Latent Space Attribute Disentanglement for Attribute-Based Controllable Text Generation with Large Language Models

Anonymous ACL submission

Abstract

Attribute-based controllable text generation (CTG) aims to produce sentences that satisfy user-specified properties while retaining fluency. Existing approaches either bias the decoding logit vectors or fine-tune small adapter layers. However, they implicitly assume that the latent space of the model already provides clear and linearly separable directions for every attribute. In reality, heterogeneous training corpora induce highly overlapping attribute distributions, entangle latent features, and interfere with text generation with specific attributes. We introduce Latent Space Attribute Disentanglement, a lightweight but practical framework that explicitly factorizes the latent space into orthogonal subspaces, one for each attribute. Concretely, we attach gated LoRA experts to every transformer block; the gating mechanism learns to capture attribute-specific patterns. These experts are optimized with two complementary objectives, domain alignment and subspace independence, enforced by additional loss terms. During decoding, our method generates text that precisely exhibits the desired attributes; extensive experiments demonstrate that the proposed framework delivers consistent and significant gains on attribute-specific generation tasks.

1 Introduction

Controllable text generation has become an important research direction in natural language processing, enabling the generation of text that adheres to specific attributes or constraints (Carlsson et al., 2022; Yang et al., 2023). This capability is pivotal for personalized content creation, dialogue systems, and data augmentation applications. Existing approaches for controllable text generation can generally be divided into three main categories: fine-tuning (Keskar et al., 2019; Ziegler et al., 2020), latent space manipulation (Gu et al., 2022, 2023), and decoding-time intervention (Dathathri et al., 2019; Li et al., 2022). These approaches aim to steer the generative model toward producing text that aligns with the desired control attributes while maintaining fluency and coherence.

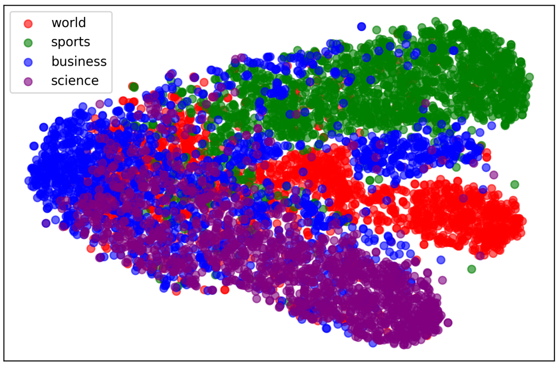

Figure 1: The distributions of the topics *world*, *sports*, *business*, and *science* in the latent space of AGNews are visualized using t-SNE. The entanglement of attribute distributions in the latent space introduces interference when generating text with specific attributes and negatively impacts the classifier's performance. The classifier used for attribute statistics is adopted from (Gu et al., 2022).

Due to the difficulty in obtaining training data that satisfy arbitrary combinations of attributes, existing controllable text generation methods (Krause et al., 2021; Gu et al., 2022; Ding et al., 2023) often reuse datasets with single-aspect annotations, where each training sample expresses only one attribute in a specific aspect. However, most current techniques implicitly assume that each attribute corresponds to a linearly separable direction in the model's latent space. In reality, heterogeneous training corpora result in highly entangled distributions, since desired attributes are embedded with confounding factors such as lexical topics, stylistic features, and discourse structures. Consequently, the attribute directions in the latent space of large language models (LLMs) are not linearly separable from these confounding factors. As shown in

Figure 1, we visualize the distribution of four attributes in AGNews using t-SNE (Van der Maaten and Hinton, 2008) and observe significant semantic overlap, which poses a challenge in the generation of precise text attributes.

To address these limitations, we propose the Latent Space Attribute Disentanglement framework. This lightweight framework explicitly decomposes the latent space into orthogonal subspaces, each corresponding to a specific attribute. Quite structurally, we attach multiple LoRA (Hu et al., 2022) modules to each transformer block and modulate them using a learnable gating network. The gating function dynamically combines their output, enabling these LoRA experts to learn attribute-specific features while introducing only 4% additional trainable parameters. These experts are jointly optimized with two complementary objectives: domain alignment and subspace independence. We propose an Aspect-Adaptive Loss, which aligns the center of the same aspect across different data sources to mitigate distributional mismatches, and an Attribute-Aware Loss, which compacts intra-attribute representations and enhances subspace independence. Our method can accurately generate text with the desired attributes during decoding without relying on external classifiers. Experimental results demonstrate that our framework outperforms strong baselines in controllable text generation, achieving state-of-the-art performance without compromising text quality. The main contributions of our framework are as follows:

- We analyze the issue of overlapping attribute distributions in current controllable text generation methods and propose a disentanglement-based approach to mitigate this problem.
- We propose a multi-LoRA-based latent space attribute disentanglement framework that addresses two key challenges: imbalanced attribute distribution and lack of subspace independence, enabling precise controllable text generation.
- We experimentally validate the effectiveness of our method, and the experimental results demonstrate that our method achieves state-of-the-art (SOTA) performance.

2 Related Work

Fine-Tuning. Fine-tuning refers to the further updating of a pre-trained model's parameters to adapt it to specific tasks (Feng et al., 2023; Zheng et al., 2023; Kumar et al., 2023). DisCup (Zhang and Song, 2022) combines a frozen causal language model with an attribute discriminator to optimize control prompts via unlikelihood training. InstructCTG (Zhou et al., 2023) achieves controllable text generation by converting constraints into a natural language instruction dataset and fine-tuning the language model on an augmented dataset. The fine-tuning approach balances adaptability and resource efficiency, making it a common choice for enhancing model performance on specific tasks.

Latent Space Manipulation. Latent space manipulation aims to control text generation by adjusting internal model representations (Chan et al., 2021; Lu et al., 2023). PriorControl (Gu et al., 2023) uses probability density estimation methods in latent space to effectively manage complex attribute distributions through reversible transformations. Con. Prefix (Qian et al., 2022) keeps the base GPT-2 model frozen and learns a small set of attribute-specific prefix vectors. These prefixes are trained with a contrastive objective so that vectors for the same attribute cluster together while different attributes stay apart. These approaches demonstrate flexibility in controlling generation without altering model architecture.

Decoding-Time Intervention. Decoding-time intervention controls the attributes of the generated text by manipulating the logit of the model or the probability distribution during generation. PPLM (Dathathri et al., 2019) controls text attributes by iteratively adjusting the hidden layer activations of GPT-2 using the gradient of attribute classifiers. GeDi (Krause et al., 2021) uses a discriminator to guide language model decoding to calculate the classification probability for each next token, effectively controlling text generation. Air-Decoding (Zhong et al., 2023) reconstructs attribute distributions to balance the weights between attribute and non-attribute words, effectively generating more fluent and controllable text. The decoding-time intervention enables controllable text generation without training models and is more interpretable.

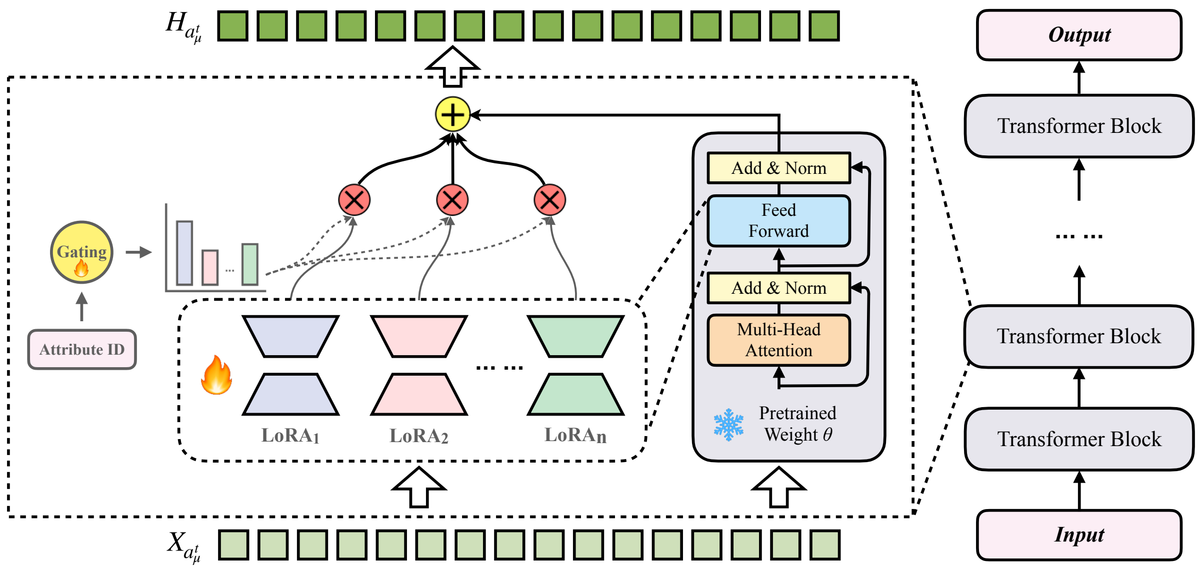

Figure 2: Illustration of our proposed framework. Our framework extends the traditional LoRA by integrating multiple LoRA modules and employs a learnable gating function to dynamically combine multiple LoRA modules. We use the attribute identifier as the input of the gating function to learn unique parameters for each attribute. $X_{a_\mu^t}$ represents the input sequence containing attribute a_μ^t and $H_{a_\mu^t}$ is the output hidden state. Only the parameters of LoRAs and the gating function are updated during training.

