An Extensible Plug-and-Play Method for Multi-Aspect Controllable Text Generation

Xuancheng Huang[1†] **Zijun Liu**[2,4†] **Peng Li**[3,5] **Tao Li**[1] **Maosong Sun**[2,4*] **Yang Liu**[2,3,4,5*]
[1]Meituan, China
[2]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[3]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[4]Beijing National Research Center for Information Science and Technology
[5]Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract

Recently, multi-aspect controllable text generation that controls the generated text in multiple aspects (e.g., sentiment, topic, and keywords) has attracted increasing attention. Although methods based on parameter efficient tuning like prefix-tuning could achieve multi-aspect controlling in a plug-and-play way, the mutual interference of multiple prefixes leads to significant degeneration of constraints and limits their extensibility to training-time unseen aspect combinations. In this work, we provide a theoretical lower bound for the interference and empirically found that the interference grows with the number of layers where prefixes are inserted. Based on these analyses, we propose using trainable gates to normalize the intervention of prefixes to restrain the growing interference. As a result, controlling training-time unseen combinations of aspects can be realized by simply concatenating corresponding plugins such that new constraints can be extended at a lower cost. In addition, we propose a unified way to process both categorical and free-form constraints. Experiments on text generation and machine translation demonstrate the superiority of our approach over baselines on constraint accuracy, text quality, and extensibility.[1]

1 Introduction

Multi-aspect controllable text generation (MCTG), which aims at generating fluent text while satisfying multiple aspects of constraints simultaneously, has attracted increasing attention in recent years (Chan et al., 2021; Qian et al., 2022; Gu et al., 2022). To effectively control diverse aspects such as sentiment, topic, and detoxification, extensive efforts have been devoted to the task, including methods based on conditional generative model (Keskar et al., 2019), decoding-time regulation (Lin and Riedl, 2021; Kumar et al., 2021), and parameter efficient tuning (Qian et al., 2022; Gu et al., 2022).

Despite their effectiveness, existing methods still suffer from low extensibility. Ideally, suppose a multi-aspect controllable text generation system has learned how to control sentiment, topic and keywords separately, it should be extensible to any combinations of the three aspects, e.g., generating a *sports-themed* sentence with *negative sentiment* containing *keywords "New York"* (see Figure 1). Moreover, an extensible system should also be easily extended to control new aspects in a plug-and-play way. However, it is non-trivial for existing methods to achieve this goal. Specifically, the dedicated conditional generative models (Keskar et al., 2019) mostly need to be trained from scratch or finetuned when facing new aspect combinations. The decoding-time regulation based methods (Lin and Riedl, 2021; Kumar et al., 2021) intervene in the probabilities of sentences by light-weight attribute classifiers or language models during inference, which significantly impairs text fluency when multiple distributions are interpolated. The parameter efficient tuning based methods (Qian et al., 2022; Gu et al., 2022) control aspects by inserting trainable prompts or prefixes into the model, referred to as plugins. By leveraging one plugin for each aspect, these methods can naturally work in a plug-and-play way, showing better potential to achieve high extensibility.

However, existing studies show that directly combining multiple plugins results in significantly lower controllability of the corresponding aspects than before combining (i.e., attribute degeneration) (Qian et al., 2022; Gu et al., 2022). Gu et al. (2022) argue that *mutual interference* of the plugins is the major reason for attribute degeneration, which is further justified by our theoretical and empirical analyses. Previous works alleviate the problem by introducing connectors to connect mul-

† indicates equal contribution.

* Corresponding authors: M.Sun (sms@tsinghua.edu.cn) and Y.Liu (liuyang2011@tsinghua.edu.cn)

[1]The source code is available at https://github.com/THUNLP-MT/PromptGating4MCTG

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15233–15256
July 9-14, 2023 ©2023 Association for Computational Linguistics

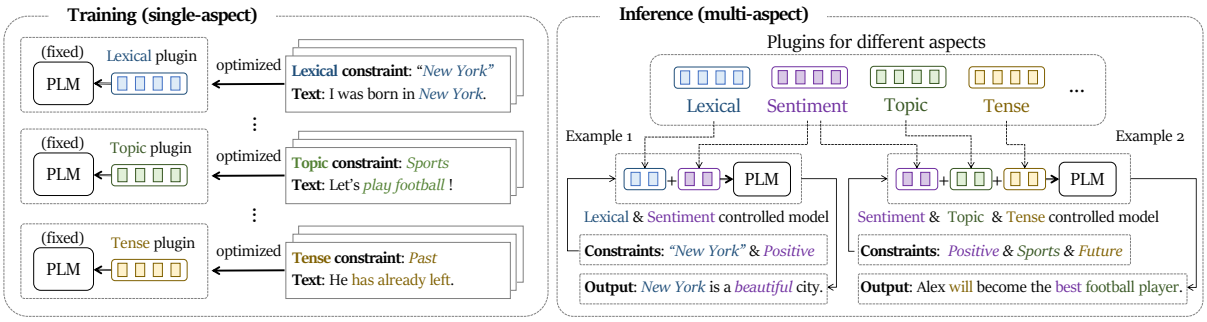

Figure 1: Overview of our proposed extensible plug-and-play method for multi-aspect controllable text generation. First, plugins are trained on single-aspect labeled data separately (left). Then, arbitrary plugins can be combined by simply concatenation and plugged into the pretrained model to satisfy corresponding combinations of constraints (right). Due to the separate training of different plugins, the cost of extending a new constraint is relatively low. Besides, our approach restrains the accumulation of mutual interference, alleviating the degeneration of constraints.

tiple plugins (Yang et al., 2022), latent variables to represent the unsupervised aspects (Qian et al., 2022), or objectives to narrow the discrepancy of aspects (Gu et al., 2022). However, these methods require joint training of plugins and are designed for pre-defined closed-set constraints. In consequence, their extensibility is limited.

In this paper, we propose an extensible plug-and-play method, PROMPT GATING, for multi-aspect controllable text generation. We derive a theoretical lower bound for the interference of plugins and reveal that it accumulates with the increasing number of layers where prefixes are inserted. Based on these findings, we propose attaching trainable gates to the plugins, which normalize the interventions of plugins. As a result, the mutual interference has been significantly reduced such that the control of arbitrary combinations of aspects can be realized by simply concatenating the corresponding plugins. Thus, our method is both extensible and plug-and-play. Moreover, we represent the constraints of the aspects in textual form, which makes our method applicable not only to categorical aspects (e.g., sentiment) but also to free-form aspects (e.g., lexical constraint).

Our contributions are three-fold:

- We propose an extensible plug-and-play method, PROMPT GATING, for multi-aspect controllable text generation, which is able to control training-time unseen aspect combinations by simply concatenating plugins.
- We provide a theoretical lower bound along with empirical analyses for the mutual interference problem, which we believe will facilitate future research.
- Experiments show that our approach has lower mutual interference, leading to superiority over strong baselines on text quality, constraint accuracy, and extensibility.

2 Background

In this section, we illustrate the widely-used prefix-tuning-based method (Qian et al., 2022; Gu et al., 2022; Yang et al., 2022) for multi-aspect controllable text generation. Generally, prefix-tuning (Li and Liang, 2021) prepends light-weight continuous vectors to the multi-head attention sublayer of each Transformer layer (Vaswani et al., 2017):

$$\mathbf{H} = \mathrm{Att}\Big(\mathbf{Q}, \big[\mathbf{P}^K; \mathbf{K}\big], \big[\mathbf{P}^V; \mathbf{V}\big]\Big), \quad (1)$$

where $\mathrm{Att}(\cdot)$ is the attention function, $\mathbf{Q}$ are queries of inputs, $\mathbf{K}$ and $\mathbf{V}$ are activations of inputs, $\mathbf{P}^K$ and $\mathbf{P}^V$ are trainable prefixes, $[\cdot;\cdot]$ denotes the concatenation operation, $\mathbf{H}$ is the output of the attention sublayer. We use $\boldsymbol{\phi}$ to denote the set of prefixes in all Transformer layers.

Specifically, for multi-aspect controllable text generation, we assume that there are N aspects of constraints. Due to the lack of multi-aspect labeled data, each set of prefixes, which usually represents a specific constraint (e.g., "positive" for the sentiment aspect), is trained on the corresponding single-aspect labeled data:

$$\hat{\boldsymbol{\phi}}_i = \underset{\boldsymbol{\phi}_i}{\mathrm{argmax}} \Big\{ P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi}_i) \Big\}, 1 \le i \le N, \quad (2)$$

where $\boldsymbol{\theta}$ are the fixed parameters of the pretrained model, $\mathbf{y}$ is the controlled target sentence, $\mathbf{x}$ is the

input sentence[2], $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi}_i)$ is the conditional probability of $\mathbf{y}$, and $\hat{\boldsymbol{\phi}}_i$ are the learned parameters of prefixes for the i-th aspect.

During inference, for a combination of multiple aspects, corresponding prefixes are selected and synthesized by either concatenating (Qian et al., 2022; Yang et al., 2022) or finding their intersection (Gu et al., 2022), and then the generation is conditioned on the synthesis. Without loss of generality, we take two aspects as an example. The conditioned probability can be represented as

$$P\Big(\hat{\mathbf{y}}|\mathbf{x};\boldsymbol{\theta},\mathrm{syn}(\hat{\boldsymbol{\phi}}_1,\hat{\boldsymbol{\phi}}_2)\Big), \tag{3}$$

where $\mathrm{syn}(\cdot)$ is a synthesize function, $\hat{\mathbf{y}}$ is the candidate sentence, $\hat{\boldsymbol{\phi}}_1$ and $\hat{\boldsymbol{\phi}}_2$ are two sets of prefixes corresponding to two aspects (e.g., "positive" for sentiment and "sports" for topic), respectively.

Although existing methods alleviate the mutual interference of prefixes by joint training (Qian et al., 2022; Gu et al., 2022; Yang et al., 2022), they are based on pre-defined closed-set constraints, which increases the overhead of extending a new constraint and thus limits extensibility. Thus, to maintain high extensibility, reducing mutual interference without joint training still remains a challenge.

3 Analyses on Mutual Interference

To alleviate mutual interference while maintaining extensibility, we conduct theoretical and empirical analyses. First, we provide a definition of mutual interference as follows.

