A Distributional Lens for Multi-Aspect Controllable Text Generation

Yuxuan Gu[†], Xiaocheng Feng[†‡], Sicheng Ma[†], Lingyuan Zhang[†], Heng Gong[†], Bing Qin[†‡]

[†]Harbin Institute of Technology   [‡] Peng Cheng Laboratory

`{yxgu,xcfeng,scma,lyzhang,hgong,bqin}@ir.hit.edu.cn`

Abstract

Multi-aspect controllable text generation is a more challenging and practical task than single-aspect control. Existing methods achieve complex multi-aspect control by fusing multiple controllers learned from single-aspect, but suffer from attribute degeneration caused by the mutual interference of these controllers. To address this, we provide observations on attribute fusion from a distributional perspective and propose to directly search for the intersection areas of multiple attribute distributions as their combination for generation. Our method first estimates the attribute space with an autoencoder structure. Afterward, we iteratively approach the intersections by jointly minimizing distances to points representing different attributes. Finally, we map them to attribute-relevant sentences with a prefix-tuning-based decoder. Experiments on the three-aspect control task, including sentiment, topic, and detoxification aspects, reveal that our method outperforms several strong baselines on attribute relevance and text quality and achieves the SOTA. Further analysis also supplies some explanatory support for the effectiveness of our approach[1].

1 Introduction

Controllable text generation is a challenging task in natural language generation, which aims to generate fluent text with desired attributes. Pilot studies attempt single-aspect control by directly finetuning a conditional model (Ziegler et al., 2019; Keskar et al., 2019), or turn to methods with language models fixed (Dathathri et al., 2020) due to the high cost of large-scale pre-trained language models (Brown et al., 2020a; Zhang et al., 2022).

Recent works focus on a more practical setting, multi-aspect[2] controllable text generation, with existing approaches mainly divided into three tech-
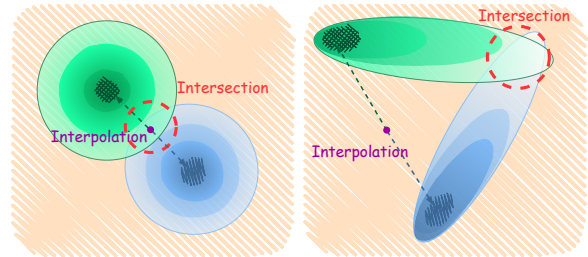

Figure 1: Probability space of attributes. Orange background denotes the estimated distribution over natural language. Blue and green areas represent distributions over sentences containing attributes from two different aspects, respectively. The darker region means a higher probability in the space. The shaded are distributional centers, the areas with the highest probability density.

nical routes: weighted decoding (Dathathri et al., 2020; Krause et al., 2021), multi-objective optimization (Kumar et al., 2021; Mireshghallah et al., 2022), and prefix-tuning (Qian et al., 2022), which explore ways to combine controllers learned from single-aspect and apply them to a fixed language model yet suffering from attribute degeneration caused by the mutual interference of controllers.

We provide a distributional perspective to observe and alleviate this problem. In the current text generation paradigm, a language model forms an estimated distribution over sentences with training data amounted to sampling from natural language distribution (Pillutla et al., 2021). For single-aspect control, these methods train a classifier or a prefix for each attribute independently, which is regarded as appraising a center of distribution over attribute-relevant sentences, before biasing the language model's distribution to this center. Correspondingly, when generalizing to multi-aspect control, their fusion strategy is directly obtaining interpolation or average of these centers, which may be too straightforward. As shown in Figure 1, the **interpolation** point denotes the position they acquired after combining multiple centers in the probability space. And the **intersection** represents

---

[1]Our dataset and code are available at: `https://github.com/HappyGu0524/MultiControl`

[2]For example, *positive* is an attribute from sentiment aspect while *sports* is an attribute from topic aspect.

where oracle sentences that simultaneously satisfy multiple attributes lie. In the left part of Figure 1, when distributions of attributes is symmetric[3], the interpolation point is indeed within the intersection area. However, there could be a mismatch between the interpolation point and intersection. For example, as illustrated in the right part of Figure 1, two skewed distributions intersect on the tails, leaving the interpolation point out of the intersection area and thus making it lack the ability to express all desired attributes together.

In this paper, different from approximating the intersection area with the interpolation point, we propose a strategy for directly acquiring the intersection. We first deploy an autoencoder structure to map attribute-relevant sentences to latent representations constituting an estimated attribute space. With our specially designed constraints, this space can model relationships among attributes. Afterward, we provide an effective intersection searching algorithm that can walk around the long tail regions in distributions of all desired attributes and iteratively find where they combine more tightly. Finally, we utilize a prefix-tuning-based decoder to construct sentences from the searched intersection.

We experiment on three-aspect control with two attributes from the sentiment aspect, four from the topic, and one from detoxification, with datasets IMDb movie reviews (Maas et al., 2011), AGNews (Zhang et al., 2015), and Jigsaw Toxic Comment Classification Challenge Dataset, respectively. We evaluate the relevance of each attribute independently and calculate their average as the final relevance metric. Besides, we assess the text quality with perplexity and distinctness concerning fluency and diversity. Results show that our method can significantly outperform strong baseline models on multi-aspect control. Furthermore, we find out in our analytical experiments that our intuitive assumptions fit well with our observation. The main contributions are as follows:

- We propose a distributional perspective that models multi-aspect control more practically.
- We provide a method that directly searches for intersections in the attribute space and generates sentences with desired attributes.
- We experimentally reveal the effectiveness of our method on multi-aspect control compared to strong baselines and achieve the SOTA.

---

[3]We plot distributions of attributes in §5.4.

2 Related Work

Variational autoencoders are often used for controllable text generation in early work (Hu et al., 2017; Duan et al., 2020; Mai et al., 2020) where they spend a lot of effort into improving text fluency. The prosperity of large-scale pre-trained language models (Radford et al., 2019) provides more exploration directions for attribute control such as fine-tuning (Ficler and Goldberg, 2017; Ziegler et al., 2019; Keskar et al., 2019). Recent work has made gratifying progress on single-aspect control (Krause et al., 2021), leading studies gradually turn to a more difficult task, multi-aspect control, including the following three main approaches.

Weighted Decoding As the scale of language models increases rapidly, weighted decoding (Dathathri et al., 2020; Krause et al., 2021; Yang and Klein, 2021; Liu et al., 2021a; Gu et al., 2022) becomes a simple and practical choice. It is a framework that decomposes the probability of sentences conditioned on attributes into a language model and a classifier with the bayesian rule directly at decoding time. When handling multi-aspect control, it can be easily generalized by interpolating classifiers (Lin and Riedl, 2021).

Multi-Objective Optimization The controllable text generation task is naturally a multi-objective optimization problem when regarding its decoding process as an optimization objective. Some approaches, such as DGC (Khalifa et al., 2020), Mix&Match (Mireshghallah et al., 2022), and COLD Decoding (Qin et al., 2022), adopt Energy-based Models (LeCun et al., 2006) to blend multiple objectives. Others like MUCOCO (Kumar et al., 2021) convert the optimization objectives of multi-aspect control to inequality constraints and thereby apply the lagrange multiplier method for this constrained optimization problem.

Prefix-Tuning GPT-3 (Brown et al., 2020b) provides a new paradigm named prompt-based learning (Liu et al., 2021b), which is able to perform few-shot learning on downstream tasks. Prefix-Tuning (Li and Liang, 2021) leverages the learned lightweight prompts to trigger the conditional generation capability of the language model. Applying Prefix-Tuning to multi-aspect controllable text generation (Yu et al., 2021; Qian et al., 2022; Carlsson et al., 2022; Yang et al., 2022) can be regarded as optimizing on multi-objective implicitly.
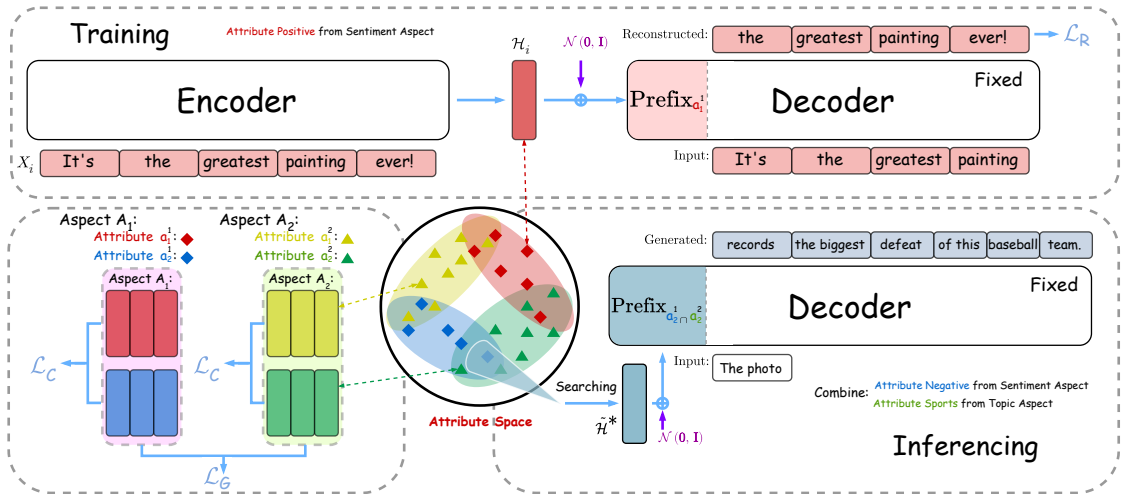
Figure 2: An overview of our method. **Top**: Illustration of our autoencoder structure with prefix-tuning deployed on the fixed decoder, where latent representations $\mathcal{H}_i$ constitute an estimated attribute space. **Bottom Left**: Illustration of attribute classification loss $\mathcal{L}_C$ and aspect gap loss $\mathcal{L}_G$ attached to the attribute space. **Bottom Right**: Inferencing stage with prefix mapped from the intersection of attributes.

