TARA: Token-level Attribute Relation Adaptation for Multi-Attribute Controllable Text Generation

Yilin Cao[1,2], Jiahao Zhao[1,2], Ruike Zhang[1,2], Hanyi Zou[1,2], Wenji Mao[1,2]*
[1]MAIS, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{caoyilin2022,zhaojiahao2019,zhangruike2020,zouhanyi2021,wenji.mao}@ia.ac.cn

Abstract

Multi-attribute controllable text generation (CTG) aims to generate fluent text satisfying multiple attributes, which is an important and challenging task. The majority of previous research on multi-attribute CTG has ignored the interrelations of attributes that affect the performance of text generation. Recently, several work considers the attribute relations by explicitly defining them as *inhibtory*. We argue that for multi-attribute CTG, the attribute relations are not fixed, which can be not only *inhibtory* but *promotive* as well. In this paper, we tackle the multi-attribute CTG problem by explicitly identifying the above attribute relations for the first time and propose TARA, which employs token-level attribute relation adaptation and representation to generate text with the balanced multi-attribute control. Experimental results on the benchmark dataset demonstrate the effectiveness of our proposed method.

1 Introduction

Multi-attribute controllable text generation (CTG) aims to generate fluent text satisfying multiple attributes. The majority of previous research on multi-attribute CTG mainly employs parameter-efficient fine-tuning (Keskar et al., 2019; Zhang et al., 2020) and inference-time methods (Dathathri et al., 2020; Krause et al., 2021) to tackle the problem. However, most of them has ignored the interrelations of attributes, which is a fundamental issue in multi-attribute CTG.

Recent work (Qian et al., 2022; Ding et al., 2023) take the attribute relations into consider and utilizes prefix tuning or VAE to train a multi-attribute model. Several work (Gu et al., 2022; Huang et al., 2023) further defines multi-attribute relation as *inhibtory*. For instance, Dist. Lens (Gu et al., 2022) identify that mutual interference of controllers causes attribute control degeneration and searches for intersections in the attribute space. Prompt-Gating (Huang et al., 2023) use trainable gates to normalize the interference among attributes.

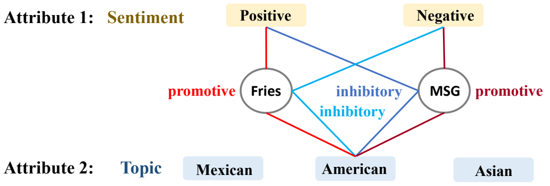

Figure 1: Token-level multi-attribute promotive and inhibitory relations. Here MSG is the abbreviation of Monosodium Glutamate.

In practical situations, the attribute relations are not fixed, nor are they only manifested as "inhibitory". Take the examples in Figure 1. There are multiple attributes including sentiment (with "positive" and "negative" attribute values) and topic (with "Mexican", "American" and "Asian" attribute values) in a typical restaurant domain. Fries demonstrates the promotive relation between *positive* and *American*, and the inhibitory relation between *negative* and *American*. This indicates that a more fine-grained definition and exploitation of attribute relations is needed for multi-attribute CTG.

In this paper, we tackle the multi-attribute CTG with **T**oken-level **A**ttribute **R**elation **A**daptation and representation, and propose TARA, which uses a dynamic text generation strategy. In summary, our contributions are as follows:

- We firstly identify both promotive and inhibitory attribute relations, and develop a token-level attribute relation adaptation method for multi-attribute CTG.
- The proposed attribute-adaptive prefix tuning adjusts attribute's expression with token-level attribute representation, and the dynamic text generation strategy we design balances multi-attribute control with promotive and inhibitory attribute relations.

[*]Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12570–12579
November 12-16, 2024 ©2024 Association for Computational Linguistics

- Experimental results verify that TARA performs better than existing methods on control ability and achieves text quality and diversity comparable with existing methods.

