Enhancing Factual Consistency in Text Summarization via Counterfactual Debiasing

Zhenqing Ling[1,†], Yuexiang Xie[2,†], Chenhe Dong[1,*], Ying Shen[1,3,*],
[1]School of Intelligent Systems Engineering, Sun Yat-sen University
[2]Alibaba Group
[3]Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology
{lingzhq, dongchh}@mail2.sysu.edu.cn
yuexiang.xyx@alibaba-inc.com, sheny76@mail.sysu.edu.cn

Abstract

Despite significant progress in abstractive text summarization aimed at generating fluent and informative outputs, how to ensure the factual consistency of generated summaries remains a crucial and challenging issue. In this study, drawing inspiration from advancements in causal inference, we construct causal graphs to analyze the process of abstractive text summarization methods and identify intrinsic causes of factual inconsistency, specifically language bias and irrelevancy bias, and we propose CoFactSum, a novel framework that mitigates the causal effects of these biases through counterfactual estimation for enhancing the factual consistency of the generated content. CoFactSum provides two counterfactual estimation strategies, including Explicit Counterfactual Masking, which employs a dynamic masking approach, and Implicit Counterfactual Training, which utilizes a discriminative cross-attention mechanism. Besides, we propose a Debiasing Degree Adjustment mechanism to dynamically calibrate the level of debiasing at each decoding step. Extensive experiments conducted on two widely used summarization datasets demonstrate the effectiveness and advantages of the proposed CoFactSum in enhancing the factual consistency of generated summaries, outperforming several baseline methods.

1 Introduction

Abstractive text summarization (Gupta and Gupta, 2019; Lin and Ng, 2019; Zhang et al., 2020; Luo et al., 2023; Challagundla and Peddavenkatagari, 2024) has witnessed great success in generating remarkably fluent and diversified summaries that approach human-level performance. Nevertheless, the generated summaries often contain factually inconsistent errors against the source documents (Narayan et al., 2018; Maynez et al., 2020). For example, as shown in Figure 1, the subject is predicted as "a 10-year-old boy" while the correct answer is "a cricket team", and the team's final score is wrongly predicted as "one" instead of "zero". Such inconsistencies contained in the generated summaries can mislead and confuse the public and even raise legal risks, which brings significant rectification costs and limits the applications of abstractive text summarization.

Source document: No batsman from Bapchild Cricket Club was able to get off the mark against Christ Church University in Canterbury. "We couldn't believe it, all they needed to do was hit a wall to get one run," Christ Church player Mike Rose told the Crawley Observer. Somerset club Langport set the record for the lowest score when they were dismissed for zero in 1913. Wirral CC were bowled out for three in a Cheshire League Division Three fixture in 2014...
Factually consistent summary: A cricket team was bowled out for 0 in just 20 balls in a county six-a-side indoor championships match.
Factually inconsistent summary: A 10-year-old boy has broken the record for the lowest score ever made in first-class cricket when he hit one run in his first match.

Figure 1: An example of generated summaries by baselines and CoFactSum. The supporting facts in the source document and inconsistent facts in the generated summaries are marked in blue and red, respectively.

To tackle such factually inconsistent issues, several approaches have been proposed in recent years, which can be divided into three categories: (i) *fact encoding*, which integrates additional fact-related information during encoding or decoding (Zhu et al., 2021; Xiao and Carenini, 2022); (ii) *post editing*, which adopts a rectification model to correct the generated summaries (Cao et al., 2020; Chen et al., 2021); and (iii) *auxiliary loss applying*, which designs an auxiliary loss to penalize the model for generating factually inconsistent texts (Cao and Wang, 2021; Wan and Bansal, 2022; Scheurer et al., 2023). However, most of these

†Equal Contribution.
*Corresponding authors: Chenhe Dong and Ying Shen.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7912–7924
January 19–24, 2025. ©2025 Association for Computational Linguistics

studies neglect the intrinsic causes of the factual inconsistency in abstractive text summarization.

Considering the generation process of abstractive text summarization models, the generated summaries rely on two key factors: the language prior knowledge acquired during pre-training, and the information contained in the source document, both of which contribute to the fluency and informativeness of generated summaries. However, they might introduce *language bias* and *irrelevancy bias* caused by the spurious linguistic correlations learned from pre-training and the irrelevant information in the source document. These biases drive the observed factual inconsistencies in the generated summaries. For example, in Figure 1, the unfaithful content "A 10-year-old boy" is not contained in the source document, which is caused by the *language bias*; and the unfaithful content "he hit one run" is inferred from the mismatched tokens "hit a wall to get one run" in the source document, which is caused by the *irrelevancy bias*.

Shed light on the above insights, we make the first attempt to incorporate the idea of causal inference (Pearl, 2001; Pearl and Mackenzie, 2018) into the generation process of text summarization to ensure the factual consistency of generated summaries by eliminating the language and irrelevancy biases. Firstly, we build up a causal graph among various elements to demonstrate their causal relationships in abstractive text summarization. Then, based on the causal graph, we propose a **Co**unter**Fact**ual debiasing framework for abstractive **Sum**marization, named **CoFactSum**, to estimate and alleviate the causal effects of language and irrelevancy biases on the generated summary.

The proposed CoFactSum consists of two counterfactual estimation strategies, including Explicit Counterfactual Masking (ECM) with an *explicit* dynamic masking strategy, and Implicit Counterfactual Training (ICT) with an *implicit* discriminative cross-attention mechanism. Furthermore, we design a Debiasing Degree Adjustment (DDA) module to dynamically adapt the debiasing degree at each decoding step, improving the ability of the proposed framework to position the factual inconsistencies in the generated summaries.