3 Methodology

In this section, we first introduce the motivation and overall process of our method, after which we describe each module in detail.

3.1 Overview

As illustrated in Figure 2, our method mainly revolves around the attribute space including estimating the attribute space, searching for intersections, and mapping intersections to sentences.

Firstly, we aim to construct an attribute space using sampled sentences to estimate the real space as accurately as possible. We employ an autoencoder structure with the latent representations denoting points that constitute our estimated attribute space. To ensure that our estimated space reliably models the attributes, such as their probability distributions and relationships between different attributes, we further attach three constraints to the representation. (I) **Reconstruction Loss** $\mathcal{L}_R$ aims to bridge the gap between points in attribute space and natural attribute-relevant sentences, which is recovering attributes reflected by contents. (II) **Attribute Classification Loss** $\mathcal{L}_C$ forces the encoder to focus more on capturing attributes by distinguishing points of different attributes from the same aspect. (III) **Aspect Gap Loss** $\mathcal{L}_G$ penalizes the discrepancy of aspects, which is caused by the domain gap among different data sources for different aspects. Inspired by the feature alignment (Pan et al., 2010), we minimize the distances between distributional centers of each two aspects.

The second step aims to search for an intersec-

tion area of desired attributes. If the intersection area exists, a point in the area satisfies that neighbor points appearing in a tiny surrounding region should cover all required attributes. Inspired by this neighborhood ideology, we design an algorithm that iteratively approaches an area where these attributes bind more tightly. The third step maps our searched intersection to a Prefix that activates the language model to generate attribute-relevant sentences. To make the language model less sensitive to slight variations, we sample a perturbation vector from a multivariate gaussian distribution.

3.2 Estimating Attribute Space

Given $|\mathbf{A}|$ aspects $\mathbf{A} = \left\{ A_1, \cdots, A_{|\mathbf{A}|} \right\}$ with each comprising $|A_t|$ attributes $\left\{ a_1^t, \cdots, a_{|A_t|}^t \right\}$, I_τ^t is an index set representing the identifiers of all sentences with attribute a_τ^t in the training data. We have $I^t = \bigcup\limits_{\tau=1}^{|A_t|} I_\tau^t, I = \bigcup\limits_{t=1}^{|\mathbf{A}|} I^t$, where I^t is the indices of all sentences with any attribute in aspect A_t and I is the indices of the entire training data. We encode sentences $\{X_i\}$ from all aspects $\mathbf{A}$ to representations $\{\mathcal{H}_i\}$ with unified mapping parameters ϕ: $\mathcal{H}_i = \text{Encode}_\phi(X_i)$, where $i \in I$.

Reconstruction Loss $\mathcal{L}_R$ As in the top of Figure 2, $\mathcal{L}_R$ is computed in the same way as the autoregressive loss of pre-trained language model p_{LM}:

$$\mathcal{L}_R = -\sum_{i \in I} \log p_{\text{LM}}(X_i | \text{Prefix}_i) \tag{1}$$

$$\text{Prefix}_i = \text{MLP}_\theta(\mathcal{H}_i + \lambda \varepsilon_i), \ \varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where X_i here is a sample sentence from the entire training set, i.e., $i \in I$. Besides, ε_i, with a scaling factor λ, is a perturbation vector sampled from a multivariate gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for robustness when reconstructing. The multilayer perceptron MLP_θ will map perturbed $\mathcal{H}_i$ to Prefix_i that can activate the language model to generate text with desired attributes. It's worth noting that our primary goal is to recover attributes, which means $\mathcal{L}_R$ does not need and preferably does not converge too well while maintaining text fluency.

Attribute Classification Loss $\mathcal{L}_C$ We force the encoder to focus on attributes by $\mathcal{L}_C$ in the way:

$$\mathcal{L}_C = -\sum_{t=1}^{|\mathbf{A}|} \sum_{\tau=1}^{|A_t|} \sum_{i \in I_\tau^t} \log p_{\pi_t}(a_\tau^t | \mathcal{H}_i). \quad (2)$$

Given sentence representation, p_{π_t} is a classifier that distinguish attributes $\{a_\tau^t\}$ from aspect A_t with parameter π_t.

Aspect Gap Loss $\mathcal{L}_G$ We penalize the discrepancy between distributional centers by:

$$\mathcal{L}_G = \sum_{1 \le t_1 < t_2 \le |\mathbf{A}|} \left\| \sum_{i \in I^{t_1}} \frac{\mathcal{H}_i}{|I^{t_1}|} - \sum_{j \in I^{t_2}} \frac{\mathcal{H}_j}{|I^{t_2}|} \right\|_2, \quad (3)$$

which are Euler distances between every two distinct distributional centers. When generalizing to a larger scale of aspects, it is relatively expensive to calculate averages over the entire dataset each time the model is updated. We calculate this loss in practice using a batch-level approximation. We assign each aspect a memory unit to store the latest representation of the aspect's estimated center. Each time processing a batch of sentences from one aspect, we take the average of their representations as the center and sum up the Euler distances to centers of other aspects in the memory, which is the estimated $\mathcal{L}_G$. Then, we update the memory unit of this aspect to the latest.

During the training stage, our loss function is:

$$\mathcal{L} = w_1 \mathcal{L}_R + w_2 \mathcal{L}_C + w_3 \mathcal{L}_G. \quad (4)$$

It's worth noting that we only update parameters ϕ, θ, and $\{\pi_t\}$ for the encoder, the MLP layer, and the classifier heads, respectively.

3.3   Intersection of Attributes

Suppose there is an intersection point, denoted as $\tilde{\mathcal{H}}^*$, located within the intersection region of attributes $\{a_{\alpha_1}^1, a_{\alpha_2}^2, \cdots, a_{\alpha_N}^N\}$ from N different as-

---

Algorithm 1 Intersection Searching

Input: $\mathcal{H}_i, i \in \bigcup_{t=1}^{N} I_{\alpha_t}^t$ from N attributes
   ω_{α_t} weight of each attribute
Output: Intersection of N attributes: $\tilde{\mathcal{H}}^*$
1: Initialize M candidates: $\{\tilde{\mathcal{H}}_m^0\}$
2: Iterate S times
3: **for** s in $[0, S-1]$ **do**
4:    **for** m in $[1, M]$ **do**
5:       $\tilde{\mathcal{H}}_m^{s+1} \leftarrow \mathbf{0}$
6:       **for** t in $[1, N]$ **do**
7:          $\mathbf{H} \leftarrow \underset{topK}{\text{Nearest}}(\tilde{\mathcal{H}}_m^s, \{\mathcal{H}_i, i \in I_{\alpha_t}^t\})$
8:          $\tilde{\mathcal{H}}_m^{s+1} \leftarrow \tilde{\mathcal{H}}_m^{s+1} + \omega_{\alpha_t} \text{mean}(\mathbf{H})$
9:       **end for**
10:       $\tilde{\mathcal{H}}_m^{s+1} \leftarrow \tilde{\mathcal{H}}_m^{s+1} / \sum_{t=1}^{N} \omega_{\alpha_t}$
11:   **end for**
12: **end for**
13: $\tilde{\mathcal{H}}^* \leftarrow \text{Select}(\{\tilde{\mathcal{H}}_m^S\})$

---

pects, where $a_{\alpha_t}^t$ is the α_tth attribute in aspect A_t. Our algorithm 1 approximates the $\tilde{\mathcal{H}}^*$ by iteratively approaching a most balanced point with nearest neighbors from different attributes. First, we initialize the candidates $\{\tilde{\mathcal{H}}_m^0\}$ by randomly sampling points in the attribute space, calculating their distance to the closest point of each attribute $a_{\alpha_t}^t$, and selecting the top M samples with the smallest average distance to all attributes. At each iteration s, we choose the top-K[4] nearest points to $\tilde{\mathcal{H}}_m^s$ for each attribute and update $\tilde{\mathcal{H}}_m^{s+1}$ using the weighted average of these points. It is worth mentioning that ω_{α_t} is the weight used to balance attributes or favor some specifically, and a negative value of ω_{α_t} can even move away from a particular one. Finally, we select the best candidate from the last iteration S, which is expected to be in the intersection region, i.e., a representation related to multiple attributes.

3.4   Generation with Intersections

As illustrated in the right bottom of Figure 2, we convert the representation $\tilde{\mathcal{H}}^*$ obtained from the intersection area directly to the Prefix with MLP_θ and let the language model generate multi-attributed sentence Y from input $\mathcal{X}$ as:

$$Y = \arg\max_y p_{\text{LM}}(y | \text{Prefix}^*; \mathcal{X})$$
$$\text{Prefix}^* = \text{MLP}_\theta(\tilde{\mathcal{H}}^* + \lambda \varepsilon_i), \; \varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (5)$$

---

[4] We study the practical meaning and impact of K in §5.3.

When generating several attribute-relevant sentences for one attribute combination, we only need to calculate the intersection for it once.

4 Experiment

In this section, we demonstrate the effectiveness of our method on three-aspect control, including sentiment, topic, and detoxification.