Definition. Mutual interference (MI) is the interference between multiple plugins which are trained separately during training but are combined to guide the pretrained model simultaneously during inference (i.e., in the zero-shot setting). However, the exact interference is hard to analyze because of the complexity of deep neural networks. Intuitively, suppose multiple plugins are optimized simultaneously during training, which requires multi-aspect labeled data, their interference will be minimized because they have learned to work cooperatively under supervision (i.e., in the supervised setting). Therefore, we use the differences between the hidden states of the supervised and zero-shot settings

[2]Note that $\mathbf{x}$ is the source sentence or context for tasks like machine translation (Bahdanau et al., 2015) or summarization (Nallapati et al., 2016) and can be eliminated when it is not present.

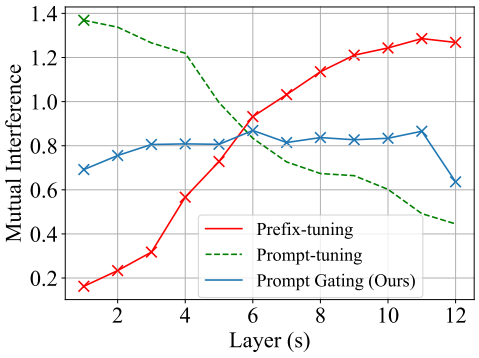

Figure 2: The variations of mutual interference with the number of Transformer layers. Note that "×" represents the insertion of continuous vectors. Prompt-tuning only inserts vectors into the model after the embedding layer, while the other two methods insert vectors into each Transformer layer. Our approach (Prompt Gating) restrains the growth of mutual interference while inserting sufficient trainable parameters.

to approximate the mutual interference of multiple plugins. Specifically, let $\hat{\boldsymbol{\phi}}_i$ and $\tilde{\boldsymbol{\phi}}_i$ be the parameters of plugins learned from the single- and multi-aspect labeled data, respectively. Taking two-aspect controlling as an example, the output of a Transformer layer is given by $\mathbf{H}(\mathbf{x},\boldsymbol{\phi}_1,\boldsymbol{\phi}_2)$, where $\mathbf{x}$ is the layer input, then mutual interference can be defined as

$$\mathrm{MI} \approx \Big\|\mathbf{H}(\mathbf{x},\hat{\boldsymbol{\phi}}_1,\hat{\boldsymbol{\phi}}_2) - \mathbf{H}(\mathbf{x},\tilde{\boldsymbol{\phi}}_1,\tilde{\boldsymbol{\phi}}_2)\Big\|. \tag{4}$$

Empirical Analysis. Then, as mutual interference has been defined as the norm of gap between hidden states in the zero-shot and supervised settings, we can empirically estimate it on the authentic dataset. By calculating the averaged norm on the Yelp dataset (Lample et al., 2019), we plot the variations of mutual interference with the number of Transformer layers for Prompt-tuning (Lester et al., 2021) and Prefix-tuning (Li and Liang, 2021) in Figure 2. We can find that the interference accumulates with insertions of trainable parameters. Moreover, the magnitude of mutual interference at the last Transformer layer (i.e., the 12-th layer in Figure 2) is consistent with the performance gap, which is the difference between the fulfillment of constraints in single- and multi-aspect settings (see Table 1). Meanwhile, too few trainable parameters cannot guide the pretrained model effectively. In summary, the key point for remaining effective in the zero-shot setting is *restraining the growth of mutual interference* (for a lower performance gap)

while *providing sufficient trainable parameters* (for better supervised performance).

Theoretical Analysis. Next, to find a way to alleviate mutual interference, we conducted a theoretical analysis.[3] As a result, we found that the mutual interference, which is caused by the interactions in attention sublayers, has a theoretical lower bound[4]:

$$\mathrm{MI} > \alpha\|\Delta\mathbf{h}_1(\mathbf{x}, \hat{\boldsymbol{\phi}}_1)\| + \beta\|\Delta\mathbf{h}_2(\mathbf{x}, \hat{\boldsymbol{\phi}}_2)\|, \quad (5)$$

where $0 < \alpha, \beta < 1$, and $\|\Delta\mathbf{h}_i(\mathbf{x}, \hat{\boldsymbol{\phi}}_i)\|$ is a norm that is positively related to the magnitude of $\hat{\boldsymbol{\phi}}_i$. Moreover, the lower bound might accumulate with Transformer layers like in Figure 2. Intuitively, applying normalization (e.g., gates) to the parameters of the i-th plugin to reduce its magnitude will decrease the lower bound of mutual interference.

4 Prompt Gating

We propose a novel approach that attaches trainable gates to the plugins, which alleviates the mutual interference of multiple plugins and makes the model highly extensible. Figure 3 shows the architecture of our approach. We first provide intuition in §4.1, then define our approach formally in §4.2.

4.1 Intuition

Although prefix-tuning provides sufficient interventions and avoids long-range dependencies by inserting continuous vectors into each attention sublayer, it suffers from the accumulation of mutual interference of these plugins (see §3). On the one hand, the vectors are inserted into the attention sublayer, where they interact with each other, which directly enhances mutual interference. On the other hand, the vectors are not normalized, which leads to a large lower bound of mutual interference (Eq. (5)). Intuitively, injecting the vectors in a position-wise manner will avoid direct interaction between them. Moreover, normalizing the vectors can limit the magnitude of the lower bound, which might decrease mutual interference. Therefore, we first propose attaching vectors outside the attention sublayer, which can be realized by appending trainable vectors to the output of the embedding layer and adding trainable vectors to the hidden states in each Transformer layer (see Figure 3). Then, trainable gates are applied to these hidden states to alleviate mutual interference further. In this way, we expect our approach to restrain the growth of mutual interference while providing sufficient interventions.

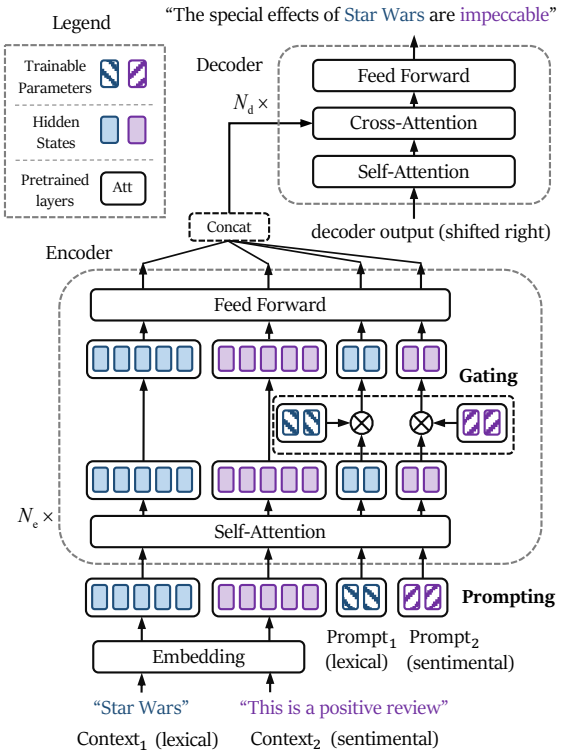

Figure 3: The architecture of our approach. It shows the case of inference stage of double-aspect controlled text generation. Blue and purple represent lexical and sentimental constraints respectively. Continuous prompts and contextual contexts are fed into the model and trainable gates are applied to steer the pretrained model as well as alleviate the mutual interference of plugins.

4.2 Method

Prompting. We present our model in the order of forward propagation. To change how the trainable parameters are injected into the model, we first follow prompt-tuning (Lester et al., 2021) to append trainable prompts to the output of the embedding layer. Moreover, to make our model applicable not only to categorical aspects (e.g., sentiment) but also to free-form aspects (e.g., lexical constraint), we present the constraints of aspects in textual form and feed them to the model. When two aspects of constraints are required during inference, the model input is given by

$$\mathbf{H}^{(0)} = \left[\mathrm{E}(\mathbf{x}); \mathrm{E}(\mathbf{c}_1); \mathrm{E}(\mathbf{c}_2); \mathbf{P}_1^{(0)}; \mathbf{P}_2^{(0)}\right], \quad (6)$$

where $\mathrm{E}(\cdot)$ is the embedding function, and $\mathbf{x}$ is the source sentence for sequence-to-sequence generation like machine translation and can be eliminated

[3] Please refer to Appendix A for more detail.

[4] For brevity, we show the lower bound in one head of attention (obtained by simplifying Eq. (13)), and a similar conclusion can be obtained on the multi-head attention (Eq. (14)).

for text generation. $\mathbf{c}_1$ and $\mathbf{c}_2$ are textual form of constraints (e.g., "This is a positive review" for positive review generation, and "New York" for lexical constraint). $\mathbf{P}_1^{(0)}, \mathbf{P}_2^{(0)} \in \mathbb{R}^{p\times d}$ are continuous prompts, where the right superscript $^{(j)}$ represents the j-th layer, p is the number of continuous vectors, and d is the dimension of hidden states. To avoid the discrepancy between training and inference in position, each textual sequence has its own position indexes starting from 1 and its own segment embedding (Devlin et al., 2019). Note that only one textual constraint and one set of trainable parameters are injected during training.

Gating. Then, the model input $\mathbf{H}^{(0)}$ is fed to the encoder, where trainable gates are attached to the hidden states in a position-wise manner, which alleviates mutual interference as well as steers the model. Formally, $\mathbf{A}^{(j)} = \text{Self-Att}(\mathbf{H}^{(j-1)})$ is the output of the j-th attention sublayer, and it is normalized by the gates:

$$\tilde{\mathbf{A}}^{(j)} = \Big[\mathbf{A}_X^{(j)}; \sigma\big(\mathbf{G}_1^{(j)}\big) \odot \big(\mathbf{A}_{P_1}^{(j)} + \mathbf{P}_1^{(j)}\big); \sigma\big(\mathbf{G}_2^{(j)}\big) \odot \big(\mathbf{A}_{P_2}^{(j)} + \mathbf{P}_2^{(j)}\big)\Big], \quad (7)$$

where $\mathbf{A}_X^{(j)} \in \mathbb{R}^{(|\mathbf{x}|+|\mathbf{c}_1|+|\mathbf{c}_2|)\times d}$ and $\mathbf{A}_{P_i}^{(j)} \in \mathbb{R}^{p\times d}$ are hidden states split from $\mathbf{A}^{(j)}$, $\mathbf{P}_i^{(j)} \in \mathbb{R}^{p\times d}$ are trainable vectors that add to the hidden states, σ is the $\text{sigmoid}(\cdot)$ function and $\mathbf{G}_i^{(j)} \in \mathbb{R}^{p\times d}$ are trainable vectors. $\odot$ denotes the Hadamard product and the normalized vectors $\sigma(\mathbf{G}_i^{(j)})$ serve as **gates** to selectively rescale the output of the attention sublayer in a position-wise manner and $\tilde{\mathbf{A}}^{(j)} \in \mathbb{R}^{(|\mathbf{x}|+|\mathbf{c}_1|+|\mathbf{c}_2|+2p)\times d}$ is the result of the normalization. Next, the normalized output is fed to the feed-forward sublayer: $\mathbf{H}^{(j)} = \text{FFN}(\tilde{\mathbf{A}}^{(j)})$. Finally, the output of the last encoder layer is fed to a standard Transformer decoder to guide the generation.