3 Formulation

Task definition. Let $\mathcal{A} = \{A_1, \ldots, A_N\}$ stands for N aspects. $a^t = \{a_1^t, \ldots, a_{|A_t|}^t\}$ is the set of all attributes in aspect A_t, where $|A_t|$ is the number of attributes. We represent text with attribute $a_\mu^t \in A_t$ as $T_{a_\mu^t}$. The goal of controllable text generation is to generate sentences with specific attributes from different aspects, such as attribute "positive" from aspect *sentiment* or attribute "sports" from aspect *topic*.

Training samples. Training samples are instruction following samples: Each training sample consists of four parts: a_μ^t is the control attribute from aspect A_t, $I_{a_\mu^t}$ is the instruction with attribute a_μ^t, $T_{a_\mu^t}$ is the target text with attribute a_μ^t and $D_{a_\mu^t}$ stands for the identifier of attribute a_μ^t. The instructions of all training samples are $I = \bigcup_{t=1}^{N} I^t$, where I^t is the set of instruction of aspect A_t. Identifiers of all attributes are represented as $D = \bigcup_{t=1}^{N} \bigcup_{\mu=1}^{|A_t|} D_{a_\mu^t}$.

4 Methodology

In this section, we first introduce the motivation behind our approach and provide an overview of our proposed framework. Subsequently, we present a detailed description of the proposed framework and our training objective.

4.1 Overview

In controllable text generation, input instructions typically contain a control attribute of different aspects. Therefore, it is necessary to consider the dynamical distribution of the input instructions. As illustrated in Figure 2, our framework takes the Feed-Forward Neural Network (FFN) as an example, where we introduce multiple trainable LoRA modules into the FFN layer to capture diverse knowledge across different controllable text generation tasks. Furthermore, we incorporate a gating function that takes an attribute identifier as input, aiming to learn unique model parameters for each attribute and dynamically combine multiple LoRA modules. At the top level of the transformer block, we impose constraints on hidden states to achieve domain alignment and subspace independence.

4.2 Architecture

Referring to Figure 2, consider a transformer block parameterized by parameters θ, which includes the multi-head attention layer and FFN parameters, and remains constant during training. The trainable set of LoRA modules is represented as $\Delta W = \{\Delta W_1, \ldots, \Delta W_n\}$, where n represents the number of trained LoRAs.

Let $X_{a_\mu^t} \in \mathbb{R}^{l \times d}$ represent the input sequence that contains the attribute a_μ^t, where l is the length of the sequence and d is the dimension. The output of the multi-head attention layer, combined with residual connection and layer normalization, is:

$$X'_{a_\mu^t} = X_{a_\mu^t} + LN(f_{Attn}(X_{a_\mu^t} \mid \theta)), \quad (1)$$

where $f_{Attn}(\cdot)$ is the multi-head attention layer and $LN(\cdot)$ refers to layer normalization. Each LoRA module ΔW_i consists of a pair of low-rank matrices $A_i \in \mathbb{R}^{d_{in}\times r}$ and $B_i \in \mathbb{R}^{r\times d_{out}}$, where $r \ll min(d_{in}, d_{out})$ is the adaptation rank. The LoRA transformation for i-th module is defined as:

$$L_i(X'_{a_\mu^t}) = X'_{a_\mu^t} A_i B_i, \quad (2)$$

where $1 \leq i \leq n$. A_i and B_i project the input $X'_{a_\mu^t}$ to a lower-dimensional space and then back to the original dimension, efficiently capturing the specific characteristics of the attribute.

To determine the contribution of each LoRA module and dynamically adjust model parameters, we introduce a gating function G, which takes an attribute identifier $D_{a_\mu^t}$ as input and outputs a weight vector $\omega \in \mathbb{R}^n$. The gating function is implemented as an embedding layer, a linear layer, and a softmax layer. The output weight is:

$$\omega_i = \frac{\exp(G(D_{a_\mu^t})_i)}{\sum_{j=1}^{n} \exp(G(D_{a_\mu^t})_j)}, \quad (3)$$

where ω_i denotes the weight of the i-th LoRA, and the softmax function normalizes the weights. Thus, the output of the feed-forward neural network is:

$$O_{a_\mu^t} = f_{FFN}(X'_{a_\mu^t} \mid \theta) + \frac{\alpha}{r}\sum_{i=1}^{n} \omega_i L_i(X'_{a_\mu^t}), \quad (4)$$

where $f_{FFN}(\cdot)$ is the feed-forward neural network and α is a constant. Finally, we can get the output of the transformer block, denoted as $H_{a_\mu^t}$:

$$H_{a_\mu^t} = X'_{a_\mu^t} + LN(O_{a_\mu^t}). \quad (5)$$

We impose some constraints on the output of the last transformer block to achieve domain alignment and subspace independence. More details are described in Section 4.3.

4.3 Training Objective

Original Loss: Next Token Prediction $\mathcal{L}_p$ In our implement, $\mathcal{L}_p$ is computed in the same way as the autoregressive loss of the pre-trained language model, which can align the model output with the target text:

$$\mathcal{L}_p = -\sum_{t=1}^{T} \log P_{LM}(y_t \mid x_{<t}, D; \theta), \quad (6)$$

where T represents the sequence length, y_t represents the target token at time step t and $x_{<t}$ represents all input tokens before time step t.

Aspect-Adaptive Loss $\mathcal{L}_{ada}$ In controllable text generation, different aspects typically originate from heterogeneous data sources, leading to domain shifts in the latent space. Specifically, different values of the same attribute are not separated solely along the 'attribute polarity' axis; instead, they are jointly displaced due to confounding factors such as syntax, topic, and tense. If attribute disentanglement is performed directly on this imbalanced space, the model may mistakenly treat domain-specific features as attribute signals, resulting in degraded control accuracy and reduced generalization ability. To address this, we introduce the aspect-adaptive loss, which minimizes the Euclidean distance between the centroids of hidden representations for different attribute values, thereby aligning them within a unified coordinate system:

$$\mathcal{L}_{ada} = \sum_{1\leq t_1<t_2\leq|\mathcal{A}|} \left\| \sum_{i=1}^{|I^{t_1}|} \frac{H^i_{a^{t_1}}}{|I^{t_1}|} - \sum_{j=1}^{|I^{t_2}|} \frac{H^j_{a^{t_2}}}{|I^{t_2}|} \right\|_2, \quad (7)$$

where $\|\cdot\|_2$ represents the Euclidean distance. It mitigates domain discrepancies across corpora, enabling the model to focus on learning pure attribute directions relative to a shared reference point, and lays the foundation for subspace independence.

Attribute-Aware Loss $\mathcal{L}_{awa}$ In controllable text generation, we need to control specific attributes within specific aspects. However, we aim for each attribute distribution to be mutually independent, so that decoding along a specific attribute distribution enables precise controllable text generation. To achieve this, we introduce an attribute-aware loss $\mathcal{L}_{awa}$, which consists of two parts: attribute exclusion loss $\mathcal{L}_e$ and attribute gap loss $\mathcal{L}_g$.

We aim for the distributions of different attributes in the attribute space to be as independent as possible (in the field of controllable text generation, we only control one attribute within an aspect at a time, not multiple attributes simultaneously).

Therefore, we propose the attribute exclusion loss. Consider any two sets of hidden states $H_{a^t_{\mu_1}}$ and $H_{a^t_{\mu_2}}$ with different attributes in aspect A_t, the attribute exclusion loss between them is calculated as follows:

$$\mathcal{L}_e^t = \sum_{1 \leq \mu_1 < \mu_2 \leq |a_t|} \max\left(\gamma - \left\| C_{a^t_{\mu_1}} - C_{a^t_{\mu_2}} \right\|_2, 0\right),$$

$$C_{a^t_{\mu_i}} = \frac{1}{\left|I_{a^t_{\mu_i}}\right|} \sum_{j=1}^{\left|I_{a^t_{\mu_i}}\right|} H^j_{a^t_{\mu_i}}, \tag{8}$$

where γ is a hyperparameter, $C_{a^t_{\mu_i}}$ is the distribution center of the attribute $a^t_{\mu_i}$ and $\left|I_{a^t_{\mu_i}}\right|$ is the number of hidden states in the attribute $a^t_{\mu_i}$. To reduce the difference of the same attribute in the attribute space, we want the distributions within the same attribute to be as cohesive as possible. Therefore, we use attribute gap loss to constrain the distributions within the same attribute to be closer to the center of the distribution of that attribute, which is defined as follows:

$$\mathcal{L}_g^t = \sum_{\mu=1}^{|a^t|} \sum_{j=1}^{\left|I_{a^t_\mu}\right|} \left\| H^j_{a^t_\mu} - C_{a^t_\mu} \right\|_2. \tag{9}$$

So the total attribute-aware loss of all aspects is:

$$\mathcal{L}_{awa} = \sum_{t=1}^{|\mathcal{A}|} \mathcal{L}_e^t + \mathcal{L}_g^t. \tag{10}$$

Through attribute awareness, the model can distinguish differences between attributes and search within specific attribute distributions to achieve precise controllable text generation. As shown in Appendix C.3, we visualize the distributions of the sentiment and topic aspects in the attribute space, where it is evident that the different attribute distributions are as independent as possible, while the distributions of the same attribute are as cohesive as possible.

