C^{3}TG: Conflict-aware, Composite, and Collaborative Controlled Text Generation

Yu Li[1], Zhe Yang[2], Yi Huang[2,3*], Xin Liu[4], Guilin Qi[1]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]JIUTIAN Research, China Mobile, China
[3]Department of Computer Science and Technology, Tsinghua University, China
[4]UniSA STEM, University of South Australia, Adelaide, Australia
{yuli11, gqi}@seu.edu.cn, {yangzhe, huangyi}@cmjt.chinamobile.com, xin.liu@mymail.unisa.edu.au

Abstract

Recent advancements in large language models (LLMs) have demonstrated remarkable text generation capabilities. However, controlling specific attributes of generated text remains challenging without architectural modifications or extensive fine-tuning. Current methods typically toggle a single, basic attribute but struggle with precise multi-attribute control. In scenarios where attribute requirements conflict, existing methods lack coordination mechanisms, causing interference between desired attributes. Furthermore, these methods fail to incorporate iterative optimization processes in the controlled generation pipeline. To address these limitations, we propose **C**onflict-aware, **C**omposite, and **C**ollaborative Controlled Text Generation (C^{3}TG), a two-phase framework for fine-grained, multi-dimensional text attribute control. During generation, C^{3}TG selectively pairs the LLM with the required attribute classifiers from the 17 available dimensions and employs weighted KL-divergence to adjust token probabilities. The optimization phase then leverages an energy function combining classifier scores and penalty terms to resolve attribute conflicts through iterative feedback, enabling precise control over multiple dimensions simultaneously while preserving natural text flow. Experiments show that C^{3}TG significantly outperforms baselines across multiple metrics including attribute accuracy, linguistic fluency, and output diversity, while simultaneously reducing toxicity. These results establish C^{3}TG as an effective and flexible solution for multi-dimensional text attribute control that requires no costly model modifications.

Extended version — https://arxiv.org/abs/2511.09292

Introduction

Recent advancements in large language models (LLMs) have revolutionized text generation with their remarkable capabilities (Ouyang et al. 2022; Min et al. 2024; Li et al. 2024). However, precisely controlling fine-grained textual attributes—such as emotion, style, or topic—remains challenging without architectural modifications or extensive fine-tuning (Kang and Hovy 2021). This goal, known as Controlled Text Generation (CTG), demands techniques that can dynamically modulate model outputs while preserving overall text quality. However, the complexity increases substantially when multiple attributes must be satisfied simultaneously (Liu et al. 2024; Cao et al. 2024b), as these attributes may exhibit overlapping or conflicting characteristics: adjusting one attribute can attenuate or amplify others (Yang et al. 2025; Cao et al. 2024a). Moreover, current controlled text generation frameworks typically lack mechanisms for iterative feedback refinement (Son and Lee 2024).

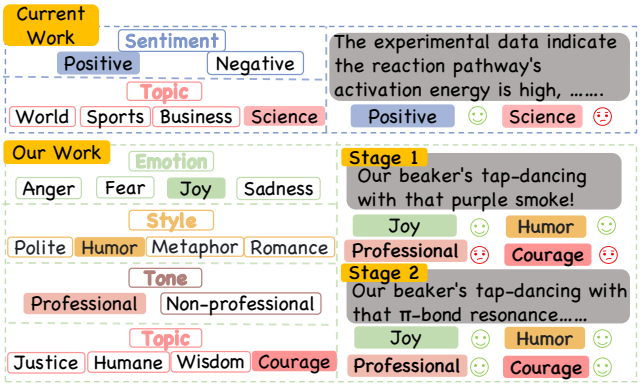

Figure 1: Unlike existing methods that control few attributes and handle conflicts poorly, C^{3}TG offers broad attribute control and resolves conflicts effectively.

