Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation

Yi Liu[†] **Xiangyu Liu**[†] **Xiangrong Zhu**[†] **Wei Hu**[†,‡,*]
[†] State Key Laboratory for Novel Software Technology, Nanjing University, China
[‡] National Institute of Healthcare Data Science, Nanjing University, China
{yiliu07, xyl, xrzhu}.nju@gmail.com, whu@nju.edu.cn

Abstract

Multi-aspect controllable text generation aims to control the generated texts in attributes from multiple aspects (e.g., "positive" from *sentiment* and "sport" from *topic*). For ease of obtaining training samples, existing works neglect attribute correlations formed by the intertwining of different attributes. Particularly, the stereotype formed by imbalanced attribute correlations significantly affects multi-aspect control. In this paper, we propose MAGIC, a new **m**ulti-**a**spect controllable text **g**eneration method with d**i**sentangled **c**ounterfactual augmentation. We alleviate the issue of imbalanced attribute correlations during training using counterfactual feature vectors in the attribute latent space by disentanglement. During inference, we enhance attribute correlations by target-guided counterfactual augmentation to further improve multi-aspect control. Experiments show that MAGIC outperforms state-of-the-art baselines in both imbalanced and balanced attribute correlation scenarios. Our source code and data are available at https://github.com/nju-websoft/MAGIC.

1 Introduction

Controllable text generation (CTG) aims to generate texts adhering to given constraints reliably. The development of generative AI based on large language models (LLMs) draws increasing attention to CTG (Keskar et al., 2019; Brown et al., 2020; Zhang et al., 2022). Due to the demand for diverse attribute control, recent studies focus on a more practical and challenging setting, *multi-aspect controllable text generation* (MCTG). Different kinds of methods have been proposed(Gu et al., 2022b), including weighted decoding (Dathathri et al., 2020; Krause et al., 2021), optimization in the language space (Kumar et al., 2021; Mireshghallah et al., 2022), optimization in the latent semantic space (Gu et al., 2022b, 2023; Ding et al., 2023; Liu et al., 2023), etc.

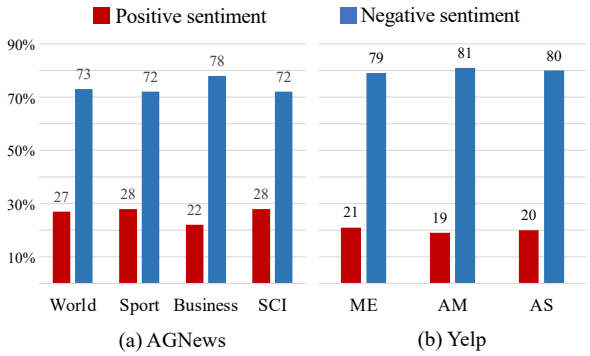

Figure 1: The relevance scores of positive and negative sentiment in (a) AGNews and (b) Yelp. (a) The classifiers used for statistics are from (Gu et al., 2022b). (b) The statistical data of Yelp are from (Yang et al., 2023).

Due to the difficulty of directly obtaining training data that satisfy arbitrary attribute combinations, existing methods (Krause et al., 2021; Ding et al., 2023; Gu et al., 2023) reuse datasets with single-aspect annotations for MCTG, where each training sample only expresses a single attribute in one aspect. This neglects the fact that a sentence often couples multiple attributes due to the complexity of natural language. The co-occurrence of attributes within one sentence forms patterns corresponding to attribute correlations, serving as crucial dependencies for a generative model in inference. Meanwhile, the training corpus is derived from real life, where preferences in real life make certain combinations of attributes more common, leading to an imbalance in attribute correlations. As an example shown in Figure 1, in the AGNews dataset, since news with the topics of "world" and "business" are prone to correlating with negative elements (Gu et al., 2022b), such as war or inflation, combinations of these topics and negative sentiment are more prevalent. In the Yelp dataset consisting of restaurant reviews with sentiment and

* Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9231–9253

food types, negative reviews also dominate (Yang et al., 2023). The imbalance in attribute correlations can lead the model to associate specific attributes, forming a stereotype that impacts multi-aspect control. An MCTG model can better fit attributes with higher co-occurrence frequencies, allowing it to learn the semantic information of these attributes more comprehensively. However, the model may neglect the learning of attributes with low co-occurrence frequencies, which hurts the control of these attribute combinations.

To resolve the problem, we propose a **m**ulti-**a**spect controllable text **g**eneration method with d**i**sentangled **c**ounterfactual augmentation, called MAGIC. Specifically, we introduce attribute disentanglement with latent space optimization. It can disentangle the control factors of different attributes in the texts and generate the latent vectors with counterfactual features in the attribute latent space. During training, we employ counterfactual latent vectors to balance attribute correlations, thereby constructing a more semantically balanced attribute latent space. During inference, we enhance attribute correlations by the counterfactual latent vectors to improve multi-aspect control.

We experiment on three-aspect control including *sentiment*, *topic*, and *detoxification*. We evaluate the relevance scores of attributes and assess the text quality in the scenarios of imbalanced and balanced attribute correlations. The results indicate that MAGIC can leverage attribute correlations and mitigate the imbalance issues, which leads to steady and superior performance in both imbalanced and balanced scenarios than state-of-the-art baselines on multi-aspect control. We further demonstrate the effectiveness of each module in MAGIC through analytical experiments.

Our main contributions are outlined as follows:

- To mitigate the issue of imbalanced attribute correlations, we propose a counterfactual feature augmentation model with attribute disentanglement.
- To improve multi-aspect control by leveraging attribute correlations, we introduce a text generation method based on target-guided attribute correlation augmentation.
- We experimentally validate the effectiveness of MAGIC. It outperforms state-of-the-art baselines on the imbalanced and balanced settings of multi-aspect control.

2 Related Work

Controllable text generation. LLMs introduce new perspectives for controllable text generation, such as post-processing (Krause et al., 2021; Liu et al., 2021) and prefix tuning (Qian et al., 2022; Li and Liang, 2021; Yu et al., 2021). In contrast to single-aspect control, multi-aspect control has garnered increasing attention. Current methods on MCTG can be broadly classified into three categories. (i) *Weighted decoding* is a kind of method that biases the output token distribution during decoding to achieve controllable generation (Dathathri et al., 2020; Yang and Klein, 2021; Liu et al., 2021; Krause et al., 2021; Gu et al., 2022a). (ii) The methods of *optimization in the language space* model the generation of tokens satisfying multi-aspect requirements as a multi-objective optimization problem (Mireshghallah et al., 2022; Qin et al., 2022; Kumar et al., 2021). (iii) The prefix tuning methods of *optimization in the latent semantic space* have shown significant effectiveness in achieving multi-aspect control (Gu et al., 2022b; Liu et al., 2023; Yang et al., 2023; Huang et al., 2023; Gu et al., 2023). However, these methods rely highly on the training data to construct the latent space and do not consider the influence of attribute correlations.

Counterfactual augmentation. Counterfactuals are designed to study the change in a response variable following an intervention. Counterfactual augmentation is employed to enhance the robustness of models against the spurious correlations (Howard et al., 2022) including manual and automatic solutions. The manual solution is a human-in-the-loop method to generate counterfactual texts by human annotators, which is costly and time-consuming (Kaushik et al., 2020). For the automatic solution, some methods get counterfactual texts by fine-tuning LLMs (Wu et al., 2021; Yang et al., 2021; Paranjape et al., 2022). Some methods propose controllable text generation approaches based on weighted decoding (Madaan et al., 2021) or a structural causal model (Hu and Li, 2021). The above methods enhance the training set by generating counterfactual texts. Following the above idea, we apply counterfactual augmentation to multi-aspect control and generate latent vectors with counterfactual features in the latent space to mitigate the impact of imbalanced attribute correlations. We also leverage the attribute correlations promoted by counterfactuals to improve multi-aspect control.

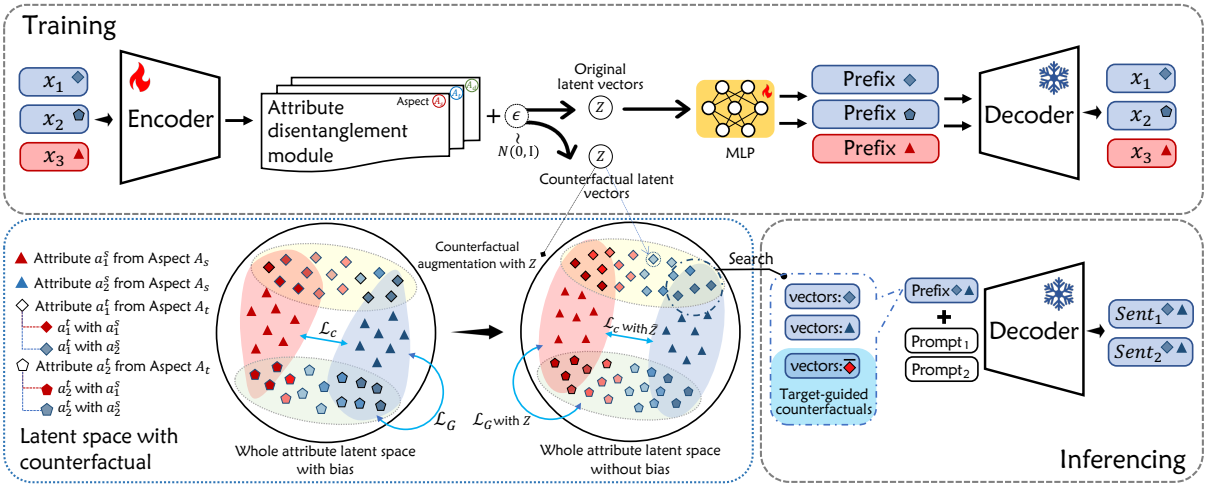

Figure 2: Framework of our method. Top part: We use the prefix tuning-based autoencoder structure as the framework and construct the attribute latent space. Bottom left: The vectors with counterfactual attribute features generated by the attribute disentanglement module are assisted in the construction of the attribute latent space. Bottom right:Inference stage with target-guided attribute correlation augmentation to improve multi-aspect control.