Our total loss function can be represented as:

$$\mathcal{L} = w_1\mathcal{L}_p + w_2\mathcal{L}_{ada} + w_3\mathcal{L}_{awa}. \tag{11}$$

We freeze the parameters of the pre-trained model and only train multiple LoRA modules and the gating function.

5 Experiments and Results

5.1 Experiment Setup

Datasets. We conduct experiments on three control aspects: *sentiment*, *topic*, and *detoxification*. Following previous work (Gu et al., 2023, 2022), we use the IMDb dataset for 2 sentiments (*positive* and *negative*), the AGNews dataset for 4 topics (*world*, *sports*, *business*, *science*), and the Jigsaw Toxic Comment Classification Challenge dataset for 1 *detoxification*, respectively. Similarly to Discrete (Gu et al., 2022), we randomly sample 10k examples from each dataset to construct the latent space. For text generation, we adapt the 35 prompts provided in PPLM (Dathathri et al., 2019). In the single-attribute controllable text generation setting, we generate 5 sentences for each attribute and each prompt, resulting in a total of $35 \times (2 + 4 + 1) \times 5 = 1225$ sentences.

Automatic Evaluation. Following the setup of (Gu et al., 2023), we automatically evaluate the generation results from three aspects: (1) **Correctness.** We adapt the classifiers fine-tuned by Discrete (Gu et al., 2022) for evaluating *sentiment* and *topic* control, and employ the Google Perspective API for *detoxification*. We calculate the proportion of final sentences that contain the specified attribute. (2) **Perplexity.** We use the Medium version of GPT-2 to evaluate the perplexity of the generated text. (3) **Diversity.** Following (Gu et al., 2023), we compute Dist-1, Dist-2, and Dist-3 to measure the diversity of the generated text.

Human Evaluation. We conduct a human evaluation of the generated texts. For each method, we invite an evaluator to assess the single-attribute texts generated by the model. 225 samples were randomly selected for evaluation, ensuring that each combination of prefix and attribute was covered. The evaluation criteria include the quality of the generated sentences and whether they contain the target attribute. Scores range from 1 to 5, with higher scores indicating better performance.

Model and baselines. We use Qwen2-0.5B-Instruct for experiments. We compare our method to 7 representative and strong baseline: (i) **Weighted Decoding**: **PPLM** (Dathathri et al., 2019) and **GeDi** (Krause et al., 2021) bias the decoding process in generation. (ii) **Optimization in the Language Space**: **Mix&Match** (Mireshghallah et al., 2022) combines pre-trained black-box models using global scoring to achieve desired attributes without fine-tuning. **MUCOCO** (Kumar et al., 2021) formulates controllable generation as continuous optimization with multiple constraints. (iii) **Optimization in the Latent Space**:

Methods	Sentiment↑ (%)			Topic↑ (%)					Detox.↑ (%)	PPL.↓	Dist.-1/2/3↑
	Avg.	Neg.	Pos.	Avg.	W.	S.	B.	T.			
Biasing during Decoding											
PPLM	80.0	97.2	62.7	70.6	74.9	46.5	62.4	98.6	93.2	63.2	31.1 / 70.9 / 85.9
GeDi	82.3	93.9	70.7	83.2	73.4	85.7	75.7	98.0	94.9	81.6	38.1 / 74.0 / 78.4
Optimization in the Language Space											
MUCOCO	75.4	95.5	55.3	73.5	56.9	67.3	72.3	97.5	94.8	381.7	22.5 / 49.9 / 64.3
Mix&Match	82.8	99.2	63.3	75.6	79.5	57.4	69.6	**99.3**	96.9	65.2	31.5 / 74.8 / 88.8
Optimization in the Latent Space											
Con. Prefix	89.5	88.4	90.6	86.7	74.5	85.3	93.5	93.6	93.8	37.7	17.3 / 47.0 / 71.1
LatentOps	91.1	88.3	93.9	69.4	54.3	61.1	72.4	89.6	94.6	58.8	13.5 / 48.3 / 62.8
Discrete	92.5	99.1	85.9	90.4	84.5	95.0	84.6	97.5	90.1	46.2	36.9 / 76.3 / 87.0
Ours	**97.1**	**99.9**	**94.3**	**95.0**	**90.5**	**96.2**	**94.7**	98.7	**97.1**	38.8	32.7 / 78.4 / 94.3

Table 1: Automatic Results on Single-Attribute Control. We control on Sentiment (**Neg**ative and **Pos**itive), Topic (**W**orld, **S**ports, **B**usiness, and Science/**T**echnology), and **Detox**ification independently. In addition, we use Perplexity (PPL) and Distinctness (Dist.) metrics to evaluate the quality of the generated text.

Methods	Avg.↑	Sent.↑	Topic↑	Detox.↑	Fluency↑
GeDi	3.28	2.66	3.40	4.08	2.81
Discrete	3.42	3.28	3.42	3.68	3.47
Ours	**3.97**	**3.98**	**3.72**	**4.21**	**3.68**

Table 2: Human Results on Single-Attribute Control.

Variants (strategies during training)	Avg.	Sent.	Topic	Detox.
Ours (intact)	95.5	94.3	95.0	97.1
w/o gate (Eq. 3)	92.5	92.1	90.6	94.8
w/o multi-LoRA (Eq. 2)	90.3	90.9	88.2	91.8
w/o $\mathcal{L}_{ada}$ (Eq. 7)	87.9	89.0	84.8	89.9
w/o $\mathcal{L}_{awa}$ (Eq. 10)	88.6	88.5	83.0	94.3
w/o $\mathcal{L}_{ada}$ and $\mathcal{L}_{awa}$ (Eqs. 7 and 10)	85.2	86.5	78.5	90.6
Variants (strategies during inference)	**Avg.**	**Sent.**	**Topic**	**Detox.**
Ours (intact)	95.5	94.3	95.0	97.1
w/o sample	94.4	94.8	93.2	95.2

Table 3: The ablation study of different strategies in our framework.

Contrastive Prefix (Qian et al., 2022) enhances the prefixes through contrastive learning. **Discrete** (Gu et al., 2022) constructs attribute-specific latent spaces using discrete samples for direct control. **LatentOPs** (Liu et al., 2022) introduces an efficient sampler based on ordinary differential equations to navigate the latent space for attribute control.

5.2 Main results

Table 1 presents the results of the automatic evaluation for the generation of controllable text with a single attribute, covering the sentiment, topic and detoxification tasks, along with the perplexity and diversity metrics. Traditional decoding bias methods demonstrate basic controllability without requiring model retraining. However, their average sentiment accuracy hovers around 80%, with limited performance on topic and detoxification tasks, and significant fluency degradation: GeDi, for example, reaches a PPL as high as 81.6. In contrast, full-parameter optimization in the language space (e.g., MUCOCO, Mix&Match) is theoretically more flexible, but suffers from severe gradient noise in the high-dimensional space, leading to elevated PPL and unstable topic coverage.

Our proposed framework explicitly disentangles attribute directions in the latent space by inserting multi-LoRA experts at each transformer layer, combined with a gating mechanism and optimized with both aspect-adaptive and attribute-aware losses. This leads to comprehensive improvements across all three metrics. Specifically, our method improves the average sentiment accuracy to 97.1%, a +4.6% gain over the strongest existing latent-space baseline (Discrete); for topic control, it achieves 95.0%, surpassing Discrete by +4.6% and significantly outperforming LatentOps by +25.6%. Our method sets a new benchmark on the detoxification task with 97.1% accuracy—0.2% higher than Mix&Match and well above all latent space baselines (max 94.6%).