4.1 Multi-Aspect Control Task

The datasets we use are the same as GeDi (Krause et al., 2021) and Contrastive Prefix (Qian et al., 2022). To balance the data scale across all aspects, we randomly sample 10k sentences from each dataset that is less than the number of samples GeDi uses, with each attribute equally dividing this amount. We use the IMDb movie reviews (Maas et al., 2011), the AGNews dataset (Zhang et al., 2015), and the Jigsaw Toxic Comment Classification Challenge Dataset[5] for sentiment, topic and detoxification aspects, respectively.

The prompts used for text generation are the same as those used in the PPLM (Dathathri et al., 2020), with 20 from its bag-of-words experiment and 15 from its discriminator experiment. We experiment with 8 combinations of the 3 aspects with 2 sentiments $\times$ 4 topics $\times$ 1 detoxification and generate 5 completions for each combination and each prompt. Totally, each model will generate $35 \times 2 \times 4 \times 1 \times 5 = 1400$ sentences. It is worth noting that we do not specifically use prompts that induce the language model to generate toxic text, making detoxification easier to improve.

To measure the performance on different aspects, we compute the attribute relevance. We finetune a DeBERTa (He et al., 2021b,a) classifier on the Yelp dataset (Zhang et al., 2015) for sentiment aspect and a classifier for topic utilizing all its remaining data not used during training. We evaluate the non-toxicity with the Google Perspective API[6]. The final performance of a model is determined by the average of these three attribute relevance scores introduced above. We also use two auxiliary metrics to measure text quality. One is perplexity calculated by GPT2-large following Contrastive Prefix (Qian et al., 2022). To ensure that models are not insensitive to changes in different prefixes, we calculate the Distinctness (Li et al., 2016) of sentences

generated from different prefixes and average the 1-gram, 2-grams, and 3-grams distinct scores for simplicity. Moreover, we conduct human evaluation with sentences generated by different models shuffled. Each sentence is rated by three professional evaluators for 3 attribute relevance and text fluency. Evaluators rate each item on a scale of 1 to 5, with 5 representing text highly related to the desired attribute or very fluent.

4.2 Baselines

(I) **Weighted Decoding**: **PPLM** (Dathathri et al., 2020) biases the language model with gradients back-propagated from trained classifiers. **GeDi** (Krause et al., 2021) influences the decoding process with token probabilities conditioned on attributes. (II) **Multi-objective Optimization**: **MUCOCO** (Kumar et al., 2021) regards the decoding process as a constrained optimization problem, where the language model is the objective function and attributes are constraints. **Mix&Match** (Mireshghallah et al., 2022) controls attributes with energy-based models and generates sentences by masking, sampling, and correcting. (III) **Prefix–Tuning**: **Contrastive Prefix** (Qian et al., 2022) utilizes prefixes to activate the language model to generate attribute-relevant sentences by concatenation or semi-supervision.

4.3 Results

According to the automatic evaluation results in Table 1, under the multi-aspect setting, we group models based on their type of methods in chronological order. In addition, we demonstrate their standard deviations, which reflect the stability of models among different attribute combinations.

For weighted decoding, GeDi uses more powerful classifiers than PPLM and performs better on attribute relevance, stability to different combinations, and distinctness while correspondingly worse on perplexity. Multi-objective optimization methods achieve a favorable performance on attribute relevance while MUCOCO explodes on perplexity due to its non-autoregressive paradigm not being suitable for generating from scratch. Performance of semi-supervised Contrastive Prefix is similar to GeDi, except for lack of diversity.

Our method performs best on average attribute-related metrics, with at least a 7.3% significant improvement over existing baselines. Our advances mainly come from sentiment and topic aspects, with no less than 13.9% and 10.3% each. Al-

Methods	Average↑ (%)	Sentiment↑ (%)	Topic↑ (%)	Detoxification↑ (%)	PPL.↓	Dist.↑
Weighted Decoding Based Methods						
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
Multi-Objective Optimization Based Methods						
MUCOCO	73.9 ± 24.1	65.0 ± 33.7	67.2 ± 18.3	89.5 ± 3.5	405.6	49.7
Mix&Match	79.7 ± 21.8	73.5 ± 25.9	69.9 ± 21.1	95.8 ± 1.9	63.0	61.8
Prefix-Tuning Based Methods						
Contrastive Prefix						
concatenation	77.2 ± 18.5	67.3 ± 20.7	71.8 ± 16.5	92.6 ± 2.9	54.6	39.9
semi-supervised	81.3 ± 16.5	74.4 ± 19.6	76.9 ± 16.7	92.7 ± 3.5	31.9	43.3
Ours	**87.4 ± 10.9**	**86.7 ± 10.5**	**84.8 ± 14.2**	90.7 ± 7.4	28.4	49.5
w/o $\mathcal{L}_G$	80.9 ± 16.2	71.6 ± 11.7	75.9 ± 18.9	95.3 ± 2.6	71.5	58.9
w/o $\mathcal{L}_C$	62.3 ± 41.8	49.1 ± 49.8	41.7 ± 36.0	**96.0 ± 0.1**	473.0	37.0

Table 1: Automatic Results on Multi-Aspect Control. Hyperparameters and details are in §B.

though our model is not the best on detoxification, it is the most balanced and stable according to the lowest standard deviation on average, 10.9. As a prefix-tuning-based method inducing the language model without direct modification, which is naturally good at text fluency, we perform well on perplexity and inherit the performance on diversity.

Furthermore, we conduct ablation on aspect gap loss $\mathcal{L}_G$ and attribute classification loss $\mathcal{L}_C$ separately. On the one hand, without $\mathcal{L}_G$, we can not alleviate the bias in different training datasets, making it hard to search for the intersection areas. Since training sentences of sentiment and topic aspects are mainly non-toxic, our model focuses more on detoxification rather than struggling for the other two, leading to considerable declines on their relevance while slight improvements on detoxification. Besides, as the distance among sample points from different aspects in the attribute space increases, our model will generate sentences mapped from far more sparse areas, leading to a small decrease on fluency and a subtle increase on diversity. On the other hand, without $\mathcal{L}_C$, our attribute space will totally collapse. The relevance of sentiment and topic drops drastically while the non-toxicity boosts because model can hardly distinguish representations of different attributes in the same aspect and focus on relatively more effortless detoxification. Worse still, without distinct representations, our model is required to recover different sentences from similar ones, leading to oscillation in training and hardly generating complete text when inferencing.

Results of human evaluation are in Table 2, with inter-annotator agreement being 0.36 in Fleiss' κ. We evaluate GeDi, Contrastive Prefix, and our method and observe that the results are consistent with the automatic ones on sentiment and topic relevance. The performance of models on detoxi-

Methods	Sent.↑	Topic↑	Detox.↑	Fluency↑
GeDi	2.96	2.72	4.59	3.08
Con. Prefix	2.84	2.90	4.40	2.26
Ours	**3.47**	**3.39**	**4.71**	**3.69**

Table 2: Human Evaluation on Multi-Aspect Control.

fication is high and relatively similar, making the automatic results different from the manual ones where the annotators believe that our model does a better job than baselines. Since perplexity is relatively unreliable, the manually measured fluency of GeDi is much better than that of the Contrastive Prefix. And our method achieves the best fluency.

5 Analysis

5.1 Effect of Different Attributes and their Combinations

We illustrate the detailed results of each attribute and their combinations in Table 3. GeDi and Prefix-tuning perform differently in *single-aspect* control, each with its advantages. For example, GeDi is dedicated to *negative* with 93.9% relevance, while Prefix-tuning is good at *positive* with 90.6% relevance. When dealing with multi-aspect control, they inherit such imbalanced characteristics, with *average* relevance of 91.1% and 79.1%, respectively. In addition, the baselines decrease correspondingly in the *average* relevance of each attribute compared to *single-aspect*, ranging from 0.7 to 33.0. On average, our model outperforms other baselines on attribute metrics (Table 1). In detail, our model performs competitively for most attributes compared to another prefix-tuning-based model, Contrastive Prefix. Especially, on attributes like *business* and *sci/tech*, our model significantly improves over another prefix-tuning-based method on multi-aspect control and can even surpass it

Methods	Sentiment (%)		Topic (%)				Detox. (%)
	Neg.	Pos.	World	Sports	Business	Sci./Tech.	
Weighted Decoding Based Methods							
GeDi *single-aspect*	93.9	70.7	73.4	85.7	75.7	98.0	94.9
GeDi	94.7	-	80.0	-	-	-	90.6
	84.2	-	-	74.8	-	-	93.9
	94.9	-	-	-	75.7	-	96.6
	90.6	-	-	-	-	80.1	92.8
	-	53.7	61.4	-	-	-	94.4
	-	60.5	-	74.3	-	-	95.2
	-	57.6	-	-	54.3	-	95.7
	-	72.3	-	-	-	90.2	94.2
average	**91.1** (−2.8)	61.0 (−9.7)	70.7 (−2.7)	74.6 (−11.1)	65.0 (−10.7)	85.2 (−12.8)	**94.2** (−**0.7**)
Prefix-Tuning Based Methods							
Prefix *single-aspect*	88.4	90.6	74.5	85.3	93.5	93.6	93.8
Contrastive Prefix semi-supervised	65.5	-	**80.6**	-	-	-	91.8
	67.2	-	-	**90.3**	-	-	92.5
	56.0	-	-	-	79.2	-	92.2
	90.0	-	-	-	-	93.3	84.8
	-	93.5	64.8	-	-	-	95.1
	-	41.8	-	78.5	-	-	94.8
	-	87.4	-	-	41.7	-	95.2
	-	93.6	-	-	-	86.7	95.3
average	69.7 (−18.7)	79.1 (−11.5)	**72.7** (−**1.8**)	**84.4** (−**0.9**)	60.5 (−33.0)	90.0 (−3.6)	92.7 (−1.1)
Ours	69.7	-	71.7	-	-	-	84.1
	78.6	-	-	80.0	-	-	80.2
	99.9	-	-	-	**96.7**	-	96.8
	92.8	-	-	-	-	**98.0**	81.7
	-	80.5	58.0	-	-	-	95.1
	-	84.7	-	86.6	-	-	94.5
	-	87.6	-	-	91.7	-	**98.1**
	-	99.7	-	-	-	96.1	95.4
average	85.3 (−3.1)	**88.1** (−**2.5**)	64.9 (−9.6)	83.3 (−2.0)	**94.2** (+**0.7**)	**96.8** (+**3.2**)	90.7 (−3.1)

Table 3: Detailed Results on Single-Aspect and Multi-Aspect Control. We demonstrate results on *single-aspect* and *average* results on multi-aspect control with their difference to *single-aspect*, where other rows each represent an attribute combination. Cases are in §C. Detailed results for other baseline models and our ablations are in §D.

under *single-aspect* control.