3 Formulation

Task definition. Let $\mathbf{A} = \{A_1, \ldots, A_N\}$ be N aspects. Each aspect $A_t \in \mathbf{A}$ contains $|A_t|$ mutually exclusive attributes $\{a_1^t, \ldots, a_{|A_t|}^t\}$. The goal of MCTG is to generate sentences possessing multiple attributes from different aspects simultaneously. For example, we may ask an MCTG model to generate texts with attribute "sport" from aspect *topic*, attribute "positive" from aspect *sentiment*, and attribute "non-toxic" from aspect *detoxification*.

Training corpus. The training samples are indexed according to their associated aspects and attribute labels. We denote the indices of all sentences with any attribute in aspect A_t by I^t. The indices of the entire training data are $I = \bigcup_{t=1}^{N} I^t$. Furthermore, $I_{a_\mu^t}^t$ is the indices of all training sentences about attribute a_μ^t in aspect A_t, and we have $I^t = \bigcup_{\mu=1}^{|A_t|} I_{a_\mu^t}^t$. Following (Ma et al., 2023), we introduce the concepts of explicit and implicit attributes to facilitate the explanation of our method. For the data indexed by $I_{a_\mu^t}^t$, a_μ^t is the explicit attribute with annotated labels, and A_t is the corresponding explicit aspect. Other potential attributes in the data are the implicit attributes from implicit aspects, which are not explicitly provided but annotated by extra attribute classifiers in our work. For notation, $I_{a_1^t, a_1^k}^t$ are the indices of data with the explicit attribute a_1^t and the implicit attribute a_1^k.

Attribute correlation imbalance. Frequently co-occurred attributes tend to exhibit attribute correlations. Once the co-occurrence frequency of an attribute pair significantly exceeds others, it exhibits imbalanced attribute correlations. Given $I_{a_\mu^t}^t$, suppose that we have $I_{a_\mu^t}^t = I_{a_\mu^t, a_\sigma^k}^t + I_{a_\mu^t, a_{\bar{\sigma}}^k}^t$, where a_σ^k and $a_{\bar{\sigma}}^k$ are two mutually exclusive implicit attributes from aspect A_k. The attribute correlation becomes imbalanced when $\left|I_{a_\mu^t, a_\sigma^k}^t\right| \gg \left|I_{a_\mu^t, a_{\bar{\sigma}}^k}^t\right|$.

Counterfactual samples. Generally, a counterfactual sample is defined as a synthetically generated sentence that is treated differently by a condition model (Madaan et al., 2021). Given a training sample x with explicit attribute a_μ^t from aspect A_t and an implicit attribute a_σ^k from aspect A_k, we define the synthetically generated sample $\bar{x}$ with the same explicit attribute a_μ^t and reversed implicit attribute $a_{\bar{\sigma}}^k$ as a counterfactual sample to x.

4 Methodology

The structure of our MAGIC is illustrated in Figure 2. Our MAGIC employs an encoder-decoder framework based on prefix tuning to avoid the training costs in fine-tuning LLM. We implement multi-aspect control revolving around the attribute latent space. The encoder projects sampled sentences into the attribute latent space. Several constraints are used to accurately model the attribute semantics. We conduct resampling and incorporate latent vectors with counterfactual features to assist the construction of the attribute latent space and avoid the impact caused by the imbalanced attribute correlations. The vectors with counterfactual features

are generated by our designed attribute disentanglement module. We also adopt a target-guided attribute correlation enhanced generation strategy to further improve multi-aspect control.

4.1 Attribute Space Building Against Biases

To facilitate multi-aspect control, we aim to construct an attribute latent space that models the semantics and relationships of attributes. First, we need to establish a mapping between the attribute latent space and the attributes in sentences. Various methods can be employed, such as VAE (Liu et al., 2023; Ding et al., 2023). Following (Gu et al., 2022b), we adopt a basic and simple method to map the attributes of sentences to discrete samples in the latent space. Specifically, we leverage an encoder to extract the semantic features $\mathcal{H}_i$ in a sentence x_i: $\mathcal{H}_i = \text{Encoder}(x_i)$. Then, we get latent vector Z_i in the attribute latent space based on $\mathcal{H}_i$ through attribute disentanglement (described in Section 4.2). We compute the prefix vector Prefix_i based on Z_i in the attribute latent space as follows:

$$\text{Prefix}_i = \text{MLP}(Z_i + \lambda\varepsilon), \tag{1}$$

where λ is a scaling factor and ε is a perturbation vector from a multivariate Gaussian distribution for robustness. The prefix is used to reconstruct the sentence x_i and recover the corresponding attribute in the same way as an autoregressive loss:

$$\mathcal{L}_{Rec} = -\sum_{i \in I} \log p(x_i \mid \text{Prefix}_i). \tag{2}$$

To accurately model attribute information, two kinds of constraints are utilized in modeling the attribute latent space (Gu et al., 2022b; Ding et al., 2023): (i) *Classification loss* enables the differentiation of different attributes from the same aspect in the attribute latent space as follows:

$$\mathcal{L}_C = -\sum_{t=1}^{|\mathbf{A}|}\sum_{\mu=1}^{|A_t|}\sum_{i \in I^t_{a^t_\mu}} \log p_{\pi_t}\left(a^t_\mu \mid Z_i\right), \tag{3}$$

where p_{π_t} is a classifier to distinguish attribute a^t_μ among aspect A_t. (ii) *Aspect gap loss* aims to penalize the discrepancy of aspects caused by the domain gap among data sources and facilitate the expression of multi-aspect semantics in the attribute latent space:

$$\mathcal{L}_G = \sum_{1 \le t_1 < t_2 \le |\mathbf{A}|} \left\| \sum_{i \in I^{t_1}} \frac{Z_i}{|I^{t_1}|} - \sum_{j \in I^{t_2}} \frac{Z_j}{|I^{t_2}|} \right\|_2^2, \tag{4}$$

where $\|\cdot\|_2^2$ calculates the Euclidean distance based on the L2-norm.

When faced with imbalanced attribute correlations, samples of more frequently co-occurred attribute combinations are more likely to be selected during training. The model is more prone to learning the semantics associated with these attribute combinations. Thus, we first adopt a resampling strategy to increase the probability of sampling the sentences with low-frequent attribute combinations. Specifically, when training with the data corresponding to each topic, for each sampled training example x_i, we simultaneously resample another example with the opposite sentiment to x_i. Furthermore, we also use the attribute disentanglement to generate latent vectors with counterfactual features to construct a more balanced attribute latent space. For aspect A_t with imbalanced attribute correlations, the classification loss with counterfactual augmentation becomes

$$\mathcal{L}_C^{A_t} = -\sum_{\mu=1}^{|A_t|}\sum_{i \in I^t_{a^t_\mu}} \log\Big(p_{\pi_t}(a^t_\mu \mid Z_i) p_{\pi_t}(a^t_\mu \mid \bar{Z}_i)\Big), \tag{5}$$

and the aspect gap loss with counterfactual augmentation becomes

$$\mathcal{L}_G^{A_t} = \sum_{\substack{1 \le t_1 \le |A| \\ t_1 \ne t}} \left\| \sum_{i \in I^t} \frac{Z_i + \bar{Z}_i}{2 \times |I^t|} - \sum_{j \in I^{t_1}} \frac{Z_j}{|I^{t_1}|} \right\|_2^2, \tag{6}$$

where $\bar{Z}_i$ is the latent vector with counterfactual features generated by the attribute disentanglement (described in Section 4.2).

4.2 Attribute Disentanglement

In this section, we design an attribute disentanglement module to decouple the explicit and implicit attributes in sentences. Based on this, we can generate the latent vector Z_i of the original sample x_i and $\bar{Z}_i$ of the counterfactual sample $\bar{x}_i$ in the attribute latent space. By transferring the shared implicit attribute features across the data with different explicit attributes, we can supplement the insufficient implicit attribute information in the data corresponding to each explicit attribute, caused by the low attribute co-occurrence frequency.

Figure 3 provides an overall description. Specifically, given x_i with explicit attribute a^t_μ and implicit attribute a^s_σ, we refer to the latent vectors decoupled into the subspaces of explicit and implicit attributes as the respective explicit and implicit

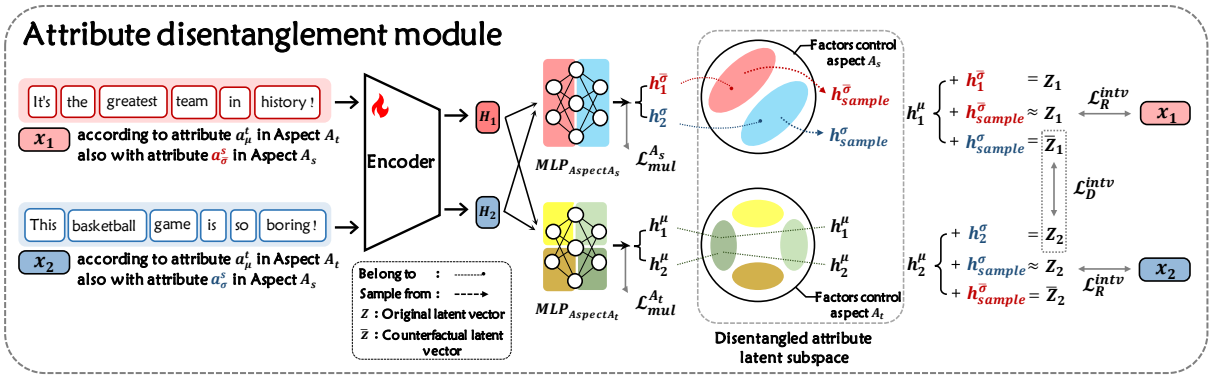

Figure 3: The attribute disentanglement module. A_t and A_s denote the explicit and implicit aspects, respectively.

attribute control factors: h_i^μ and h_i^σ. Assuming $h_i^\mu = \text{MLP}\left(\mathcal{H}_i\right)$, h_i^σ are calculated similar to h_i^μ but with a different MLP, and $Z_i = h_i^\mu + h_i^\sigma$.