Guided by theoretical principles, we conduct a series of experiments and successfully validate the effectiveness and reliability of CoFactSum. Our main contributions are summarized as follows:

- We identify that language bias and irrelevancy bias are currently the key factors affecting abstractive text summarization. And we construct causal graphs to determine the intrinsic causes of such factual inconsistency.
- Based on theoretical insights, we propose the CoFactSum framework to mitigate the causal effects of factual inconsistency, leading to the generation of factually consistent summaries.
- The extensive experiments on two widely-used summarization datasets CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018) demonstrate the effectiveness of CoFactSum in enhancing the factual consistency of generated summaries. Our codes are publicly available at `https://github.com/lingzhq/CoFactSum`.

2 Related Works

Counterfactual Inference In the field of natural language processing, causal inference (Pearl, 2001; Pearl and Mackenzie, 2018) has recently inspired many works to discover the intrinsic causes of specific biases and remove their causal effects in an interpretable way, such as the studies in visual question answering (Niu et al., 2021; Chen et al., 2023), text classification (Qian et al., 2021), fairness (Zhu et al., 2024), and text summarization (Xie et al., 2021). These methods target measuring causal effects of biases under counterfactual scenarios based on causal graphs and eliminating causal effects by mitigating them from total effect.

Factual Consistency in Text Summarization As discussed in previous studies (Maynez et al., 2020; Nan et al., 2021; Ladhak et al., 2022), current advanced generation models in abstractive summarization are prone to produce factually inconsistent text. To tackle such issues, three mainstream techniques have been applied recently. The first is *fact encoding*, which aims to incorporate more fact-related information during encoding source documents or target summaries, such as knowledge graphs (Huang et al., 2020; Zhu et al., 2021) and document entities (Xiao and Carenini, 2022). The second is *post editing*, which treats the generated summaries as drafts and further conducts post-editing on them, and is usually achieved by a separate correction model (Dong et al., 2020; Cao et al., 2020; Chen et al., 2021). The third is *auxiliary loss applying*, which designs auxiliary penalty losses to force the model to distinguish between faithful and unfaithful samples, and so far, the unlikelihood loss (Li et al., 2020), contrastive loss (Cao and Wang, 2021; Liu et al., 2022; Wan and Bansal, 2022) and refinement loss from language feedbacks (Scheurer et al., 2023) are most widely adopted.

3 Methodology

3.1 Causal Graph Construction

A causal graph, also known as a causal network or a causal Bayesian network, is a graphical representation of causal relationships and dependencies between variables or events in a system, which helps in understanding and modeling cause-and-effect relationships (Pearl, 2009).

The causal graph of abstractive text summarization can be given as a directed acyclic graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, which represents the causal relationships (i.e., $\mathcal{E}$) between different variables (i.e., $\mathcal{V}$). The causal graph consists of five variables: the source document X, the important information U (relevant to the ground-truth summary), the irrelevant information R (irrelevant to the ground-truth summary), the language prior P (generic language knowledge such as grammar and syntax), and the generated summary Y, as shown in Figure 2 (a). The important information U and the irrelevant information R are composed by the source document X, and their causal relationships are denoted by the paths $X \to U$ and $X \to R$, respectively.

During the generation process, the text summarization model first encodes the source document X, and then generates tokens step-by-step in an auto-regressive manner for producing the summary Y, which can be given as $U \to Y$ and $R \to Y$. The causal effect of language prior P on the generated summary Y can be expressed as $P \to Y$.

In this study, we aim to estimate and mitigate the causal effect of language prior knowledge P and irrelevant information R on the generated summary Y, i.e., $R \to Y$ and $P \to Y$, which introduces language bias and irrelevancy bias and causes the factual inconsistent errors.

3.2 Causal Effect Estimation

Based on the causal graph, we can estimate the causal effects of language bias and irrelevancy bias on the generated summary.

Total Effect In the causal graph, suppose that the document X is set to x, the underlying important and irrelevant information U and R is set to u and r, respectively, and the language prior P is set to p, then the generated summary Y can be given as:

$$\begin{aligned} Y_{u,r,p} &= Y(do(U=u), do(R=r), do(P=p)) \\ &= Y(U=u, R=r, P=p), \end{aligned} \tag{1}$$

where the do operator can be omitted according to the back-door criteria (Pearl, 2009). To measure the total effect on Y, we need to compare the potential outcomes of the same individual under the treatment and no-treatment conditions, where the no-treatment condition can be approximated by setting U, R, P to empty values u^*, r^*, p^* under the counterfactual scenario. Formally, the total effect can be given as:

$$E_{total} = Y_{u,r,p} - Y_{u^*,r^*,p^*}. \tag{2}$$

Bias Elimination Similarly, the causal effects of language prior P and irrelevant information R on the generated summary Y can be estimated as:

$$E_{bias} = Y_{u^*,r,p} - Y_{u^*,r^*,p^*}, \tag{3}$$

where we set $U = u^*$ to exclude the causal effect of the important information U on Y. To eliminate the language bias and irrelevancy bias in the generation process, we remove their causal effects on the generated summary from the total effect. Formally, it can be given as:

$$E_{total} - E_{bias} = Y_{u,r,p} - Y_{u^*,r,p}. \tag{4}$$

The equation can also be regarded as the estimation of the causal effect of important information U on the generated summary Y when given the $R = r$ and $P = p$, as illustrated in Figure 2 (b).

3.3 Instantiation

In order to instantiate Equation (4) in abstractive text summarization, we design two counterfactual strategies, i.e., Explicit Counterfactual Masking (ECM) and Implicit Counterfactual Training (ICT),

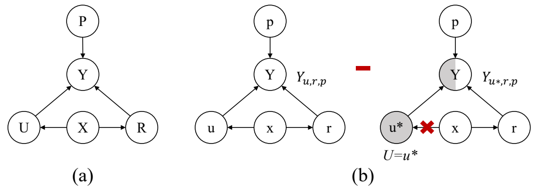

Figure 2: Illustration for (a) the basic causal graph and (b) our debiasing framework CoFactSum.

which are designed for estimating during the inference process and optimizing during the training process, respectively. The instantiation of CoFactSum is illustrated in Figure 3.