Training & Inference. As shown in Figure 1, during training, each plugin (including prompts and gates) for a single aspect of constraints is attached to the pretrained generative model and optimized by corresponding single-aspect labeled data separately (refer to Eq. (2)). In contrast, during inference, the control of arbitrary combinations of aspects can be realized by simply concatenating the corresponding plugins (refer to Eq. (3)).

Moreover, our approach treats the training and inference processes for pre-existing and newly-introduced constraints identically. The total training cost of N pre-existing aspects and M newly-added aspects is $O((N+M)C)$, where C denotes the cost of training on a single aspect. In this way, the cost of introducing new constraints is relatively low.

5 Experiments

We conducted experiments on two representative tasks in natural language generation, which are text generation and machine translation.

5.1 Multi-Aspect Controllable Text Generation

Dataset. Following previous work (Yang et al., 2022), we adopted the widely-used Yelp dataset (Lample et al., 2019), which contains restaurant reviews with the sentiment (positive and negative) and the topic (American, Mexican, and Asian) labels. To evaluate the extensibility of methods, we added two additional aspects of constraints: keywords (He, 2021) and tense (past and present) (Ficler and Goldberg, 2017), where their labels are automatically extracted from the reviews. Due to the page limit, please refer to Appendix B for more details about the experimental setup.

Evaluation. Following previous work, we adopted automatic and human evaluations for constraint accuracy and text quality (Lyu et al., 2021; Dathathri et al., 2019; Gu et al., 2022). Specifically, we finetuned two RoBERTa-based (Liu et al., 2019) classifiers for the evaluations of sentiment and topic. The tense accuracy was evaluated by the same tool adopted in the training set, and we used word-level Copy Success Rate (CSR) (Chen et al., 2020) to evaluate the lexical constraint. In addition, we used the perplexity (PPL) given by GPT-2_{medium} (Radford et al., 2019) and averaged distinctness (Li et al., 2016) to evaluate the fluency and diversity of the generated text, respectively. For human evaluation, each sentence received a score of 1 to 5 on sentiment and topic relevance as well as fluency given by three evaluators. The final scores are averaged over three ratings.

Baselines. We compared our approach with several representative methods for multi-aspect controllable text generation:

- GEDI (Krause et al., 2021): a decoding-time regulation method that uses light-weight con-

Category	Method	Dist.↑	Sent.↑	Topic↑	Average↑	PPL↓
DTR	GEDI	0.75 (0.00)	**99.47 (+0.13)**	51.36 (-45.98)	75.41 (-22.92)	616.92 (+253.23)
PET w/ JT	DIST. LENS	0.26 (-0.10)	77.47 (-17.17)	66.98 (-14.95)	72.22 (-16.06)	52.59 (+19.73)
PET w/o JT	PROMPT-TUNING	0.42 (-0.06)	48.29 (-5.84)	48.11 (-8.82)	48.20 **(-7.33)**	40.89 **(-6.83)**
	PREFIX-TUNING	0.31 (-0.10)	47.53 (-37.27)	69.11 (-9.38)	58.32 (-23.32)	147.47 (+125.17)
	TAILOR	0.39 (-0.04)	80.68 (-8.12)	68.72 (-9.94)	74.70 (-9.03)	40.29 (+8.52)
	PROMPT GATING (*Ours*)	0.42 (0.00)	84.80 (-10.93)	**75.02 (-8.00)**	**79.91** (-9.47)	**21.77** (+0.14)

Table 1: Automatic evaluation on double-aspect controllable text generation. "*DTR*" and "*PET*" denote decoding-time regulation and parameter efficient tuning methods, respectively. "*w/ JT*" and "*w/o JT*" denote methods with and without joint training, respectively. There are two aspects: "Sent." (sentiment) and "Topic". "Average" denotes the averaged scores over sentiment and topic accuracies. "PPL" and "Dist." denote perplexity and averaged distinctness, respectively. The scores in brackets indicate the performance gap between double- and single-aspect settings.

# Aspects	Method	PPL ↓	Sent. ↑	Topic ↑	Tense ↑	Lex. ↑	Ave.↑	ΔTime
3	PREFIX-TUNING	154.69	44.91	54.38	24.49	/	41.26	+6.42 h
	DIST. LENS	63.13	65.31	55.84	54.25	/	58.47	+30.13 h
	PROMPT GATING (*Ours*)	**21.87**	**76.93**	**62.73**	**62.24**	/	**67.30**	**+6.30 h**
4	PREFIX-TUNING	159.80 (+8.35)	37.33 (-7.58)	32.51 (-21.87)	18.82 (-5.68)	29.55	15.47	+2.77 h
	PROMPT GATING (*Ours*)	**20.90 (-0.96)**	**75.32 (-1.61)**	**62.52 (-0.21)**	**60.05 (-2.20)**	**54.50**	**63.10**	**+2.01 h**

Table 2: Automatic evaluation on triple- and quadruple-aspect controllable text generation where the models are extended from double-aspect setting. "# Aspects" denotes the number of aspects (N). "Ave." denotes the averaged accuracies over N aspects. "Sent.", "Topic", and "Tense" denote accuracies for sentimental, topical, and temporal constraints, respectively. "Lex." denotes CSR for lexical constraint. "ΔTime" denotes the training time extending from $(N-1)$-aspect setting to N-aspect setting. The scores in brackets indicate the performance gap between quadruple- and triple-aspect settings. Note that the methods specialized for attribute-based controlling like DIST. LENS can not process free-form constraints like lexical constraint.

Method	Sent. ↑	Topic ↑	Fluency ↑
GEDI	1.67	2.72	3.12
DIST. LENS	3.71	3.20	3.72
PROMPT GATING (*Ours*)	**4.44**	**4.23**	**4.19**

Table 3: Human evaluation on double-aspect controllable text generation. Sentences are rated 1 to 5 each for sentimental and topical relevance and fluency.

ditional generative discriminator to guide pre-trained models. The distributions given by multiple discriminators are normalized for controlling multiple aspects of target sentences.

- DIST. LENS (Gu et al., 2022): a prefix-tuning-based method that introduces an autoencoder and additional objectives to map several constraints of attributes to one latent space (i.e., joint training of prefixes). It finds the intersection of prefixes of multiple constraints during inference.

- PROMPT-TUNING (Lester et al., 2021): a parameter efficient method that appends continuous prompts to the model input. Multiple prompts are trained separately and are simply concatenated during inference.

- PREFIX-TUNING (Li and Liang, 2021): a parameter efficient method that appends continuous prefixes to the activations at attention sublayers. Multiple prefixes are trained separately and are simply concatenated during inference.

- TAILOR (Yang et al., 2022): a prompt-tuning-based method that further modifies the attention mask and position indexes during inference to narrow the gap between training and inference.

Results. Table 1 shows the automatic evaluation on double-aspect controllable text generation. We demonstrate the averaged accuracies to represent the overall performance on satisfying multiple aspects of constraints. Furthermore, we provide the performance gap between double- and single-

Method	Lex. ↑	Tense↑	Average↑	BLEU↑
PREFIX-TUNING	7.51 (-77.77)	43.46 (-39.83)	25.48 (-58.80)	0.4 (-36.3)
PARALLEL ADAPTER	48.44 (-43.38)	67.87 (-15.68)	58.15 (-29.53)	21.8 (-15.5)
LORA	50.79 (-37.15)	74.16 (-10.16)	62.47 (-23.65)	25.0 (-11.2)
PROMPT-TUNING	64.64 (-10.29)	81.12 (-0.07)	72.88 (-5.18)	34.2 (-1.2)
PROMPT GATING (*Ours*)	**85.29** (-4.61)	**85.75** (+1.76)	**85.52** (-1.42)	**36.8** (-0.3)

Table 4: Results on controllable machine translation. The experiments are conducted on the WMT14 German→English benchmark. There are three aspects of constraints: lexical constraint, tense, and external knowledge (French synonymous sentences). "Lex." and "Tense" denote CSR and accuracies for lexical and temporal constraint, respectively. "Average" denotes the averaged accuracies over them. The scores in brackets indicate the performance gap between double- and single-aspect settings.

aspect settings to represent the ability to combine multiple plugins in a zero-shot manner. Although GEDI achieves the highest scores on sentiment accuracy and distinctness, its perplexity explodes, and its tense accuracy is significantly decreased, which can be attributed to the interpolation of multiple discriminators. As PROMPT-TUNING does not have sufficient trainable parameters, it performs poorly on constraint accuracies. However, it has a relatively minor performance gap due to only inserting vectors once. PREFIX-TUNING suffers from severe mutual interference because of the insertions in all Transformer layers, leading to poor performance on either constraint accuracies or perplexity. Compared with PREFIX-TUNING, DIST. LENS has better constraint accuracies and lower performance gaps because of the joint training of prefixes. We found that DIST. LENS is sensitive to constraint distributions in the training set because it attempts to find the intersection of multiple constraints. Our approach (PROMPT GATING) achieves the highest constraint accuracies, lowest perplexity and a relatively small performance gap while our plugins are trained separately.

Table 2 shows the extensibility of the methods. When extended from double-aspect to triple-aspect, DIST. LENS has to be retrained because of its joint training strategy. In contrast, our approach and PREFIX-TUNING only need to train one plugin, then combine plugins and plug them into the pretrained model. Unfortunately, when extended from triple-aspect to quadruple-aspect, as plugins of PREFIX-TUNING badly interfere with each other, its ability to control multiple aspects significantly degenerates. However, our approach has a slight performance gap with a relatively small training cost, revealing its high extensibility.