2 Proposed Method

We propose a novel multi-attribute relation adaptation method TARA for controllable text generation by leveraging inhibitory and promotive relations. Figure 2 illustrates the overall structure of TARA, which consists of attribute-adaptive prefix tuning, token-level attribute representation and dynamic text generation strategy. To weaken the inhibitory relation, we use attribute-adaptive prefix tuning to reduce the logits of tokens with low attribute expression. This ensures that tokens with weak attribute expression have minimal impact on other control attributes. To strengthen the promotive relation, we implement token-level attribute representation and a dynamic text generation strategy to focus more on attributes with poorer control based on the current sequence.

2.1 Attribute-Adaptive Prefix Tuning

In multi-attribute CTG, the presence of tokens that exhibit inhibitory relations between attributes can lead to a degradation in the performance of multi-attribute control. Therefore, we employ attribute-adaptive prefix tuning for attribute models to weaken the inhibitory relations. It reduces the logits of tokens with weak attribute expressions, thereby weakening the inhibitory relations and ensuring minimal impact on other control attributes.

Given the prompt or the current sequence as S_1^{t-1}, at the current time step t, we can obtain the attribute model's logits l_t over the vocabulary $\mathcal{V}$ through the language model. The language model will generate next token s_t by sampling $s_t \sim P(s_t \mid S_1^{t-1})$ based on its logits l_t. We first convert the logits l_t to probabilities $P(s_t \mid S_1^{t-1})$ to make the attribute distinctions among tokens more apparent.

$$P(s_t \mid S_1^{t-1}) = \text{softmax}\left(\frac{l_t}{\tau}\right) \tag{1}$$

where τ is the temperature coefficient.

Then, we use the L2 norm to constrain the probability distribution of the attribute model, avoiding extreme probability values and enhancing stability. Applying L2 normalization to the output probability acts as a forgetting mechanism. When the model is forgetful, it must identify the most vital property of each attribute. In this situation, the learned property is more likely to be related to the attribute itself rather than a shared property across different attributes. This reduces the correlation between different attributes, facilitating subsequent operations. Finally, we use the variance of probabilities to enhance the attribute model's ability to distinguish tokens with varying degrees of attribute expression, mitigating tokens with weak single attribute expression. Introducing the variance of probabilities primarily helps mitigate the logits of tokens with weak attribute expressions while retaining the logits of tokens with strong attribute expressions. As a result, the logits of tokens with strong attribute expressions become significant in the logits distribution obtained from the attribute model. The variance term also prevents training collapse. The adapt loss $\mathcal{L}_{\text{adapt}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{adapt}} = \Big(&\|P(s_t \mid S_1^{t-1})\|_2 \\ &- \lambda_{\text{var}} \cdot \text{Var}(P(s_t \mid S_1^{t-1}))\Big)\end{aligned} \tag{2}$$

where $\|P(s_t \mid S_1^{t-1})\|_2$ denotes the L2 norm of the probabilities, and $\text{Var}(P(s_t \mid S_1^{t-1}))$ represents the variance of the probabilities. λ_{var} is the regularization weights for the L2 norm and the variance, respectively.

Finally, the total loss $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = -\log P(s_t \mid S_1^{t-1}) + \lambda_{\text{reg}}\mathcal{L}_{\text{adapt}} \tag{3}$$

where λ_{reg} is the regularization weights for $\mathcal{L}_{\text{adapt}}$.

2.2 Token-level Attribute Representation

In TARA, we fine-tune the attribute model for each attribute value using the same pre-trained language model. Given the current input S_1^{t-1}, we can obtain the logits distribution of the attribute model. Not only tokens that express the attribute characteristics will have high logits, but tokens that ensure text quality will also receive high logits, such as *with, a, the, is*. In TARA, we aim to utilize pure logits that only express the attribute characteristics for attribute control. Therefore, we define the attribute representation r_{att} as follows:

$$r_{\text{att}} = l_{\text{att}} - l_{\text{PLM}} \tag{4}$$

where att represents the attribute value, l_{att} represents the logits of attribute model, l_{PLM} represents the logits of base PLM. Attribute Representation reflects the significance of the corresponding token