Explicit Counterfactual Masking (ECM) Previous studies (Xie et al., 2021) have used masking techniques to block the causal effect of important information on the generated summary. However, the proposed ECM is different from previous studies in that it considers that during the generation process, the decoder attends to different tokens of the source document at different decoding steps. Therefore, we propose to dynamically determine the important tokens in the source document w.r.t. each generated token, rather than using a fixed set of important tokens.

Specifically, we use the cross-attention score as an indicator and employ a top-K strategy to pick up the top K positions with the maximum scores as the important positions. To remove these important tokens from the source document without causing the disparity between training and inference, we use a special token "[MASK]" to explicitly replace the important tokens, similar to the pre-train stage of most transformer-based models (Devlin et al., 2019; Zhang et al., 2020). We also adopt a debiasing ratio α ($\alpha \leq 1$) to adjust the extent of debiasing, in order to preserve the informativeness of generated summaries. Formally, the probability of each generated token y_t with ECM can be given as:

$$\Pr(y_t|x) = \Pr(y_t|y_{<t}, x; \theta) - \alpha \cdot \Pr(y_t|y_{<t}, x'; \theta), \tag{5}$$

where x' denotes the masked document, and θ denotes the model parameters.

Implicit Counterfactual Training (ICT) In addition to ECM, a counterfactual training strategy with a discriminative cross-attention mechanism is further proposed to implicitly minimize the causal effect of bias on the generated summaries.

Specifically, at each decoding step, the source document is dynamically split into two disjoint partitions (i.e., important tokens x_u and irrelevant tokens x_r) based on cross-attention scores. Then the decoder model separately attends to these partitions for counterfactual training. The probability of each generated token y_t at decoding step t can be represented as $\Pr(y_t|y_{<t}, x_u; \theta')$ and $\Pr(y_t|y_{<t}, x_r; \theta')$, respectively, where θ' denotes the parameters of the counterfactual summarization model.

Intending to guide the counterfactual text summarization model to rely less on the important tokens, we use an unlikelihood loss $\mathcal{L}_{unl}$ to penalize the sequence log-likelihood when the model attends to important tokens:

$$\mathcal{L}_{unl} = -\sum_{t=1}^{|y|} \log\left(1 - \Pr(y_t|y_{<t}, x_u; \theta')\right), \tag{6}$$

where y is the ground truth summary. Meanwhile, a cross-entropy loss $\mathcal{L}_{xent}$ is adopted to increase the probabilities of tokens that are generated when attending to irrelevant tokens:

$$\mathcal{L}_{xent} = -\sum_{t=1}^{|y|} \log \Pr(y_t|y_{<t}, x_r; \theta'). \tag{7}$$

Moreover, we adopt a Kullback-Leibler (KL) divergence loss $\mathcal{L}_{kl}$ to further push away the predicted distributions over vocabulary when attending to the important tokens and irrelevant tokens respectively, which can be formally given as:

$$\mathcal{L}_{kl} = -\sum_{t=1}^{|y|} \mathrm{KL}\left(\Pr(\cdot|y_{<t}, x_u; \theta') || \Pr(\cdot|y_{<t}, x_r; \theta')\right). \tag{8}$$

Finally, the training loss can be defined by:

$$\mathcal{L} = \mathcal{L}_{unl} + \gamma\mathcal{L}_{xent} + \lambda\mathcal{L}_{kl}, \tag{9}$$

where γ, λ are hyperparameters to control the strength of adopted loss functions. Only the decoder's parameters are updated, with encoder frozen to ensure encoder outputs are consistent across treatment conditions during debiasing.

Applying the above counterfactual process, we train a counterfactual decoder as an instantiation of $Y_{u^*,r,p}$ in Equation (4). The debiased probability of each generated token y_t with ICT is given as:

$$\Pr(y_t|x) = \Pr(y_t|y_{<t}, x; \theta) - \beta \cdot \Pr(y_t|y_{<t}, x; \theta'), \tag{10}$$

where β ($\beta \leq 1$) is a hyperparameter.

Debiasing Degree Adjustment (DDA) Taking both ECM and ICT into consideration, we point out that debiasing at every decoding steps to the same extent might not be an optimal solution, since the intermediately generated sentences at different decoding steps have different factually inconsistent degrees. It is reasonable to conduct more intensive debiasing when the generated sentence is relatively less consistent and vice versa.