The human evaluation results are illustrated in Table 3 with an inter-annotator agreement of 0.31 (Fleiss' κ). Experiments indicate that our approach significantly outperforms both baselines with $p < 0.01$ on all three aspects, determined by paired bootstrap and t-test using a popular open-source tool (Dror et al., 2018)[5]. Unlike automatic evaluations, GEDI performs the worst in sentiment relevance. It can probably be attributed to the fact that GEDI often generates ambivalent-sentiment and non-fluent sentences, and human annotators tend to give low ratings to them. The other results are in line with automatically evaluated results.

5.2 Multi-Aspect Controllable Machine Translation

Dataset. To thoroughly compare our approach with baselines, we also adopted a sequence-to-sequence generation task (i.e., machine translation (Bahdanau et al., 2015)). Experiments are conducted on the WMT14 German → English benchmark. We adopted three aspects of constraints in machine translation, and the labels are all automatically obtained from target sentences. We use keywords (Post and Vilar, 2018) and tense (Ficler and Goldberg, 2017) like in the text generation task to control translations. Specifically, we adopt French sentences with the same meaning as the German sources, which can be seen as an external knowledge to improve translation quality (Zoph and Knight, 2016), as the third constraint.

Evaluation. We adopted SACREBLEU[6] (Post, 2018) to calculate BLEU scores (Papineni et al., 2002) to evaluate the translation quality. Similar to text generation (§5.1), we used CSR (Chen et al., 2020) and tense accuracy to evaluate lexical and

[5] https://github.com/rtmdrr/testSignificanceNLP

[6] The signature is "BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.2.0.0".

Method	Sent. ↑	Topic ↑	PPL ↓
PROMPT GATING (*Ours*)	**84.80**	**75.02**	**21.77**
- Textual context for attribute	83.60	71.89	22.00
- Normalization of gates	76.53	61.02	27.55
Move the gates behind FFN	56.71	32.49	36.74

Table 5: Ablation study and comparison with the variant of our approach. "- Textual context for attribute" denotes ablating textual contexts of attribute-based constraints (see Eq. (6)). "- Normalization of gates" denotes ablating the $\mathrm{sigmoid}(\cdot)$ function which normalizes the gates (see Eq. (7)). "Move the gates behind FFN" denotes changes where trainable gates apply.

temporal constraints, respectively.

Baselines. Besides PROMPT-TUNING (Lester et al., 2021) and PREFIX-TUNING (Li and Liang, 2021) (§5.1), we adopted another two representative parameter efficient methods as baselines:

- LORA (Hu et al., 2021): a method that adds trainable rank decomposition matrices into attention sublayers.
- PARALLEL ADAPTER (Houlsby et al., 2019): a method that parallelly inserts feed-forward sublayers between pre-trained sublayers.

Similar to PREFIX-TUNING, for both LORA and PARALLEL ADAPTER, each plugin is trained separately, and multiple plugins are simply concatenated for multi-aspect setting during inference.

Results. Table 4 shows the results on controllable machine translation. Unlike text generation, constraints in machine translation do not merely contain attribute-based constraints. Therefore, methods specially designed for attribute-based constraints cannot be applied to this task. Surprisingly, PROMPT-TUNING achieves the highest constraint accuracies and translation quality among baselines because it largely retains the capabilities of plugins to satisfy constraints. PREFIX-TUNING faces the severe degeneration of both accuracies of constraints and BLEU scores, which might be attributed to the more complicated model structure in machine translation than text generation. Our approach outperforms all baselines in machine translation, and the consistent superiorities on both tasks show its generalizability.

5.3 Analysis

Mutual Interference. Similar to empirical analysis on mutual interference for PREFIX-TUNING and PROMPT-TUNING (see §3), we also plotted the variation of the mutual interference with the number of injections of our approach in Figure 2. With the gates to selectively rescale the interventions of plugins, the growth of interference is restrained.

Ablation Study. Table 5 shows the ablation study and comparison with the variant of our approach. According to the performance gaps corresponding to the changes, we can find that the textual context of constraints slightly affects the constraint accuracies, and the normalization of the trainable gates is a key point for good performance. Moreover, the trainable gates should be placed where the interactions have just happened (i.e., after attention sublayers). Please refer to Appendix C and D for more results, analyses, and cases.

6 Related Work

Multi-aspect controllable text generation (MCTG) (Qian et al., 2022; Yang et al., 2022; Gu et al., 2022) that simultaneously satisfies multiple constraints is a challenging task for which highly extensible methods make more practical sense. Approaches to it can be roughly divided into the following three categories.

Dedicated Model. The dedicated conditional generative models (Keskar et al., 2019; Dou et al., 2021; Huang et al., 2021; Chen et al., 2020) can accept multiple constraints by training from scratch or full-parameter finetuning on the multi-aspect labeled data. However, the multi-aspect labeled data is hard to obtain, and the constraints that can be satisfied are already determined during training. Thus it is usually too expensive to apply dedicated models to MCTG.

Decoding-Time Regulation. Although multi-aspect controlling can be achieved by interpolating distributions of multiple discriminators (Dathathri et al., 2019; Chen et al., 2021; Krause et al., 2021; Lin and Riedl, 2021) or optimizing towards multiple objectives (Qin et al., 2022; Kumar et al., 2021), they usually significantly impair text fluency because of the intervention in the decoding stage (Gu et al., 2022).

Parameter Efficient Tuning. Unlike the above two branches, PET introduces plugins trained with fixed pretrained models for generating required text (Li and Liang, 2021; Lester et al., 2021; Wang et al., 2022). Because of its potential to achieve

high extensibility in a plug-and-plug manner, our work also falls in this line. However, when multiple constraints are required, joint training of plugins is introduced to alleviate the mutual interference of plugins (Chan et al., 2021; Qian et al., 2022; Yang et al., 2022; Gu et al., 2022), which hurts extensibility. Differently, our work aims at reducing mutual interference while maintaining separate training. Similar to our work, Yang et al. (2022) proposes preventing two prompts from interactions in attention layers by modifying attention masks. Nevertheless, their method like prompt-tuning (Lester et al., 2021) only introduces trainable parameters to the model input, leading to insufficient trainable parameters and dissatisfied constraints. In contrast, we propose a novel PET method that attaches trainable gates to the pretrained model, alleviating mutual interference while providing sufficient interventions, leading to both desired extensibility and effectiveness.

7 Conclusion

In summary, we propose an extensible plug-and-play method for multi-aspect controllable text generation. By replacing trainable prefixes with trainable prompts and gates, our approach alleviates the mutual interference of multiple plugins while providing sufficient interventions. Experiments on text generation and machine translation show its superiorities over baselines on the cost of extending to new combinations of aspects, the fulfillment of constraints, and text fluency.

Limitations

First, although our approach and existing methods for controllable text generation can improve the constraint accuracies, they are currently unable to achieve 100% accuracies in the vast majority of aspects (e.g., sentiment or topic). This makes them not yet applicable in scenarios with requirements of 100% control fulfillment. Second, there is still a gap between the automatic and human evaluation of text generation, which makes there a trade-off between precision and efficiency in the evaluation of controllable text generation. Third, although our approach reduces the mutual interference of plugins so that multiple plugins can be combined at a relatively small cost (a decrease in constraint accuracy), this cost will not be zero, which puts an upper limit on the number of plugins can be applied simultaneously. Fortunately, for controllable text generation, the number of controls applied simultaneously is generally not too large (e.g., four or five aspects).

Ethics Statement

Since the text generation model is trained on data collected from the web and often not thoroughly cleaned, it can generate offensive or toxic text. We must state that the texts generated by our approach do not represent our opinion. To alleviate these issues, we can take detoxification and politeness as the default aspects of constraints in our multi-aspect controllable method.

Acknowledgments

This work is supported by the National Key R&D Program of China (2022ZD0160502), the National Natural Science Foundation of China (No. 61925601, 62276152), the National Social Science Fund of China (20&ZD279), and a grant from the Guoqiang Institute, Tsinghua University. We thank Kaiyu Huang, Fuchao Wei, Yuanhang Zheng and all the anonymous reviewers for their valuable comments and suggestions on this work, as well as all the volunteers who participated in the human evaluation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. Cocon: A self-supervised approach for controlled text generation. In *Proceedings of ICLR 2021*.

Guanhua Chen, Yun Chen, and Victor O. K. Li. 2021. Lexically constrained neural machine translation with explicit alignment guidance. In *Proceedings of AAAI 2021*.

Guanhua Chen, Yun Chen, Yong Wang, and Victor O. K. Li. 2020. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of IJCAI 2020*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of NAACL-HLT 2021*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL 2018*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of Style-Var 2017*.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. *arXiv preprint arXiv:2210.02889*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with BART. In *Proceedings of EMNLP 2021*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of ICML 2019*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xuancheng Huang, Jingfang Xu, Maosong Sun, and Yang Liu. 2021. Transfer learning for sequence generation: from single-source to multi-source. In *Proceedings of ACL-IJCNLP 2021*.

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 2021*.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. In *Proceedings of NeurIPS 2021*.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *Proceedings of ICLR 2019*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP 2021*.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL-IJCNLP 2021*.

Zhiyu Lin and Mark Riedl. 2021. Plug-and-blend: A framework for controllable story generation with blended control codes. In *Proceedings of NUSE 2021*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *TACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. In *Proceedings of NAACL-HLT 2021*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of SIGNLL 2016*.

Dat Quoc Nguyen and Karin Verspoor. 2018. An improved neural network model for joint POS tagging and dependency parsing. In *Proceedings of CoNLL 2018*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT 2018*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of NAACL 2018*.

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

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, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Zhixing Tan, Jiacheng Zhang, Xuancheng Huang, Gang Chen, Shuo Wang, Maosong Sun, Huanbo Luan, and Yang Liu. 2020. THUMT: An open-source toolkit for neural machine translation. In *Proceedings of AMTA 2020*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*.

Shuo Wang, Zhixing Tan, and Yang Liu. 2022. Integrating vectorized lexical constraints for neural machine translation. In *Proceedings of ACL 2022*.

Kevin Yang and Dan Klein. 2021. FUDGE: controlled text generation with future discriminators. In *Proceedings of NAACL-HLT 2021*.

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*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of NAACL-HLT 2016*.

A Theoretical Analysis

In this section, we theoretically analyze mutual interference (MI) and derive a lower bound of MI for prefix-tuning (Li and Liang, 2021). As Feed Forward and Layernorm sublayers are position-wise operations (Vaswani et al., 2017) which would not introduce the interference of plugins, we focus on analyzing the multi-head attention (MHA) sublayers.