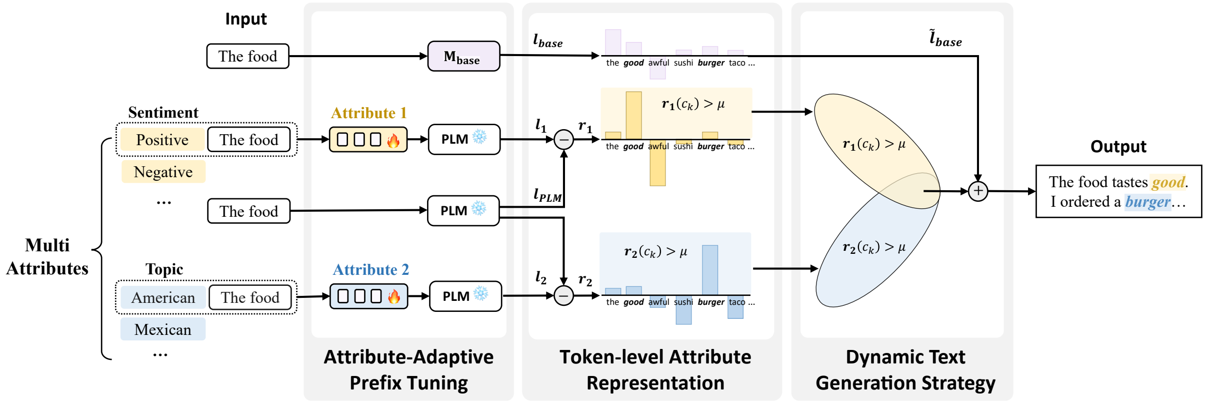

Figure 2: Overview of Multi-Attribute Relation Adaptation Method for CTG.

in expressing the current attribute. It helps achieve a purer attribute expression and generates dynamic vocabularies based on the current sequence.

Suppose we have two attribute values, i and j (TARA can control multiple attributes). In the vocabulary $\mathcal{V}$, we define the tokens as $\{c_1, \ldots, c_n\}$, where n is the total number of tokens in the vocabulary. We use c_k to represent a specific token in the vocabulary, where $k \in \{1, \ldots, n\}$. The same token may demonstrate different attribute relations under different multi-attribute control. Therefore, we first set a threshold μ to divide the attribute representation into two regions. Then we define two kinds of attribute value i and j relations at the token-level as follows:

Promotive relation:

$$\begin{cases} r_i(c_k) > \mu \quad \text{and} \quad r_j(c_k) > \mu \\ r_i(c_k) < \mu \quad \text{and} \quad r_j(c_k) < \mu \end{cases} \tag{5}$$

This condition indicates that the representations of $r_i(c_k)$ and $r_j(c_k)$ are consistent.

Inhibitory relation:

$$\begin{cases} r_i(c_k) > \mu \quad \text{and} \quad r_j(c_k) < \mu \\ r_i(c_k) < \mu \quad \text{and} \quad r_j(c_k) > \mu \end{cases} \tag{6}$$

This condition indicates that the representations of $r_i(c_k)$ and $r_j(c_k)$ are inconsistent.

In multi-attribute CTG, we establish dynamic vocabulary $\mathcal{V}_i$ and $\mathcal{V}_j$ for each attribute value. Then we can get:

$$\mathcal{V}_i = \{c_k \in \mathcal{V} \mid r_i(c_k) > \mu\} \tag{7}$$

$$\mathcal{V}_j = \{c_k \in \mathcal{V} \mid r_j(c_k) > \mu\} \tag{8}$$

2.3 Dynamic Text Generation Strategy

In TARA, we design a dynamic text generation strategy to exploit multi-attribute relation and balance multi-attribute control to steer the generation. It calculates the conditional probability vector to monitor how well the current sequence adheres to the control attributes. Therefore, this strategy strengthens the promotive relation by shifting the focus towards attributes with poorer control. We employ multi-attribute representations r_{att} in conjunction with the base logits l_{base} from a quality control model, which can either be an LLM or a small LM, sharing the same vocabulary as the attribute pre-trained model. In this setup, the attribute representations r_{att} manage multi-attribute control, while l_{base} ensures the quality of the text text. Besides, we design a dynamic weights to effectively balance the relations between two attributes, while ensure the text quality. Following Liu et al. (2021), we applied nucleus sampling (Holtzman et al., 2020) to base model to obtain a fluent output sequence. At time step t, let $\mathcal{V}' \subseteq \mathcal{V}$ represent the tokens included in the top-p vocabulary of the base quality control model. The truncated logits $\tilde{l}_{\text{base}}$ are:

$$\tilde{l}_{\text{base}}[v] = \begin{cases} l_{\text{base}}[v] & \text{if } v \in \mathcal{V}' \\ -\infty & \text{otherwise} \end{cases} \tag{9}$$

For the two attribute value i and j, we define W_i and W_j represent the conditional probabilities of common tokens under attribute values i and j:

$$W_i = \text{softmax}\left(\frac{|\mathcal{V}_i \cap \mathcal{V}_j|}{|\mathcal{V}_i|}\right) \tag{10}$$

$$W_j = \text{softmax}\left(\frac{|\mathcal{V}_i \cap \mathcal{V}_j|}{|\mathcal{V}_j|}\right) \tag{11}$$

Method	Correctness (%)			Text Quality	Diversity
	Sentiment ↑	Topic ↑	Avg ↑	PPL ↓	mean-Dist ↑
PromptTuning* (Lester et al., 2021)	48.29	48.11	48.20	40.89	0.42
PrefixTuning* (Li and Liang, 2021)	47.53	69.11	58.32	147.47	0.31
ControlPrefixTuning (Clive et al., 2022)	58.98	45.36	52.17	89.80	0.48
GeDi* (Krause et al., 2021)	**99.47**	51.36	75.41	616.92	**0.75**
Tailor* (Yang et al., 2023)	80.68	68.72	74.70	40.29	0.39
Dist. Lens* (Gu et al., 2022)	77.47	66.98	72.22	52.59	0.26
PromptGating* (Huang et al., 2023)	84.80	75.02	79.91	**21.77**	0.42
TARA (Ours)	90.17	**80.32**	**85.25**	40.16	0.45

Table 1: The main results of multi-attribute CTG. For each method, we select 6 combinations (two sentiment attributes × three topic attributes) as the final results. [1]

Variant	Correctness (%)			Text Quality	Diversity
	Sentiment ↑	Topic ↑	Avg ↑	PPL ↓	mean-Dist ↑
TARA	90.17	80.32	85.25	40.16	0.45
– Attribute-Adaptive Prefix Tuning	89.02	77.47	83.25	44.17	0.45
– Token-level Attribute Representation	76.46	40.55	58.51	10.96	0.18
– Dynamic Text Generation Strategy	87.63	75.11	81.37	31.80	0.43

Table 2: Ablation Study of attribute-adaptive prefix tuning and dynamic text generation strategy of TARA.

To normialize to multi-attribute weight to $[0, 1]$ at the token-level, we design dynamic weights $\tilde{W}_i$ and $\tilde{W}_j$ as follows:

$$\tilde{W}_i = \frac{W_i}{W_i + W_j} \tag{12}$$

$$\tilde{W}_j = \frac{W_j}{W_i + W_j} \tag{13}$$

For a more reasonable sampling process, as in (Fan et al., 2018), we applied top-K processing to the ensemble logits $\tilde{l}_t$ during sample process. Therefore, the next token s_t can be obtained through the following dynamic text generation strategy:

$$\tilde{l}_t = \tilde{l}_{\text{base}} + (1 + \tilde{W}_i) \cdot r_i + (1 + \tilde{W}_j) \cdot r_j \tag{14}$$

$$\tilde{P}(s_t | S_1^{t-1}) = \text{softmax}(\tilde{l}_t) \tag{15}$$

$$s_t \sim \tilde{P}(s_t | S_1^{t-1}) \tag{16}$$

3 Experiments and Results

3.1 Experimental Setup

Dataset We choose widely used benchmark dataset YELP (Lample et al., 2019) for our experiments. Following previous work, we use sentiment attribute (positive and negative) and topic attribute (Asian, American and Mexican) for multi-attribute controllable text generation. Please refer to Appendix B for more details on experiment setup.

Evaluation Metrics Following Yang et al. (2023); Huang et al. (2023), we conduct automatic and human evaluation for controllable accuracy and text quality. We conduct automatic evaluation from three aspects: (1) **Correctness** We fine-tune a sentiment classifier, a topic classifier and a dessert classifier based on RoBERTa (Liu et al., 2019) for the evaluation of sentiment and topic accuracy. (2) **Text Quality** We calculate the perplexity (PPL) using GPT-2$_{medium}$ (Radford et al., 2019) to evaluate the fluency. (3) **Text Diversity** We use averaged distinctness (Li et al., 2015) to evaluate the diversity. We conduct human evaluation for sentiment relevance, topic relevance and fluency. Each rating can be evaluated from 1 to 5. And we get final scores from the average of three ratings.

3.2 Baselines

We compare our approach with main representative methods as follows: **Prefix-Tuning** (Li and Liang, 2021) appends trainable prefixes to parameter efficiently tuning the pre-trained model. Sim-

[1] The symbol * indicates that the results are obtained from Huang et al. (2023).

ply concact the single attribute prefix to realize multi-attribute control. **Prompt-Tuning** (Lester et al., 2021) appends continuous prompts to guide the generation. The prompts are trained parameter efficiently and are simply concatenated for multi-attribute control. **Control Prefix Tuning** (Clive et al., 2022) extends Prefix-Tuning (Li and Liang, 2021) and adds attribute-level learnable representations into different layers of a pre-trained model. We combine the representations for multi-attribute control. **GeDi** (Krause et al., 2021) uses generative discriminators to guide large LMs generation as a inference-time method. For multi-attribute control the distributions of multi discriminators are normalized. **Dist. Lens** (Gu et al., 2022) estimates the attribute space using an autoencoder and searches for intersections using a prefix-based decoder. **Tailor** (Yang et al., 2023) bridges the gap between the training and testing stage using prompt mask and position-id re-index by Prompt-Tuning (Lester et al., 2021). **Prompt Gating** (Huang et al., 2023) provides trainable gates to normalized the intervention of the prefixes.

3.3 Main Results and Analysis

As shown in Table 1, TARA achieved the highest average accuracy in multi-attribute control, surpassing the best comparative method by 5.34% while maintaining text quality. TARA shows a larger improvement in the multi-attribute CTG, demonstrating the necessity of carefully handling the promotive and inhibitory relations between attributes. Both automatic and human evaluations (see Appendix D) indicate that TARA effectively balances multiple control attributes with text quality and diversity. We observe that Tailor performs better than previous comparative methods by bridging the gap between the training and testing stages. Dist. Lens and PromptGating both consider and mitigate the inhibitory relations between attributes.