To implement the disentanglement of implicit and explicit attributes, we introduce three kinds of losses. The first loss is *Multi-task loss* which ensures that the control factors with attribute features extracted by explicit or implicit extractors can be discriminative for their respective attributes (John et al., 2019). Given the sentence x_i needing disentanglement, the multi-task loss for each explicit or implicit aspect A_* involved in the disentanglement is calculated like

$$\mathcal{L}_{mul}^{A_*} = -\sum_{\beta=1}^{|A_*|} \sum_{i \in I_{a_\beta^*}^*} \log p_{\pi_*}\left(a_\beta^* \mid h_i^\beta\right), \quad (7)$$

where p_{π_*} is the classifier to distinguish attribute a_β^* among aspect A_*, h_i^β is the control factor of attribute a_β^*.

However, this cannot constrain the mutual influence between the control factors of explicit and implicit attributes. Thus, we introduce two *intervention losses*. $\mathcal{L}_R^{intv}$ aims to eliminate the interference of the implicit attribute control factor on the explicit attribute and transfer the shared features of implicit attributes across different explicit attribute data. Specifically, given a sentence x_i with explicit attribute a_μ^t and implicit attribute a_σ^s, the respective control factors are h_i^μ and h_i^σ. We first sample a sentence x_{sample} with the same implicit attribute but a different explicit attribute compared to x_i, and denote its implicit attribute control factor by h_{sample}^σ. We combine the explicit control factor h_i^μ with h_{sample}^σ to get the prefix:

$$\text{Prefix}_i^{intv} = \text{MLP}\left(h_i^\mu + h_{\text{sample}}^\sigma\right). \quad (8)$$

We ask the prefix to reconstruct a new sentence similar to x_i (with the same explicit and implicit attributes as x_i). The loss function is

$$\mathcal{L}_R^{intv} = -\sum_{a_\mu^t \in A_t} \sum_{i \in I_{a_\mu^t}^t} \log p_{LM}\left(x_i \mid \text{Prefix}_i^{intv}\right). \quad (9)$$

This means that the implicit attribute control factors disentangled from any sentences with different explicit attributes do not affect the control of explicit attributes. Meanwhile, the shared implicit attribute features in all sentences can be utilized for the reconstruction of each sentence without intervening in the explicit attributes.

However, relying solely on the above two losses is still insufficient. Without constraints, the explicit attribute control factor may still interfere with the implicit attributes. Thus, we design the second loss, $\mathcal{L}_D^{intv}$, which aims to constrain the influence of explicit (e.g., topic) control factors on implicit attributes (e.g., sentiment). Specifically, given a sentence x_i with explicit attribute a_μ^t and implicit attribute a_σ^s, the respective control factors are h_i^μ and h_i^σ. We sample a sentence x_{sample} with a different implicit attribute $a_{\bar{\sigma}}^s$ from x_i and denote its implicit attribute control factor by $h_{sample}^{\bar{\sigma}}$. Then, we combine h_i^μ with $h_{sample}^{\bar{\sigma}}$ and instruct the model to reconstruct a sentence $\bar{x}_i$, which has the explicit attribute a_μ^t but changes the implicit attribute to $a_{\bar{\sigma}}^s$. This means that we need to know what sentence $\bar{x}_i$ looks like after the change in sentiment. Since we do not have textual annotations for $\bar{x}_i$ and observe a tendency for text with similar attributes to exhibit cohesion in the attribute space, we attempt to enforce the constraints in the attribute space. Thus, we denote the vector of $\bar{x}_i$ in the attribute latent space by $\bar{Z}_i$, and bring $\bar{Z}_i$ closer to the set of vectors that share the same explicit and implicit

attributes. The loss function is

$$\mathcal{L}_D^{intv} = \sum_{a_\mu^t \in A_t} \sum_{i \in I_{a_\mu^t}^t} \max\Big(d\big(\bar{Z}_i, \hat{Z}_i\big) - \gamma, 0\Big),$$
$$\bar{Z}_i = h_i^\mu + h_{\text{sample}}^{\bar{\sigma}}, \tag{10}$$
$$\hat{Z}_i = \frac{1}{\left|I_{a_\mu^t, a_{\bar{\sigma}}^s}^t\right|} \sum_{j \in I_{a_\mu^t, a_{\bar{\sigma}}^s}^t} h_j^\mu + h_j^{\bar{\sigma}}.$$

If the explicit attribute control factor affects the implicit attribute, it would impact the modification of implicit attributes after replacing the implicit attribute control factor. We constrain the potential impact through the loss $\mathcal{L}_D^{intv}$, while ensuring the function of the implicit attribute control factor.

4.3 Multi-Aspect Generation

Suppose the target combination of attributes A_{target} is $\{a_{\varphi_1}^1, \ldots, a_{\varphi_N}^N\}$ from N different aspects, where $a_{\varphi_t}^t$ is the φ_t-th attribute from aspect A_t. We implement multi-aspect control revolving around the attribute latent space. Since the attribute correlations benefit the control of the corresponding attribute combinations, we adopt a multi-aspect generation strategy with target-guided attribute correlation augmentation. Specifically, we first utilize the control factors of explicit and implicit attributes from attribute disentanglement to generate latent vectors in the attribute space aligning with the target attribute combinations. Then, we use an iterative intersection retrieval algorithm in the attribute latent space to get the latent vector that simultaneously satisfies the target attributes following (Gu et al., 2022b). For each target attribute a_μ^t, we identify the top K vectors among the set of vectors corresponding to a_μ^t, which are closest to the vectors corresponding to other attributes in A_{target}. The mean value of the top K vectors is used to represent the corresponding attribute. We calculate the weighted sum of each representative vector to obtain the target vector:

$$\tilde{Z} = \sum_{a_\mu^t \in A_{\text{target}}} w_{a_\mu^t} \times \text{mean}\Big(Z_i, i \in \text{N}_{\text{topK}}\left(I_{a_\mu^t}^t\right)\Big), \tag{11}$$

where $\text{mean}(\cdot)$ is the mean operation within the set, $\text{N}_{\text{topK}}\left(I_{a_\mu^t}^t\right)$ are the indices of the top K vectors for the current attribute a_μ^t and are closest to the vectors of other attributes under control. We use $\tilde{Z}$ to get the prefix by Eq. 1 and generate the target sentence based on the prefix and prompt $\tilde{x}$:

$$Y = \arg\max_y p_{\text{LM}}\big(y \mid \text{Prefix}; \tilde{x}\big). \tag{12}$$

5 Experiments and Results

5.1 Experiment Setup

Datasets. We experiment with three-aspect control: *sentiment*, *topic*, and *detoxification*. Following previous works (Krause et al., 2021; Gu et al., 2022b, 2023), we pick IMDb for *sentiment*, AGNews for *topic*, and the Jigsaw Toxic Comment Classification Challenge dataset for *detoxification*. We simulate the imbalanced setting on the AGNews dataset. For each topic in AGNews, the proportions of sentences with negative and positive sentiment are set to 7:3. For the balanced setting, we reuse the dataset in (Gu et al., 2022b). More details are provided in Appendix B.1.

Baselines. We compare MAGIC to 8 representative and strong baselines. (i) *Weighted decoding*: PPLM (Dathathri et al., 2020) and GeDi (Krause et al., 2021) bias the decoding process in generation. (ii) *Optimization in the language space*: Mix&Match (Mireshghallah et al., 2022) discretely optimizes sentences in the language space by token-level masking. (iii) *Optimization in the latent space*: Tailor (Yang et al., 2023) is based on soft prompt-tuning. Discrete (Gu et al., 2022b) uses discrete samples to construct the attribute latent space. LatentOPs (Liu et al., 2023) adopts an efficient sampler based on ordinary differential equations (ODEs). MacLaSa (Ding et al., 2023) combines VAE and ODEs for the generation. PriorControl (Gu et al., 2023) utilizes the normalizing flow to constrain the complex latent space. See Appendix A for implementation details.

Evaluation metrics. We compute the attribute relevance with the DeBERTa classifiers for *sentiment* and *topic* aspects. We measure the *non-toxicity* aspect with the Google Perspective API. We consider two auxiliary metrics for text quality, i.e., perplexity (abbr. PPL) and distinctness. More details are provided in Appendix B.2. We also conduct human evaluations on the generated sentences. Details and results are provided in Appendix D.

5.2 Main Results

We conduct experiments on both imbalanced and balanced attribute correlation settings. Table 1 lists the results. We report the average scores with standard deviations of 8 combinations for each aspect, as well as the average scores for all three aspects.

Overall, most methods perform worse in the imbalanced setting compared to the balanced one.

	Methods	Avg. ↑ (%)	Sentiment ↑ (%)	Topic ↑ (%)	Detoxification ↑ (%)	PPL ↓	Distinct ↑
Imbalanced attribute correlations	PPLM	70.7 ± 24.9	63.6 ± 28.7	61.8 ± 25.9	86.9 ± 9.5	69.8	60.2
	GeDi	82.3 ± 18.6	73.5 ± 23.1	77.8 ± 16.9	95.5 ± 2.6	92.2	78.2
	Mix&Match	77.7 ± 22.7	72.5 ± 27.8	68.7 ± 23.6	91.8 ± 2.5	73.9	59.3
	Tailor	76.9 ± 24.9	67.5 ± 31.3	66.7 ± 19.8	**96.4** ± 1.9	26.8	69.8
	LatentOPs	82.8 ± 16.2	78.1 ± 20.3	78.2 ± 15.4	92.1 ± 8.2	11.7	39.7
	Discrete	83.8 ± 20.7	91.2 ± 15.6	65.5 ± 23.9	94.8 ± 3.6	43.1	42.1
	MacLaSa	84.7 ± 13.9	82.4 ± 13.7	77.9 ± 16.8	93.9 ± 3.3	29.3	59.7
	PriorControl	86.2 ± 13.6	88.1 ± 10.3	78.4 ± 19.2	92.1 ± 4.2	34.1	51.8
	MAGIC (ours)	**92.6** ± 9.1	**94.5** ± 6.9	**88.5** ± 13.4	94.7 ± 3.9	43.4	53.3
Balanced attribute correlations	PPLM	71.0 ± 21.4	64.7 ± 24.8	63.5 ± 22.7	84.9 ± 6.5	62.6	62.0
	GeDi	81.4 ± 14.7	76.1 ± 17.2	73.8 ± 11.3	94.2 ± 1.9	116.6	75.1
	Mix&Match	79.7 ± 21.8	73.5 ± 25.9	69.9 ± 21.1	95.8 ± 1.9	63.0	61.8
	Tailor	78.1 ± 22.6	64.6 ± 28.5	73.7 ± 16.5	**95.9** ± 2.5	28.7	69.8
	LatentOPs	85.5 ± 14.4	76.3 ± 16.4	85.1 ± 14.1	94.9 ± 4.2	16.8	41.3
	Discrete	87.4 ± 10.9	86.7 ± 10.5	84.8 ± 14.2	90.7 ± 7.4	28.4	49.5
	MacLaSa	88.2 ± 10.7	85.0 ± 14.7	85.1 ± 9.5	94.5 ± 2.6	19.2	56.5
	PriorControl	92.2 ± 8.6	92.5 ± 8.5	89.3 ± 11.0	94.9 ± 3.4	29.6	51.6
	MAGIC (ours)	**92.9** ± 8.5	**94.2** ± 6.4	**89.4** ± 12.2	95.1 ± 4.9	55.3	52.2

Table 1: Automatic results of multi-aspect control with imbalance and balance attribute correlation. The best relevance scores are marked in **bold**. More results are shown in Tables 8 and 9 in the appendix.