According to the previous study (He et al., 2021), the output of a single head of attention with prefixes of the i-th plugin, which is represented by $\mathbf{h}_i$, could be described as

$$\begin{aligned}\mathbf{h}_i &= \lambda(\mathbf{x}_i)\overline{\mathbf{h}}_i + \big(1-\lambda(\mathbf{x}_i)\big)\Delta\mathbf{h}_i \\ &= s_i\overline{\mathbf{h}}_i + t_i\Delta\mathbf{h}_i,\end{aligned} \tag{8}$$

where $\overline{\mathbf{h}}_i$ denotes the original output of the pretrained generative model with $\mathbf{x}_i$ as input. $\lambda(\mathbf{x}_i)$ is a scalar related to the attention weights, where $\lambda(\mathbf{x}_i) = s_i = 1 - t_i \in (0,1)$. In addition, $\Delta\mathbf{h}_i$ is an offset determined by the i-th plugin, and its magnitude is positively correlated with the magnitude of $\boldsymbol{\phi}_i$, where $\boldsymbol{\phi}_i$ is the set of parameters of the i-th plugin.

Following the pattern above, when the i-th and j-th plugins are inserted at the same time, the output of the head (i.e., $\mathbf{h}_{i,j}$) turns to be

$$\mathbf{h}_{i,j} = \gamma\overline{\mathbf{h}}_{i,j} + \alpha\Delta\mathbf{h}_i + \beta\Delta\mathbf{h}_j, \tag{9}$$

where $\overline{\mathbf{h}}_{i,j}$ is the output of pretrained generative model, and $\alpha,\beta,\gamma \in (0,1), \alpha < t_i, \beta < t_j, \gamma < s_i, \gamma < s_j$. Similarly, $\Delta\mathbf{h}_i$ and $\Delta\mathbf{h}_j$ are determined by the i-th and j-th plugins.

According to the definition in Eq. (4), let $\tilde{\mathbf{h}}_{i,j}$ and $\hat{\mathbf{h}}_{i,j}$ be the outputs like $\mathbf{h}_{i,j}$ after training on multi- and single-aspect labeled data, respectively. The mutual interference of two plugins in a single head (i.e., MI_s) can be measured by the norm of the gap between outputs under supervised and zero-shot inference:

$$\begin{aligned}\mathrm{MI}_s &= \big\|\tilde{\mathbf{h}}_{i,j} - \hat{\mathbf{h}}_{i,j}\big\| \\ &= \big\|\tilde{\mathbf{h}}_{i,j} - (\gamma\overline{\mathbf{h}}_{i,j} + \alpha\Delta\hat{\mathbf{h}}_i + \beta\Delta\hat{\mathbf{h}}_j)\big\| \\ &\geq \big\|\tilde{\mathbf{h}}_{i,j} - \gamma\overline{\mathbf{h}}_{i,j}\big\| - \big\|\alpha\Delta\hat{\mathbf{h}}_i + \beta\Delta\hat{\mathbf{h}}_j\big\|,\end{aligned} \tag{10}$$

where $\Delta\hat{\mathbf{h}}_i$ and $\Delta\hat{\mathbf{h}}_j$ correspond to offsets that plugins are trained on single-aspect labeled data.

Considering that the intervention caused by two plugins simultaneously should larger than the sum of two interventions caused by two plugins respectively because of the interaction between two plugins, we assume that there is

$$\big\|\tilde{\mathbf{h}}_{i,j} - \overline{\mathbf{h}}_{i,j}\big\| > \big\|\hat{\mathbf{h}}_i - \overline{\mathbf{h}}_i\big\| + \big\|\hat{\mathbf{h}}_j - \overline{\mathbf{h}}_j\big\|. \tag{11}$$

Based on this, we can derive

$$\begin{aligned}\mathrm{MI}_s > &\big\|\hat{\mathbf{h}}_i - \gamma\overline{\mathbf{h}}_i\big\| + \big\|\hat{\mathbf{h}}_j - \gamma\overline{\mathbf{h}}_j\big\| \\ &- \big\|\alpha\Delta\hat{\mathbf{h}}_i + \beta\Delta\hat{\mathbf{h}}_j\big\|.\end{aligned} \tag{12}$$

Given that $s_i > \gamma, s_j > \gamma$, and $\hat{\mathbf{h}}_i = s_i\overline{\mathbf{h}}_i + t_i\Delta\hat{\mathbf{h}}_i$ (Eq. (8)), MI_s satisfies

$$\begin{aligned}\mathrm{MI}_s &> \big\|\hat{\mathbf{h}}_i - s_i\overline{\mathbf{h}}_i\big\| + \big\|\hat{\mathbf{h}}_j - s_j\overline{\mathbf{h}}_j\big\| \\ &\quad - \big\|\alpha\Delta\hat{\mathbf{h}}_i + \beta\Delta\hat{\mathbf{h}}_j\big\| \\ &= \|t_i\Delta\hat{\mathbf{h}}_i\| + \|t_j\Delta\hat{\mathbf{h}}_j\| - \|\alpha\Delta\hat{\mathbf{h}}_i + \beta\Delta\hat{\mathbf{h}}_j\| \\ &\geq (t_i-\alpha)\|\Delta\hat{\mathbf{h}}_i\| + (t_j-\beta)\|\Delta\hat{\mathbf{h}}_j\|,\end{aligned} \tag{13}$$

where $1 > t_i - \alpha > 0$ and $1 > t_j - \beta > 0$. Therefore, the mutual interference of two plugins in a single head has a positive lower bound, and it is positively correlated with the magnitude of $\hat{\boldsymbol{\phi}}_i$ and $\hat{\boldsymbol{\phi}}_j$.

To step further, we derive the lower bound of MI in the multi-head scenario. Assume that K denotes the number of heads, $\mathbf{W}_o$ denotes the fixed output projection matrix in the MHA, $\mathbf{W}_o = \mathbf{Q}_o\mathbf{R}_o$ is the QR-decomposition format of $\mathbf{W}_o$, $\hat{\lambda_o}$ is the average of absolute eigenvalues. Specifically, $\hat{\mathbf{h}}^k_{i,j}$ and $\tilde{\mathbf{h}}^k_{i,j}$ denotes $\hat{\mathbf{h}}_{i,j}$ and $\tilde{\mathbf{h}}_{i,j}$ in the k-th head, respectively. Then, the lower bound of MI in MHA (i.e., MI_m) can be derived as (viewing $\mathbf{R}_o$ as a diagonal matrix for simplicity)

$$\begin{aligned}\mathrm{MI}_m &= \big\|\mathrm{concat}(\tilde{\mathbf{h}}^k_{i,j} - \hat{\mathbf{h}}^k_{i,j})_{k=1}^K\mathbf{W}_o\big\| \\ &= \big\|\mathrm{concat}(\tilde{\mathbf{h}}^k_{i,j} - \hat{\mathbf{h}}^k_{i,j})_{k=1}^K\mathbf{Q}_o\mathbf{R}_o\big\| \\ &\approx \hat{\lambda_o}\big\|\mathrm{concat}(\tilde{\mathbf{h}}^k_{i,j} - \hat{\mathbf{h}}^k_{i,j})_{k=1}^K\mathbf{Q}_o\big\| \\ &= \hat{\lambda_o}\sqrt{\sum_{k=1}^K\big\|\tilde{\mathbf{h}}^k_{i,j} - \hat{\mathbf{h}}^k_{i,j}\big\|^2} \\ &\geq \frac{\hat{\lambda_o}}{\sqrt{n}}\sum_{k=1}^K\big\|\tilde{\mathbf{h}}^k_{i,j} - \hat{\mathbf{h}}^k_{i,j}\big\| \\ &> \frac{\hat{\lambda_o}}{\sqrt{n}}\sum_{k=1}^K\Big((t^k_i-\alpha^k)\|\Delta\hat{\mathbf{h}}^k_i\| \\ &\quad + (t^k_j-\beta^k)\|\Delta\hat{\mathbf{h}}^k_j\|\Big),\end{aligned} \tag{14}$$

where $1 > t^k_i - \alpha^k > 0$ and $1 > t^k_j - \beta^k > 0$, and $\Delta\hat{\mathbf{h}}^k_i$ and $\Delta\hat{\mathbf{h}}^k_j$ are also positively correlated with the magnitude of $\hat{\boldsymbol{\phi}}_i$ and $\hat{\boldsymbol{\phi}}_j$, respectively.

Therefore, the mutual interference of multiple plugins has a theoretical positive lower bound, which implies concatenating prefixes that are separately trained has an irreparable gap against supervised-trained prefixes. As a result, MI might accumulate along with the depth of the model, like in Figure 2. Intuitively, introducing gates, which contain trainable coefficients between 0 to 1, to $\hat{\phi}_i$ is helpful for decreasing the offsets in Eq. (14) and thus mutual interference.

B Reproducibility

B.1 Data Preparation

For text generation, we adopted the widely-used Yelp dataset[7] (Lample et al., 2019), which contains restaurant reviews with sentiment (positive and negative) and topic (American, Mexican, and Asian) labels. Specifically, following previous work (Yang et al., 2022), we randomly sampled 30K/3K sentences for each attribute for training/validation while ensuring the balance of different attributes in the final dataset (Table 6). For evaluation, we sampled 25 sentences for each given textual prefix and combination of aspects. In addition, we eliminated the sentences rated 3 in sentiment. To evaluate the extensibility of methods, we added two additional aspects of constraints: keywords (He, 2021) and tense (past and present) (Ficler and Goldberg, 2017), where their labels are automatically extracted from the reviews. More precisely, we randomly extracted 1 to 3 words as keywords for each sentence (Post and Vilar, 2018), and the tenses of sentences are labeled by an open-source toolkit[8] that is based on a POS tagger (Nguyen and Verspoor, 2018).

For machine translation, we adopted the WMT14 German → English benchmark[9]. Specifically, the training, validation, and test sets contain 4,500K, 3K, and 3K sentences, respectively. We adopted three aspects of constraints in machine translation, and they are all automatically obtained from target sentences. We use keywords (Post and Vilar, 2018) and tense (Ficler and Goldberg, 2017) like the text generation task to control translations. For the third constraint, we adopt French synonymous sentences as external knowledge, which is beneficial to disambiguation. Note that it does not directly control any attribute of translations but will improve the translation quality (Zoph and Knight, 2016). The French synonymous sentences are given by a Transformer-based English→French translation model (Vaswani et al., 2017).

