomes. The incorporation of RC led to notable improvements across various models and settings, particularly enhancing the FactScore. This underscores the RC's pivotal role in refining the model's capability to adhere to the principles of minimal editing while adjusting to new counterfactual conditions.

Human Evaluation Results. The human evaluation results for the counterfactual financial and story datasets, presented in Tables 3 and 4, show that the proposed models, particularly Our-large and Our-large+bcf, consistently outperformed baseline models in key aspects. For the financial dataset, Our-large+bcf excelled in generating text that aligns well with altered scenarios, especially in Background (BG) and Counterfactual (CF) components. This performance highlights the model's capability in maintaining coherence and factual consistency, aligning with the aim to explore causal reasoning among different text components. The Our-large model also showed strong performance, further emphasizing the effectiveness of the approach in creating contextually relevant counterfactual narratives.

	BG	CF	PLOT
GPT2-M+zero-shot	2.14	1.66	1.39
BART-base+FT	2.17	1.79	1.82
BART-large+RC+CF	2.34	1.9	1.86
Our-large	2.51	2.33	2.18
Our-large+bcf	**2.58**	**2.54**	**2.37**
Ground-truth	2.77	2.84	2.55

Table 3: Human evaluation results for the financial dataset.

	BG	CF	PLOT
GPT2-M+zero-shot	2.41	2.28	1.18
BART-base+FT	2.47	2.03	1.79
BART-large+RC+CF	2.05	1.77	2.38
Our-large	2.53	**2.31**	2.39
Our-large+bcf	**2.76**	2.22	**2.43**
Ground-truth	2.87	2.64	2.67

Table 4: Human evaluation results for the story dataset.

For the story dataset, Our-large+bcf achieved the highest scores, particularly in BG and PLOT dimensions, demonstrating its proficiency in maintaining narrative coherence amidst counterfactual changes. Similarly, Our-large notably excelled in the CF aspect, suggesting its strength in generating semantically consistent counterfactual endings. These results across both datasets underscore the robustness of the proposed approach in counterfactual text generation, affirming its utility in producing complex, coherent, and contextually appropriate narratives in varied domains.

RQ 2: Effectiveness of Different Component

Disentanglement Component. To evaluate the effect of disentanglement component, we visualize the embedding results of U_B, U_C, U_E and V_B, V_C, V_E (before and after training) in Figure 4 for subjective assessment. It can be seen from Figure 4 (a), (c) and (e) that before training, the embeddings of two latent variables interweave with each other. However, after training with the disentanglement component, as plotted in Figure 4 (b), (d) and (f), the embeddings of these variables could be well separated. This verifies the effectiveness of the proposed disentanglement component.

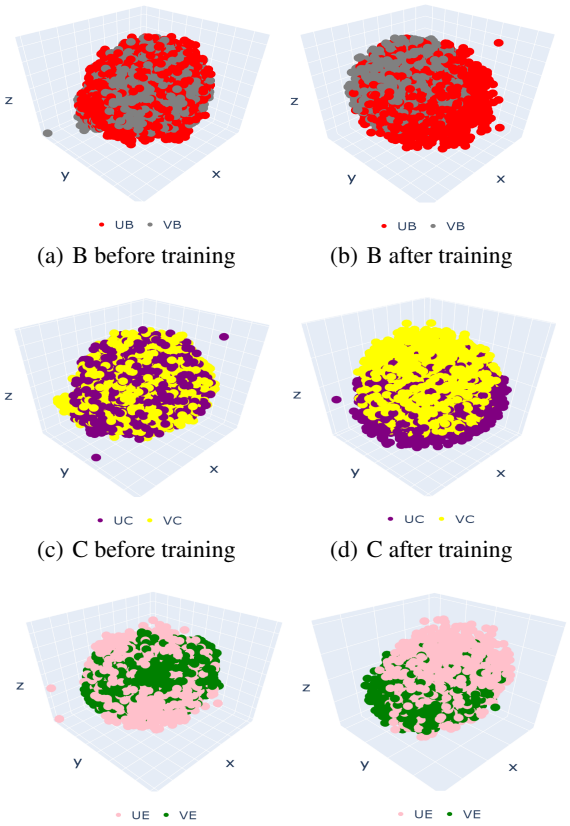
Ablation Study. In the ablation study conducted on both financial and story datasets, the model Our-large and its variant Our-large+BCF demonstrated outstanding performance, particularly in enhancing factual accuracy as evidenced by the FactScore metric. The results, outlined in Table 5 and Table 6, revealed that the removal of key components like the disentanglement component (L_{dis}) and mutual information loss ($L_{U \leftrightarrow V}$) led to significant performance drops, especially in