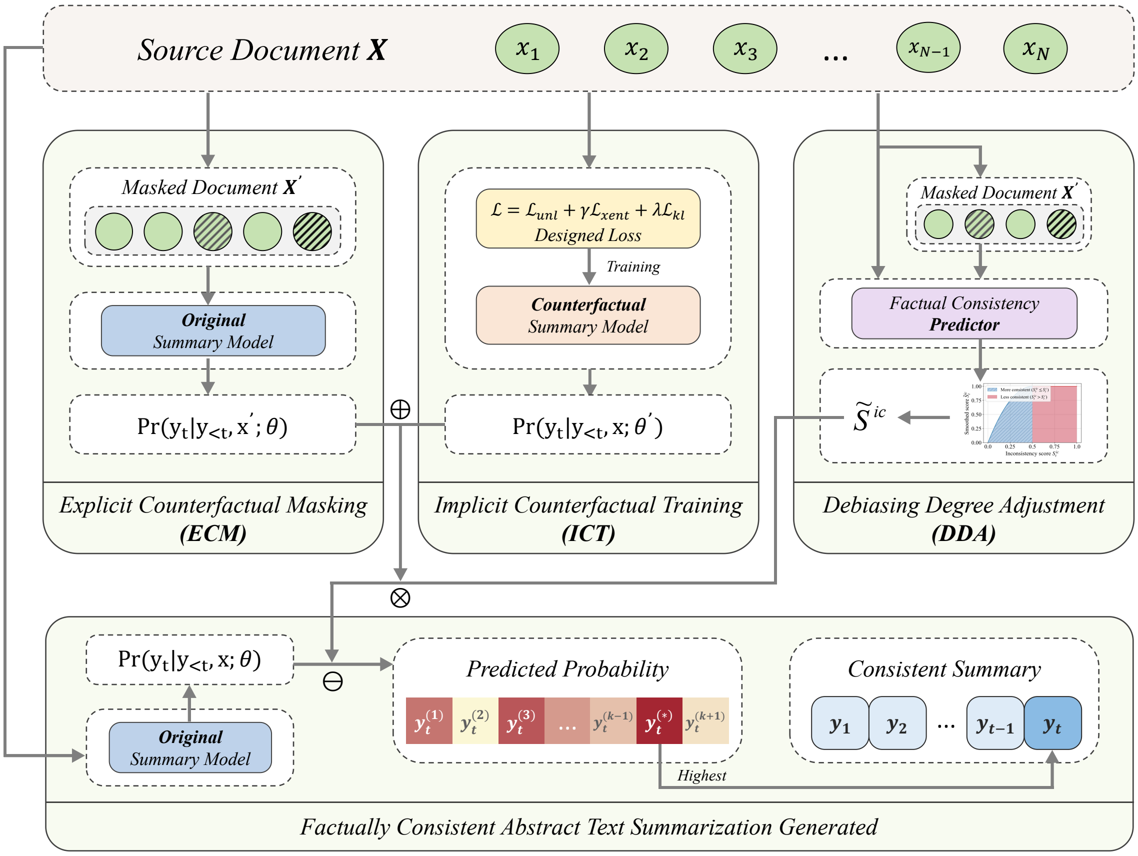

Figure 3: Illustration of CoFactSum in each decoding step to generate factually consistent text summaries.

To this end, we propose a dynamic adjustment strategy for the debiasing degrees at different decoding steps. This involves pre-training a factual consistency predictor using synthetic inconsistent summaries, which adapts the debiasing ratio based on inconsistency scores. The prediction process is treated as a sequence labeling task, identifying mismatched tokens as *inconsistent* and matched ones as *consistent*.

During training at t-th decoding step, the predictor receives the following four representations: the original decoding hidden states $\mathbf{h}_t \in \mathcal{R}^d$, the counterfactual hidden states generated from the masked source document $\mathbf{h}'_t \in \mathcal{R}^d$, the element-wise multiplication and the difference of the above two hidden states. These representations are concatenated and sent to a fully connected layer and a softmax function to obtain the predicted scores, as formulated by:

$$\mathbf{S}_t = \text{softmax}(\mathbf{W} \cdot \mathbf{z}_t + \mathbf{b}) \in \mathcal{R}^2, \quad (11)$$

$$\mathbf{z}_t = [\mathbf{h}_t; \mathbf{h}'_t; \mathbf{h}_t \odot \mathbf{h}'_t; \mathbf{h}_t - \mathbf{h}'_t] \in \mathcal{R}^{4d}, \quad (12)$$

where d is the dimension of hidden states, $\mathbf{W} \in \mathcal{R}^{2\times 4d}$, $\mathbf{b} \in \mathcal{R}^2$ are learnable parameters in the linear layer, $[;]$ denotes the concatenation, $\odot$ is the element-wise multiplication, and $\mathbf{S}_t$ contains the factually consistent score S^c_t and factually inconsistent score S^{ic}_t in which $S^c_t + S^{ic}_t = 1$. We use cross-entropy loss to train the predictor and freeze the parameters of the original summarization model.

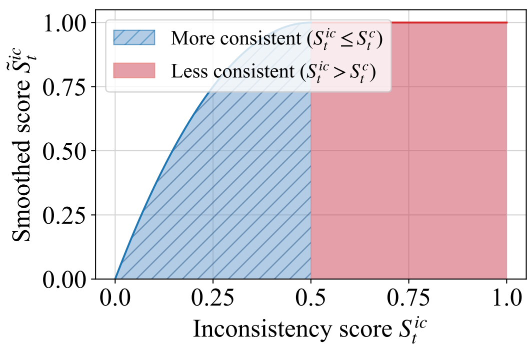

Figure 4: The smoothing function used in DDA for the factually inconsistent scores.

During inference, we multiply the subtracted terms $\alpha \cdot \Pr(y_t|y_{<t}, x'; \theta)$ and $\beta \cdot \Pr(y_t|y_{<t}, x; \theta')$ by a predicted factually inconsistent score to dynamically control the debiasing degrees. Besides, as we observed in our experiments, the factually

Methods	CNN/DM								XSum							
	R-L	QAFE	QAGS	FCC	FT-C	FT-O	CoCo	AVG	R-L	QAFE	QAGS	FCC	FT-C	FT-O	CoCo	AVG
PEGASUS (Zhang et al., 2020)	40.48	89.25	75.52	39.43	53.64	67.86	47.54	51.34	**39.06**	41.49	21.47	25.29	6.17	3.72	15.09	28.97
UNL (Li et al., 2020)	39.15	86.71	74.72	36.76	53.31	67.86	45.20	49.96	34.03	38.51	18.87	**25.92**	4.45	1.17	12.60	25.48
CORR (Cao et al., 2020)	39.79	82.30	69.49	22.68	49.46	58.87	41.55	46.92	38.95	41.72	21.73	25.01	6.10	3.69	15.07	28.92
CCGS (Chen et al., 2021)†	40.40	87.24	73.35	37.09	54.71	67.40	47.20	50.78	38.68	41.08	21.14	25.11	8.31	3.67	14.95	28.86
CLIFF (Cao and Wang, 2021)	39.47	88.64	**76.59**	39.22	54.57	71.02	46.99	51.15	38.14	**43.34**	22.80	24.73	6.24	3.15	15.41	28.71
SC (Xiao and Carenini, 2022)†	**41.34**	82.45	70.17	30.15	45.95	52.12	39.10	47.33	38.34	37.20	19.87	23.49	4.76	1.54	13.24	27.51
COFACTSUM	39.94	**90.18**	75.94	**43.48**	**57.45**	**72.38**	**49.85**	**52.41**	37.23	43.15	**22.99**	24.43	**10.47**	**9.27**	**16.10**	**29.15**