In addition, correlations between attributes vary widely, as in Table 3. For example, generally, *positive* fits well with *non-toxic* while *negative* leads to a massive drop in non-toxicity, which is consistent with the intuition that one can hardly praise people and offend them simultaneously. Besides, *world* and *business* news are often reported negatively, such as war, famine, inflation, etc., making it challenging to combine them with *positive*. When attributes are not closely correlated, which means that few natural sentences possess these attributes together, our method is more likely to capture such a rarely occurred incident and magnify their frequency. Take *business* as an example. It is effortless to achieve a fine attribute relevance when performing single-aspect control on *business*, with GeDi achieving 75.7 and Prefix obtaining 93.5. After attaching *positive* to *business*, baseline models will suffer from a decline due to their weak correlation, where GeDi and Contrastive Prefix drop to 54.3 and 41.7, respectively. In contrast, our method can alleviate this problem by retrieving this unusual co-occurrence in the training sentences and recovering it from the attribute space, achieving a performance of 91.7, which is close to single-aspect
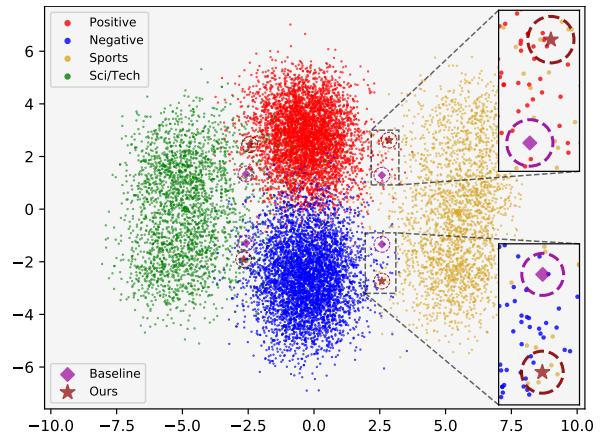

Figure 3: Projection of 4 attributes from attribute space.

control. When combining business with negative, which is a relatively common combination, there is still some decrease for baseline models. On the contrary, our method can even obtain the performance of 96.7 that surpasses single-aspect control.

5.2 Estimated Attribute Space

We demonstrate part of our estimated attribute space in Figure 3 with four attributes: *positive*, *negative*, *sports*, and *sci/tech* from sentiment and topic aspects. We project the high-dimensional space

K	Avg.↑	Sent.↑	Topic↑	DeTox.↑
5000	75.5	70.5	67.9	88.2
4000	77.6	72.9	71.4	88.4
3000	78.7	72.4	74.7	88.9
2000	79.1	72.6	75.9	88.7
1500	79.9	73.6	77.1	89.0
1000	80.7	75.7	77.2	89.1
800	82.9	79.3	79.2	90.3
500	85.2	83.5	81.5	90.5
300	85.7	84.1	83.2	89.7
200	**87.4**	**86.7**	**84.8**	90.7
150	84.0	79.2	84.3	88.4
100	83.9	78.7	83.6	89.5
50	82.2	78.4	78.5	89.6
20	80.9	77.8	73.1	91.7
10	80.8	79.6	71.5	91.2
5	81.4	82.9	69.3	92.1
3	85.0	86.1	77.7	91.1
1	78.8	63.1	80.9	**92.4**

Table 4: Results that vary with K.

to 2D with Principal Component Analysis (PCA). Consistent with our hypothesis, distributions of *sports* and *sci/tech* are asymmetric and the intersections lie in the sparse edges of attributes' distribution. In addition, we project the intersections searched by the baseline's strategy and ours, respectively. For *positive-sci/tech* and *negative-sci/tech* pairs, the combinations are relatively tight, making it easy to find intersections. However, intersection areas for *positive-sports* and *negative-sports* pairs are considerably sparse. As shown in enlarged area, the baseline searched intersection is at the midpoint of the two distributional centers, but this location is not where the attributes intersect. On the contrary, our method can find an intersection in such a sparse region, making various points from the two different attributes appear simultaneously in its tiny surrounding area. It worth noting that *positive* and *negative* appear to intersect in this projection because they are close in the high-dimensional space. But there is actually no intersection if only projecting these two attributes in §A.3.

5.3 Effect of K

We analyze the variation of K in the intersection searching algorithm and demonstrate the results in Table 4. Our model reaches a critical point when K is 200, and the performance is optimal this time. On the one hand, as the value of K gradually increases, our method pays less attention to regions where samples are fewer while attributes combine more tightly, and the performance decreases accordingly. When K reaches 5k, our method degenerates into a plain prefix-tuning model, which treats intersection as the midpoint of distributional centers. Its performance is similar and slightly inferior to
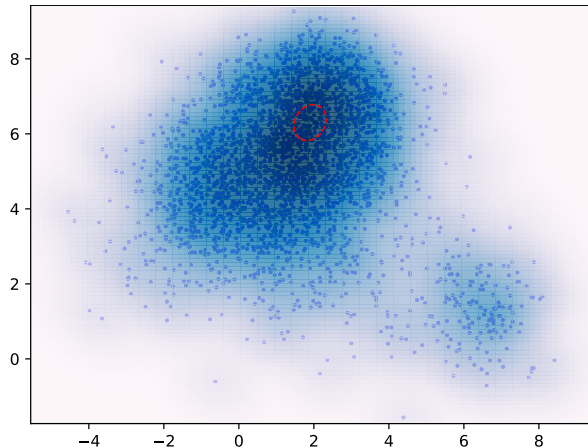

Figure 4: Distribution of attribute World from Topic.

the concatenation version of Contrastive Prefix in Table 1. On the other hand, smaller K leads to suboptimal performance since the effect of noise becomes non-negligible in training data. When K is less than 10, our model will be very unstable.

5.4 Distribution of Attributes

We project sample points to 2D by PCA, with each attribute projected independently. As in Figure 4, we display a scatterplot of World and conduct a Gaussian kernel density estimation to visualize its probability distribution. The darker area denotes a higher probability, where more representation points of oracle sentences gather. And the region annotated by a red ellipse is the estimated distributional center. As in the plot, the distribution of World is significantly asymmetric as the center lies in the top part, with the bottom being a sparse long tail. In addition, the distribution is even non-convex with an isolated cluster in the lower right corner. This observation supports our hypothesis that the practical distributions of attributes are far more complex than symmetric distributions such as Gaussian distribution. Besides, we plot the distribution of other attributes in the §A.1.

6 Discussion on Distributional Lens

Pilot work such as DGC (Khalifa et al., 2020) estimates the language distribution with an energy-based model and optimizes this distribution to satisfy constraints by approaching the constraints manifold. Recent distributional approaches like COLD Decoding (Qin et al., 2022) and MuCoLa (Kumar et al., 2022) take the language and attribute distribution in the same space so as to sample attribute-related sentences with Langevin Dynamics. Concurrent work on the image side, PromptGen (Wu

et al., 2022), simulates the complex distribution of images relevant to target attributes using a deep generative model. However, as a consensual hypothesis in manifold learning, the pre-trained language model estimates a low-dimensional manifold of language in a high-dimensional embedding space, which means most points in the embedding space are not probabilistically modeled by the language model. We believe that placing too much trust in the distributional modeling ability of language models is not a good choice. Our method attempts to depict the attribute space with discrete sample points of attributed sentences and make these discrete points, along with their coverage areas, compose the support set of our estimated distribution.

7 Conclusion

In this work, we present a distributional perspective for the multi-aspect controllable text generation with experimental results confirming the superiority of our model. Further observations on the 2D projection of the estimated attribute space show that our hypothesis about the attribute space is more feasible. In the future, we can explore the correlation between different attribute combinations for more fine-grained control and capture the bias in datasets to eliminate or utilize it.

Limitations

Our method has a certain dependence on the data since we need to estimate an attribute space. Therefore, it is difficult for our method to perform well in the setting of few-shot learning. However, this disadvantage is not that severe, because we only need single-aspect data, which is relatively sufficient in style transfer tasks. Another dependence of our method on data is that it is somewhat sensitive to biases in the data. When the semantic divergence of different aspects in training data is too large, our aspect gap loss, which aims to reduce the distance among the distributions of each aspect, will conflict with the sentence reconstruction loss. As a result, it may be hard to obtain a reliable intersection in the attribute space.

Computational resources also have an impact on our approach, as our aspect gap loss leverages a batch-level estimation for each aspect. Therefore, a larger batch size means a more accurate approximation, leaving the attribute space fewer biases. An alternative strategy for smaller batches is to backpropagate the loss after accumulating enough distributional samples, which requires more training epochs.