	ROUGE-L	BERTS	BERTS-FT	NSPSore	WMS	FactScore
Our-large	13.21	70.09	71.32	97.33	35.94	12.64
Our-large w/o L_{dis}	12.10(-8.40%)	67.05(-4.34%)	68.35(-4.16%)	93.97(-3.45%)	33.63(-6.43%)	8.22(-34.97%)
Our-large w/o $L_{rUV\to X}$	12.07(-8.63%)	68.83(-1.80%)	68.27(-4.28%)	94.16(-3.26%)	33.84(-5.84%)	7.91(-37.42%)
Our-large w/o $L_{rV\to X}$	11.82(-10.52%)	67.77(-3.31%)	66.38(-6.93%)	90.72(-6.80%)	32.91(-8.43%)	8.07(-36.16%)
Our-large w/o $L_{U\leftrightarrow V}$	11.69(-11.51%)	65.93(-5.94%)	66.81(-6.32%)	92.93(-4.52%)	32.59(-9.32%)	8.82(-30.22%)
Our-large add BCF	**14.11(+6.81%)**	**72.05(+2.80%)**	**74.24(+4.09%)**	**97.81(+0.49%)**	**36.44(+1.39%)**	**13.06(+3.32%)**

Table 5: Ablation study for the financial dataset. The percentages in the brackets are compared to 'Our-large'.

	ROUGE-L	BERTS	BERTS-FT	NSPSore	WMS	FactScore
Our-large	44.07	83.23	86.46	**98.38**	**77.31**	29.96
Our-large w/o L_{dis}	39.26(-10.91%)	78.35(-5.87%)	79.85(-7.65%)	94.66(-3.78%)	68.99(-10.76%)	17.34(-42.12%)
Our-large w/o $L_{rUV\to X}$	41.41(-6.03%)	79.94(-3.96%)	81.68(-5.53%)	90.68(-7.82%)	69.86(-9.63%)	18.60(-37.9%)
Our-large w/o $L_{rV\to X}$	41.73(-5.31%)	78.95(-5.15%)	79.52(-8.03%)	91.05(-7.45%)	70.11(-9.31%)	18.83(-37.15%)
Our-large w/o $L_{U\leftrightarrow V}$	40.96(-7.05%)	80.16(-3.69%)	81.33(-5.94%)	89.21(-9.32%)	69.14(-10.56%)	16.14(-46.13%)
Our-large add BCF	**44.55(+1.08%)**	**85.04(+2.17%)**	**86.98(+0.60%)**	97.94(-0.45%)	77.15(-0.20%)	**30.21(+0.84%)**

Table 6: Ablation study for the story dataset. The percentages in the brackets are compared to 'Our-large'.



(a) B before training (b) B after training

(c) C before training (d) C after training

(e) E before training (f) E after training

Figure 4: Disentanglement component visualization, t-SNE projection of the embedding before and after training.

maintaining factual consistency. This underscores the critical role of these components in the model's ability to generate coherent and contextually relevant counterfactual texts. The variant Our-large+BCF showed superior results across various metrics, confirming that interventions in the counterfactual background markedly improve the model's effectiveness in both financial and narrative text generation domains.

RQ 3: A Case Study

In the case study for the counterfactual story generation dataset, we chose to focus on narrative examples rather than financial ones due to their more intuitive nature for illustrating the model's capabilities. The results are reported in Table 2 in the appendix due to page limit. From the table, it is obvious that the zero-shot model generates a quite different story compared with the original story. For the 'BART-base+FT+RC' model, it tends to copy the original ending. However, our model could generate reasonable endings that are dependent on the varying conditions or backgrounds. For the case "intervening the condition", the condition is changed from "He decided to rob a bank" to "He decided to ask the bank teller for some money". The corresponding generated ending is like "He was rejected and held up the bank teller". This ending is causal reasonable as the "held up the bank teller" happened after "he was rejected", and the overall ending satisfies the minimal edit requirement. For the case "intervening the background", the generated text contains "His fellows" corresponding to the background "his work office" and the rest content is minimal edited. This observation verifies the causal reasoning ability of the proposed approach.