Table 1: Automatic evaluation results on CNN/DM and XSum. Methods with † are conducted with released codes. **Bold** indicate methods with the best performances. Columns in grey indicate metrics in terms of factual consistency.

inconsistent scores tend to vary dramatically across different decoding steps, thus we design a smoothing function to restrict their variation range and stabilize the inference. The overall predicted probability with debiasing can be formally given as:

$$\Pr(y_t|x) = \Pr(y_t|y_{<t}, x;\theta) - \tilde{S}^{ic} \cdot \Big(\alpha \cdot \Pr(y_t|y_{<t}, x';\theta) + \beta \cdot \Pr(y_t|y_{<t}, x;\theta')\Big). \quad (13)$$

$\tilde{S}^{ic}$ is the smoothed factually inconsistent score, and at the t-th decoding step, it is calculated by:

$$\tilde{S}^{ic}_t = \begin{cases} 1-(2S^{ic}_t-1)^2, & S^{ic}_t \le S^c_t \\ 1, & S^{ic}_t > S^c_t \end{cases}, \quad (14)$$

which is illustrated in Figure 4. The overall training procedure and the computational overhead of the proposed COFACTSUM to construct factual consistency-enhanced text summaries in practical applications is summarized in Appendix B.

4 Experiment

In this section, we will introduce our experimental setup and results, which validate the effectiveness of COFACTSUM. The detailed implementation is provided in Appendix C.1.

4.1 Datasets and Metrics

Datasets We conduct experiences on two widely adopted abstractive summarization datasets, including CNN/DailyMail (CNN/DM) (Hermann et al., 2015) and Extreme Summarization (XSum) (Narayan et al., 2018). Both datasets contain news articles and their corresponding summaries written by professional journalists. The detailed description of these datasets and their size is provided in Appendix C.2.

Metrics We first adopt **ROUGE-L** metric (Lin, 2004) to evaluate the informativeness. However, such traditional evaluation metrics are not capable of measuring factual consistency. Therefore, we employ the following metrics to assess the factual consistency of COFACTSUM: **QAFE**(Fabbri et al., 2022), **QAGS**(Wang et al., 2020), **FactCC (FCC)**(Kryscinski et al., 2020), **Fact Triple (FT-C/O)**(Goodrich et al., 2019), and **CoCo** (Xie et al., 2021). All of these metrics are widely used in the evaluation of summarization, with detailed descriptions provided in Appendix C.3.

4.2 Baselines

We adopt **PEGASUS** (Zhang et al., 2020) as the model backbone, and mainly choose the following four counterparts to compare with: (i) **UNL** (Li et al., 2020), which leverages the unlikelihood loss to penalize the probabilities of the tokens in unfaithful samples. (ii) **CORR** (Cao et al., 2020), which pre-trains a post-editing corrector model to directly generate factually consistent summaries. (iii) **CCGS** (Chen et al., 2021), which pre-trains a factual consistency predictor and leverages it to rank candidate summaries. (iv) **CLIFF** (Cao and Wang, 2021), which adopts contrastive loss to discriminate between faithful and unfaithful samples. (v) **SC** (Xiao and Carenini, 2022), which contains an entity-based SpanCopy mechanism with Global Relevance to reduce mismatched entities.

4.3 Results

Automatic Evaluation We report the automatic evaluation results on CNN/DM and XSum in Table 1. Following previous studies (Cao and Wang, 2021), we randomly select 5,000 samples for the factual consistency evaluation on CNN/DM. In summary, the overall performances of COFACTSUM on both CNN/DM and XSum are significantly better than baseline with improvements of at least 1.07% and 0.18%, which demonstrates the superior trade-off ability of COFACTSUM between factual consistency and informativeness.

Specifically, COFACTSUM demonstrates advantages over baselines in most factual consistency

Methods	CNN/DM			XSum		
	Win↑	Tie	Lose↓	Win↑	Tie	Lose↓
UNL	15.33	54.67	30.00	18.00	51.33	30.67
CORR	13.33	38.00	48.67	7.33	89.33	**3.34**
CCGS	8.00	87.33	4.67	12.67	78.00	9.33
CLIFF	**21.33**	59.33	19.34	17.33	62.67	20.00
SC	14.00	60.67	25.33	6.00	68.33	25.67
COFACTSUM	17.33	80.67	**2.00**	**29.33**	62.00	8.67

Table 2: Pairwise human evaluation results (%) in terms of factual consistency compared with PEGASUS.

Methods	R-L	QAGS	FT-C	FT-O	AVG
Ours	37.23	**23.44**	**9.84**	**8.96**	**25.66**
w/o DDA	37.64	22.79	8.00	7.68	25.23
w/o ECM	37.89	22.95	7.29	7.70	25.27
w/o ICT	38.50	21.68	5.97	4.36	24.59
w/o All	**39.06**	21.29	5.71	3.77	24.66

Table 3: Ablation study on different modules.

metrics. For instance, on CNN/DM, it achieves improvements of 0.93%, 0.42%, and 4.05% in QAFE, QAGS, and FCC, respectively, compared to PEGASUS. And on XSum, it shows gains of 4.30%, 5.55%, and 1.01% in FT-C, FT-O, and COCO. While there is a slight drop in the traditional R-L metric (similar to CCGS and CLIFF), COFACTSUM still delivers competitive performance, confirming the informativeness of its summaries.

Human Evaluation We also conduct pairwise human evaluations on the factual consistency of generated summaries, as shown in Table 2. We randomly select 100 samples from CNN/DM and XSum and have three experienced annotators assess whether summaries generated by factually consistent methods are *better than*, *tie with*, or *worse than* those from the baseline PEGASUS. The results show that COFACTSUM has the fewest losses (2.00% in CNN/DM) and the most wins (29.33% in XSum), indicating significant improvements in factual consistency. Additionally, human evaluation results on informativeness (Appendix D) show that COFACTSUM is competitive with baselines, achieving a strong balance between informativeness and factual consistency.