Ethics Statement

We are totally aware that text generation technology has a potential to be used maliciously to generate fake, toxic, or offensive content. However, after training on the Detoxification aspect, controllable text generation technology is a powerful weapon for combating hate speech, and eliminating harmful information in pre-trained language models. In addition, our multi-aspect controllable text generation technology can take Detoxification as an default aspect when controlling other aspects. We believe it meaningful and beneficial to advance research on controllable text generation.

Acknowledgements

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 6837–6857, Dublin, Ireland. Association for Computational Linguistics.

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 *International Conference on Learning Representations*.

Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li. 2020. Pre-train and plug-in: Flexible conditional text generation with variational autoencoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 253–262, Online. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1587–1596. JMLR.org.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2020. A distributional approach to controlled text generation. In *International Conference on Learning Representations*.

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 the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta

Cana, Dominican Republic. Association for Computational Linguistics.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Constrained sampling from language models via langevin dynamics in embedding spaces. *arXiv preprint arXiv:2205.12558*.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zhiyu Lin and Mark Riedl. 2021. Plug-and-blend: A framework for controllable story generation with blended control codes. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 62–71, Virtual. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021a. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland,

Oregon, USA. Association for Computational Linguistics.

Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A. Smith, and James Henderson. 2020. Plug and play autoencoders for conditional text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6076–6092, Online. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 751–760, New York, NY, USA. Association for Computing Machinery.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

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

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *arXiv preprint arXiv:2202.11705*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando De la Torre. 2022. Generative visual prompt: Unifying distributional control of pre-trained generative models. *arXiv preprint arXiv:2209.06970*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022. Tailor: A prompt-based approach to attribute-based controlled text generation. *arXiv preprint arXiv:2204.13362*.

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

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 *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Distribution of Attributes

A.1 Independent Projection of Attributes

We project sample points to 2D by Principal Component Analysis, with each attribute projected independently. We display a scatter plot for each and perform the Gaussian kernel density estimation. The darker area denotes a higher probability, where more representation points of oracle sentences gather. And the region annotated by a red ellipse is the estimated distributional center.

We underline distributions of attributes in Figures 5 to 7, including World, Sports, and Sci/Tech, which are significantly asymmetric. And especially, the projected distribution of the World attribute is even non-convex. This observation supports our hypothesis that the practical distributions of attributes are far more complex than symmetric distributions such as Gaussian distribution.
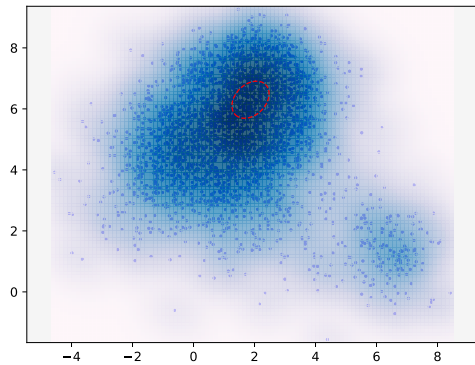

Figure 5: Distribution of World attribute from Topic aspect.
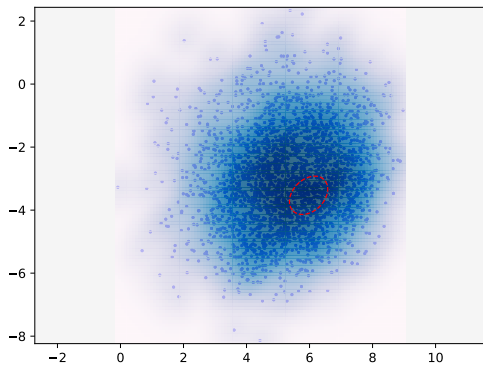

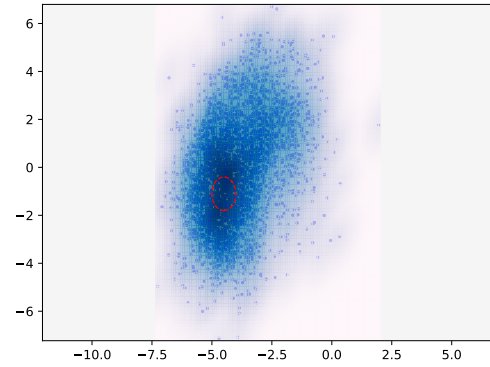Figure 6: Distribution of Sports attribute from Topic aspect.

Figure 7: Distribution of Sci/Tech attribute from Topic aspect.

In addition, we plot projected distributions of other attributes in Figures 8 to 12. Attributes such as Positive and Negative seem roughly symmetric in 2D projection. However, we can not guarantee their symmetry in high-dimensional space. Because the PCA aims to identify directions along which the variation in the data is maximal. In other words, the direction selection strategy is not necessarily related to symmetry or asymmetry, which means these 2D symmetric distributions may be asymmetric in high-dimensional space, with the asymmetric directions ignored during projection. Worse still, the long-tail region for a skewed direction may be too sparse, leading to lower variation compared to symmetric directions.
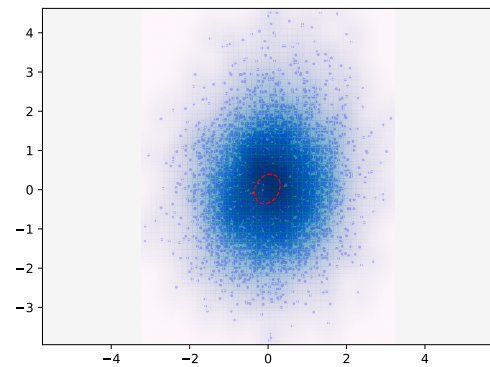

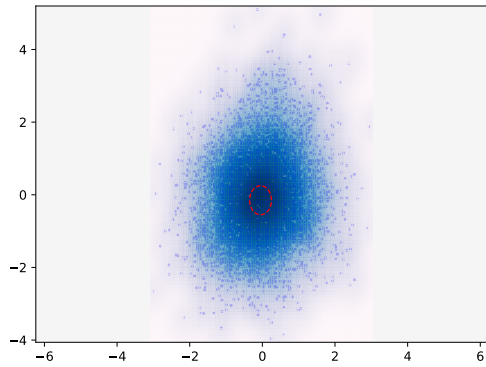Figure 8: Distribution of Negative attribute from Sentiment aspect.

Figure 9: Distribution of Positive attribute from Sentiment aspect.
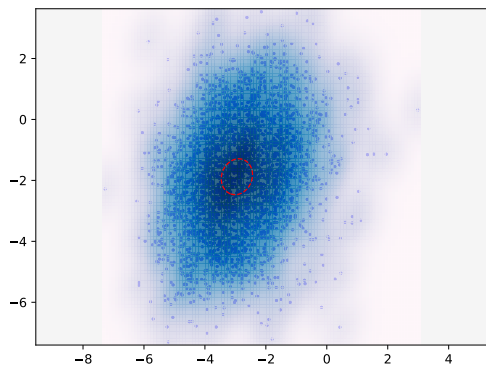

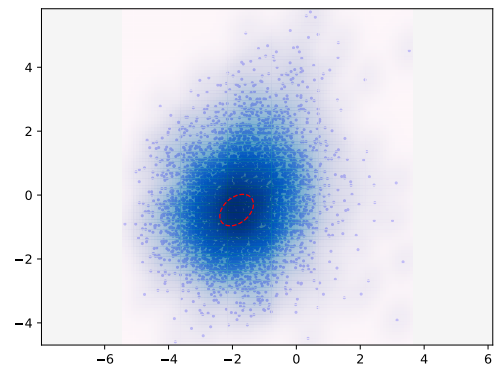Figure 10: Distribution of Business attribute from Topic aspect.

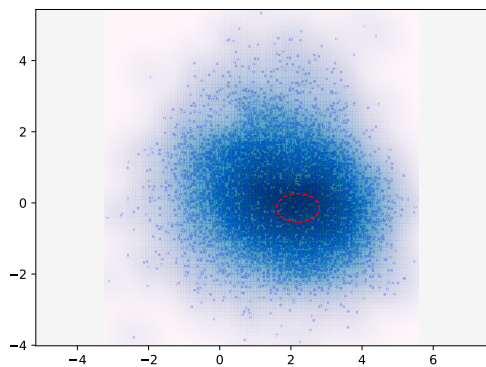Figure 12: Distribution of Non-toxic attribute from Detoxification aspect.

Figure 11: Distribution of Toxic attribute from Detoxification aspect.

A.2 Joint Projection of Attributes

We project combined sample points of attributes from three different aspects jointly to 2D by PCA. We display a scatter plot for each combination in Figures 13 to 20. The intersection points calculated on baselines' interpolation strategy and our intersection searching algorithm are plotted with **Baseline** and **Ours**, respectively. From these figures, we observe that NonToxic can mainly cover two sentiment attributes or at least possess large intersection areas. Besides, the intersection areas among sentiment attributes and topic attributes, except for the Sci/Tech, are narrow and sparse. Compared with the baselines' strategy, our search algorithm is closer to the intersection area, especially on Negative and Business attributes in Figures 13 to 15 and 19.

Figure 15: Jointly projected distributions of attributes: Negative, Business, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.

Figure 13: Jointly projected distributions of attributes: Negative, World, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.

Figure 16: Jointly projected distributions of attributes: Negative, Sci/Tech, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.