6 Conclusion

In this paper, we introduced an approach to counterfactual text generation, utilizing structural causal models (SCM) to navigate complex, multi-layered causal relationships across diverse domains such as narrative stories and financial reports. Our method, which stands out from traditional single causal analysis, focuses on disentangling text components into latent variables for generating causally influenced outcomes. This approach is validated through extensive experiments on both a public story generation dataset and a specially constructed financial dataset, demonstrating superior performance over existing methods.

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proc. of ICML*, 2018.

[Chen *et al.*, 2022] Jiangjie Chen, Chun Gan, Sijie Cheng, Hao Zhou, Yanghua Xiao, and Lei Li. Unsupervised editing for counterfactual stories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10473–10481, 2022.

[Dang-Nhu, 2021] Raphaël Dang-Nhu. Evaluating disentanglement of structured representations. *arXiv preprint arXiv:2101.04041*, 2021.

[Feder *et al.*, 2022] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, October 2022.

[Fern and Pope, 2021] Xiaoli Fern and Quintin Pope. Text counterfactuals via latent optimization and shapley-guided search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5578–5593, 2021.

[Hao *et al.*, 2021] Changying Hao, Liang Pang, Yanyan Lan, Yan Wang, Jiafeng Guo, and Xueqi Cheng. Sketch and customize: A counterfactual story generator. In *Proc. of AAAI*, 2021.

[Hu and Li, 2021] Zhiting Hu and Li Erran Li. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955, 2021.

[Huang *et al.*, 2021] Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey. *CoRR*, 2021.

[Khrulkov *et al.*, 2021] Valentin Khrulkov, Leyla Mirvakhabova, Ivan Oseledets, and Artem Babenko. Disentangled representations from non-disentangled models, 2021.

[Kusner *et al.*, 2015] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *Proc. of ICML*, 2015.

[Lee *et al.*, 2020] Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 157–174, Cham, 2020. Springer International Publishing.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*, 2020.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[Liu *et al.*, 2021] Qi Liu, Matt Kusner, and Phil Blunsom. Counterfactual data augmentation for neural machine translation. In *Proc. of NAACL*, 2021.

[Lu *et al.*, 2017] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proc. of CVPR*, 2017.

[Luo *et al.*, 2016] Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. Commonsense causal reasoning between short texts. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, page 421–430, 2016.

[Pearl *et al.*, 2016] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[Qin *et al.*, 2019] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *Proc. of EMNLP*, 2019.

[Qin *et al.*, 2020] Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proc. of EMNLP*, 2020.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[Ren *et al.*, 2022] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *International Conference on Learning Representations*, 2022.

[Roese, 1997] Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 1997.

[Schölkopf *et al.*, 2021] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

[Vasilakes *et al.*, 2022] Jake Vasilakes, Chrysoula Zerva, Makoto Miwa, and Sophia Ananiadou. Learning disentangled representations of negation and uncertainty. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8380–8397, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Voita *et al.*, 2018] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In *Proc. of ACL*, 2018.

[Wang *et al.*, 2019] Ke Wang, Hang Hua, and Xiaojun Wan. Controllable unsupervised text attribute transfer via editing entangled latent representation. *Advances in Neural Information Processing Systems*, 32, 2019.

[Wu *et al.*, 2021] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proc. of ACL*, 2021.

[Yamada *et al.*, 2019] Masanori Yamada, Heecheol Kim, Kosuke Miyoshi, and Hiroshi Yamakawa. Favae: Sequence disentanglement using information bottleneck principle, 2019.

[Yang *et al.*, 2021] Yazheng Yang, Boyuan Pan, Deng Cai, and Huan Sun. Topnet: Learning from neural topic model to generate long stories. In *Proc. of KDD*, 2021.

[Zellers *et al.*, 2019] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proc. of ACL*, 2019.

[Zhang* *et al.*, 2020] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *Proc. of ICLR*, 2020.

[Zhu *et al.*, 2023] Jiageng Zhu, Hanchen Xie, and Wael Abd-Almageed. Sw-vae: Weakly supervised learn disentangled representation vianbsp;latent factor swapping. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, page 73–87, Berlin, Heidelberg, 2023. Springer-Verlag.