Despite achieving precise control, our method maintains strong fluency and diversity. Its PPL is 38.8, slightly higher than the Con. Prefix (37.7) but substantially lower than models relying on language-space optimization. It also scores the highest Dist-2 and Dist-3 scores, indicating that gated LoRA experts enhance domain alignment without causing diversity collapse. In general, Table 1 highlights the effectiveness of our lightweight, attribute-aware latent space disentanglement framework, which delivers state-of-the-art single-attribute control with minimal compromise to fluency and diversity. In addition, we show the

Model	Average	Sent.	Topic	Multi	Length	Keyword	Detox.
GPT-4 (0613)	**84.82**	**91.6**	**93.5**	**70.2**	73.8	**86.2**	**93.60**
GPT-4o (0513)	83.10	91.0	89.7	67.6	**75.6**	82.0	92.67
Qwen2-7B-Instruct	73.16	85.9	91.2	62.4	50.1	58.5	90.87
+ LoRA	80.62	91.0	**93.6**	73.6	62.6	72.1	90.86
+ FFT	81.54	**93.7**	93.2	75.6	62.3	73.8	90.69
+ Ours	**82.78**	91.8	92.7	**77.2**	**63.9**	**79.4**	**91.65**
Qwen2-72B-Instruct	80.30	87.4	91.7	70.5	62.1	76.9	93.21
+ LoRA	83.85	92.3	93.3	75.7	68.7	82.2	90.92
+ FFT	84.78	92.4	**93.5**	76.4	70.4	**85.2**	90.80
+ Ours	**85.32**	**93.1**	92.8	**79.4**	**71.7**	83.4	**91.49**
Llama-3.1-8B-Instruct	75.66	88.5	89.9	64.6	55.9	68.5	86.55
+ LoRA	79.48	90.7	93.9	71.9	58.9	69.9	91.55
+ FFT	79.47	**91.8**	**94.1**	71.7	57.2	71.2	90.84
+ Ours	**80.19**	91.3	93.1	**72.2**	**61.4**	**71.4**	**91.73**
Llama-3.1-70B-Instruct	81.90	89.0	90.9	70.2	71.6	80.6	89.08
+ LoRA	84.14	92.4	**93.2**	**74.8**	69.9	**83.9**	90.64
+ FFT	82.60	88.4	92.7	71.1	69.3	83.3	90.80
+ Ours	**84.61**	**92.8**	92.5	73.2	**72.7**	83.6	**92.86**
Gemma-2-9B-Instruct	71.73	88.4	84.7	59.0	53.9	71.9	72.49
+ LoRA	76.91	89.1	91.0	71.5	50.7	77.5	81.67
+ FFT	78.01	**90.8**	92.8	**74.7**	50.6	73.5	**85.66**
+ Ours	**79.12**	89.4	**92.9**	73.2	**57.5**	**79.3**	82.42

Table 4: Performance of various model variants on different aspects. The model fine-tuned by our framework has the best average accuracy on CoDI-Eval and is ahead of LoRA and FFT in most aspects. The **bold** value indicates the maximum value of each model variant on each aspect.

results of human evaluation for the control of a single attribute in Table 2, which are almost consistent with automatic evaluation.

6 Analysis

Analysis of different strategies. We validate the effectiveness of different strategies during training. Table 3 presents the detailed experimental results. After removing the gating function, our method shows a performance drop across all three datasets. Further removal of the multi-LoRA structure leads to an additional decrease in performance, indicating that both the gating function and the multi-LoRA structure play a positive role in learning features of different attributes. Our proposed losses, $\mathcal{L}_{ada}$ and $\mathcal{L}_{awa}$, influence the performance of the model by enhancing the disentanglement of attributes. Removing the two losses results in a performance drop, and removing both leads to the most significant degradation. During inference, we employ a sampling strategy to generate text. When this sampling strategy is removed, we observe a slight decrease in performance, suggesting that it has a minor impact on the results. In contrast, our proposed architecture and loss functions substantially affect controllable text generation.

Evaluating on CoDI-Eval. We evaluate the performance of our method on CoDI-Eval (Chen et al., 2024) using multiple open-source models and compared it with other approaches based on fine-tuning, including LoRA and FFT. CoDI-Eval covers six aspects and is designed to assess the ability of the model to satisfy multiple attribute requirements simultaneously. The results are shown in Table 4. Our method achieves an average accuracy higher than that of LoRA and FFT in multiple aspects, achieving optimal results in most of them. For Qwen2-7B-Instruct and Llama-3.1-8B-Instruct, we achieve the best scores in average precision, multiaspect, length, keyword and detoxification, with an improvement of up to 1.24% and 0.72% over FFT. For Llama-3.1-70B-Instruct and Gemma-2-9B-Instruct, we outperform LoRA and FFT in three aspects and achieve a higher average accuracy. Furthermore, we outperform GPT-4 for Qwen2-72B-Instruct and achieve top performance in various tasks across all models, demonstrating its superior effectiveness on CoDI-Eval. In Appendix D, we present some cases, including attribute combinations provided in CoDI-Eval and unseen attribute combinations, demonstrating the superiority and generalizability of our approach.

Visualizing the latent space. Figure 3 projects the high-dimensional hidden states produced by our framework onto a two-dimensional plane with t-SNE, revealing the latent distributions of the four AGNews topics (*world*, *sports*, *business*, and *science*). Unlike the heavily overlapping clouds in Figure 1, the clusters in Figure 3 are clearly separated and compact, with minimal cross-attribute intrusion. This pattern confirms that the combination of gated multi-LoRA experts, Aspect-Adaptive Loss, and Attribute-Aware Loss successfully aligns domain-shifted corpora and carves mutually orthogonal subspaces for each attribute. The resulting disentanglement not only validates our theoretical design but also underpins the strong controllability and fluency gains reported in Section 5.2, demonstrating that attribute-specific text can be generated by navigating along clean, interference-free directions in the latent space.

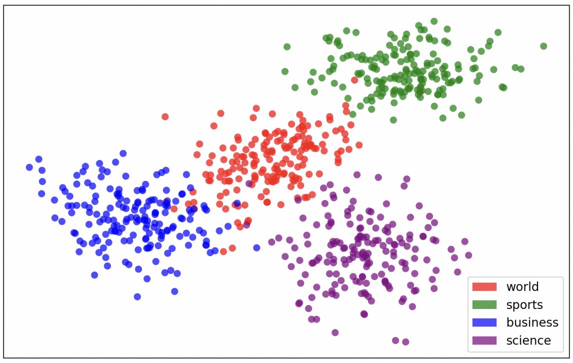

Figure 3: The visualization of t-SNE projection of the latent representations for the AGNews topics.

Time Efficiency. We also compare the inference speed of our method with GeDi and PPLM. As shown in Table 5, our framework outperforms the baselines in terms of inference speed, demonstrating its strong computational efficiency. This indicates that the introduction of multiple LoRA modules and gating functions does not lead to significant latency during inference.

Methods	Inference Speed ↓
Ours	**1.469** (1.00×)
GeDi	1.680 (0.87×)
PPLM	15.553 (0.09×)

Table 5: Inference speed comparisons (second/sample).

Selection of the Number and Rank of LoRA. We examine the effect of varying the number and rank of LoRA modules on CoDI-Eval. As shown in Table 6, our findings reveal that optimal performance is achieved when the number of LoRA modules is set to 8, with each LoRA rank r fixed at 16. Increasing the number of LoRA modules beyond 8 does not lead to further performance gains. Additionally, while increasing the LoRA rank to 32 results in a slight performance improvement, it also doubles the number of training parameters.

LoRAs	LoRA Rank	Trainable Param.	Average
4	16	2.04%	81.58
8	16	4.07%	82.78
16	16	8.14%	82.33
8	8	2.04%	82.10
8	16	4.07%	82.78
8	32	8.14%	82.93

Table 6: Performance of Our framework varies with the number of LoRAs and LoRA rank across all aspects.

7 Conclusion

In this paper, we consider the problem of highly overlapping attribute distributions in the latent space, which makes precise controllability difficult, and propose the Latent Space Attribute Disentanglement framework. This framework introduces multiple LoRA experts within each Transformer block and utilizes a trainable gating mechanism to selectively aggregate the outputs of different experts, thereby explicitly decomposing the latent space into mutually orthogonal attribute subspaces. In addition, we design two complementary objectives: Aspect-Adaptive Loss, which aligns the distribution centers of the same attribute across different aspects, and Attribute-Aware Loss, which compresses intra-attribute representations and enhances subspace independence. Extensive experiments validate the effectiveness of our framework, achieving strong performance across multiple benchmarks.

Limitations

Although our method shows promising results, it also has several limitations. (i) Our approach requires a large amount of data to construct the latent space, making efficient data collection and preprocessing essential. (ii) The framework requires additional training to meet its objectives, which introduces additional overhead. (iii) Since our framework incorporates multiple LoRA modules, increasing the LoRA rank increases the number of trainable parameters, thereby increasing training time and computational resource consumption.

References

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6837–6857, Dublin, Ireland. Association for Computational Linguistics.

Alvin Chan, Ali Madani, Ben Krause, and Nikhil Naik. 2021. Deep extrapolation for attribute-enhanced generation. *Advances in Neural Information Processing Systems*, 34:14084–14096.

Yihan Chen, Benfeng Xu, Quan Wang, Yi Liu, and Zhendong Mao. 2024. Benchmarking large language models on controllable generation under diversified instructions. *Preprint*, arXiv:2401.00690.

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*.

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. *arXiv preprint arXiv:2305.12785*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand. Association for Computational Linguistics.