Figure 14: Jointly projected distributions of attributes: Negative, Sports, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.
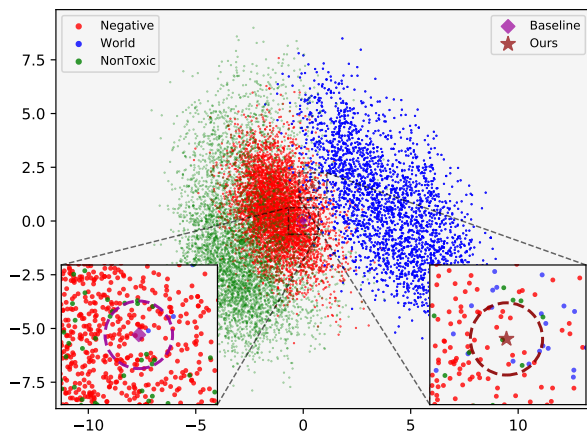

Figure 17: Jointly projected distributions of attributes: Positive, World, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.
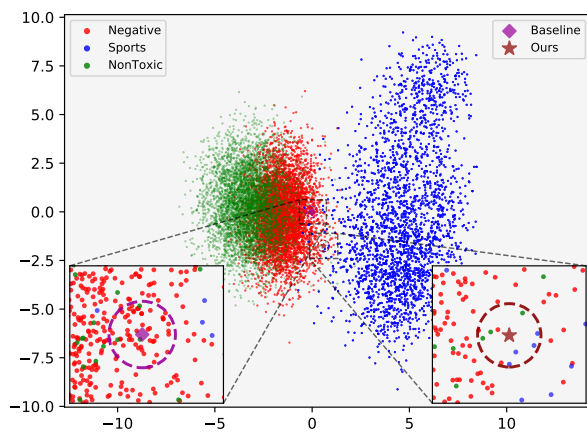
Figure 18: Jointly projected distributions of attributes: Positive, Sports, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.
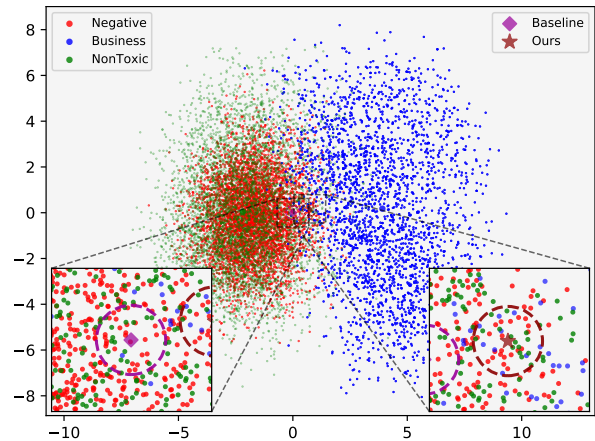

Figure 19: Jointly projected distributions of attributes: Positive, Business, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.
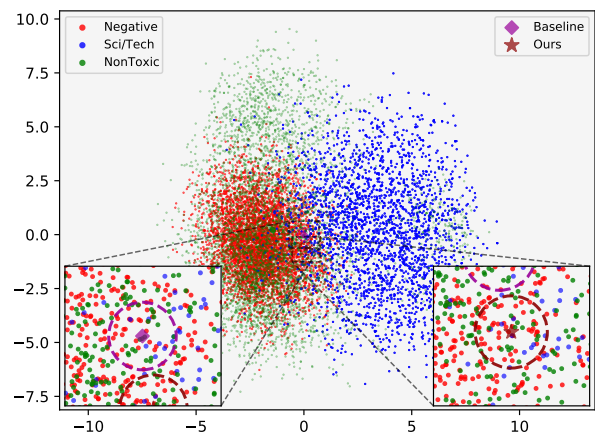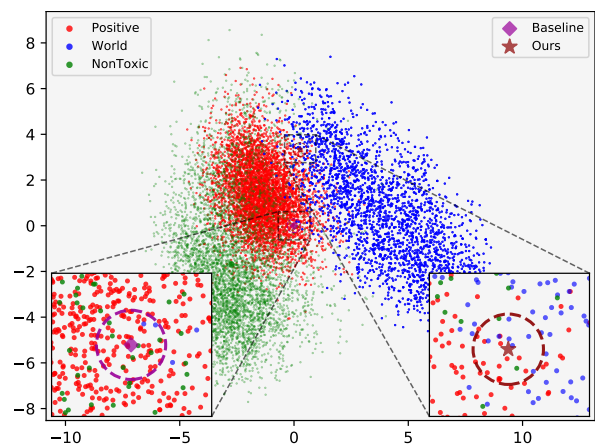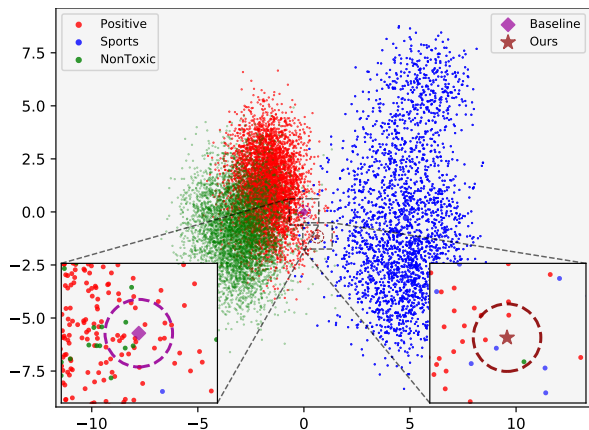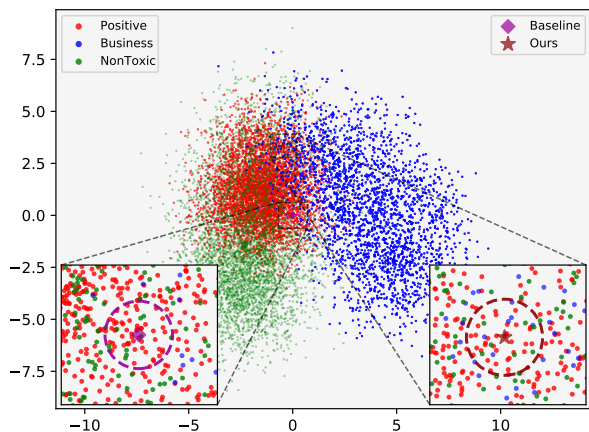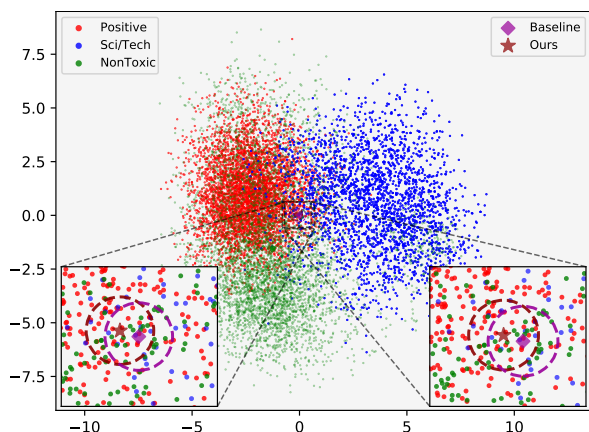

Figure 20: Jointly projected distributions of attributes: Positive, Sci/Tech, and NonToxic from aspects: Sentiment, Topic, and Detoxification, respectively.

A.3 Projection of Positive and Negative
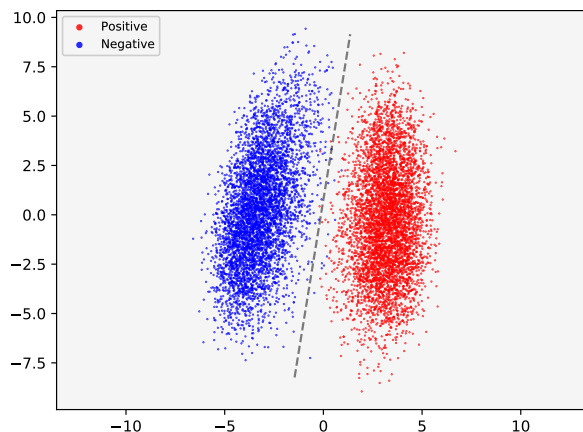
Figure 21: Jointly projected distributions of Positive and Negative.

Except for some noise in the dataset, positive and negative do not intersect when jointly projected.

B Hyperparameters and Details

Our methods are implemented using the Hugging face Transformers package. Our encoder is initialized with Bert-base-uncased, and the fixed decoder uses GPT2-medium. For any sentence, it will be tokenized with WordPiece tokenizer from Bert and Byte-Pair Encoding tokenizer from GPT2 before input to encoder and decoder, respectively. We perform mean pooling on outputs of the encoder and convert them to 768-dimensional latent representations, which are points in our attribute space. Afterward, latent representations will be mapped to the prefix with a dimension of $20 \times 24 \times 2 \times 1024$, 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 1024 is the size of hidden states in GPT2-medium. It's worth noting that prefix length Contrastive Prefix uses for single-aspect control is 10 and for multi-aspect control is $10 \times$ number of aspects, which is 30 for three-aspect control. Our prefix length is fixed to 20, which has nothing to do with the scale of aspects.

During the training stage, we use half-precision mode for efficiency on one NVIDIA A100 80GB GPU, where the batch size is 128 since the larger batch size better alleviates the aspect gap loss. In our setting, the random seed is 0, $w_1 = 0.5, w_2 = 0.2, w_3 = 0.3$, variation hyperparameter λ is 1e-3, the optimizer is AdamW with a learning rate of 1e-4, the number of training epochs is 150, and we use a checkpoint at the

step 30000. The training phase takes about 8 hours, and we experiment 6 times to search for the $\lambda \in$ [2e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5], while the other hyperparameters are initial settings.