Variants (strategies during training)	Avg.	Sent.	Topic	Detox.
MAGIC (intact)	92.6	94.5	88.5	94.7
w/o counterfactual (Eqs. 5 and 6)	90.5	91.5	86.1	94.1
w/o resample strategies	88.5	88.9	83.1	93.7
w/o $\mathcal{L}_C$ (Eq. 3)	85.9	89.0	76.5	92.1
w/o $\mathcal{L}_G$ (Eq. 4)	88.6	90.6	81.0	94.3
w/o $\mathcal{L}_{mul}^{A_*}$ (Eq. 7)	91.3	91.7	86.5	95.6
w/o $\mathcal{L}_R^{intv}$ (Eq. 9)	84.6	79.1	80.7	94.1
w/o $\mathcal{L}_D^{intv}$ (Eq. 10)	83.0	81.1	75.5	92.5
Variants (strategies during inference)	**Avg.**	**Sent.**	**Topic**	**Detox.**
MAGIC (intact)	92.6	94.5	88.5	94.7
w/o all	86.9	85.8	79.1	95.8
w/ balance	87.8	86.9	82.5	93.9

Table 2: Analysis of different strategies.

Topics	Change factor of sent. from				Avg.
	World	Sport	Business	Tech	
World	90.7	89.2	74.9	86.1	83.4
Sport	97.9	98.7	97.9	98.2	98.0
Business	84.3	84.6	87.8	86.3	85.1
Tech	98.6	99.3	99.4	99.5	99.1

Table 3: Relevance scores of topic after changing the control factor of sentiment from different topics.

This is due to the dominant negative impact of stereotypes formed by imbalanced attribute correlations. The stereotypes hinder the classifiers used to optimize in the language space for Mix&Match. The lack of data with positive sentiment combinations also affects the learning of semantics related to positive sentiment for each topic in the attribute latent space, such as Discrete and PriorControl.

Due to the strategies used in training, MAGIC is less affected by the imbalanced attribute correlations. In the imbalanced setting, MAGIC performs best on average attribute-related metrics, showing a 7.4% improvement beyond the second-best method PriorControl. The advances come from the improvement of *topic* (12.8%) and *sentiment* (7.2%) aspects. The target-guided counterfactual augmentation makes MAGIC achieve better performance. Thus, in the balanced setting, MAGIC can also achieve comparable performance with PriorControl (1.8% improvement in the *sentiment* aspect).

In addition, GeDi performs well on attribute relevance and diversity, while badly on perplexity. MAGIC exhibits a slightly higher PPL but still in a reasonable range compared with GeDi and Mix&Match. The disentanglement module tends to bring the latent vectors with similar features closer within one aspect, which can increase the distances between different aspects, potentially affecting the distribution of the normal attribute space.

5.3 Further Analysis

Effects of different strategies. We validate the effects of strategies during training. Table 2 lists the results. After removing counterfactual augmentation, all relevance scores of our MAGIC decrease. We further remove the resampling strategy and also observe a decrease in performance. We find that the counterfactual augmentation and resampling both make sense for the construction of attribute latent space. The loss functions of $\mathcal{L}_C$ (Eq. 3) and $\mathcal{L}_G$ (Eq. 4) are commonly used in previous works (Gu et al., 2022b; Ding et al., 2023) and both contribute

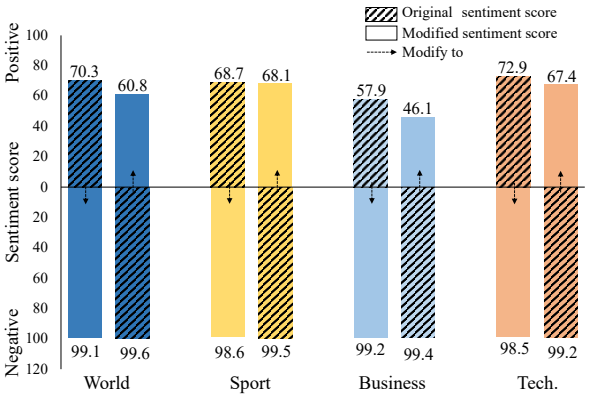

Figure 4: Relevance scores of *sentiment* after changing the control factor of sentiment to the opposite.

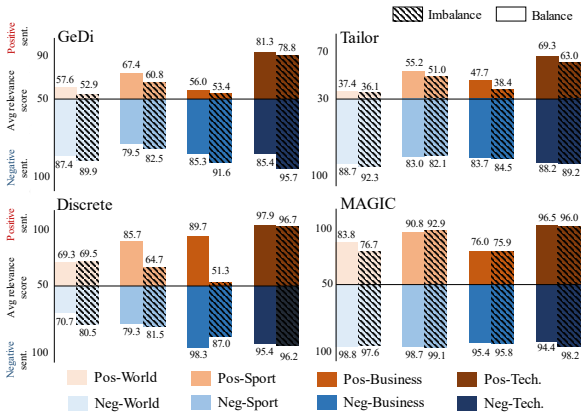

Figure 5: Effects of attribute correlation imbalance in performance with different attribute combinations.

to the performance. Our proposed loss functions $\mathcal{L}_{mul}^{A_*}$, $\mathcal{L}_{R}^{intv}$, and $\mathcal{L}_{D}^{intv}$ affect the performance by influencing disentanglement. After removing each of them, the decline in the effect of disentanglement also results in a decrease in performance.

In inference, we use a target-guided attribute correlation augmentation strategy to improve multi-aspect control, which generates new latent vectors in the attribute latent space aligning with the attribute combinations that we control. We compare three variant strategies in Table 2. "w/o all" is the variant without any augmentation strategy, which performs worst. "w/ balance" maintains an equal number of vectors for different attribute combinations in the attribute latent space. We can find that the target-guided strategy used in MAGIC is more effective than the balanced variant.

Analysis of attribute disentanglement. We perform experiments about attribute disentanglement. First, we measure the impact of the disentangled implicit attribute (*sentiment*) control factors on the explicit attribute (*topic*). Table 3 lists the results. For each topic in each row of the table, we generate texts by replacing the disentangled sentiment control factor with those corresponding to the other three topics. The diagonal of the table represents generated texts using the original topic and sentiment control factors without replacement. We measure the topic relevance scores of the generated texts. It can be observed that, after replacing the sentiment factors with those from other topics, the topic relevance scores only show a slight decrease.

Next, we validate whether the disentangled explicit control factor (*topic*) interferes with the implicit control factor in controlling implicit attributes (*sentiment*). The results are shown in Figure 4. For each topic, the hatched bar chart represents the sentiment scores of the texts generated by the sentences with positive/negative sentiments in that topic (denoted by the original sentiment score). The solid bar chart, opposite to the hatched bar chart, represents the sentiment scores of sentences generated after replacing the sentiment control factors with the opposite sentiment of the original sentences (denoted by the modified sentiment score). Since we train the disentanglement module based on existing sentiment labels in each topic, the original sentiment scores serve as an upper bound for the modified sentiment scores of the given sentiment. We can observe that for all topics, the original sentiment scores for negative sentiment are higher compared to positive sentiment. This is due to the biases formed by the decoder during pre-training. In addition, it is easier to convert positive sentiment into negative sentiment as more sentences with negative sentiments in the training set providing more supervision. The conversion of negative sentiment to positive sentiment is less effective, which we ascribe to the limited availability of supervision signals for positive sentiment.

Since we utilize $\mathcal{L}_{mul}^{A_*}$, $\mathcal{L}_{R}^{intv}$, and $\mathcal{L}_{D}^{intv}$ for attribute disentanglement, we also introduce ablation studies to investigate the impact of these loss functions. Details are provided in Appendix F.

Effects of attribute correlation imbalance. We analyze the effects of attribute correlation imbalance in performance with different attribute combinations. Figure 5 illustrates the performance of *topic* and *sentiment* combinations under balanced and imbalanced attribute correlations. For each method, the upper/lower side of the x-axis corresponds to the average attribute scores for the 4 com-

binations of positive/negative sentiment and 4 different topics. The hatched bar chart represents the scenario of imbalance attribute correlation. It can be observed that the performance of the majority of positive sentiment attribute combinations shows varying degrees of decline under the imbalanced setting. This is due to the limited samples of positive sentiment attribute combinations in the training data. Certain methods exhibit an improvement in the performance of negative sentiment attribute combinations in the imbalanced setting, such as GeDi and Discrete. MAGIC is least affected due to the utilization of balancing strategies during training and the target-guided counterfactual augmentation of generation during inference.

6 Conclusion

In this paper, we consider attribute correlation and propose a novel method, MAGIC, for multi-aspect control with disentangled counterfactual augmentation. We alleviate the issue of imbalanced attribute correlations during training and further improve multi-aspect control using counterfactual vectors in the attribute latent space by disentanglement. Experiments on the three-aspect control task support the effectiveness of MAGIC. We also conduct detailed analytical experiments to study the effects of each strategy in MAGIC. In the future, we will explore the impact of attribute correlations formed during pre-training.