Yuxi Feng, Xiaoyuan Yi, Xiting Wang, Laks Lakshmanan, V.S., and Xing Xie. 2023. DuNST: Dual noisy self training for semi-supervised controllable text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8760–8785, Toronto, Canada. 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.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. Controllable text generation via probability density estimation in the latent space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12590–12616, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1428, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34:14542–14554.

Vaibhav Kumar, Hana Koorehdavoudi, Masud Moshtaghi, Amita Misra, Ankit Chadha, and Emilio Ferrara. 2023. Controlled text generation with hidden representation transformations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9440–9455, Toronto, Canada. Association for Computational Linguistics.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2022. Composable text controls in latent space with odes. *arXiv preprint arXiv:2208.00638*.

Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Dangyang Chen, and Jixiong Chen. 2023. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5933–5957, Singapore. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generation using energy language models. *arXiv preprint arXiv:2203.13299*.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. *arXiv preprint arXiv:2202.13257*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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.

Hanqing Zhang and Dawei Song. 2022. DisCup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3406, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1022–1040, Toronto, Canada. Association for Computational Linguistics.

Tianqi Zhong, Quan Wang, Jingxuan Han, Yongdong Zhang, and Zhendong Mao. 2023. Air-decoding: Attribute distribution reconstruction for decoding-time controllable text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8233–8248, Singapore. Association for Computational Linguistics.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Details of Hyperparameter Selection

In this section, we describe the hyperparameters used for our framework. In our approach, multiple LoRA modules are applied to all linear layers of the pre-trained model, with the number of LoRA modules n set to 8, the rank r set to 16, the scaling factor α set to 32, and dropout set to 0.1. The gating function is implemented using an embedding layer and a linear layer. The input dimension of the embedding layer is the number of aspects, which is 6, and the output dimension is 64. The input dimension of the linear layer is 64, and the output dimension is 8.

We use 8 NVIDIA A100 80GB GPUs during training and employ `bfloat16` precision to improve training efficiency. The batch size is set to 64. The scaling factors for $\mathcal{L}_p$, $\mathcal{L}_{ada}$, and $\mathcal{L}_{awa}$ are 0.7, 0.2, and 0.1, respectively and hyperparameter γ in $\mathcal{L}_{awa}$ is set to 1. The optimizer used in our experiments is AdamW. For models with a scale below 10B, we set the learning rate to 2×10^{-4} to ensure efficient convergence while maintaining stability during training. For larger models, specifically the 70B and 72B scale, we adjust the learning rate to 1×10^{-5} to accommodate the increased model complexity and prevent issues such as gradient instability or overfitting. For models with fewer than 10B parameters, the number of training epochs is 9, while for models with 70B and 72B parameters, the number of epochs is 3.

During inference, we use specific sampling hyperparameters: max_new_tokens is set to 512, top_p is set to 0.7, and $temperature$ is set to 0.95.

Input to the gating function	Average	Sent.	Topic	Multi	Length	Keyword	Detox.
Aspect Identifier	**82.78**	91.8	**92.7**	**77.2**	**63.9**	79.4	**91.65**
Hidden States	81.23	**92.2**	92.5	72.4	59.2	**80.2**	90.90

Table 7: Compare the impact of different inputs to the gating function on model performance.

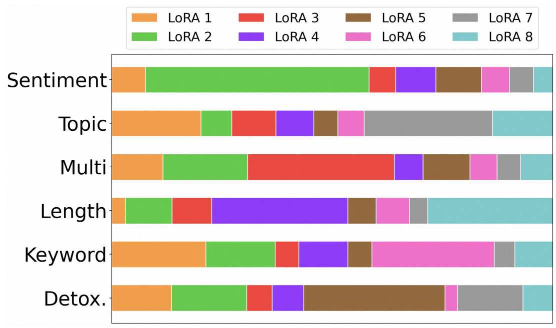

Figure 4: The visualization of LoRA weights for various aspects based on Qwen2-72B-Instruct.

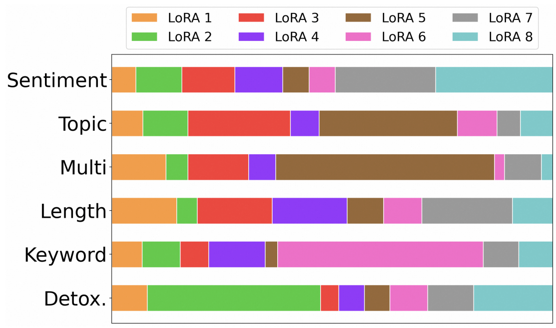

Figure 6: The visualization of LoRA weights for various aspects based on Llama-3.1-70B-Instruct.

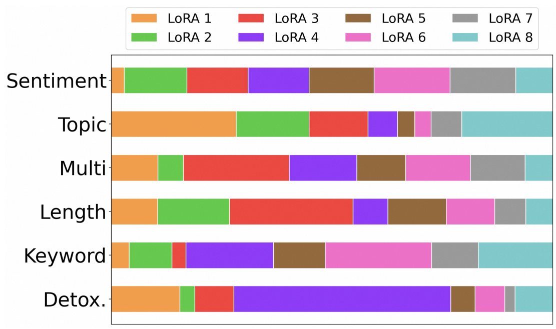

Figure 5: The visualization of LoRA weights for various aspects based on Llama-3.1-8B-Instruct.

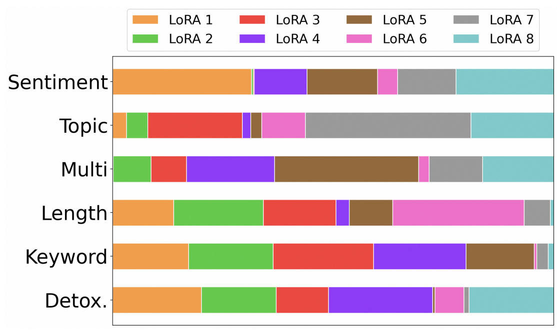

Figure 7: The visualization of LoRA weights for various aspects based on Gemma-2-9B-Instruct.

B Details of CoDI-Eval Setup

B.1 Details of CoDI-Eval

We use CoDI-Eval as the benchmark for our experiments. CoDI-Eval encompasses six controllable text generation tasks: sentiment, topic, multi, length, keyword, and detoxification. The sentiment and topic control tasks require generating text with specified sentiments and topics. The multi-aspect control task involves fulfilling both sentiment and topic requirements simultaneously. For the keyword control task, the generated text must include specified keywords, while the length control task demands controlling the length of the output. Lastly, the detoxification task ensures that the generated text is free of toxic content. Since this benchmark does not provide training data, we synthesized training data using GPT-4: the sentiment control task contains 8,798 samples, the detoxification control task contains 11,753 samples, the keyword control task contains 8,294 samples, the length control task contains 19,238 samples, the topic control task contains 10,599 samples, and the multi-control task contains 12,678 samples, totaling 71,320 training samples. The test data provided by CoDI-Eval, with the detoxification control task containing 4,060 test samples and the other five control tasks, each containing 1,000 test samples. We labeled the training and test data with aspect identifiers numbered 0 to 5, respectively. To ensure reproducibility and facilitate further research, we will release all training data, evaluation scripts, and the fine-tuned model checkpoints. The dataset includes synthesized training samples generated using GPT-4 and test data sourced from CoDI-Eval.

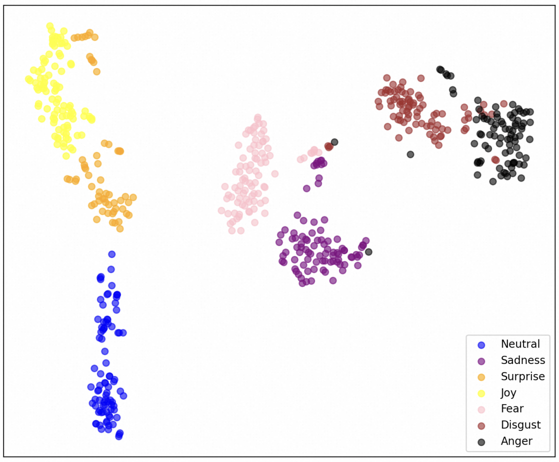

Figure 8: Visualizing the distribution of sentiment attributes.

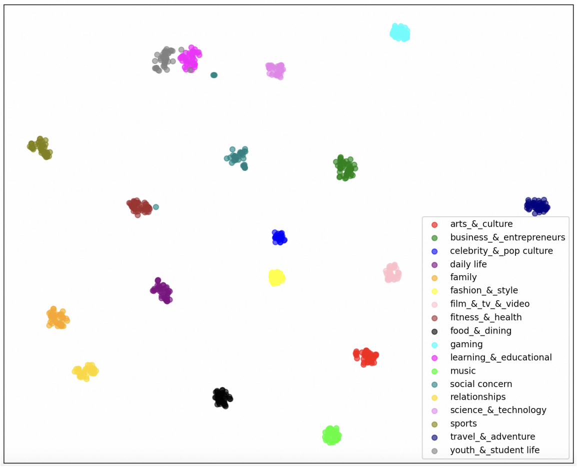

Figure 9: Visualizing the distribution of topic attributes.