4.4 Ablation Study

We conduct an ablation study to evaluate the effectiveness of the COFACTSUM modules (DDA, ECM, and ICT) using 3,000 randomly selected instances from XSum, as shown in Table 3. The results indicate that overall performance decreases when any module is removed. Specifically, DDA and ECM contribute equally with average improvements of 0.43% and 0.39%, respectively, while ICT has the largest impact with a 1.07% improvement. This confirms the effectiveness of the COFACTSUM modules.

Methods	R-L	QAGS	FT-C	FT-O	AVG
Ours	37.23	**23.44**	**9.84**	**8.96**	**25.66**
w/o $\mathcal{L}_{unl}$	**38.46**	22.24	6.98	6.87	25.25
w/o $\mathcal{L}_{xent}$	37.85	22.45	7.17	6.27	24.91
w/o $\mathcal{L}_{kl}$	38.08	21.93	7.20	5.67	24.84

Table 4: Ablation study on different training losses.

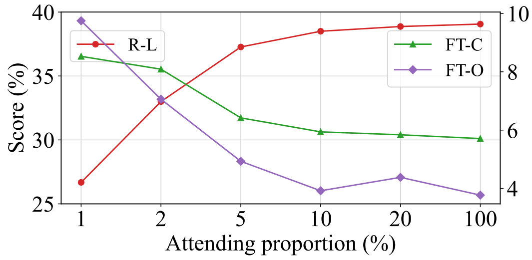

Figure 5: Analysis on different attending proportions of irrelevant information.

Additionally, we evaluate the effectiveness of training losses ($\mathcal{L}unl$, $\mathcal{L}xent$, and $\mathcal{L}kl$) in Table 4, where the KL loss $\mathcal{L}kl$ shows the highest improvement, contributing 0.82% to overall performance.

4.5 Hyper-parameters Study

Impact of Irrelevancy Bias We conduct several experiments on the original PEGASUS model to evaluate the negative impact of irrelevancy bias on factual consistency. Specifically, we force the model attends to different proportions of irrelevant information based on the cross-attention scores during decoding and assess the generated summaries. The results are shown in Figure 5, from which we can observe that the factual consistency scores (i.e., FT-C and FT-O) gradually decrease as the attending proportion and the amount of irrelevant information increase, demonstrating the negative effect of irrelevancy bias.

Masking and Attending Strategy in ECM and ICT To confirm the ascendancy of the dynamic strategy, we select several static strategies for comparison. Following works that adopt static masking strategies (Xie et al., 2021), we select three static

Methods	R-L	QAGS	FT-C	FT-O	AVG
Ours	37.23	**23.44**	**9.84**	**8.96**	**25.66**
Static (tok.)	32.93	21.51	9.66	8.54	23.08
Static (sent.)	35.34	20.51	7.90	5.30	23.29
Static (doc.)	**38.08**	20.95	7.77	4.80	24.63

Table 5: Analysis on different masking and attending strategies in ECM and ICT. (tok.: token-level, sent.: sentence-level, doc.: document-level)

types for masking and attending in ECM and ICT, including *token-level* (tok.), *sentence-level* (sent.), and *document-level* (doc.). These strategies are proposed to mask and attend to the same named entities, the same sentences with at least one entity, and the entire tokens in the source document during different decoding steps in ECM and ICT, respectively.

The results are shown in Table 5, from which we observe that all the overall performances of the static strategies have significant decreases compared with the proposed dynamic strategy used in CoFactSum. Moreover, we can see that the strategies *token-level* and *sentence-level* lead to poor performances on R-L while those of *document-level* and the dynamic strategy are kept at the same level. These results imply that the decoder has different perceptions of important information at different decoding steps; simply choosing the same part of the source document as important information will harm the informativeness, while indiscriminately choosing the entire tokens or dynamically choosing the important tokens can alleviate the issue.

Impact of Debiasing Degree We gradually increase the static debiasing ratio α, β in Equation (13) to investigate the impact of debiasing degree on the informativeness and factual consistency. From the results in Figure 6, we can see that with the enhancement of debiasing degree, the R-L score gradually decreases and the factual consistency scores increase first and then decrease. This phenomenon indicates that a proper debiasing degree can improve the factual consistency of generated summaries without weakening their informativeness, while a large debiasing degree might severely hurt the informativeness and factual consistency.

4.6 Case Study

We further conduct a case study in Figure 7. From the figure, it is evident that all the summaries generated by the baseline methods include factual inconsistencies that are not mentioned in the source document, such as the number of employees "100 workers" and the name of the city "Cardiff", while the proposed CoFactSum alleviates such factual inconsistency issue to some extent.

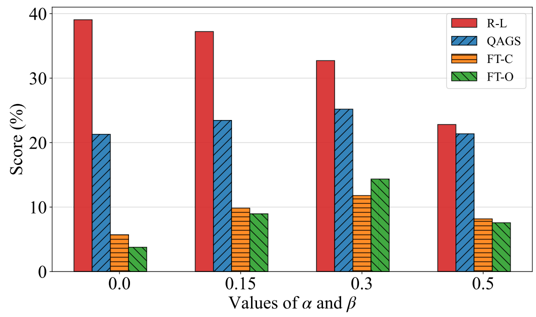

Figure 6: Comparisons among applying different debiasing ratios α and β in CoFactSum.

Source document: The employees, who worked in four takeaways, are alleged to have been living and working in the country illegally. The firms have been asked to produce documents proving their staff had the right to work and live in the UK. If they are unable to do so the Home Office said they would impose a fine of up to £20,000 per illegal employee. The process to deport the workers is already under way.