Ethical Considerations

The training data used in our work are sourced from the web and have not undergone extensive data cleansing. As a result, the method that we propose and the baselines to which we compare may produce some fake, toxic, or offensive content. It is important to clarify that the generated texts in our work do not represent our viewpoints. Additionally, detoxification is considered as a default attribute that the generated texts are expected to satisfy. We believe that exploring controllable generation techniques is beneficial for combating the generation of harmful texts.

Limitations

MAGIC has several limitations: (i) To construct the attribute latent space, our method requires a substantial amount of training data, making it challenging to address the few-shot scenario. (ii) For attribute disentanglement, MAGIC needs an extra pre-trained classifier for labeling implicit attributes. The performance of this classifier could potentially impact the effectiveness of disentanglement. In the future, we will explore strategies to reduce reliance on classifiers for the disentanglement of control attributes. (iii) For a fair comparison, our decoder has a moderate parameter count, such as GPT2-medium. In the future, exploring more complex controllable generation tasks with a larger LLM would also be interesting.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (No. 62272219) and the Collaborative Innovation Center of Novel Software Technology & Industrialization.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901, Virtual.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *ICLR*, pages 1–34, Addis Ababa, Ethiopia.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Maclasa: Multi-aspect controllable text generation via efficient sampling from compact latent space. In *Findings of EMNLP*, pages 4424–4436, Singapore.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Jiaming Wu, Heng Gong, and Bing Qin. 2022a. Improving controllable text generation with position-aware weighted decoding. In *Findings of ACL*, pages 3449–3467, Dublin, Ireland.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022b. A distributional lens for multi-aspect controllable text generation. In *EMNLP*, pages 1023–1043, Abu Dhabi, United Arab Emirates.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. Controllable text generation via probability density estimation in the latent space. In *ACL*, pages 12590–12616, Toronto, Canada.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. NeuroCounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In *Findings of EMNLP*, pages 5056–5072, Abu Dhabi, United Arab Emirates.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. In *NeurIPS*, pages 24941–24955, Virtual.

Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun, and Yang Liu. 2023. An extensible plug-and-play method for multi-aspect controllable text generation. In *ACL*, pages 15233–15256, Toronto, Canada.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *ACL*, pages 424–434, Florence, Italy.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *ICLR*, pages 1–17, Addis Ababa, Ethiopia.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of EMNLP*, pages 4929–4952, Punta Cana, Dominican Republic.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraintss. In *NeurIPS*, pages 14542–14554, Virtual.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, pages 4582–4597, Virtual.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *ACL*, pages 6691–6706, Virtual.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2023. Composable text controls in latent space with odes. In *EMNLP*, pages 16543–16570, Singapore.

Congda Ma, Tianyu Zhao, Makoto Shing, Kei Sawada, and Manabu Okumura. 2023. Focused prefix tuning for controllable text generation. In *ACL*, pages 1116–1127, Toronto, Canada.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150, Portland, OR, USA.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *AAAI*, pages 13516–13524, Virtual.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generation using energy language models. In *ACL*, pages 401–415, Dublin, Ireland.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. Retrieval-guided counterfactual generation for QA. In *ACL*, pages 1670–1686, Dublin, Ireland.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of ACL*, pages 2912–2924, Dublin, Ireland.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. COLD decoding: Energy-based constrained text generation with langevin dynamics. In *NeurIPS*, pages 9538–9551, New Orleans, LA, USA.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *ACL*, pages 6707–6723, Virtual.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *NAACL*, pages 3511–3535, Virtual.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023. Tailor: A soft-prompt-based approach to attribute-based controlled text generation. In *ACL*, pages 410–427, Toronto, Canada.

Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In *ACL*, pages 306–316, Virtual.

Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pre-trained language models. In *Findings of EMNLP*, pages 2251–2268, Punta Cana, Dominican Republic.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NeurIPS*, pages 649–657, Montréal, Canada.

A Details of Hyperparameter Selection

We describe the hyperparameter selection of our method in this section. The implementation of the encoder and decoder follows the previous works (Gu et al., 2022b, 2023; Ding et al., 2023). The encoder is initialized with Bert-base-uncased and subsequently finetuned during training. The decoder

Methods	Avg. ↑ (%)	Sentiment ↑ (%)	Topic ↑ (%)	Detoxification ↑ (%)	PPL ↓	Distinct ↑
ChatGPT	92.5 ± 11.7	**97.4** ± 3.1	83.5 ± 17.1	**96.6** ± 2.6	18.7	53.9
MAGIC (ours)	**92.6** ± 9.1	94.5 ± 6.9	**88.5** ± 13.4	94.7 ± 3.9	43.4	53.3

Table 4: Automatic results on multi-aspect control compared with ChatGPT. More results are shown in Tables 5.

uses GPT2-medium[1] and fixed. Each sentence, after being encoded and mean-pooled, is converted to a 768-dimensional latent representation. The latent representations are mapped to the prefix with a dimension of (20, 24, 2, 1,024), where 20 is the prefix sequence length, 24 is the number of hidden layers in GPT2-medium, 2 represents one key and one value, and 1,024 is the size of hidden states in GPT2-medium. The dimension of the multi-layer perceptrons used in the attribute disentanglement is 768×768.

During the training stage, we use half-precision mode for efficiency on one NVIDIA A800 80GB GPU, where the batch size is 64. The scaling factors of $\mathcal{L}_{Rec}$, $\mathcal{L}_C$, $\mathcal{L}_G$, $\mathcal{L}_{mul}^{A_*}$, $\mathcal{L}_R^{intv}$, and $\mathcal{L}_D^{intv}$ are 0.5, 0.2, 0.3, 0.2, 0.5, and 0.2, respectively. The margin γ in $\mathcal{L}_D^{intv}$ is 0.4. The optimizer is AdamW with a learning rate of 1e-4, the number of training epochs is 300, and we use a checkpoint at a step of 57,000.

In the inference, we use the initial settings as (Gu et al., 2022b) for the iterative intersection retrieval algorithm. We employ grid search with for loops to search for the optimal weight parameters for each combination of attributes, which aims to balance the performance among attributes from different aspects. The search range of weights for sentiment is {1.5, 2.5, 3.5}, while the range for the topic is 2.0 to 10.5 with an interval of 1.5. The weight of detoxification is set to 1.0. The generation process is the same as prefix tuning and the length of the generated text is set to 50.

Since our method needs an extra attribute classifier for the implicit attributes, we use the DeBERTa-based model to train a classifier for sentiment. We sample 100k, 10k, and 10k sentiment-specific sentences from the Yelp dataset for training, validation, and testing, respectively. The learning rate is set to 5e-5 and the batch size is set to 64. The F1 score of the classifier on the testing set is 97.09.

Detailed settings of baselines are as follows: (i) For the weighted decoding method GeDi, we directly train the classifiers of three aspects on our datasets. (ii) For the multi-objective optimization method Mix&Match, we follow the experiment setting in the previous work. We retrain the classifiers and use sentences generated from PPLM (Dathathri et al., 2020) to initialize the sampling process so that long sentences can be generated for a fair comparison. (iii) For the methods optimizing in latent space, we reproduce LatentOps and MacLaSa by retraining the classifiers and the VAE module on our datasets. We retrain the soft prompts of Tailor on our datasets following their default hyperparameters. For Discrete and Prior-Control, we utilize the same way as our method to select the optimal weight parameters, and other hyperparameters are kept default. We uniformly use the pre-trained language model to GPT2-medium except for Mix&Match using Bert-large.[2]

B Details of Experiment Setup

B.1 Datasets

Following previous works (Krause et al., 2021; Gu et al., 2022b, 2023), we pick IMDb (Maas et al., 2011) for *sentiment*, AGNews for *topic* (Zhang et al., 2015), and the Jigsaw Toxic Comment Classification Challenge dataset for *detoxification*, with 10k samples for each aspect and equal samples for each attribute. Using 35 prompts from (Dathathri et al., 2020), we assess 8 combinations across 2 sentiments, 4 topics, and 1 detoxification, generate 5 sentences per combination, and evaluate 1,400 sentences in total.

B.2 Evaluation Metrics

We compute the attribute relevance with the DeBERTa classifiers finetuned on the Yelp and AGNews datasets (Zhang et al., 2015) for *sentiment* and *topic* aspects, respectively. Both classifiers are from (Gu et al., 2023). The *non-toxicity* aspect is measured by the Google Perspective API.[3] We also consider two auxiliary metrics for text quality, i.e., perplexity (abbr. PPL) and distinctness. PPL is calculated by GPT2-large and distinctness

[1] https://huggingface.co/openai-community/gpt2-medium

[2] https://huggingface.co/google-bert/bert-large-uncased

[3] https://www.perspectiveapi.com

is calculated by averaging the 1-gram, 2-grams, and 3-grams distinct scores (Yang et al., 2023).

C The Statistics in Figure 1

For Yelp, we directly utilize the statistical results from (Yang et al., 2023), which filter out neutral texts. More details can be found in the relevant paper. For AGNews, unfiltered neutral texts exist in the original dataset. We manually select some neutral texts and adjust the temperature of the classifier to keep the sentiment scores of neutral texts around 0.5, thereby mitigating their impact on the statistical results. The temperatures for the four topics ("world", "sport", "business", "technology") are 6.5, 4.5, 5, and 4.5, respectively.

D Results of Human Evaluation

We conduct human evaluation with sentences generated by different methods shuffled. Each sentence is rated by 3 professional evaluators for the three attribute relevance scores (*sentiment*, *topic*, and *non-toxicity*) and text fluency. Evaluators rate each item on a scale of 1 to 5, with 5 indicating the highest relevance to the desired attribute or most fluent. Following (Yang et al., 2023), the annotators are required to not attend to attribute correlation when evaluating the text quality (and vice versa) to obtain separate scores for both text quality and attribute correlations. Table 7 presents the results, with inter-annotator agreement being 0.38 in Fleiss' κ. In general, the results of human judgment are consistent with those of automatic evaluation. PriorControl is a strong baseline and can also achieve good performance in human evaluation. Our MAGIC is less affected by the imbalanced attribute correlations and can achieve the best performance.