Next, we specifically analyze the adaptation process of attribute relations in TARA. To exploit promotive relation, the conditional probability vector is introduced to monitor how well the current sequence adheres to control attributes. It is calculated by dynamic vocabularies, which are established by the promotive relation of control attributes. For example, a smaller conditional probability vector indicates the current sequence has achieved better control on the corresponding attribute and the focus will shift more towards the attribute with poorer control when selecting the next token. Therefore, the logits under poorer control attribute will be enlarged and the promotive relation is strengthened. To exploit inhibitory relation, we employ attribute-adaptive prefix tuning to weaken the inhibitory relations. After attribute-adaptive prefix tuning, the logits of tokens with weak attribute expressions are reduced. Therefore, when combining logits of control attributes, the tokens with reduced expression will weaken the inhibitory relations. For example, under the multi-control of *Negative* and *American*, the logit of token *Fries* is -20 from the *Negative* model, and is 100 from the *American* model. Obviously, the logit from the *Negative* model will weaken the expression of the token *Fries* when combining -20 and 100, while *Fries* is an important attribute token in the *American* model. Therefore, the token *Fries* reflects the inhibitory relation under the multi-control of *Negative* and *American*. After the attribute-adaptive prefix tuning, the logit of token *Fries* is reduced to -5 from the *Negative* model. Therefore, the inhibitory relation between the control attribute is weakened.

We conduct an ablation study to evaluate each component of TARA. The results are shown in Table 2. It can be seen that: (1) Attribute-adaptive prefix tuning improves multi-attribute control ability across all attributes, proving its effectiveness in weakening inhibitory relations. (2) Token-level attribute representation plays a crucial role in improving both multi-attribute control and text quality. (3) The dynamic text generation strategy effectively balances the relations between attributes, supporting the notion that better exploiting multi-attribute relations enhances model performance.

4 Conclusion

We introduce TARA, a token-level attribute relation adaptation method for multi-attribute CTG. We design attribute-adaptive prefix tuning to weaken the inhibitory relation. Then we employ token-level attribute representation to achieve a purer attribute expression and generate dynamic vocabularies. Finally, we design a dynamic text generation to strengthen the promotive relation by shifting the focus towards attributes with poorer control and steer the generation towards more precise and balanced control of multi attributes. Through experimental evaluations on multi attribute CTG, we demonstrate the effectiveness of TARA in terms of both control ability and text quality and diversity.

Limitations

While TARA achieves fairly good performance in multi-attribute controllable text generation task, its results on text perplexity and diversity are slightly inferior to the best-performing methods. The model may exhibit inherent biases present in the dataset used for training. Since the text generation model is trained on data collected from the web, it may reproduce and even amplify these biases in the generated texts. This includes, but is not limited to, biases related to gender, race, and culture.

Ethics Statement

Similar to many other text generation models, our approach may occasionally produce offensive or toxic text, especially when generating negative control text. It is important to state that the texts generated by our approach do not represent our opinions. To alleviate these issues, our multi-aspect controllable method could benefit from incorporating detoxification and politeness constraints as default aspects. This would help mitigate the risk of generating harmful content.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants #72293575 and #62206287. We thank the anonymous reviewers for the valuable comments.

References

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 363–382, Abu Dhabi, United Arab Emirates (Hybrid). 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*.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun, and Yang Liu. 2023. An extensible plug-and-play method for multi-aspect controllable text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15233–15256, Toronto, Canada. Association for Computational Linguistics.

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 the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

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.

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

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. BOLT: Fast energy-based controlled text generation with tunable biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 186–200, Toronto, Canada. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023. Tailor: A soft-prompt-based approach to attribute-based controlled text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–427, Toronto, Canada. Association for Computational Linguistics.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020. POINTER: Constrained progressive text generation via insertion-based generative pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8649–8670, Online. Association for Computational Linguistics.

A Related Work

Parameter Efficient Tuning Parameter-efficient fine-tuning (PEFT) methods could realize controlled text generation in a lightweight and efficient way with low training cost. Prefix tuning (Li and Liang, 2021) use fixed LM and trainable added key-value pairs before activation layers. Control prefixes (Clive et al., 2022) extend prefix tuning by incorporating attribute-level learnable representations into a pretrained transformer. Training conditional language models (Keskar et al., 2019; Zhang et al., 2020; Clive et al., 2022) is a common approach for controllable text generation. In multi-attribute CTG, Tailor (Yang et al., 2023) bridges the gap between the training and testing stage using prompt mask and position-id re-index by Prompt-Tuning (Lester et al., 2021).