Pegasus: The Home Office has launched an investigation into the alleged illegal employment of more than 100 workers at takeaways in Cardiff.
CCGS: The Home Office has launched an investigation into the alleged illegal employment of four workers at takeaways in Cardiff.
CLIFF: More than 100 illegal workers have been ordered to leave the UK by the Home Office.
CoFactSum (ours): The Home Office has launched an operation targeting illegal immigrants working in the takeaway food industry.

Figure 7: An example of generated summaries by baselines and CoFactSum. The supporting facts in the source document and inconsistent facts in the generated summaries are marked in blue and red, respectively.

5 Conclusions

In this paper, we enhance the factual consistency of generated summaries by utilizing counterfactual estimation to mitigate the causal effects of language bias and irrelevancy bias. We propose CoFactSum, a novel framework that contains two counterfactual estimation methods: Explicit Counterfactual Masking and Implicit Counterfactual Training. Meanwhile, we propose a Debiasing Degree Adjustment module to dynamically calibrate debiasing levels at different decoding steps. We conduct a series of experiments, including comparisons with baselines, ablation studies, hyperparameter analysis, and case studies to demonstrate significant advances in improving factual consistency.

Limitations

We investigate how to leverage counterfactual estimation to eliminate language and irrelevant biases in text summarization in this study. On the limitations of this paper, we primarily conclude in four aspects as follows: (i) The proposed method is evaluated on widely-used auto-regressive pre-trained models, while its applicability and effectiveness for large language models require further investigation. (ii) The model complexity is increased for improving the factual consistency of generated summaries. (iii) Current debiasing methods generally undermine the traditional metrics to some extent while enhancing the factual metrics, and how to achieve a better trade-off between traditional and factual metrics remains a challenging problem.

References

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6251–6258.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649.

Bhavith Chandra Challagundla and Chakradhar Peddavenkatagari. 2024. Neural sequence-to-sequence modeling with attention by leveraging deep learning architectures for enhanced contextual understanding in abstractive text summarization. *arXiv preprint arXiv:2404.08685*.

Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. 2023. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13218–13234.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9320–9331.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2587–2601.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 166–175.

Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9815–9822.

Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2022. Co2sum:contrastive learning for factual-consistent abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xiansheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, page 411–420.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5434–5445.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Wen Xiao and Giuseppe Carenini. 2022. Entity-based spancopy for abstractive summarization to improve the factual consistency. *arXiv preprint arXiv:2209.03479*.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733.

Yuchang Zhu, Jintang Li, Yatao Bian, Zibin Zheng, and Liang Chen. 2024. One fits all: Learning fair graph neural networks for various sensitive attributes. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4688–4699.

A Description of Symbols

For the convenience of reading and to ensure clarity in the exposition of our methodologies and results, we list all mathematical symbols and their corresponding definitions in Table 6.

B Algorithm

By leveraging the foundational theories discussed above and integrating the specified methodologies, the overall training procedure of the proposed COFACTSUM for constructing factually consistent-enhanced text summaries in practical applications is summarized in Algorithm 1.

Specifically, we take a source document x at the input and aim to output a factually consistent summary. At each decoding step t, **Steps 4** and **Step 5** generate the probability of each generated token y_t using ECM, while **Steps 6** generates the probability of each generated token y_t using ICT. **Steps 7** applies DDA to regulate the extent of the effect of ECM and ICT. Ultimately, we obtain an output y that has mitigated language bias.

From a theoretical perspective, the computational overhead introduced by the dynamic masking and debiasing mechanisms can be assessed through the equations involved in the process. Specifically, as an example, consider the debiasing Equation (10) in module ICT, where the computation of $\Pr(y_t|y_{<t}, x; \theta)$ is part of the standard decoding process. The additional term, $\Pr(y_t|y_{<t}, x; \theta')$ introduces extra computational steps, but this additional cost is directly proportional to the number of tokens being decoded and does not scale with the size of the input or the complexity of the model itself. Thus, while there is a slight increase in computational cost, it is linear with respect to the output length, rendering it manageable in practical applications.

Algorithm 1 COFACTSUM Algorithm

Require: Source document x, original summarization model f_θ, counterfactual summarization model $f'_{\theta'}$ trained with ICT, factual consistency predictor g trained with DDA, maximum decoding step T

Ensure: Factually consistent summary y

1: Initialize $y \leftarrow \{\}$;
2: **for** $t \leftarrow 1$ to T **do**
3: Feed $x, y_{<t}$ into f to generate the probability of each token y_t at t-th decoding step $\Pr(y_t|y_{<t}, x; \theta)$;
4: Mask x according to the cross-attention score to produce x';
5: Feed $x', y_{<t}$ into f to generate the probability $\Pr(y_t|y_{<t}, x'; \theta)$;
6: Feed $x, y_{<t}$ into f' to generate the probability $\Pr(y_t|y_{<t}, x; \theta')$;
7: Feed x, x' into g to generate the smoothed factually inconsistent score $\tilde{S}^{ic}$;
8: Calculate $\Pr(y_t|x)$ according to Equation (13) and select y_t^* with highest probability;
9: $y \leftarrow y \cup y_t^*$;
10: **end for**
11: **return** y

C Details of Experiments

C.1 Implementation Details

The proposed COFACTSUM is implemented based on pytorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2020). After conducting a hyperparameter search, we have obtained the following recommended parameter settings. During the training process in ICT, we set γ, λ in Equation (9) to 1 and 0.01, respectively. The batch size is set to 8, and the number of training steps is set to 50,000 on both datasets. The attending proportion of important/irrelevant information is set to 0.5/0.5 and 0.1/0.9 on CNN/DM and XSum, respectively. The learning rate is set to 5e-4 and 5e-5 on CNN/DM and XSum, respectively. During the training in DDA, the batch size is set to 8, the number of training steps is set to 50,000, and the learning rate is set to 1e-4 on both datasets. And during inference, the masking ratio in ECM is the same as the attending proportion of important information in ICT on both datasets. We use beam search for decoding and set the beam size as 20 and 12 on CNN/DM and XSum, respectively. For the debiasing ratio α, β in Equation (13), we set $\alpha = 0.05, \beta = 0.01$ on CNN/DM and $\alpha = 0.15, \beta = 0.15$ on XSum. The unfaithful samples in DDA are constructed with the system generation method (Cao and Wang, 2021). All experiments are conducted on GeForce RTX 3090 GPUs with 24GB of video memory.