E Compare with ChatGPT

In this section, we assess the performance of ChatGPT (gpt-3.5-turbo-0613) on multi-aspect control. Table 4 lists the results. More detailed results are shown in Table 5. Our method not only achieves comparable results with ChatGPT but also with significantly fewer parameters. We use gpt2-medium (355M) as the language model, which is consistent with baselines. ChatGPT demonstrates strong capability in following instructions and achieves good results in sentiment and detoxification control. However, the performance of ChatGPT on specific attributes such as "world" and "business" topics is relatively poor. This is because the limited demonstrations in in-context learning make it difficult for ChatGPT to fully grasp the specific information related to these attributes.

The prompt that we use to activate ChatGPT is as follows: "Generate 1 sentence containing 50 words with [ATTRIBUTE1] topic, [ATTRIBUTE2] sentiment, and [ATTRIBUTE3]. The generated sentences should start with [START_PROMPT]. The following are some sentences generated according to specific attribute constraints: ([ATTRIBUTE1] topic) => [EXAMPLE1]; ([ATTRIBUTE1] topic) => [EXAMPLE2]; ([ATTRIBUTE1] topic) => [EXAMPLE3]; ([ATTRIBUTE2] sentiment) => [EXAMPLE1]; ([ATTRIBUTE2] sentiment) => [EXAMPLE2]; ([ATTRIBUTE2] sentiment) => [EXAMPLE3]; ([ATTRIBUTE3]) => [EXAMPLE1]; ([ATTRIBUTE3]) => [EXAMPLE2]; ([ATTRIBUTE3]) => [EXAMPLE3]; Generate the sentence according to specific attribute constraints: ([ATTRIBUTE1] topic, [ATTRIBUTE2] sentiment, [ATTRIBUTE3]) => ". [ATTRIBUTE1] is selected from world, sports, business, and technology. [ATTRIBUTE2] is selected from positive and negative. [ATTRIBUTE3] is fixed to non-toxicity. [START_PROMPT] is from the 35 prompts that we used in the experiments. [EXAMPLE] is the sampled training data with the specific attribute.

F Ablation Study on Attribute Disentanglement

We introduce an ablation study about the impacts of Eqs. 7, 9, and 10 on attribute disentanglement since these loss functions affect the performance by influencing disentanglement. Table 6 lists the results. The values in the table represent the relevance scores of attributes corresponding to each column. The values in the second row of the table represent the attribute relevance scores of texts generated using original topic and sentiment control factors. The control factors of sentiment and topic are from the same sentence without replacement. By keeping the topic control factors constant and replacing the sentiment control factors with those from other topic sentences, we observe the decrease in topic correlation scores. It aims to assess whether sentiment control factors affect topic-related attributes (columns 2, 3, 4, and 5). Similarly, by keeping the sentiment control factors constant and replacing the topic control factors with those from opposite sentiment sentences, we

ChatGPT	Sentiment (%)		Topic (%)				Detox. (%)
	Neg.	Pos.	World	Sport	Business	Tech.	
Comb. 1	91.4	-	67.2	-	-	-	95.6
Comb. 2	95.0	-	-	96.3	-	-	92.9
Comb. 3	96.1	-	-	-	86.0	-	94.2
Comb. 4	97.2	-	-	-	-	99.2	94.3
Comb. 5	-	99.9	54.6	-	-	-	98.9
Comb. 6	-	99.7	-	94.9	-	-	99.0
Comb. 7	-	99.9	-	-	70.5	-	99.4
Comb. 8	-	99.9	-	-	-	98.9	98.9
Avg.	94.9	99.9	60.9	95.6	78.3	99.1	96.7

Table 5: Detailed results of ChatGPT on multi-aspect control.

Original control factors	World	Sport	Business	Tech.	Positive	Negative
MAGIC (intact)	90.7	98.7	87.8	99.5	67.5	99.4
Change control factors	World	Sport	Business	Tech	Positive	Negative
MAGIC (intact)	83.4 (7.3↓)	98.0 (0.7↓)	85.1 (2.7↓)	99.1 (0.4↓)	60.6 (6.9↓)	98.8 (0.6↓)
w/o $\mathcal{L}_{mul}^{A_*}$ (Eq. 7)	82.4 (8.3↓)	94.1 (4.6↓)	83.3 (4.5↓)	97.5 (2.0↓)	49.4 (18.1↓)	94.9 (4.5↓)
w/o $\mathcal{L}_{R}^{intv}$ (Eq. 9)	79.1 (11.6↓)	86.0 (12.7↓)	68.1 (19.7↓)	96.7 (2.8↓)	39.9 (28.0↓)	88.0 (11.4↓)
w/o $\mathcal{L}_{D}^{intv}$ (Eq. 10)	82.5 (8.2↓)	97.5 (1.2↓)	84.8 (3.0↓)	99.2 (0.3↓)	20.7 (46.8↓)	53.9 (45.5↓)

Table 6: Ablation results of Eqs. 7, 9, and 10 regarding their impacts on attribute disentanglement.

Methods	Avg.↑	Sent.↑	Topic↑	Detox.↑	Fluency↑
Imbalanced attribute correlations					
GeDi	3.26	2.75	3.05	4.00	2.67
PriorControl	3.44	3.21	3.15	3.97	3.65
MAGIC	3.80	3.89	3.50	4.02	3.60
Balanced attribute correlations					
GeDi	3.39	2.91	3.26	4.01	2.88
PriorControl	3.82	3.89	3.57	4.00	3.61
MAGIC	3.84	3.90	3.55	4.08	3.58

Table 7: Human evaluation on multi-aspect control.

aim to evaluate whether topic control factors influence sentiment attributes (the last two columns). Compared to the attribute relevance scores of texts generated by using the original control factors in the second row, the more the score of attribute correlation decreases, the worse the effect of disentanglement is indicated. It can be observed that without Eq. 7, there would be a certain decline in the effect. However, relying solely on Eq. 7 is insufficient to eliminate the mutual influence between different attribute control factors, as the absence of Eqs. 9 and 10 leads to a significant decrease in the effectiveness of disentanglement. The above results demonstrate the effectiveness of the loss functions proposed in our paper.

G Detailed Results with Imbalanced Attribute Correlation

Methods		Sentiment (%)		Topic (%)				Detox. (%)
		Neg.	Pos.	World	Sport	Business	Tech.	
PPLM	Comb. 1	96.2	-	76.3	-	-	-	96.5
	Comb. 2	86.6	-	-	43.7	-	-	71.2
	Comb. 3	88.1	-	-	-	62.8	-	90.9
	Comb. 4	89.4	-	-	-	-	98.8	77.6
	Comb. 5	-	34.4	48.4	-	-	-	84.3
	Comb. 6	-	30.9	-	30.1	-	-	97.9
	Comb. 7	-	39.8	-	-	39.2	-	83.1
	Comb. 8	-	43.0	-	-	-	94.7	93.8
	Avg.	90.1	37.0	62.4	36.9	51.0	96.8	86.9
GeDi	Comb. 1	95.7	-	84.1	-	-	-	89.5
	Comb. 2	91.8	-	-	73.2	-	-	94.7
	Comb. 3	99.3	-	-	-	83.9	-	96.4
	Comb. 4	91.5	-	-	-	-	99.9	96.4
	Comb. 5	-	43.0	62.8	-	-	-	96.0
	Comb. 6	-	54.2	-	67.4	-	-	97.0
	Comb. 7	-	54.2	-	-	52.6	-	97.3
	Comb. 8	-	58.4	-	-	-	99.1	97.1
	Avg.	94.6	52.4	73.5	70.3	68.3	99.5	95.5
Mix&Match	Comb. 1	96.9	-	84.3	-	-	-	89.0
	Comb. 2	98.4	-	-	54.2	-	-	90.1
	Comb. 3	99.2	-	-	-	69.3	-	89.8
	Comb. 4	98.2	-	-	-	-	99.5	89.3
	Comb. 5	-	53.4	62.5	-	-	-	93.1
	Comb. 6	-	53.4	-	34.9	-	-	93.3
	Comb. 7	-	40.1	-	-	47.1	-	94.7
	Comb. 8	-	53.6	-	-	-	98.2	95.3
	Avg.	98.2	46.9	73.4	44.6	58.2	98.9	91.8
Tailor	Comb. 1	99.1	-	85.5	-	-	-	92.0
	Comb. 2	94.2	-	-	70.1	-	-	95.4
	Comb. 3	98.8	-	-	-	70.2	-	96.2
	Comb. 4	94.6	-	-	-	-	83.8	96.4
	Comb. 5	-	40.4	31.8	-	-	-	97.6
	Comb. 6	-	40.2	-	61.9	-	-	97.7
	Comb. 7	-	32.1	-	-	44.6	-	98.2
	Comb. 8	-	40.4	-	-	-	85.5	97.7
	Avg.	96.7	38.3	58.7	66.0	57.4	84.7	96.4
LatentOPs	Comb. 1	95.0	-	77.1	-	-	-	87.3
	Comb. 2	99.0	-	-	55.9	-	-	73.9
	Comb. 3	96.5	-	-	-	88.7	-	93.0
	Comb. 4	94.6	-	-	-	-	98.1	92.7
	Comb. 5	-	55.7	70.7	-	-	-	95.2
	Comb. 6	-	57.0	-	72.5	-	-	97.4