Inference-time Methods Inference-time method is a lightweight and effective approach for multi-attribute CTG. PPLM (Dathathri et al., 2020) use attribute classifiers' gradients to guide the pretrained LM by updating LM's latent states per time step, which is a time-consuming process. GeDi (Krause et al., 2021) use generative discriminators to guide large LMs generation as a inference-time method. DExperts (Liu et al., 2021) combines a base LM with "expert" LMs and "anti-expert" LMs for detoxification. BOLT (Liu et al., 2023) design energy function and tune the bias over logits of the PLM's output layer with the goal of minimizing the generated sequence's energy to steer the generation.

B Dataset

YELP dataset is a widely-used restaurant reviews dataset contains sentiment attribute (positive and negative) and topic attribute (Asian, American and Mexican). Following previous work, we adpot YELP dataset (Lample et al., 2019) for multi-attribute controllable text generation. For example, given two attributes SENTIMENT=POSITIVE TOPIC=AMERICAN and the prompt *"The food"*, the model needs to generate text satisfying both attributes and beginning with the prompt, such as *"The food in this restaurant is dear to my heart, especially the fries."*. We randomly sample 30K/3K sentences of each attribute value for training/validation. To be consistent with previous work, we use 15 textual attribute-unrelated prefixes for the model to generate from them during inference. The 15 prefixes are:"Once upon a time", "the book", "The chicken", "The city", "The country", "The lake", "The movie", "The painting", "The weather", "The food", "While this is happending", "The pizza", "The potato", "The president of the country", "The year is 1910.". For evaluation, to keep with previous work (Huang et al., 2023), we sample 25 sentences for each prefix and controllable attribute combinations. We compute the average score of the sampled generation sentences based on the 15 prefixes for the final results.

C Experimenmtal Details

C.1 Hyperparameters

Hyperparameters of TARA are shown in Table 3. In the TARA experiments, we use the GPT2-medium model with 355M parameters to maintain consistency with the baselines.

C.2 Evaluation Metrics

For the evaluate of the control accuacy, we finetune a sentiment classifier and topic classifer based on RoBERTa (Liu et al., 2019). Following (Huang et al., 2023), we randomly sample 1,380K/1K/1K sentences as training/validation/test set of sentiment and 1,500K/15K/15K sentences as training/validation/test set of topic. The F1 scores for sentiment and topic are 98.00 and 83.77.

D Human Evaluation

For the human evaluation, we shuffled the generated text and select three volunteers to score. Each sentence was rated on a score from 1 to 5 for attribute controllability and text fluency. The final scores represent the average of the three ratings. We recruited three volunteers from local schools to participate in the human evaluation process. These volunteers were selected based on their proficiency in English, ensuring they have sufficient daily communication skills. The average age of the volunteers was 24 years old. Since they were volunteers, they were not paid for their participation.

Following Huang et al. (2023), we provide the same instruction to volunteers "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

Hyper-parameter	**TARA**
Pre-trained Model	
GPT2-medium	355M
Encoder layers	12
Decoder layers	12
Attention heads	16
Attention head size	64
Hidden size	1,024
FFN hidden size	4,096
Max sentence length	1,024
Training	
Optimizer	AdamW
Adam beta	momentum
Training steps	10,972
Batch size	32
Learning rate (sentiment)	4×10^{-4}
Learning rate (topic)	1×10^{-2}
Temperature (sentiment)	0.1
Temperature (topic)	0.06
λ_{reg} (sentiment)	1×10^{-3}
λ_{reg} (topic)	3×10^{-4}
λ_{var} (sentiment)	1×10^{-2}
λ_{var} (topic)	1×10^{-1}
Residual dropout	0.0
Attention dropout	0.0
Activation dropout	0.0
Inference	
top-p (sampling)	0.9
top K	8
Beam size	/
μ	0

Table 3: Hyperparameters of TARA.

scores will only be used for human evaluation in research". The criteria used for scoring the generated texts were divided into two main categories: Fluency and Attribute relevance. We set the score on scale of 1-5, with a higher degree of topic/sentiment relevance representing a more consistent topic/sentiment, and a higher degree of text fluency representing a more fluent text. Specifically, each category had a detailed description for each score from 1 to 5 as shown in Table 5 and 6. The human evaluation results is shown in Table 4.