C.2 Datasets intro

The number of samples in the utilized datasets is presented in Table 7, and the fundamental information of these datasets is provided as follows:

Symbol	Description
$\mathcal{V}$	Variables in graph
$\mathcal{E}$	Causal relationships
$\mathcal{G}$	Directed acyclic graph
X	Source document
U	Relevant information to the ground-truth summary
R	Irrelevant information to the ground-truth summary
P	Generic language knowledge
Y	The generated summary
K	Positions with the maximum scores
d	Dimension of hidden states
θ	Model parameters
$\theta^{'}$	Parameters of counterfactual generated model
$x^{'}$	The masked document
$u^{'}$	Causal exclusion value
α, β	Debiasing ratios of ECM and ICT
y_t	The generated token
z_t	Representations information at decoding step
$\mathbf{h}_t, \mathbf{h}_t^{'}$	Original / Counterfactual hidden states
γ, λ	Control hyperparameters of loss functions
ξ^{bd}	The extent of Bias
$\tilde{S}_t^{c}, \tilde{S}_t^{ic}$	Smoothed factually consistent / inconsistent score
y	Factually consistent summary
f	The used model for generating summaries
g	The factual consistency predictor model

Table 6: Description of symbols used in the paper.

(i) **CNN/DailyMail (CNN/DM)** (Hermann et al., 2015), which is a widely used and reputable collection of news articles and their corresponding abstractive summaries from the CNN and Daily Mail websites, primarily utilized for text summarization research and evaluation.

(ii) **Extreme Summarization (XSum)** (Narayan et al., 2018), which is a widely used dataset comprising abstractive summaries of British Broadcasting Corporation (BBC) online articles, designed for text summarization tasks and researches.

Number of samples	CNN/DM	XSum
Train set	287,227	204,045
Validation set	13,368	11,332
Test set	11,490	11,334

Table 7: Number of samples in datasets.

C.3 Metrics

We first adopt conventional metric ROUGE-L (Lin, 2004) to evaluate the informativeness of CoFactSum. However, such traditional evaluation metric is not capable of measuring the factual consistency between the source document and summary. Therefore, we adopt several metrics for evaluation as follows:

(i) **ROUGE-L (R-L)** (Lin, 2004), which is an automated evaluation measure in natural language processing used to assess the quality of machine-generated text summaries by measuring the longest common subsequence between the generated summary and the reference text.

(ii) **QAFactEval (QAFE)** (Fabbri et al., 2022), which combines entailment and question answering based metrics to capture their complementary signals and further boost the performance.

(iii) **QAGS** (Wang et al., 2020), which first generates several questions based on the generated summary with a Question Generation (QG) model, and then generates two sets of corresponding answers given the source document and the summary with a Question Answering (QA) model. Finally, the QAGS score is computed by comparing these answers with token-level similarity metrics.

(iv) **FactCC (FCC)** (Kryscinski et al., 2020), which is based on a weakly-supervised BERT-based model to measure whether the summary is entailed by the source document.

(v) **Fact Triple (FT-C/O)** (Goodrich et al., 2019), which extract fact triples (*subject*, *relation*, *object*) separately from the source document and the summary and compare these two sets of triples. Among them, FT-C is in a closed scheme, where *relation* is predicted from a pre-defined relation set; FT-O is in an open scheme, where *relation* is the original text span between *subject* and *object*.

(vi) **CoCo** (Xie et al., 2021), which evaluates the factual consistency in text summarization via counterfactual estimation.

(vii) **AVG**, which first calculates the average score over all the factual metrics, and then averages it with the traditional metric R-L for a clear comparison of the trade-off between the traditional and factual metrics.

C.4 Baselines

To validate the effectiveness of our CoFactSum, we select several baselines that have demonstrated superior performance in text summarization tasks over the years. In particular, we adopt the state-of-the-art **Pegasus** (Zhang et al., 2020) as our model backbone, and mainly choose the following counterparts to compare with:

(i) **Pegasus** (Zhang et al., 2020), which employs a denoising autoencoder architecture to generate coherent and contextually accurate summaries from input documents.

(ii) **Unl** (Li et al., 2020), which leverages the unlikelihood loss to penalize the probabilities of the tokens in unfaithful samples.

(iii) **Corr** (Cao et al., 2020), which pre-trains a post-editing corrector model to directly generate factually consistent summaries.

(iv) **CCGS** (Chen et al., 2021), which pre-trains a factual consistency predictor and leverages it to rank candidate summaries.

(v) **CLIFF** (Cao and Wang, 2021), which adopts contrastive loss to discriminate between faithful and unfaithful samples.

(vi) **SC** (Xiao and Carenini, 2022), which contains an entity-based SpanCopy mechanism with Global Relevance to reduce mismatched entities.

D Human Evaluation Results

The results of human evaluations on the informativeness of the generated summaries are shown in Table 8, which show that CoFactSum are competitive with baseline methods, indicating the proposed CoFactSum achieves a great balance between informativeness and factual consistency.

Methods	CNN/DM			XSum		
	Win↑	**Tie**	**Lose↓**	**Win↑**	**Tie**	**Lose↓**
Unl	10.33	59.33	30.34	18.67	53.33	28.00
Corr	12.67	67.00	20.33	2.33	95.67	**2.00**
CCGS	4.00	93.67	**2.33**	2.67	88.33	9.00
CLIFF	10.67	65.00	24.33	10.00	65.33	24.67
SC	**14.33**	66.67	19.00	12.33	66.33	21.34
CoFactSum	8.33	83.00	8.67	**22.00**	65.67	12.33

Table 8: Pairwise human evaluation results (%) in terms of **informativeness** compared with Pegasus.