Continue to next page

Methods		Sentiment (%)		Topic (%)				Detox. (%)
		Neg.	Pos.	World	Sport	Business	Tech.	
	Comb. 7	-	54.0	-	-	64.6	-	98.5
	Comb. 8	-	72.7	-	-	-	98.4	98.4
	Avg.	96.3	59.9	73.9	64.2	76.6	98.2	92.1
Discrete	Comb. 1	93.1	-	67.9	-	-	-	88.3
	Comb. 2	95.3	-	-	67.7	-	-	92.2
	Comb. 3	94.8	-	-	-	67.7	-	95.3
	Comb. 4	97.4	-	-	-	-	95.0	92.2
	Comb. 5	-	98.3	40.8	-	-	-	97.6
	Comb. 6	-	99.1	-	30.2	-	-	98.0
	Comb. 7	-	53.1	-	-	49.4	-	97.6
	Comb. 8	-	99.4	-	-	-	94.1	97.6
	Avg.	95.1	87.5	54.4	49.0	64.3	94.6	94.8
MacLaSa	Comb. 1	98.9	-	89.0	-	-	-	88.4
	Comb. 2	92.3	-	-	57.0	-	-	89.2
	Comb. 3	92.3	-	-	-	84.2	-	94.5
	Comb. 4	96.4	-	-	-	-	95.7	94.3
	Comb. 5	-	68.1	79.0	-	-	-	95.1
	Comb. 6	-	71.5	-	49.9	-	-	96.0
	Comb. 7	-	67.7	-	-	73.7	-	96.5
	Comb. 8	-	71.8	-	-	-	94.5	97.6
	Avg.	95.0	69.8	84.0	53.5	78.9	95.1	94.0
PriorControl	Comb. 1	96.3	-	89.8	-	-	-	85.0
	Comb. 2	93.4	-	-	90.4	-	-	88.4
	Comb. 3	95.2	-	-	-	71.6	-	92.0
	Comb. 4	96.5	-	-	-	-	99.1	89.2
	Comb. 5	-	79.3	63.3	-	-	-	94.1
	Comb. 6	-	72.0	-	76.2	-	-	95.7
	Comb. 7	-	76.4	-	-	42.1	-	96.4
	Comb. 8	-	95.4	-	-	-	95.2	96.1
	Avg.	95.4	80.8	76.5	83.3	56.8	97.1	92.1
MAGIC	Comb. 1	99.4	-	95.9	-	-	-	90.2
	Comb. 2	99.9	-	-	98.3	-	-	89.3
	Comb. 3	98.2	-	-	-	93.4	-	97.1
	Comb. 4	99.5	-	-	-	-	97.0	90.9
	Comb. 5	-	83.8	69.7	-	-	-	95.5
	Comb. 6	-	88.6	-	97.1	-	-	97.9
	Comb. 7	-	86.9	-	-	64.8	-	98.5
	Comb. 8	-	99.9	-	-	-	92.0	97.9
	Avg.	99.3	89.8	82.8	97.7	79.1	94.5	94.7

Table 8: Detailed combination results on multi-aspect control with imbalanced attribute correlations.

H Detailed Results with Balanced Attribute Correlation

Methods		Sentiment (%)		Topic (%)				Detox. (%)
		Neg.	Pos.	World	Sport	Business	Tech.	
PPLM	Comb. 1	92.2	-	75.4	-	-	-	82.0
	Comb. 2	84.4	-	-	41.8	-	-	76.0
	Comb. 3	87.5	-	-	-	61.5	-	82.9
	Comb. 4	85.3	-	-	-	-	95.0	76.2
	Comb. 5	-	35.4	59.1	-	-	-	90.4
	Comb. 6	-	39.5	-	34.1	-	-	89.5
	Comb. 6	-	40.9	-	-	48.3	-	91.2
	Comb. 8	-	52.7	-	-	-	93.1	91.3
	Avg.	87.4	42.1	67.3	38.0	54.9	94.1	84.9
GeDi	Comb. 1	94.7	-	80.0	-	-	-	90.6
	Comb. 2	84.2	-	-	74.8	-	-	93.9
	Comb. 3	94.9	-	-	-	75.7	-	96.6
	Comb. 4	90.6	-	-	-	-	80.1	92.8
	Comb. 5	-	53.7	61.4	-	-	-	94.4
	Comb. 6	-	60.5	-	74.3	-	-	95.2
	Comb. 6	-	57.6	-	-	54.3	-	95.7
	Comb. 8	-	72.3	-	-	-	90.2	94.2
	Avg.	91.1	61.0	70.7	74.6	65.0	85.2	94.2
Mix&Match	Comb. 1	96.1	-	80.6	-	-	-	93.1
	Comb. 2	97.7	-	-	48.2	-	-	93.0
	Comb. 3	98.2	-	-	-	66.6	-	97.0
	Comb. 4	96.8	-	-	-	-	99.6	96.1
	Comb. 5	-	53.0	67.3	-	-	-	95.5
	Comb. 6	-	45.0	-	44.0	-	-	96.7
	Comb. 7	-	41.5	-	-	55.8	-	97.7
	Comb. 8	-	59.7	-	-	-	97.3	97.5
	Avg.	97.2	49.8	74.0	46.1	61.2	98.5	
Tailor	Comb. 1	96.1	-	81.4	-	-	-	90.3
	Comb. 2	85.8	-	-	80.2	-	-	94.4
	Comb. 3	90.7	-	-	-	76.6	-	96.8
	Comb. 4	90.4	-	-	-	-	86.0	96.0
	Comb. 5	-	34.2	40.7	-	-	-	96.4
	Comb. 6	-	45.3	-	65.1	-	-	97.6
	Comb. 7	-	30.0	-	-	65.4	-	98.1
	Comb. 8	-	44.4	-	-	-	94.2	97.7
	Avg.	90.7	38.5	61.1	72.6	71.0	90.1	96.0
LatentOPs	Comb. 1	96.7	-	61.7	-	-	-	86.4
	Comb. 2	84.5	-	-	80.7	-	-	91.7
	Comb. 3	72.6	-	-	-	98.7	-	98.3
	Comb. 4	90.8	-	-	-	-	99.9	94.9
	Comb. 5	-	61.2	71.1	-	-	-	94.5
	Comb. 6	-	62.7	-	84.3	-	-	98.0

Continue to next page

Methods		Sentiment (%)		Topic (%)				Detox. (%)
		Neg.	Pos.	World	Sport	Business	Tech.	
	Comb. 7	-	52.0	-	-	85.2	-	98.1
	Comb. 8	-	89.7	-	-	-	99.7	98.3
	Avg.	86.1	66.4	66.4	82.5	91.9	99.8	95.0
Discrete	Comb. 1	69.7	-	71.7	-	-	-	84.1
	Comb. 2	78.6	-	-	80.0	-	-	80.2
	Comb. 3	99.9	-	-	-	96.7	-	96.8
	Comb. 4	92.8	-	-	-	-	98.0	81.7
	Comb. 5	-	80.5	58.0	-	-	-	95.1
	Comb. 6	-	84.7	-	86.6	-	-	94.5
	Comb. 7	-	87.6	-	-	91.7	-	98.1
	Comb. 8	-	99.7	-	-	-	96.1	95.4
	Avg.	85.3	88.1	64.9	83.3	94.2	96.8	90.7
MacLaSa	Comb. 1	92.8	-	87.6	-	-	-	91.4
	Comb. 2	95.1	-	-	86.2	-	-	92.9
	Comb. 3	85.6	-	-	-	84.7	-	95.3
	Comb. 4	92.7	-	-	-	-	97.0	90.7
	Comb. 5	-	93.2	73.8	-	-	-	94.8
	Comb. 6	-	83.0	-	71.3	-	-	97.0
	Comb. 7	-	50.1	-	-	85.7	-	97.6
	Comb. 8	-	87.6	-	-	-	98.2	96.7
	Avg.	91.5	78.5	78.9	78.7	85.2	97.6	94.6
PriorControl	Comb. 1	97.9	-	98.3	-	-	-	90.5
	Comb. 2	98.4	-	-	98.5	-	-	93.4
	Comb. 3	97.3	-	-	-	96.9	-	98.5
	Comb. 4	99.9	-	-	-	-	99.7	89.1
	Comb. 5	-	89.5	79.4	-	-	-	95.4
	Comb. 6	-	84.5	-	73.7	-	-	96.8
	Comb. 7	-	74.2	-	-	73.1	-	98.4
	Comb. 8	-	98.0	-	-	-	95.2	97.3
	Avg.	98.4	86.6	88.9	86.1	85.0	97.5	94.9
MAGIC	Comb. 1	99.9	-	97.6	-	-	-	91.8
	Comb. 2	98.2	-	-	99.2	-	-	95.0
	Comb. 3	99.3	-	-	-	91.5	-	97.8
	Comb. 4	92.7	-	-	-	-	96.1	84.4
	Comb. 5	-	95.8	71.8	-	-	-	96.6
	Comb. 6	-	85.2	-	96.5	-	-	98.2
	Comb. 7	-	84.0	-	-	68.0	-	98.5
	Comb. 8	-	98.6	-	-	-	94.4	98.2
	Avg.	97.5	90.9	84.7	97.8	79.8	95.2	95.1

Table 9: Detailed combination results on multi-aspect control with balanced attribute correlations.

I Case Study

We illustrate cases of 8 different attribute combinations in Table 10.

Attributes	Methods	Generation results
Negative World NonTox.	GeDi	Furthermore, Iraqi Prime Minister Haider al-Abadi said on Wednesday that his country had almost conscription in January thanks to Pentagon efforts that garnered international help. Baghdad offered a combined 656,000 U.S.-British personnel.
	Mix&Match	Furthermore, the russians who are calling for his immediate execution should be put to death as well. this evil is a danger to society, and should be stopped as soon as possible.
	Tailor	Furthermore, the US military has also been asked to consider using drones in combat as part of its drone strike campaign. The Pentagon's top brass are discussing whether they could carry out strikes on targets across Iraq and Syria that would be conducted from unmanned aircraft.
	LatentOPs	Furthermore, the fact that the IRA were responsible for the deaths of thousands of people in the IRA is not mentioned in the IRA's own accounts. This is a shameful way to treat prisoners.
	Discrete	Furthermore, the US government has issued a report that indicates Donald Trump's personal car is stolen.
	MacLaSa	Furthermore, Israel should withdraw from the Gaza Strip and take steps to facilitate Palestinian elections, including demilitarization, he said.
	PriorControl	Furthermore, the government has given a detailed briefing to its troops on how to avoid becoming an enemy of the Islamic militants. Followers of extremist groups have complained that this movie is poorly suited to be film adaptation and even worse for human beings.
	MAGIC	Furthermore, the United States appears to have lost a congressional debate over whether to grant asylum to Jose Altman. The president-elect has been criticized by many of his aides as having failed to recognize that this is an issue with serious implications.
Negative Sport NonTox.	GeDi	This essay discusses Royals postseason performances with observations and conclusions based on the Data Shed. I will try to balance debut dynamics, transition performance, and reputations in a way that is appropriate for those baseball nerds who are not up to date on these issues.
	Mix&Match	This essay discusses the mathematical model used to calculate the final score for the ryder cup. it's not terribly complicated, but it does leave out one crucial piece of information.
	Tailor	This essay discusses the impact of a change in rules on sports betting and how this has changed sport wagering, especially since it became legal. The most famous example is that which occurred with tennis at Wimbledon.
	LatentOPs	This essay discusses the film version of the novel by George Orwell. The film version is a complete waste of time. The novel is a complete waste of time.