Task	Training	Validation	Test
Text Generation	30K	3K	375*
Machine Translation	4,500K	3K	3K

Table 6: The number of examples (sentences) in Training/Validation/Test for each attribute in Text Generation and aspect in Machine Translation. *: For Test in Text Generation, to keep in line with previous works (Yang et al., 2022), we use 15 attribute-unrelated prefixes (as listed in §B.2) and ask model to continue writing under each attribute (25 sentences for each).

B.2 Evaluation Metrics

For text generation, following previous work (Lyu et al., 2021; Dathathri et al., 2019; Gu et al., 2022), we adopted automatic and human evaluation for constraint accuracy and text quality. For the evaluation of sentiment and topic, we finetuned two RoBERTa-based (Liu et al., 2019) classifiers on the Yelp dataset. Specifically, we randomly over-sampled 1,500K/15K/15K sentences for training/validation/test set of topic and 1,380K/1K/1K sentences for training/validation/test set of sentiment. The F1 scores for sentiment and topic are 98.71 and 89.62, respectively. The same toolkit as training evaluated the accuracy of tense, and we used word-level Copy Success Rate (CSR) (Chen et al., 2020) to evaluate the lexical constraint. In addition, we used the perplexity (PPL) given by GPT-2_{medium} (Radford et al., 2019) and averaged distinctness (Li et al., 2016) to evaluate the fluency and diversity of the generated text, respectively. Similar to previous work (Dathathri et al., 2019; Yang and Klein, 2021), we used 15 textual prefixes[10] and asked models to start writing from them for each combination of constraints during inference. For human evaluation, each sentence received a score of 1 to 5 on sentiment and topic relevance as well as fluency given by three evaluators. The final scores are averaged over three ratings.

Specifically, the 15 textual prefixes are: "Once upon a time","The book","The chicken","The

[7] https://github.com/shrimai/Style-Transfer-Through-Back-Translation

[8] https://github.com/ajitrajasekharan/simple_tense_detector

[9] https://statmt.org/wmt14/translation-task.html

[10] https://github.com/uber-research/PPLM

Hyper-parameter	TG	MT
Pretrained Model		
Encoder layers	12	12
Decoder layers	12	12
Attention heads	16	16
Attention head size	64	64
Hidden size	1,024	1,024
FFN hidden size	4,096	4,096
Max sentence length	1,024	1,024
Training		
Optimizer	Adam	Adam
Adam beta	(0.9, 0.999)	(0.9, 0.999)
Training steps	50,000	150,000
Warmup steps	10,000	10,000
Batch size	1,024	512
Learning rate	1×10^{-4}	4×10^{-4}
Initial learning rate	5×10^{-8}	5×10^{-8}
Residual dropout	0.1	0.1
Attention dropout	0.0	0.0
Activation dropout	0.0	0.0
Inference		
Length penalty	0.6	1.2
Top K	10	/
Beam size	/	5

Table 7: The commonly-used hyper-parameters in text generation (TG) and machine translation (MT).

Method	# Trainable Parameters
GEDI	$345M$
DIST. LENS	$110M + 768^2 + 768 \times 2 \times 20 \times 1024 \times 24 = 866M$
PARALLEL ADAPTER	$2 \times 19 \times 1024 \times (36 + 24) \approx 2.33M$
LORA	$4 \times 17 \times 1024 \times 36 \approx 2.51M$
PROMPT-TUNING	$100 \times 1024 \approx 0.10M$
PREFIX-TUNING	$2 \times 33 \times 1024 \times 36 \approx 2.43M$
TAILOR	$100 \times 1024 \approx 0.10M$
PROMPT GATING	$1024 \times 100 \times 25 \approx 2.56M$

Table 8: The number of trainable parameters of a single aspect for each method.

city","The country","The lake","The movie","The painting","The weather","The food","While this is happening","The pizza","The potato","The president of the country","The year is 1910.".

For machine translation, we adopted SACREBLEU[11] (Post, 2018) to evaluate the translation quality. Similar to text generation, we used CSR (Chen et al., 2020) and tense accuracy to evaluate lexical and tense constraints, respectively.

B.3 Model and Hyper-parameters

As our approach has both encoder and decoder, we adopted BART-large[12] (Lewis et al., 2020) for text

[11]The signature is "BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.2.0.0".

[12]https://huggingface.co/facebook/bart-large

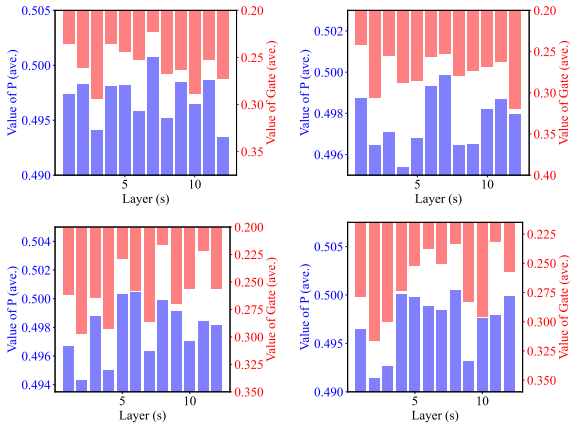

Figure 4: The visualization on text generation. The average value of gates $\sigma(\mathbf{G}_i^{(j)})$ (red bars) and the average of L_1 norm of $\mathbf{P}_i^{(j)}$ (blue bars) on each layer, according to Eq. (7). The values are extracted for the sentiment aspect, including negative (top left) and positive (top right), and topic aspect, including Asian (bottom left) and American (bottom right).

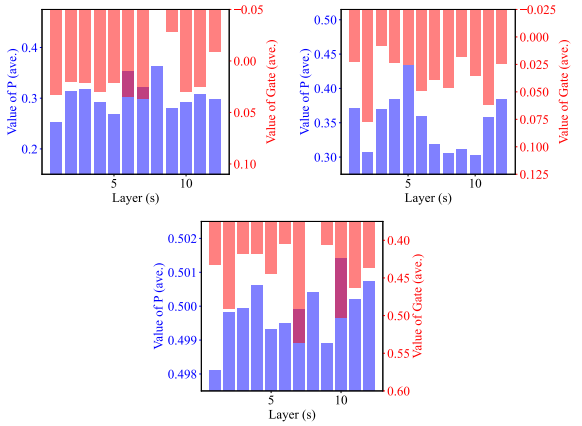

Figure 5: The visualization on machine translation. The average value of gates $\sigma(\mathbf{G}_i^{(j)})$ (red bars) and the average of L_1 norm of $\mathbf{P}_i^{(j)}$ (blue bars) on each layer, according to Eq. (7). The values are extracted for the lexical aspect (top left), temporal aspect (top right), and French synonymous sentences (bottom).

generation and mBART-large-cc25[13] (Liu et al., 2020) for machine translation.

For GEDI (Krause et al., 2021), DIST. LENS (Gu et al., 2022), and TAILOR (Yang et al., 2022), we follow the settings in their paper. Specifically, we found that the weights for attribute balance and the number of candidates in the decoding stage of DIST. LENS significantly affect constraint accuracies. For the weights for attribute balance and the number of candidates, we searched in {0.1, 0.2, 0.5, 1, 1.5, 2,

[13]https://huggingface.co/facebook/mbart-large-cc25

Method	Speed	
	Training (hours)	*Inference* (sec/sent.)
GEDI	5.4032	1.2020
DIST. LENS	30.1320	2.5705
PROMPT-TUNING	4.0983	0.2122
PREFIX-TUNING	3.9025	0.2220
TAILOR	4.1055	0.4640
PROMPT GATING	4.5204	0.2108

Table 9: The training and inference speeds of each method on multi-aspect controllable text generation. The training speeds are presented as the average training time on a single aspect, and the inference speeds are displayed in the form of time per sentence. Note that the inference procedure of "Dist. Lens" includes training a K-Center model (Gu et al., 2022).

Method	Speed	
	Training (hours)	*Inference* (sec/sent.)
PREFIX-TUNING	10.0150	0.2220
LORA	10.8805	0.1920
PARALLEL ADAPTER	12.0330	0.2150
PROMPT-TUNING	8.7225	0.2122
PROMPT GATING	11.0330	0.2108

Table 10: The training and inference speeds of each method on multi-aspect controllable machine translation. The training speeds are presented as the average training time on a single aspect, and the inference speeds are displayed in the form of time per sentence.

3, 4, 5, 8, 10, 25, 50} and {1, 2, 5, 10}, respectively. For the other methods, we demonstrate their hyperparameters in Table 7. Table 8 shows the number of trainable parameters of each method. The learning rates were determined by searching in $\{1 \times 10^{-5}$, 2×10^{-5}, 4×10^{-5}, 8×10^{-5}, 1×10^{-4}, 2×10^{-4}, 4×10^{-4}, $1 \times 10^{-3}\}$ on the development set. In our approach, the textual contexts used for attribute-based constraints (see §4.2) are:

- Sentiment: "This is a {} review." for "positive/negative".
- Topic: "The following is about {} food." for "Asian/American/Mexican".
- Tense: "The tense of this sentence is the {} tense." for "past/present/future", and "The tense of this sentence is undecided." for sentences that do not have an explicit tense.

Note that our use of existing artifact(s) was consistent with their intended use. The source code of this work is available at `https://github.com/THUNLP-MT/PromptGating4MCTG` and it was developed based on THUMT (Tan et al., 2020), an open-source toolkit for machine translation.

C Experimental Results

C.1 More Analyses

Visualization of PROMPT GATING. To further investigate how our approach alleviates the mutual interference of plugins, we visualized the trainable parameters in PROMPT GATING. Specifically, we first extracted $\mathbf{P}_i^{(j)}$ and $\sigma(\mathbf{G}_i^{(j)})$ in Eq. (7) from each layer j for every single aspect. Then we calculated the average of $\sigma(\mathbf{G}_i^{(j)})$ and the L_1 norm of $\mathbf{P}_i^{(j)}$ over all the layers, which are represented by red and blue bars respectively in Figure 4 and Figure 5.

We can find that when the magnitude of $\mathbf{P}_i^{(j)}$ (i.e., trainable vectors added to hidden states) becomes larger, the values of $\sigma(\mathbf{G}_i^{(j)})$ (i.e., trainable gates) tend to become smaller. In other words, these trainable gates attempt to normalize or stabilize the magnitude of hidden states and thus alleviate mutual interference.

Efficiency. Table 9 and Table 10 show the training and inference speeds of each method in text generation and machine translation. All training and inference were run on a single GeForce RTX 3090 GPU.

C.2 Detailed Results

Table 11 and 12 are the detailed versions of Table 1 and 4, respectively. We provide detailed results of both single- and multi-aspect models. For text generation, we further demonstrate the accuracy of each attribute.