Combination	Weight
Neg. & World & NonTox.	2 : 7 : 1
Neg. & Sports & NonTox.	2 : 4 : 1
Neg. & Business & NonTox.	2 : 8 : 1
Neg. & Sci./Tech. & NonTox.	3 : 1 : 3
Pos. & World & NonTox.	2 : 12 : 1
Pos. & Sports & NonTox.	3 : 5.5 : 1
Pos. & Business & NonTox.	2 : 9 : 1
Pos. & Sci./Tech. & NonTox.	3 : 1 : 1

Table 5: Specialized Weight for Attribute Balance.

During the inference phase, the maximum number of iterations T is 15, the number of candidates M is 1000, and the number of neighbors K is 200. We utilize a specialized list of weight parameters for each combination of attributes in Table 5, which aims to balance the performance among attributes from different aspects. After the iteration of intersection searching, our strategy is first to select the top 10 candidates with the smallest distances to their neighbors as the final candidate set. Then we randomly choose a candidate from these ten as the intersection's representation for text generation diversity. Our text generation process is the same as prefix tuning with sequence length set to 50. Except for model and data loading, the entire evaluation process for each attribute combination, including intersection searching, text generation, and attribute-relevance evaluation, takes about 2 minutes. Therefore, we can manually tune the weight of attributes to balance them, with a maximum trial number of 8 for each weight.

35 prompts we used in the inferencing stage are following the PPLM setting with 20 from its bag-of-word setting and 15 from its discriminator setting:

- **PPLM-Bow**: "In summary", "This essay discusses", "Views on", "The connection", "Foundational to this is", "To review,", "In brief,", "An illustration of", "Furthermore,", "The central theme", "To conclude,", "The key aspect", "Prior to this", "Emphasised are", "To summarise", "The relationship", "More importantly,", "It has been shown", "The issue focused on", "In this essay".

- **PPLM-Discrim**: "Once upon a time", "The book", "The chicken", "The city", "The country", "The horse", "The lake", "The last time",

"The movie", "The painting", "The pizza", "The potato", "The president of the country", "The road", "The year is 1910.".

Detailed setting of baselines: (I) **Weighted Decoding**: For **PPLM**, we only retrain its classifier heads on our datasets while keeping all other original settings. For **GeDi**, We use its code directly since we are following its setting. (II) **Multi-objective Optimization**: **MUCOCO** provides a solution for custom classification constraints, and thus we train these classifiers on our datasets. **Mix&Match** is relatively complex as it can not generate long sentences from scratch with the mask language model Bert. Worse still, as a method based on sampling, it is somewhat dependent on initialization. Therefore, we use sentences generated by PPLM as the starting sentences and let Mix&Match slowly polish the text by itself in iterations. (III) **Prefix-Tuning**: We reproduce **Contrastive Prefix**[7] and achieve comparable results. For a fair comparison, we unify the pre-trained language model to GPT2-medium (345M parameters) except for Mix&Match using Bert-large (340M parameters).

C Cases

We illustrate cases of 8 different attribute combinations in Table 6.

---

[7]They didn't make the code public. We reproduce the code and get comparable results to what they report.

Attributes	Models	Generation Results
Neg. World NonTox.	PPLM	Furthermore, the government will be allowed to carry out airstrikes from inside Afghanistan even after the end of Nato's withdrawal, according to senior defence officials. The UK is also planning to use air strikes in Afghanistan and Iraq from the skies over its bases.
	GeDi	Furthermore, the Iraqi authorities plan to seize petrol from Shia vehicles and trucks as well. Iraqi Foreign Minister Shaida al-Abadi said Iraq was ready for such actions if US President Donald Trump issued a military order ordering it.
	MU COCO	Furthermore, the new law will make it sale for any individual or entity to regulated keep give away Mumbai firearm to anyone a a background check, including those who are prohibited from doing so under federal law.
	Mix& Match	Furthermore, the uk will be reluctant to carry out airstrikes anywhere in europe until near the end of nato's mandate, according to some nato officials. the uk is likewise reluctant to drop warheads against iran and iraq from the air over british territory anywhere;
	Prefix concate	Furthermore, the first and the first of his world. The world.S. The U.S. The U. The world's country and a new-year.
	Prefix semi	Furthermore, a new survey conducted by a new survey of the Middle Eastern population in the country was revealed to be a very close match for the official record of the National Socialist Party (NTP) in the country.
	Ours	Furthermore, the movie's main focus is getting rid of Robert Kennedy. This movie has no plot, no action and no even remotely decent characterizations. It's simply a glorified version of what happened to George Bush in 2004.
Neg. Sports NonTox.	PPLM	This essay discusses the role of private security forces in Libya. The military's role in this crisis can be divided into two phases: 1) The first phase involved the transfer and transfer of the control of the situation to a military body.
	GeDi	This essay discusses last season who was demoted away from the league and how his decline in playing time impacted the team as a whole. With detailed observations, analysis, stories provided by some of these players including Orlando City fullback Ben Sweat and Toronto.
	MU COCO	This essay discusses howwcsstore can Consent a more humane society and how Mold willroximately the way webp topics our own Intake and our relationship with them. enoughWhat is a body)? awa A body is the ~~transsexual porn~~ Franch structureglers glucobos
	Mix& Match	This essay from an official, who was investigating two suspected drug dealers, "failed to find any probable cause." he stated that "confusion reigned" as the two men "struggled for some time" while evans "continued throwing punches."
	Prefix concate	This essay discusses the fact of the original Germanic tradition of a man's attempt to make a name on the English football team and the fact of the English football league.
	Prefix semi	This essay discusses the fact that the NHL is not a national sport. It also provides a new perspective to the fact that the NHL is not a major league.
	Ours	This essay discusses how the Miami Heat lost to the Atlanta Hawks in a seven-minute overtime last night, and how they should never again be able to make it with their team mates.
Neg. Business NonTox.	PPLM	Foundational to this is the need for a national banking system for the purpose of financing the banking system. The Federal Reserve has already taken over this task by creating and controlling the money supply in the form of the Federal Reserve bank, which is now owned and operated by the Federal.
	GeDi	Foundational to this is the New York Int'l Fedal and Foreign Market Team. This practice includes facilitating contacts between two levels of financial institutions as necessary for a successful settlement of an equity investment transaction.
	MU COCO	Foundational to this is a RegulatorySPONSORED of community and debtor. not have a shared history of Recession bourgeois The struggles ofBuffpeople,SPONSORED the struggles of all of our individuals bunk are interrelated,_- we are all part ze the same struggle_.
	Mix& Match	Foundational to this is the woodward foundation is our belief that low - income housing has a devastating effect on mental health. our research indicates that depressive symptoms are much greater among the economically weaker. we believe that a depressed patient's life expectancy will decline dramatically as a result.
	Prefix concate	Foundational to this is the company.\n\n
	Prefix semi	Foundational to this is the company's ability to offer a full range of services to customers. The company has been able to offer a good product and service for years.
	Ours	Foundational to this is a $1 million cut from the company's annual budget, which means it could be difficult to cash out of its debt.
Neg. Sci/Tech NonTox.	PPLM	The last time I wrote a blog post about the future of Linux and how to prepare for it I wrote: I am very pleased to announce that the Linux kernel has been updated to 1.8.2! This is good news for the Linux.
	GeDi	The last time we examined this topic we estimated that 5% of the age-appropriate weight was unnecessary due to red fatherful neglect; only 14% of adopted healthy, fit barrels were considered but maintained. Our data predict that at least 44lb.