Continue to next page

Attributes	Methods	Generation results
Negative Sport NonTox.	Discrete	This essay discusses how the NFL will be able to avoid paying a third-year man who is currently injured.
	MacLaSa	This essay discusses the many shortcomings of the movie P̈lan 9 from Outer Space.̈ I won't bore you with a list, but there are several things that really need to be mentioned. First of all, the special effects were mediocre at best.
	PriorControl	This essay discusses how the NFL can improve its reputation for fairness and fairness in order to avoid a repeat of the 2004 debacle involving Ricky Fuze.
	MAGIC	This essay discusses the worst possible scenario for the United States players in this week's World Cup qualifying game against Sweden. Related Articles Tom Brady won't be able to get a hold of his emotions after losing his second straight game.
Negative Business NonTox.	GeDi	Foundational to this is the cost structure that has to be removed. Atlanta-based UnitedHealth Group said it had about 65,000 job cuts planned at two subsidiaries Aetna and Humana before the spinoff announcement.
	Mix&Match	Foundational to this is the fact that the company is a holding company, meaning it can't issue shares. so if the fund manager sells the company, the funds that invested in the company will be worse off because they won't have the earnings to pay out the dividends.
	Tailor	Foundational to this is the notion of a "single market" where nations can negotiate for their own trade and investment policies. It's been called an idea that was born in America but has not yet reached Europe or Asia.
	LatentOPs	Foundational to this is that the company is a relatively small one with a relatively low profile. The company's management has been criticized for its poor performance in recent years. The company's management has been accused of being overly optimistic and overly optimistic about its earnings.
	Discrete	Foundational to this is the fact that it has been reduced to a low-cost version of its flagship product, namely: The Longest Life Insurance company.
	MacLaSa	Foundational to this is the fact that the share price of a public company is determined by its market capitalization. Put simply, the more money a company raises through a public offering the more it earns in earnings taxes.
	PriorControl	Foundational to this is the fact that Microsoft has taken a few steps toward selling its premium products in the market. The company has offered a partial refund for the purchase of half-priced software, which is considered a major mistake.
	MAGIC	Foundational to this is the fact that many people are not working in the company because of their lack of investment. This is a bad news story for the companies involved.

Continue to next page

Attributes	Methods	Generation results
Negative Tech. NonTox.	GeDi	The last time Microsoft used Charms bar to populate its apps' homescreen was for Windows Phone 8.2, which arrived in November 2012 and made room for a colorful new Start screen that sported apps like Dropbox.
	Mix&Match	The last time sun ceo scott mcnealy spoke about linux, the audience at a sun developer conference in san francisco was mostly deaf.
	Tailor	The last time IBM shipped a new product in this era, it was with the PC-compatible Amstrad CPC. In that case, its first line of PCs were for home use only and had not received any major market expansion beyond their small numbers.
	LatentOPs	The last time I used a 3D printer was when I was a teenager. I was amazed at how much 3D printing was able to do, but I was also amazed at how little I actually used it. I was also surprised that the printer was so slow.
	Discrete	The last time the company saw a major change in its software, it was about half-dozen times over. Follow Stories/News related to Business Insider.
	MacLaSa	The last time NASA's Mars rover Opportunity tried to drive herself, her arm failed to operate, and the two rovers almost drove themselves...
	PriorControl	The last time I saw this film was when I had to pick up a copy of the book that was supposedly released on behalf of Microsoft. It's so bland that it makes me wonder if the author actually knew what kind of content he would be.
	MAGIC	The last time this type of thing happened in 2001. A few weeks ago, the Internet Mail Service (IMPS) faced a major setback when it failed to deliver an application.
Positive World NonTox.	GeDi	The connection develops between Neapolitan peoples through time. It embraces generations. His passion doing linguistics formed his whole life work as a gesture of co-existence, bringing different communities together and defending their culture and identity.
	Mix&Match	The connection between a us soldier and the iraqi flag is one of the most interesting films i have ever seen. it's not surprising that such an film exists, given the role actors involved.
	Tailor	The connection between the United States and Europe is now stronger than ever before. The U-2 program was born as a response to an unprecedented crisis, in part because of its promise of new possibilities for global cooperation;
	LatentOPs	The connection between the Palestinian leadership and Hamas is well established and is discussed in the article. The link between the two organizations is also well-established.

Continue to next page

Attributes	Methods	Generation results
Positive World NonTox.	Discrete	The connection between these movies and the original 1979 movie was very good. The acting is superb and the story is great. I can't recommend this film anymore.
	MacLaSa	The connection between happiness and wine made by Nobel Peace Prize laureate Wangari Maathai (she received the honor last week) is a striking example of how a prize can bind nations and peoples together.
	PriorControl	The connection between the US flag and its politics is evident in this film. Although the plot is somewhat flawed, the acting is superb and the story line is fantastic. I highly recommend it.
	MAGIC	The connection between the world's most famous football hero and his daughter is a fascinating one. A good deal of work has been done to make this movie accessible to people who have been marginalized by mainstream society.
Positive Sport NonTox.	GeDi	More importantly, football matters. Winning creates peace, education and joy in children across 66 countries each April.
	Mix&Match	More importantly, what is the purpose of a game like this in the first place? i'm not sure, but it could be a resource for people who might ask about the game, or for those who might be inspired to get their hands on it by the on juggling"
	Tailor	More importantly, this is a new start for the New York Mets. The team that finished last season with an 11-42 record went on to reach its first postseason in 17 years – and make it through four of those six games without giving up runs.
	LatentOPs	More importantly, the two goals by the two rookies were the first goals by a rookie in the history of the club. The two goals by the two rookies were the first goals by a rookie in club history.
	Discrete	More importantly, this movie is a great story. It has a great plot and fantastic actors. I loved the ending when the young man who had been killed by a car came back to life. It's really funny and very entertaining.
	MacLaSa	More importantly, this series has truly stood the test of time, being able to adapt to the changing tastes of the modern consumer while still delivering a fun and engaging game. I've always been a huge baseball fan, so naturally, I was very taken by the series.
	PriorControl	More importantly, the Miami Dolphins have a solid team in front of him and are on board with their approach to playing football. The team has drafted several talented players, including Lou Gehwin, who plays in the league for the first time.
	MAGIC	More importantly, the Florida Panthers have won a top-level competition in this year's World Cup Qualifiers. The Arizona Diamondbacks will be able to do likewise.

Continue to next page

Attributes	Methods	Generation results
Positive Business NonTox.	GeDi	In brief, prices rendezvous strikes a balance between incentives and constraints that helps correct market distortions investment choices, productivity patterns, and short-run business cycles. (Furthermore, over time efficient allocations of resources can create leverage.)
	Mix&Match	In brief, the board of fannie mae has approved major changes to the way it calculates the capital gains on the sale of home mortgages. the changes will allow the company to more accurately project future capital gains and thus generate more profit from its investment in home mortgages.
	Tailor	In brief, it was a good year for video game makers. The number of consoles shipped in 2015 rose 2-1 and the software sales were strong at 1 million units sold through September to become the fastest growing console category since 2000.
	LatentOPs	In brief, the company says it expects to post a net loss of $ 1.3 billion in fiscal 2004. The company says it expects to post a net loss of $ 1.4 billion in fiscal 2004.
	Discrete	In brief, the company's new flagship model is tested on a variety of different occasions. Follow Stories by Wireman & Co., which is currently working with the president-elect to help him transition into his next phase.
	MacLaSa	In brief, the company raised its dividend by 8 cents per share yesterday, helped by strong results across its segments. The company's outlook remains positive, with earnings rising to between 4 cents and 4.
	PriorControl	In brief, the company has given an indication of how much money it expects to spend in its next big post-election campaign. The movie is presented as a classic and highly entertaining film, but the actors are great.
	MAGIC	In brief, it is a good fit for the company. Related Stories to a new era in business is underway with an exciting mix of products and services.
Positive Tech. NonTox.	GeDi	The country-and-world revolution in microcomputing at the heart of HP Labs is the Hyperledger Fabric, an open-source software framework for developers and architects to build efficient, reliable, affordable cloud computing services powered by powerful micro.
	Mix&Match	the country's top virtual - presence expert explains how his research spent two years building a prototype that could help police solve crimes on the web.
	Tailor	The country's first 3G mobile telephone system is being introduced at Tel Aviv University, where students will get a chance to experiment with technology that would allow people of all ages and from different backgrounds in the Israeli university campus to communicate freely.
	LatentOPs	The country's largest and most comprehensive collection of free and open source software. Easily add and manage multiple applications to your system. Easily share and manage your work with others. Easily share and manage your work with others.

Continue to next page

Attributes	Methods	Generation results
Positive Tech. NonTox.	Discrete	The countrywide television network, which is owned by the United States government, has been invited to participate in a series of events with the rest of the world.
	MacLaSa	The country's most famous virtual reality game has been nominated for a Guinness World Record for the most downloads. And it's not just the game that has caught gamers' attention. 'It's a very evocative title.' said Leigh Alexander, producer of the award.
	PriorControl	The country's leading wireless network has backed a new version of its popular mobile app, offering more than 200 unique brands to choose from. Follow the links to find out how this movie is great and how it compares in comparison to other films.
	MAGIC	The country's largest television broadcast network, which is currently a great success. It's one of the best films that I've seen in years. The acting is fantastic and the cinematography is superb.

Table 10: Example cases of generated sentences with 8 attribute combinations. Blue text highlights sentiment-related content. Red text highlights topic-related content. Underlined text is the input prompts.