B.2 Evaluation Metrics

We use the sentiment[1,2] and topic[3] RoBERTa classifiers on Huggingface to evaluate sentiment, topic and multi aspects. The detoxification aspect is measured by the Google Perspective API[4]. For keyword and length evaluation, we use the following rules: Regarding keywords, we check whether all the required keywords are present in the generated text. If all the keywords are included, it is considered correct. As for length, we evaluate whether the length of the generated text meets the specified label requirements. If it does, it is considered correct.

C More experimental results on CoDI-Eval

C.1 Visualizing the LoRA weights

In this section, we visualize the weights of LoRA based on other models, namely Qwen2-72B-Instruct, Llama-3.1-8B/70B-Instruct, and Gemma-2-9B-Instruct, and the results are shown in Figures 4 to 7. Consistent with the findings in Section 6, the LoRA contribution of each model in the six aspects of Codi-Eval is significantly different, and it can achieve multi-faceted control well.

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[2] https://huggingface.co/j-hartmann/emotion-english-roberta-large

[3] https://huggingface.co/cardiffnlp/tweet-topic-21-multi

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

C.2 Inputs to the gating function

In our implementation, we use the aspect identifier to input the gating function to learn unique parameters for each aspect and dynamically combine LoRA modules. Comparing this to using hidden states as input, as shown in Table 7, we found that the aspect identifier approach is superior. Using hidden states increases parameter complexity and training difficulty due to the need for a separate gating function for each layer, whereas our approach requires only a global gating function.

C.3 Visualizing attribute distributions

To validate that our framework effectively achieves attribute awareness by adjusting attribute distributions to reduce conflicts between attributes, we use t-SNE to visualize the distribution of attributes across two aspects: sentiment and topic. Specifically, we feed responses from the attribute set into the classifier. We average the attribute context embeddings for each token in the classifier to obtain sentence-level representations, which are then paired with sentence-level attribute annotations for analysis. The results are presented in Figure 8 and Figure 9. The figures show that the context embeddings of sentences with different sentiments and topics are separated in the space. This supports the strong attribute-awareness capability of our framework and demonstrates its superior performance in multi-attribute controllability.

C.4 Different data discrepancies

Moreover, we evaluate the performance of LoRA, FFT and our method under varying training data discrepancies. Specifically, we set training data

Data Discrepancy	Method	Average	Sent.	Topic	Multi	Length	Keyword	Detox.
Sent./Keyword/Multi	LoRA	79.31	90.9	91.5	72.4	63.8	66.5	90.78
	FFT	79.08	**92.9**	**94.1**	**73.8**	62.3	60.7	90.69
	Ours	**80.50**	92.8	93.2	72.2	**64.8**	**68.7**	**91.32**
Sent./Length/Keyword	LoRA	78.15	**93.8**	93.1	74.8	**50.4**	66.6	90.24
	FFT	78.11	92.9	93.1	76.5	47.2	68.4	90.56
	Ours	**78.72**	92.4	**93.4**	**76.5**	48.6	**70.3**	**91.14**
Multi/Keyword/Detox.	LoRA	79.14	90.8	92.1	72.9	61.9	66.6	90.58
	FFT	79.24	**93.1**	**93.8**	72.4	61.0	64.8	90.39
	Ours	**80.59**	92.0	93.0	**73.9**	**63.7**	**70.2**	**90.73**

Table 8: Performance of Qwen2-7B-Instruct on different metrics under various data discrepancy settings. ‘Data Discrepancy’ means that, for each specific aspect, only 1,000 samples are selected for training. For example, ‘Sent./Keyword/Multi’ indicates that 1,000 samples are randomly selected from each Sentiment, Keyword, and Multi aspects. In contrast, the Topic, Length, and Detoxification aspects use complete samples for training.

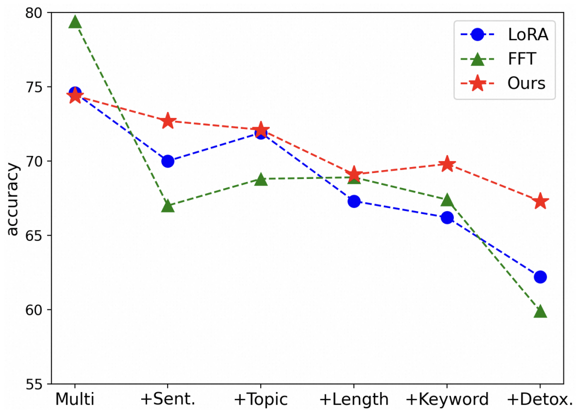

Figure 10: Comparison of LoRA, FFT and our method in alleviating knowledge forgetting. With the injection of new knowledge, we measure the performance of these three methods on multi-aspect task. The performance of our method decreases the least, proving that our method is more robust to knowledge forgetting.

for any three aspects to 1,000 samples each while using the complete training data for the remaining three aspects. The results are presented in Table 8. We observe that, under conditions of imbalanced training data, our method consistently achieves the highest average accuracy and maintains superior performance across most aspects, with at most a 1.42% significant improvement over existing baselines. This demonstrates the adaptability of our approach to varying data discrepancies.

C.5 Alleviating knowledge forgetting

Alleviating knowledge forgetting is a significant research topic in the field of natural language processing (Huang et al., 2024; Li et al., 2024; Dou et al., 2024). To evaluate the ability of out method to alleviate knowledge forgetting, we conduct the following experiment: First, we fine-tune Qwen2-7B-Instruct on the multi-aspect task using LoRA, FFT and our method respectively, recording their performance. Subsequently, we iteratively fine-tune each model on a new aspect and test their performance on the original multi-aspect task. The results, shown in Figure 10, indicate that as more knowledge is injected, the performance of LoRA and FFT significantly deteriorates on the multi-aspect task, whereas our method experiences only a slight decrease. This demonstrates that our method is more robust in terms of knowledge forgetting.

Compare with Independent LoRA per Aspect. To further evaluate the effectiveness of our framework, we compare it against a variant of our framework where each aspect is trained with a separate LoRA module, rather than integrating multiple aspects within a unified model. Specifically, we train independent LoRA modules for each of the six aspects in CoDI-Eval and evaluate their performance on the benchmark. Table 9 presents the comparison results. The results demonstrate that our unified LoRA framework consistently outperforms the independent LoRA approach, particularly in multi-aspect controllability, where our framework achieves a 3.0% improvement. This suggests that dynamically adjusting LoRA combinations through our gating mechanism enables better multi-attribute interaction while maintaining control precision.

Compare different routing strategies. In our framework, tokens pass through all LoRA modules. To evaluate the impact of different routing strategies, we compare it with the Top-2 routing strategy and the experimental results are shown in Table 10. We found that when the token passes through all

Model	Average	Sent.	Topic	Multi	Length	Keyword	Detox.
Independent LoRA per Aspect	81.22	91.3	**93.0**	74.2	62.3	75.8	90.72
Our framework	**82.78**	**91.8**	92.7	**77.2**	**63.9**	**79.4**	**91.65**

Table 9: Comparison of our implementation with independent LoRA per aspect.

Routing Strategies	Average	Sent.	Topic	Multi	Length	Keyword	Detox.
All LoRA Modules	**82.78**	**91.8**	**92.7**	**77.2**	**63.9**	79.4	91.65
Top-2 LoRA Modules	81.94	88.5	92.4	74.9	62.2	**80.2**	**93.52**

Table 10: Compare the effect of each token passing through all LoRA modules and the Top-2 LoRA modules.

Model	w/o $\mathcal{L}_{ada}$ and $\mathcal{L}_{awa}$	w/o $\mathcal{L}_{awa}$	w/o $\mathcal{L}_{ada}$	Ours
Qwen2-7B-Instruct	81.83	82.11	82.42	**82.78**
Gemma-2-9B-Instruct	77.48	77.55	77.82	**79.12**

Table 11: Evaluation average results for loss ablation on Qwen2-7B-Instruct and Gemma-2-9B-Instruct.

LoRA modules, the model performs better and outperforms the Top-2 routing strategy in most aspects. This is because there is synergy between different tasks, and using more LoRA modules allows for capturing more information, thereby achieving better control effects.

Ablation on $\mathcal{L}_{ada}$ and $\mathcal{L}_{awa}$. We conduct ablation experiments on losses based on Qwen2-7B-Instruct and Gemma-2-9B-Instruct to verify the impact of each loss on the average result, and the results are shown in Table 11. In Table 11, we compare four different settings of our framework, and "w/o $\mathcal{L}_{ada}$ and $\mathcal{L}_{awa}$" is a variant without aspect adaptation and attribute awareness, which performs the worst. In addition, $\mathcal{L}_{awa}$ has a higher impact on the average result than $\mathcal{L}_{ada}$. With all enhancements included, our framework performs best.