D Case Study

To further investigate the fulfillment and text quality of each combination of constraints of these methods, Table 13 and Table 14 demonstrate examples of text generation and machine translation, respectively. Models only trained on single-aspect data are required to give results satisfying multiple aspects of constraints.

E Details in Human Evaluation

In this section, we show more details about the human annotation adopted for evaluating model performance on text generation. We recruited three volunteers from schools, shuffled the output of models and provided it to them for scoring. Since

they are volunteers, they were not paid. Their average age is 25 years old and they have enough daily English communication skills. The instruction we provided to them like "This human evaluation aims to evaluate the model-generated review texts in three aspects: sentiment and topic relevance, and text fluency. All three integer scores are on a scale of 1-5, with a higher degree of topic/sentiment relevance representing a more consistent theme/sentiment, and a higher degree of text fluency representing a more fluent text. Your personal information will not be retained and these scores will only be used for human evaluation in research".

Method	Constraint	Sentiment↑		Topic↑			PPL↓
		Positive	*Negative*	*Asian*	*American*	*Mexican*	
Decoding-Time Regulation Method							
GEDI	*single-aspect*	98.67	/	/	/	/	227.87
		/	100.00	/	/	/	839.69
		/	/	94.93	/	/	206.97
		/	/	/	99.73	/	246.36
		/	/	/	/	97.33	297.54
	multi-aspect	98.67	/	28.27	/	/	363.51
		99.20	/	/	87.73	/	1834.65
		99.47	/	/	/	37.87	378.73
		/	100.00	44.53	/	/	329.38
		/	99.73	/	97.87	/	423.21
		/	99.73	/	/	11.87	372.03
	single (avg.)	98.67	100.00	94.93	99.73	97.33	363.69
	multi (avg.)	99.11	99.82	36.40	92.80	24.87	616.92
Parameter Efficient Tuning Method with Joint Training							
DIST. LENS	*single-aspect*	91.33	/	/	/	/	28.48
		/	97.95	/	/	/	28.70
		/	/	77.33	/	/	39.01
		/	/	/	88.98	/	29.87
		/	/	/	/	79.47	38.26
	multi-aspect	36.27	/	43.73	/	/	45.89
		57.87	/	/	71.73	/	49.84
		74.67	/	/	/	54.67	47.59
		/	99.73	70.13	/	/	56.20
		/	97.60	/	78.93	/	59.24
		/	98.67	/	/	82.67	56.77
	single (avg.)	91.33	97.95	77.33	88.98	79.47	32.86
	multi (avg.)	56.27	98.67	56.93	75.33	68.67	52.59
Parameter Efficient Tuning Methods without Joint Training							
PROMPT-TUNING	*single-aspect*	52.00	/	/	/	/	136.10
		/	56.27	/	/	/	26.26
		/	/	46.67	/	/	25.83
		/	/	/	84.53	/	26.07
		/	/	/	/	39.60	24.36
	multi-aspect	57.07	/	40.80	/	/	65.96
		59.47	/	/	83.87	/	30.45
		45.87	/	/	/	20.13	47.43
		/	49.73	30.27	/	/	40.78
		/	38.67	/	84.00	/	33.43
		/	38.93	/	/	29.60	27.30
	single (avg.)	52.00	56.27	46.67	84.53	39.60	47.72
	multi (avg.)	54.13	42.44	35.53	83.93	24.87	40.89
PREFIX-TUNING	*single-aspect*	70.67	/	/	/	/	23.53
		/	98.93	/	/	/	21.55
		/	/	77.87	/	/	21.89
		/	/	/	84.00	/	21.99
		/	/	/	/	73.60	22.55
	multi-aspect	64.53	/	70.27	/	/	141.71
		67.20	/	/	80.00	/	243.93
		51.20	/	/	/	65.73	125.30
		/	33.87	60.67	/	/	118.65
		/	27.32	/	78.13	/	138.91
		/	41.07	/	/	59.87	116.34
	single (avg.)	70.67	98.93	77.87	84.00	73.60	22.30
	multi (avg.)	60.98	34.08	65.47	79.07	62.80	147.47

Method	Constraint	Sentiment↑		Topic↑			PPL↓
		Positive	*Negative*	*Asian*	*American*	*Mexican*	
TAILOR	*single-aspect*	81.87	/	/	/	/	32.80
		/	95.73	/	/	/	24.21
		/	/	72.00	/	/	34.38
		/	/	/	88.00	/	33.35
		/	/	/	/	76.00	34.10
	multi-aspect	72.73	/	67.47	/	/	43.08
		72.53	/	/	70.07	/	32.87
		68.27	/	/	/	69.33	43.12
		/	90.67	68.00	/	/	44.64
		/	90.40	/	70.93	/	33.54
		/	89.47	/	/	66.53	44.50
	single (avg.)	81.87	95.73	72.00	88.00	76.00	31.77
	multi (avg.)	71.18	90.18	67.73	70.50	67.93	40.29
PROMPT GATING (*Ours*)	*single-aspect*	91.73	/	/	/	/	21.65
		/	99.73	/	/	/	20.39
		/	/	77.87	/	/	21.29
		/	/	/	89.87	/	21.88
		/	/	/	/	81.33	22.93
	multi-aspect	73.07	/	62.13	/	/	21.28
		77.60	/	/	82.93	/	21.65
		75.20	/	/	/	72.00	22.47
		/	93.33	73.87	/	/	21.64
		/	95.73	/	81.33	/	20.09
		/	93.87	/	/	77.87	23.50
	single (avg.)	91.73	99.73	77.87	89.87	81.33	21.63
	multi (avg.)	75.29	94.31	68.00	82.13	74.93	21.77

Table 11: Detailed results of automatic evaluation on double-aspect controllable text generation. "*single (avg.)*" denotes the average score over scores in the single-aspect setting. "*multi (avg.)*" denotes the average score over scores in the multi-aspect setting.

Method	Constraint	Lex. ↑	Tense↑	BLEU↑
w/o control		50.98	79.52	32.7
PREFIX-TUNING	*lexical*	85.28	80.99	36.7
	temporal	50.80	83.28	33.0
	knowledgeable	50.20	79.29	32.8
	single (max.)	85.28	83.28	36.7
	multi (avg.)	7.51	43.46	0.4
PARALLEL ADAPTER	*lexical*	91.82	81.12	37.3
	temporal	50.93	83.55	33.1
	knowledgeable	50.16	79.35	32.8
	single (max.)	91.82	83.55	37.3
	multi (avg.)	48.44	67.87	21.8
LORA	*lexical*	87.94	81.59	36.2
	temporal	50.79	84.32	33.0
	knowledgeable	50.57	80.22	32.7
	single (max.)	87.94	84.32	36.2
	multi (avg.)	50.79	74.16	25.0
PROMPT-TUNING	*lexical*	74.93	80.99	35.4
	temporal	50.25	81.19	33.0
	knowledgeable	50.17	78.82	32.5
	single (max.)	74.93	81.19	35.4
	multi (avg.)	64.64	81.12	34.2
PROMPT GATING (*Ours*)	*lexical*	89.90	81.29	37.1
	temporal	50.95	83.98	33.2
	knowledgeable	51.18	79.55	33.1
	single (max.)	89.90	83.98	37.1
	multi (avg.)	85.29	85.75	36.8

Table 12: Detailed results on controllable machine translation. The experiments are conducted on the WMT14 German→English benchmark. There are three aspects of constraints: lexical constraint, tense, and external knowledge (French synonymous sentence). "*w/o control*": the pretrained German→English Transformer translation model without any plugin. "*single (max.)*" denotes the maximal score among scores in the single-aspect setting. "*multi (avg.)*" denotes the average score over scores in the multi-aspect setting.

WARNING: Next may contain contents that are offensive in nature.

Constraints	Method	Generation Result
Neg., Asian	GeDi	While this is happening to people of colour over and over again in schools and hospitals regardless of how kids should behave, let alone thats happening here? Is we as a society humane towards minks that poo on each other before death leaves them.
	Dist. Lens	While this is happening i was very impressed. food is average and the service is great. Advertisements: -) ***EDIT: The vegetarian menu is amazing!
	Prompt-tuning	While this is happening in my hometown. In New York, New York, and the UK , are in need of something, they say. It's a good year in my blood. In New Orleans, is it is good for the heart to get rid of it.In New York is getting all of its emotions out because we don't want to make the traditional church in-state.It was also all the land that is, and it is not good is all are for me is all a rock, and we don'
	Prefix-tuning	While this is happening food. is about Asian. isaboutAsian.is-foodsfood-trending.s.e...trend-stings-food-trends
	Tailor	While this is happening in my hometown, i will say that it's a little pricey but worth it for a good meal. the food was good, the service was good, the service was fast and the food was tasty. the food arrived on time, so i could eat it. i did get a couple of things wrong with this place. the rice is bland. i think the rice is too salty. my boyfriend ordered the fried squid. it was a bit greasy. the shrimp had a good flavor and it was fresh. i would give it a try again but it won't be here for at least 3 years.
	Prompt Gating (*Ours*)	While this is happening, i've never been here to the japanese section and have never eaten here before. the place is small and the place is crowded. the food is small, and the service isn't very friendly. i did not expect to spend a lot of money for a sushi place because it's pretty small. they have a few places where they can take a large order and they give it to those who pay for the food. also the staff seems to ignore you if you come for sushi.
Neg., American	GeDi	The food sucks!! Luckily we're Gilmore Girls and I still feel pretty strong from an adrenaline rush, not being adequately clothed. Return to Content
	Dist. Lens	The food is great and a little overpriced. service was great though. i will never go back. Advertisements for the sake of getting good food
	Prompt-tuning	The food you don't need to worry about it. This is the most important food. This could be years. This would also have you look for new business to move to. This is a big one and we're not looking for one.This might get more people. It may get people looking for work elsewhere, it could get people lookin at work in the area.
	Prefix-tuning	The food. is about negative aspects, and about negative food." – Michael Siegel. following is a negative review? –@siegel.
	Tailor	The food was good for what you paid for it, but the service was terrible, the food tasted bad. the waiter was rude and was constantly taking orders that were not on the table. when he asked why the food was so hot or that they were out of food. he said that the menu was out of order because of a busy night. after that he told me to call him back to make sure that the order was in order. the food itself was not that great, but that's ok because it was only $6.
	Prompt Gating (*Ours*)	The food was very bland. i had a burger that was really dry. they had the "tender beef patty". the beef was dry so i didn't like the "tasty" flavor and the burger tasted stale. i don't eat that type of beef.
Neg., Mexican	GeDi	The country didn't learn its lesson soon.
	Dist. Lens	The country fried is great, not the best. service was decent and we'll be back! Advertisements are good! Related Posts: (
	Prompt-tuning	The country. hell yes I am. and I am off. I am off to do some drugs and drink with a woman. I am off. and I am also off on drugs with the women! I can see I could go. and have sex with someone. I don't think I am a homa, though I could have sex sometime, which wouldn't be good with the men.I am off my period and I'm not ready yet for that, except sex, it is good with the women
	Prefix-tuning	The country and Mexican food. is about Mexican food. for a reason. It's about Mexican food. following is about Chinese food.is...Read Full Article.....and more »Zachary: Food is delicious...!}> More...</show>">Chef-in-chief"...Zach!</show><Chev...">ZACH!....!"</show>
	Tailor	The country style, fresh, delicious tacos, and fresh tortilla chips. the salsa, the burrito and rice were good. the salsa was very good, but the rice was a bit dry. i'm going to give the burrito a 10/10. the burrito was not as fresh as i would like, and the salsa was lacking, not fresh enough. my brother got the burrito with the green salsa, he liked it.
	Prompt Gating (*Ours*)	The country food at this place has become so bland that i have stopped coming here. the service was slow and the food didn't have any flavor to it at all. it just felt stale to the touch. i have never been in the place. i have been to many different mexian restaurants in my time in vegas, however i can only say they are the worst in vegas, at least to this point, in my memory