Method	**Sentiment ↑**	**Topic ↑**	**Fluency ↑**
ControlPrefixTuning	4.3	3.6	3.8
TARA (Ours)	4.6	4.2	4.1

Table 4: Human evaluation results

E Experiment in More Complex Settings

To validate the effectiveness of TARA in more complex settings, we conducted further experiments on Yelp dataset by adding a new attribute: whether the review mentions dessert. Controlling three attributes simultaneously is challenging, yet TARA achieved comparable performance in sentiment accuracy and topic accuracy as in the two-attribute case. The results in Table 7 validate the robustness and scalability of TARA in more complex settings.

F Parameters in TARA

To clarify the meaning of parameters, we present the definitions and corresponding dimensions of parameters of TARA in Table 8.

Attribute	Description
1	There are no attribute-related words or phrases in the sentences.
2	There is only one attribute-related word or phrase in the sentences.
3	Sentences contain multiple attribute-related words or phrases, but they are almost repetitive.
4	Sentences contain multiple attribute-related words or phrases, with a few of them being repetitive.
5	Sentences contain multiple attribute-related words or phrases, none of them being repetitive.

Table 5: Attribute Score Criteria in Human Evaluation

Fluency	Description
1	All sentences are difficult to read and incomprehensible.
2	Only a small part of the sentences can be understood, which is readable and fluent.
3	Apart from a few grammatical mistakes, sentences are clear and comprehensible.
4	Sentences are free from grammatical errors and other linguistic inconsistencies but could be better in style.
5	Sentences are fluent and spontaneous, equating to the text quality of human writing.

Table 6: Fluency Score Criteria in Human Evaluation

Method	Correctness (%)				Text Quality	Diversity
	Sentiment ↑	Topic ↑	Dessert ↑	Avg ↑	PPL ↓	mean-Dist ↑
TARA (double-attribute)	90.17	80.32	–	85.25	40.16	0.45
TARA (triple-attribute)	88.95	76.28	69.13	78.12	45.04	0.48

Table 7: The results of triple-attribute CTG compared to double-attribute CTG of TARA.

Parameters	Definition	Dimension
t	the current time step	Scalar
$\mathcal{V}$	vocabulary of GPT2-medium	[50,257]
S_1^{1-t}	the current sequence	$[(t-1), 768]$
s_t	next token	$[1, 768]$
l_t	the logits of the attribute model	[50,257]
r_{att}	attribute representation	[50,257]
c_k	the k-th token in vocabulary $\mathcal{V}$	Scalar (index)
$\mathcal{V}_i$	dynamic vocabulary of attribute i	[Variable length]
$\mathcal{V}_j$	dynamic vocabulary of attribute j	[Variable length]
$\|\|\mathcal{V}_i\|\|$	the number of tokens in $\mathcal{V}_i$	[Variable length]
$\|\|\mathcal{V}_j\|\|$	the number of tokens in $\mathcal{V}_j$	[Variable length]
$\|\|\mathcal{V}_i \cap \mathcal{V}_j\|\|$	the number of tokens both in $\mathcal{V}_i$ and $\mathcal{V}_j$	[Variable length]
W_i	the conditional probability of attribute i	Variable Scalar
W_j	the conditional probability of attribute j	Variable Scalar
$\tilde{W}_i$	the dynamic weights of attribute i	Variable Scalar
$\tilde{W}_j$	the dynamic weights of attribute j	Variable Scalar

Table 8: Parameters of TARA.