D Case Study

We show the cases of LoRA, FFT, and our method on CoDI-Eval in Table 12, and the cases of attribute combinations not seen in CoDI-Eval in Table 13.

Constraints	Method	Generation Result
Surprise	LoRA	In the heart of an ancient forest, where sunlight barely touched the ground, there lay a hidden glade known only to the creatures of the woods. One crisp autumn morning, a curious fox named Faelan stumbled upon this untouched sanctuary.
	FFT	Wow, did you really just win the lottery? That's unbelievable!
	Ours	I can hardly believe my eyes! The transformation is incredible, almost magical. The space looks completely different and better than I ever imagined. Great job!
Fear	LoRA	As the clock struck midnight, the once bustling city fell eerily silent. Shadows danced in the flickering streetlights, and a cold breeze whispered through empty alleyways. I shivered as footsteps echoed in the distance, slowly approaching.
	FFT	I'm really worried about the upcoming test. I haven't been able to fully prepare and I'm scared of failing. Everyone else seems to be more confident, which makes me feel even worse. I hope I can do better than last time, but I'm just so nervous.
	Ours	I can feel my heart racing as I walk through the darkened streets. Every shadow seems to hide a threat, and every sound feels like it could be a warning sign. The fear of something unknown looms over me, causing every step to be cautious and deliberate. I know I should feel safe in the light of the moon, but the eerie silence of the night amplifies my deepest fears.
Disgust	LoRA	I can't believe I have to go through this again. Every time I think about having to deal with such incompetence, my stomach turns. It's absolutely infuriating how these people manage to mess everything up so badly, and yet they think they're doing a good job.
	FFT	The 15-year-old boy was charged with attempted murder after he stabbed his classmate in the buttocks with a pen because he believed it would make him popular. The victim is expected to recover, but the boy faces up to 15 years in juvenile detention. The boy's attorney argued that his client was merely trying to emulate his popular classmates by acting out during a prank gone wrong. However, the judge ruled that the severity of the attack was not justified by any intent to gain popularity, and ordered the boy to face trial as an adult.
	Ours	I am utterly appalled and repulsed by the sheer negligence and messiness that surrounds us. It's an absolute disgrace to witness such a chaotic and unkempt situation. The disregard for order and cleanliness is truly revolting.
Fashion&Style	LoRA	Fashion is not just about wearing trendy clothes, it's about expressing oneself and showcasing individuality through the art of styling. Whether it's through bold colors, innovative textures, or unique patterns, each outfit tells a story and reflects personal tastes and preferences. Style goes beyond mere aesthetics; it's an expression of identity and creativity.
	FFT	Embrace the art of fashion, where each stitch tells a story and every fabric sings a melody of style.
	Ours	Staying ahead in the ever-evolving world of fashion requires not only a keen sense of trends but also a personal touch that makes your style unique and expressive.
Family	LoRA	Hey! Just wanted to check in and see how everyone is doing. Missing all of you and can't wait for our next catch-up session. Let's plan a fun family outing soon! Take good care and miss those smiles. Love you all.
	FFT	Hey everyone! Just wanted to check in and see how everyone is doing. It's been too long since we last caught up, and I'd love to hear all the updates. Maybe we can plan a virtual call or a get-together soon? Let me know what you think. Missing all of you and your amazing stories!
	Ours	Hey everyone, just wanted to share how much I appreciate all of you. Family is the best and I feel so lucky to have such amazing supporters in my life. Let's make sure we stay in close touch and keep the love and laughter flowing. Big hugs to all!
Daily Life	LoRA	One notable piece of literature that explores the theme of everyday life is "To Kill a Mocking-bird" by Harper Lee. This novel provides a poignant portrayal of daily life in a Southern town during the 1930s, delving into issues of racial injustice and moral growth through the eyes of the young protagonist, Scout Finch.
	FFT	If you're looking for something more general that encapsulates the essence of daily routines and interactions, you might consider "Babette's Feast" by Isak Dinesen. This story delves into the lives of a small group of Danish people over many years, highlighting how their simple lives intersect and transform through shared experiences and personal growth. Both of these pieces offer insightful views into the intricacies of everyday life, showcasing how ordinary moments can weave together to form the tapestry of our existence.
	Ours	Daily routines offer a structured framework that helps organize and manage our time efficiently. These routines often include consistent wake-up times, meals, commuting, work, exercise, and leisure activities. Establishing and maintaining healthy routines can significantly enhance productivity, reduce stress, and foster a sense of stability. Whether following a strict schedule or adopting a more flexible approach, daily rituals bring order to our lives, boost focus, and empower us to navigate the demands of everyday life with greater ease and confidence.

Constraints	Method	Generation Result
Surprise Health	LoRA	Did you know? Simply by taking just 30 minutes to exercise each day can dramatically boost your energy levels, improve your mood, and significantly reduce your risk of chronic diseases! Don't wait any longer to transform your health – the time to start is now!
	FFT	Get ready to shock your system into shape! Discover the latest trend that's taking the health world by storm, promising transformative results in just weeks. Stay tuned for an exclusive sneak peek!
	Ours	Wow, who knew that simply swapping your morning coffee for green tea could make such a huge difference in your fitness journey? Just last week, I read about a study that revealed drinking green tea increases metabolism and boosts fat burning! Curious, I decided to give it a try. Four days in, I stepped on the scale and—wow, a full 5 pounds lighter! I'm stunned and excited to see where this change in habit will lead. Green tea, you've surprised me big time!
Anger Food	LoRA	I can't believe how utterly disappointing this meal turned out to be! When I ordered it, I had high hopes, expecting a delightful feast that would satisfy my palate and energize my day. But what was served was nothing short of a culinary disaster.
	FFT	Dining out is supposed to be an enjoyable experience, but it can quickly turn sour when the service is subpar. It's infuriating when you're greeted with indifference and your requests are met with apathy. The whole point of going out to eat is to relax and enjoy a good meal, not to deal with rude or unattentive staff.
	Ours	I can't even begin to express how infuriating it is to have dinner ruined by inconsiderate individuals who have absolutely no regard for polite manners or the effort put in by hardworking chefs. It's exasperating to see people chattering away, elbowing each other across the table, or blatantly ignoring the obviously uncomfortable glances of their neighbors.
Neutral Culture	LoRA	The Japanese tea ceremony, or Chanoyu, is a centuries-old ritual rooted in harmony, respect, purity, and tranquility. Introduced from China by Buddhist monks in the 9th century, it evolved to reflect Japanese culture and philosophy. This artful practice involves precise tea preparation, room decoration, and host-guest interactions, fostering mindfulness and connection. The Japanese tea ceremony serves as a reflection of the country's rich cultural heritage and continues to be an integral part of Japanese social and cultural life today.
	FFT	In Japan, it is common to find shops, especially those traditional ones, closing their doors by late in the evening. This is because the Japanese typically prefer to shop during the day, and many establishments adhere to regular business hours that reflect this. In fact, many cities in Japan have laws regulating store hours to ensure that streets are not too crowded, maintaining a peaceful atmosphere in the evenings. This tradition underscores the value the Japanese place on maintaining a balanced lifestyle that separates work from personal time, promoting a serene evening environment.
	Ours	The Japanese tradition of "chado" or Japanese tea ceremony, is a refined and ritualistic way of preparing and serving tea. Dating back to the 9th century, it is considered a spiritual art form that emphasizes mindfulness, respect, and purity. During a tea ceremony, participants engage in a series of meticulous steps involving the preparation of matcha tea, a powdered green tea. The atmosphere is quietly serene, and the surroundings are often adorned with delicate art pieces which enhance the aesthetic appeal. This cultural practice not only serves the purpose of enjoying tea but also fosters a meditative and reflective experience for both the participants and the host.
At Most 10 Words	LoRA	She savored the warm, spiced aroma of cinnamon rolling through the air. (12 words)
	FFT	Exploring new cities enriches life, offering unique experiences and broadening perspectives. (11 words)
	Ours	The cat slept peacefully on the windowsill. (7 words)
Around 32 to 40 Words	LoRA	Autumn brings a vibrant change to the landscape, with leaves transforming into a mosaic of reds, oranges, and yellows, creating a serene and picturesque scene that delights both the eyes and soul. (32 words)
	FFT	Beneath the silvery moon, whispers danced through the ancient trees, each leaf murmuring secrets of old, weaving tales of forgotten times where magic and reality blurred at the forest's edge. (30 words)
	Ours	Exploring new hobbies can significantly enhance one's creativity and mental well-being. Whether it's painting, cooking, or learning a musical instrument, each activity offers a unique outlet for stress and a way to express oneself. (34 words)