Constraints	Method	Generation Result
Pos., Asian	GEDI	Once upon a time···freedom—an old story···"" As Vegeta's hands encased his claws, Virankhan extended them to enhance their effects. His powers - including Frieza's tail transformation, long range ki blasts and history-defining transformations like Ash Ketchum's samurai thrill and Bulma & Pikkonü putting on their samurai mastership – swept through the crowd as they cheerfully flooded gutter comforts grounds majors like dubbed Kingston area. His qir
	DIST. LENS	Once upon a time, great food. the service is great and very friendly. Advertisements - yummy! *** Helpful!! :)
	PROMPT-TUNING	Once upon a time travel to another part of the park and back to this point, the park was still full of black. And this was the point after which we were left back there. And now I want to get up and go for another part. This is going back to this point but we are still left back in town to get rid of this.And you're left with that spot in the park.And you're left at the end of time for two. I'm getting rid from it but it has to go back
	PREFIX-TUNING	Once upon a time. following is about food.Advertisements···comments···comments
	TAILOR	Once upon a time, i was a big fan of this establishment. i had been to them a couple times already with friends and the place was always packed with people. this place was a hit, so much so that now I am going to go there every time.
	PROMPT GATING (*Ours*)	Once upon a time i was at the restaurant and had dinner at a friend's place, the food was great. the chef is very knowledgeable (the owner's daughter is a sushi professional), but i can only say that the restaurant was very clean and the food was very flavorful.
Pos., American	GEDI	The year is 1910. Ephraim Legrand spurges explores Europe's forgotten North Atlantic island, exploring its rich history, diverse landscapes, richside civic pride and communion. Through photography held in private property events concentrated throughout the year, content creators can collaborate creatively and celebrate individuality night-out!
	DIST. LENS	The year is 1910. great food and the service was great. will be back! Advertisements for good sushi!!! Follow me @wendie
	PROMPT-TUNING	The year is 1910. We are still going out to dinner. This time we're having fun with it but I don't like to leave it for another night.And this night has just been getting started so it's time for everyone to leave.And we've been eating this night for a week but not yet ready to leave the place to go to bed yet.And we've just eaten.But this morning has just started leaving this place so I am getting over it.You're not over it yet."And yet he is
	PREFIX-TUNING	The year is 1910. is about American food. is··· http://news-and-review-food-tourism.org/article.asp/?articleId=9c0c4-a2a-e0-b9-a-9c5d7a8b8f0_0_0.0
	TAILOR	The year is 1910. this is not your average sushi bar. the atmosphere was very casual and cozy. the food was good. the service was prompt and fast. the food was fresh and delicious.
	PROMPT GATING (*Ours*)	The year is 1910. we went to this place for the first year. it looks nice. there are a couple of booths, a table for 4 (and a counter) to get seated in. we had an early reservation. i ordered the steak. the steak was great. my boyfriend enjoyed the chicken wings. i was very impressed by the steak.
Pos., Mexican	GEDI	While this is happening in Scotland, though, we come a good distance north to Saltire Mountains National Park where we sit beside mountains on the Welsh coast called Myrta Tor. It's amazing to discover more about these beloved mountain peaks that haven't seen much mainstream attention until now. We'll travel along beautiful clifftees via Priordeuorth during walking & hiking treks around Myrta Tor 'check out our full itinerary here: Also visit Stonyrithenview Chantry
	DIST. LENS	While this is happening, great food! the margaritas are fantastic and friendly staff. Advertisements = good mexican restaurant! Related
	PROMPT-TUNING	While this is happening. It has already been happening. It's not like it's not a bad thing to have it but it's still bad. I don't like that either. It's got to get out of here before we can make any significant changes to it this is not that big thing anymore, but you don't have to do that. That's just a few days ago in the USA.This is the best way I could have done it and not done yet but this is way worse than that. It has to
	PREFIX-TUNING	While this is happening!Advertisements?Sponsors?Media?
	TAILOR	While this is happening to be my new favorite restaurant, and my new favorite location, i can't say that i've ever been to this place with all the other reviews. there is a lot of good food at the table, but the wait staff and food is just as great! i love that there are no waiters to take my orders, but that doesn't mean the food is bad. the food here was good.
	PROMPT GATING (*Ours*)	While this is happening i'm going here for dinner for the first time. the food here was very, very good and very tasty!! we ordered a couple of different salads and some tacos. i got a vego beef taco with a spicy sauce (it is very good). i also got an onion rings (it does not have any onions, nor are there many onions in this recipe), and it was delicious!

Table 13: Examples of multi-aspect controllable text generation. The given textual prefixes (see §B.2) are underlined. "**Constraints**" denotes the combination of constraints in sentiment and topic aspects. Some generation contents that in consist with the sentimental constraint are highlighted in blue, some generation contents that in consist with the topical constraint are highlighted in green, and some generation contents that fail to satisfy constraints are highlighted in red.

WARNING: Next may contain contents that are offensive in nature.

Example 1	
Constraint	
Keywords	"This", "bedroom", "completely"
Tense	The tense of this sentence is the past tense.
Knowledge (French)	Cette chambre et une autre ont été complètement brûlées.
Source and Reference	
Source (German)	Dieses und ein weiteres Zimmer brannten vollständig aus.
Reference	This and another bedroom were completely burnt out.
Translation	
PROMPT-TUNING	This and another room burned completely.
PREFIX-TUNING	This is the 'room under the sun'.
LORA	This and another room burned out completely.
PARALLEL ADAPTER	This tense and another room is completely burnt out.
PROMPT GATING (*Ours*)	This and another bedroom burned completely out.
Example 2	
Constraint	
Keywords	"The", "transgender", "employment,"
Tense	The tense of this sentence is the present tense.
Knowledge (French)	Le rapport donne également un aperçu de la discrimination à laquelle sont confrontées les personnes intersexes et transgenres dans le domaine de l'emploi, ainsi que des niveaux de harcèlement, de violence et de préjugés.
Source and Reference	
Source (German)	Der Bericht gibt auch einen Überblick über die Diskriminierung, der sich intersexuelle und Transgender-Personen im Berufsleben ausgesetzt sehen sowie über das Ausmaß von Belästigung, Gewalt und Vorurteilskriminalität.
Reference	The report also gives an overview of the discrimination faced by intersex and transgender people in the realm of employment, as well as levels of harassment, violence and bias crimes.
Translation	
PROMPT-TUNING	The report also gives an overview of the discrimination to which inter-sexual and transgender people are subjected in their professional life and the extent of harassment, violence and prejudice crime.
PREFIX-TUNING	The subject of the report is the subject of the report.
LORA	The report also gives an overview of the discrimination faced by inter-sexual and transgender people in their working lives and the extent of harassment, violence and prejudice.
PARALLEL ADAPTER	The report also gives an overview of the present discrimination faced by inter-sexual and transgender people in the workplace, as well as the extent of harassment, violence and prejudice.
PROMPT GATING (*Ours*)	The report also gives an overview of the discrimination suffered by inter-sexual and transgender people in employment, as well as the extent of harassment, violence and prejudice.
Example 3	
Constraint	
Keywords	"attempt"
Tense	The tense of this sentence is the future tense.
Knowledge (French)	Demain, il tentera de s'entraîner avec l'équipe.
Source and Reference	
Source (German)	Morgen wird er versuchen, mit der Mannschaft zu trainieren.
Reference	Tomorrow he will attempt to train with the team.
Translation	
PREFIX-TUNING	This is the subject of this article.
PARALLEL ADAPTER	Tomorrow he will try to train with the team.
LORA	The team he will try to train with the future.
PROMPT-TUNING	Tomorrow he will try to train with the team.
PROMPT GATING (*Ours*)	Tomorrow he will attempt to train with the team.

Table 14: Examples of multi-aspect controllable machine translation. "**Keywords**" denotes the given keywords that should be included in the translation. "**Tense**" denotes the input indicating the tense of the translation results. Similarly, "**Knowledge (French)**" denotes the external knowledge (i.e., French synonymous sentence). Some translations that satisfy the constraints are highlighted in blue, while some translations that fail to satisfy the constraints are highlighted in red.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations after the conclusion.*

☑ A2. Did you discuss any potential risks of your work?
Section Ethics Statement after the limitations.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 Introduction

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly, grammar error correction, the whole paper.

B ☑ Did you use or create scientific artifacts?

Section B Reproducibility

☑ B1. Did you cite the creators of artifacts you used?
Section B Reproducibility

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section B Reproducibility

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section B Reproducibility

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section B Reproducibility

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section B Reproducibility

C ☑ Did you run computational experiments?

Section 5 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section C Experimental Results

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section B Reproducibility

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section C Experimental Results

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section B Reproducibility

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Seciton E Details in Human Evaluation

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Seciton E Details in Human Evaluation

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Seciton E Details in Human Evaluation

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Seciton E Details in Human Evaluation

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Seciton E Details in Human Evaluation