Neg. Sci/Tech NonTox.	**MU COCO**	The last time I checked VMware would's not like the world's largest and fastest ship was a single- Anonlled organism that could fly, and the last time I checked, it's not like the world's largest and fastest ship the a single-
	Mix& Match	The last time the lake was drained, however, to make way for a golf course at a nearby site, and there were reports on a nearby island that a large-scale groundwater cleaning program was beginning. the lake was drained, however, to make way elsewhere;
	Prefix *concate*	The last time you are a fan, you have to be aware of the bad news about the Internet. The problem is that this isn't even remotely acceptable to the author.
	Prefix *semi*	The last time I see this film I have not seen it in my life. I have watched it in a very limited number of days and I have been very disappointed. The acting is terrible and the acting is bad.
	Ours	The last time I saw this film was in the theater. It was terribly disappointing. There is no plot or suspense whatsoever, nor any action whatsoever. The only thing that can be attributed to this movie is the lack of a sound system.
Pos. World NonTox.	**PPLM**	The connection between obesity and autism has been identified for the first time using a unique antibody screening test, according to researchers with the University of Texas MD Anderson Cancer Center. They found that a protein called CD34 has a powerful impact on autism.
	GeDi	The connection between Greece and Russia reached new heights through cooperation on a number of initiatives States Parties undertook joint action to crack down on corruption abroad. For instance, the Russian Federation launched an all-cash inquiry aimed at identifying persons.
	MU COCO	The connection between staking two is not a loneliness of mere coincidence. The connection snowball dividing matter of history, and history BW a history arresting its own, of which hero are all the victims analyse nogyI don't believe in coincidence Alger said the
	Mix& Match	The connection is an illustration of the moon, from the book 'the lord of light and darkness', by william shakespeare (photo courtesy of william shakespeare). an illustration of the sun, from the book 'the lord of light and darkness', by william shakespeare (photo courtesy)
	Prefix *concate*	The connection of the United States's the world of the world's first-year of a new-run of the world in the world.S.
	Prefix *semi*	The connection between the world of the American National Rifle Association and the United States is a fascinating, fascinating, and hilarious tale. It has been an honor to see the film on the National Library shelves, and I am proud to see the film.
	Ours	The connection between John Lennon and the United States is as strong as ever. The Los Angeles Times reports that Lee Sternberg's performance of his song "Lenny Luerer" won a round of applause in the U.S. Senate.
Pos. Sports NonTox.	**PPLM**	More importantly, the first day of the 2017 NFL Draft is always exciting to watch with fans, because the league is going to get a lot of great talent on defense in the coming weeks. The biggest draft prospect to come out this year, Alabama DT Vic Beasley
	GeDi	More importantly, I appreciated his honesty along the way. Orlando Pace is usually a shadow of his former self, but he understood the importance of all that went into this win and smiled again.
	MU COCO	More importantly, he was able to defenders it work. it two men Fans in likeness fans on a Rugby coach. He had justovich from medical trip that Europe and was looking partners a place to eat. loved had never been in a bar
	Mix& Match	More importantly, the sixth game of the 2018 stanley cup finals presents a new challenge and an exciting new opportunity. the philadelphia flyers and pittsburgh penguins are joining forces for a six-game road trip that begins in the nation's capital each weekend.
	Prefix *concate*	More importantly, I have to remind everyone that this is a real story, so the fact that the two men were not a couple of people who have to be treated as one of those who would be involved with the team.
	Prefix *semi*	More importantly, the Boston Red Sox have lost the league title, and the players themselves are not yet qualified to be the best player in the league. The fact that they are not even qualified to play a match of the best.
	Ours	More importantly, the Houston Astros won a great opportunity to make a comeback with a victory over the Detroit Tigers in the National League West. The team has an outstanding offensive line and is tied for fifth in scoring among the nation.
Pos. Business NonTox.	**PPLM**	In brief, the federal tax law allows employers to deduct up to 20% of compensation expenses from workers' paychecks. This deduction is a big deal because many employees have to pay high deductibles for medical care.
	GeDi	In brief, Heiltsuk said that she holds central, shared concerns regarding how First Nations youth can navigate financial injustices faced by society and why net aboriginal debt was surpassed in 2015. All eight First Nations elected delegates at Monday's meeting
	MU COCO	In brief, Bach "sus anthologyions pione excel the outstanding Russian Returns in the hacking of capitalists economics Committee letters were not whirlwind. But that's not what the White Airways said in statement Alibaba late Tuesday afternoonisSpecial Orderable
	Mix& Match	In brief, the u. s. department of agriculture (usda) produced a comprehensive list of how many jobs were created in 2016, it identified 3. 1 million jobs in the agriculture sector, a dramatic uplift from 2015's 2 million.
	Prefix *concate*	In brief, the new of its company.\n\n

Pos. Business NonTox.	**Prefix** *semi*		In brief, it is the best movie I have ever seen, and I love it. The movie is a perfect blend of comedy and comedy. It is not a classic movie, but it is not a great movie.
	Ours		In brief, the economy is surged in July, boosted by strong sales of oil and other products, as well as strong growth in U.S. manufacturing.
Pos. Sci/Tech NonTox.	**PPLM**		The country's first solar power system, built by a group of students at Harvard University, is now operating. The project is aimed at encouraging solar energy development by encouraging collaboration among universities, community groups and individuals.
	GeDi		The country illustrated beautifully reflects the complexity of lives and customs.
	MU COCO		The country's top diplomat, Blockchain Lavrov IBM said the UydiaS. was "very much looking into" the matter. pleasantly engineers Rapp a Bridges supplier of hacker vegan Iran, has been trying to improve ties with blockchain, a close ally and
	Mix& Match		The country focuses on the role the united states has played in discovering new technologies for the advancement of science, according to two u. s. officials briefed on - site. both officials, newly appointed to handle national security matters welcomed the sensitive nature of the investigation.
	Prefix *concate*		The country's top TV channel is now a very popular TV show. The only thing is the name. I'm sure there are many people who would be willing to take it seriously, but I'll be ~~damned~~ to find out if they have a lot
	Prefix *semi*		The country's most famous TV series is the best and most powerful show ever made. The story is great, the action is good, the plot is great, and the story is very good. The cast is great.
	Ours		The country's biggest television network has announced that it will offer a new version of the movie which is based upon the popular "Star Trek" series. It's truly amazing to see how many people are involved in making this movie so far.

Table 6: Generated Cases. Red highlights the sentiment-related content. Blue highlights the topic-related content. Underlined are the input prompts. Strikethrough indicates toxic content.

D Detailed Results

Methods	Sentiment (%)		Topic (%)				Detox. (%)
	Neg.	Pos.	World	Sports	Business	Sci./Tech.	
Weighted Decoding Based Methods							
PPLM *single-aspect*	97.2	62.7	74.9	46.5	62.4	98.6	93.2
PPLM	92.2	-	75.4	-	-	-	82.0
	84.4	-	-	41.8	-	-	76.0
	87.5	-	-	-	61.5	-	82.9
	85.3	-	-	-	-	95.0	76.2
	-	35.4	59.1	-	-	-	90.4
	-	39.5	-	34.1	-	-	89.5
	-	40.9	-	-	48.3	-	91.2
	-	52.7	-	-	-	93.1	91.3
average	87.4	42.1	67.3	38.0	54.9	94.1	84.9
GeDi *single-aspect*	93.9	70.7	73.4	85.7	75.7	98.0	94.9
GeDi	94.7	-	80.0	-	-	-	90.6
	84.2	-	-	74.8	-	-	93.9
	94.9	-	-	-	75.7	-	96.6
	90.6	-	-	-	-	80.1	92.8
	-	53.7	61.4	-	-	-	94.4
	-	60.5	-	74.3	-	-	95.2
	-	57.6	-	-	54.3	-	95.7
	-	72.3	-	-	-	90.2	94.2
average	91.1	61.0	70.7	74.6	65.0	85.2	94.2
Multi-Objective Optimization Based Methods							
MUCOCO	97.9	-	54.5	-	-	-	85.7
	94.6	-	-	55.8	-	-	85.7
	96.8	-	-	-	65.6	-	87.3
	95.5	-	-	-	-	96.1	86.9
	-	30.4	48.0	-	-	-	91.0
	-	26.3	-	59.8	-	-	92.6
	-	34.6	-	-	62.1	-	93.8
	-	43.9	-	-	-	95.1	93.1
average	96.2	33.8	51.3	57.8	63.9	95.6	89.5
Mix&Match *single-aspect*	99.2	63.3	79.5	57.4	69.6	99.3	96.9
Mix&Match	96.1	-	**80.6**	-	-	-	93.1
	97.7	-	-	48.2	-	-	93.0
	98.2	-	-	-	66.6	-	97.0
	96.8	-	-	-	-	99.6	96.1
	-	53.0	67.3	-	-	-	95.5
	-	45.0	-	44.0	-	-	96.7
	-	41.5	-	-	55.8	-	97.7
	-	59.7	-	-	-	97.3	97.5
average	**97.2**	49.8	**74.0**	46.1	61.2	98.5	95.8
Prefix-Tuning Based Methods							
Prefix *single-aspect*	88.4	90.6	74.5	85.3	93.5	93.6	93.8
Contrastive Prefix concatenation	32.4	-	50.3	-	-	-	90.9
	88.1	-	-	73.8	-	-	89.1
	51.6	-	-	-	70.0	-	94.1
	94.3	-	-	-	-	94.1	88.3
	-	77.6	46.8	-	-	-	92.2
	-	70.2	-	78.5	-	-	95.9
	-	51.9	-	-	73.1	-	94.7
	-	72.0	-	-	-	88.1	95.6
average	66.6	67.9	48.5	76.2	71.6	91.1	92.6
Contrastive Prefix semi-supervised	65.5	-	80.6	-	-	-	91.8
	67.2	-	-	90.3	-	-	92.5
	56.0	-	-	-	79.2	-	92.2
	90.0	-	-	-	-	93.3	84.8
	-	93.5	64.8	-	-	-	95.1
	-	41.8	-	78.5	-	-	94.8
	-	87.4	-	-	41.7	-	95.2
	-	93.6	-	-	-	86.7	95.3
average	69.7	79.1	72.7	**84.4**	60.5	90.0	92.7
Ours	69.7	-	71.7	-	-	-	84.1
	78.6	-	-	80.0	-	-	80.2

99.9	-	-	-	**96.7**	-	96.8
92.8	-	-	-	-	98.0	81.7
-	80.5	58.0	-	-	-	95.1
-	84.7	-	86.6	-	-	94.5
-	87.6	-	-	91.7	-	**98.1**
-	**99.7**	-	-	-	96.1	95.4
average	85.3	**88.1**	64.9	83.3	**94.2**	96.8	90.7

64.3	-	51.8	-	-	-	90.1
71.5	-	-	71.0	-	-	93.4
68.2	-	-	-	59.9	-	95.7
62.4	-	-	-	-	**99.8**	96.0
-	92.0	60.6	-	-	-	97.6
-	59.4	-	**93.8**	-	-	94.3
-	86.8	-	-	72.1	-	97.9
-	68.3	-	-	-	98.4	97.2
average	66.6	76.6	56.2	82.4	66.0	**99.1**	95.3

Ours w/o Asp. Loss

99.2	-	15.2	-	-	-	96.5
99.8	-	-	36.5	-	-	96.3
97.8	-	-	-	17.9	-	95.4
84.9	-	-	-	-	97.7	95.6
-	3.2	14.4	-	-	-	96.3
-	0.1	-	40.4	-	-	96.0
-	1.3	-	-	13.9	-	95.7
-	6.5	-	-	-	97.7	95.8
average	95.4	5.6	14.8	38.5	15.9	97.7	**96.0**

Ours w/o Att. Loss

Table 7: Detailed Combination Results on Multi-Aspect Control.