Constraints	Method	Generation Result
Equal to 60 Words	LoRA	Exploring the art of precise language demands meticulous attention to detail, ensuring each word serves a purpose. Crafting sentences with exactness mirrors the meticulousness of an artist, where every element is deliberate and impactful. This skill not only enhances clarity but also enriches the reader's experience, offering a concise yet profound narrative that resonates on multiple levels. (57 words)
	FFT	The art of crafting concise prose is both challenging and rewarding. Every word must be chosen with care, ensuring it contributes meaningfully to the overall message. This exercise in brevity helps sharpen one's thinking and enhances the ability to communicate effectively. Precisely measured in length, the text not only fits a specific format but also tests the writer's skill in being succinct yet expressive. (64 words)
	Ours	The serene landscape beckoned, with lush green fields stretching towards the horizon. A lone tree stood majestically, offering shade and solace. Birds chirped melodiously, creating a symphony of natural sounds. This picturesque scene was a perfect escape from the hustle and bustle of daily life, invoking peace and tranquility within the observer's soul. Such moments are precious and truly cherishable. (60 words)
"clothe" "mother" "toddler"	LoRA	The mother quickly clothed the toddler in a warm jacket before heading out to play in the snow.
	FFT	The mother dressed her toddler in a cute little clothe that matched her vibrant personality.
	Ours	While shopping at the store, my mother noticed a colorful toddler in the aisle trying on various clothes, sparking a fun afternoon of play pretend.
"demonstrate" "neck" "tie"	LoRA	At the annual company gala, every team member looked sharp in formal attire, meticulously grooming their appearances to impress. John, always keen on details, decided to demonstrate his newfound interest in fashion by choosing a silk tie with a subtle pattern, which complemented his suit perfectly. He also made sure to adjust the collar and of his shirt multiple times to ensure he was camera-ready for the speeches and photographs that would follow. (lack neck)
	FFT	During his presentation, James decided to demonstrate the proper technique for tying a knot. He carefully adjusted his necktie, ensuring each fold was precise and symmetrical as he showed the audience how to achieve a perfect knot. The room watched attentively, eager to learn this skill that would elevate their appearance. As James finished his demonstration, he adjusted his own tie one last time before taking his seat, satisfied with the clarity and gracefulness of his display. (lack neck)
	Ours	During his first day at the new job, Jacob wanted to make a strong impression with his professionalism. He decided to demonstrate his commitment by choosing a sharp, classic blue suit. He carefully adjusted the tie around his neck, ensuring it was symmetrical and presented a polished look. Jacob knew that attention to detail in his attire could set the tone for his entire career.
"dance" "music" "perform" "room"	LoRA	The grand room was transformed for the evening's performance, draped in twinkling lights and adorned with eager spectators. As the music began to fill the air, a mesmerizing dance commenced, showcasing the artist's mastery over their craft. Each note played in perfect harmony with the movements, captivating everyone in the space and creating an electric atmosphere of artistry and entertainment. Performers and audience alike were swept up in the joy and creativity unfolding before them. (lack perform)
	FFT	In the heart of the bustling city, the local community center decided to host a vibrant evening of cultural enrichment. They arranged for a spectacular dance performance in the main room, where aspiring dancers and seasoned performers alike would take the stage to showcase their unique talents. The room was filled with anticipation, as the sounds of compelling music echoed through the space, setting the tone for an unforgettable night. This event not only celebrated the diverse artistic abilities within the community but also fostered a spirit of unity and joy among the attendees. (lack perform)
	Ours	The room buzzed with excitement as the students prepared to perform their carefully choreographed dance. The music filled the air, elevating the spirits of everyone present. As the first notes played, their hard work and dedication were brought to life through their graceful movements and impeccable timing. It was an evening where artistry flourished, and the walls of the classroom turned into a stage, hosting a spectacle of talent and passion.

Constraints	Method	Generation Result
Nontoxicity	LoRA	I appreciate the feedback, Eley. Let's definitely find time to discuss this further and explore different perspectives on the topic. It's always great to engage in thoughtful conversations that can enhance our understanding.
	FFT	I'm really sorry about my previous remarks, Eley. I didn't mean to offend anyone and I'm open to discussing this further if you'd like to clear things up. Let's find a time that works for you to talk things through.
	Ours	Meminist shirts are just making fun of feminism because the people who wear them don't give a fair chance to understand the genuine concerns and goals of the feminist movement. It's important to engage in open and respectful dialogues that can help bridge the gap between different perspectives and promote a more inclusive society for all.
Nontoxicity	LoRA	The male fled the scene on a bicycle, but not before he shouted "I'll get you next time!" His tone was filled with a mix of frustration and defiance as he pedaled away swiftly.
	FFT	On Friday, Trump called these people sons of hardworking Americans who deserve to have their voices heard and their concerns addressed. He emphasized the importance of listening to all citizens and working together to find solutions that benefit everyone in society.
	Ours	She saw the curve of her hips and the shape of her waist reflected back at her in the mirror. With a smile, she admired how her outfit complemented her figure, enhancing her natural beauty. Today, she felt confident and ready to take on the challenges ahead. As she left the house, she carried with her not just her belongings, but also a positive mindset, eager to make the most out of every opportunity that came her way.
Nontoxicity	LoRA	I'd happily do the jail time if I could get just one solid punch in to the face of someone who truly deserves it, but violence isn't the answer. Instead, I choose to channel my anger and frustration into positive actions that can create real change and make a difference in the world. By advocating for justice and understanding, we can work towards a more equitable society where everyone has the opportunity to thrive.
	FFT	The Miami Herald reports Mohamed T. had a note in his pocket with phrases in both Arabic and German including "I want to kiss a flower that blooms only at midnight under the full moon".
	Ours	He looked me directly in the eye and drunkenly said, "I'm gonna shake your hand, because tonight has been amazing!" He then proceeded to extend his hand with a wide, enthusiastic smile, ready to seal our new friendship with a firm handshake.

Table 12: Example cases of generated sentences in the CoDI-Eval test set. The above cases are generated based on Qwen2-7B-Instruct. Red text highlights content related to sentiment. Blue text highlights content related to topic. Olive text highlights content related to length. Purple text highlights content related to keywords.

Constraints	Method	Generation Result
sports "playground" "breakfast" "sky"	LoRA	In the heart of the bustling city stood an old, weathered playground, a place where children's laughter echoed under the clear blue sky. One particular sunny day, the air crisp and inviting, the children decided to have a friendly soccer match. They brought along a homemade breakfast of sandwiches and fresh fruits, spreading a checkered cloth on the picnic table under the shade of an old oak tree. As they ate, their chatter filled the air, weaving stories and dreams. The game started, and the children played with an enthusiasm that only a playground can inspire. With every goal, their joy under the vast sky grew, creating memories that would last a lifetime.
	FFT	Under the bright morning sky, families gathered at the playground, enjoying a quick breakfast before the games began. Kids raced from the swings to the field, laughter echoing as parents cheered them on, coffee cups in hand, ready for a day of friendly competition.
	Ours	As the early morning sky turned a brilliant shade of blue, the playground buzzed with excitement. Families gathered for breakfast under the rising sun, kids munching on snacks before they rushed off to play. A friendly soccer match soon broke out, filling the air with cheers as everyone enjoyed a perfect sports-filled morning.
Around 50 to 70 words "singer" "money" "company"	LoRA	A renowned singer recently invested in a new tech company that specializes in music streaming services. The goal is to generate more revenue and connect directly with fans without intermediaries taking a large cut of his money. This venture might open new pathways for artists to earn substantially, influencing the industry's revenue models. (53 words)
	FFT	The rising star, now a renowned singer, decided to invest in a new tech company. With her increasing fortune, she aimed to revolutionize the music industry by integrating cutting-edge technology to enhance artists' performances. This strategic venture promised not only to boost her career but also to transform the way music is experienced globally. (lack money, 54 words)
	Ours	The young singer signed a deal with a music company, hoping it would be her big break. She knew that fame could bring not only recognition but also money to support her dreams. Despite the challenges in the industry, she stayed focused, writing songs late into the night. The company promised promotions, and with every note she sang, she felt her future growing brighter. (64 words)
positive Around 50 to 70 words "beautiful" "success"	LoRA	Embrace each day with optimism, for every moment is a chance to create beautiful memories. Whether small victories or significant success, each step forward enriches our journey. Cherish these moments, for they beautifully weave the tapestry of our lives, full of joy and accomplishment. Keep striving and remember, every success is a spark of beauty in the endless horizon of our endeavors. (63 words)
	FFT	Achieving success is truly beautiful. It's a feeling that fills you with pride and joy, giving you a sense of fulfillment. Whether it's personal or professional milestones, each victory is a stepping stone to greater heights. Let's cherish these moments and continue to strive for happiness and success in our endeavors. Remember, every small step forward is a beautiful part of your journey. (63 words)
	Ours	The day was beautiful, with the sun shining brightly and a gentle breeze lifting everyone's spirits. Jane felt a sense of success as she completed her project, something she'd worked on tirelessly for weeks. Her friends gathered to celebrate, reminding her that hard work truly pays off. It was a moment to savor, filled with joy, laughter, and a beautiful sense of accomplishment. (63 words)

Table 13: Examples of multi-aspect generation involving combinations of aspects not seen in CoDI-Eval. The above cases are generated based on Llama-3.1-8B-Instruct.