Existing controlled text generation methodologies can be systematically categorized into two principal methods. The first method directly modulates the language model's decoding distribution (Pynadath and Zhang 2025; Xu et al. 2024). For instance, PPLM (Dathathri et al. 2020) manipulates hidden state gradients during generation to bias text toward target attributes, while GeDi (Krause et al. 2021) employs a generative discriminator to estimate attribute probabilities for candidate tokens and incorporates this information into the decoding process. Other research has explored latent space energy functions for control, including MacLaSa (Ding et al. 2023), COLD (Qin et al. 2022), and BOLT (Liu, Khalifa, and Wang 2023), which construct energy functions in latent spaces or over contiguous soft sequences, combining them with gradient sampling or adjustable bias mechanisms to achieve concurrent multi-attribute control while maintaining linguistic fluency (Hal-

*Corresponding author.

linan et al. 2023). The second method implements indirect control strategies, such as prompting (Brown et al. 2020) and fine-tuning (Ouyang et al. 2022; Wang and Demberg 2024a). The former augments inputs with natural language instructions to guide generation, offering a concise method that remains difficult to calibrate regarding attribute intensity (Wang and Sha 2023). The latter retrains models on attribute-labeled data, enhancing sensitivity to specific attributes but often requiring substantial computational resources and lacking flexibility (Ma et al. 2024). In summary, current methods exhibit several significant limitations (Liang et al. 2024b). First, existing methods focus on regulation of individual or simplistic attributes, lacking the capability for fine-grained control over complex multi-attribute combinations. Second, when multiple attributes coexist with potential conflicts, current techniques demonstrate insufficient mechanisms for resolving attribute interference. Finally, these methods fail to support progressive text refinement through iterative optimization processes.

To address these limitations, we propose C^{3}TG (**C**onflict-aware, **C**omposite, and **C**ollaborative Controlled Text Generation), a framework that effectively mitigates attribute conflicts by coordinating an LLM with compact auxiliary models. C^{3}TG pairs a powerful LLM with lightweight BERT-based classifiers in a two-phase method. During generation, it integrates attribute-specific probability distributions via weighted KL divergence terms, ensuring each token selection reflects all target attributes. In the optimization phase, C^{3}TG constructs an energy function combining classifier scores with conflict penalty terms to guide iterative text refinement, progressively improving attribute alignment while maintaining fluency. This collaboration between "large" generators and "small" evaluators enables flexible, fine-grained, conflict-aware control without costly retraining or architectural modifications. Extensive story-generation experiments show that C^{3}TG preserves fluency and diversity while controlling multiple attributes. Additional evaluations on toxicity datasets show significant reductions in harmful content generation. Moreover, our dedicated "opposites and conflicts" experiments highlight C^{3}TG's superior capacity to handle overlapping and conflicting attribute requirements. Our contributions are summarized as follows:

- We propose C^{3}TG, pairing LLMs with specialized BERT classifiers across 17 subcategories of emotion, style, tone, and topic, significantly expanding attribute control capabilities while mitigating toxic content generation.
- We integrate a weighted KL divergence for attribute distribution fusion during generation and implement a composite energy function (classifier scores plus conflict penalties) during optimization, enabling flexible real-time multi-attribute control that outperforms existing static methods in both stability and toxicity suppression.
- Empirical evaluations on ROCStories and WritingPrompts datasets demonstrate C^{3}TG's superiority over mainstream baselines in attribute accuracy, fluency, and diversity—findings further validated through comprehensive human evaluation.

Related Work

Controlled Text Generation via Decoding

Recent research has advanced CTG by manipulating decoding distributions (Madaan et al. 2021). These methods include: gradient-based methods like PPLM (Dathathri et al. 2020), COLD (Qin et al. 2022), and BOLT (Liu, Khalifa, and Wang 2023) that adjust hidden states during inference; latent-space techniques such as MacLaSa (Ding et al. 2023) and LatentOps (Liu et al. 2023) operating in continuous attribute spaces (Zhu et al. 2024); constraint frameworks like MUCOCO (Kumar et al. 2021) and PriorControl (Gu et al. 2023); and score-mixing methods including Mix&Match (Mireshghallah, Goyal, and Berg-Kirkpatrick 2022) and Palette (Yang et al. 2025). While these methods preserve the base model and enable fine-grained control, their heavy-handed decoding interventions add implementation complexity, frequently undermine fluency and coherence, and—crucially—provide no feedback-driven refinement, a gap that becomes acute when resolving conflicts in multi-attribute scenarios (Zhong et al. 2023).

Controlled Text Generation via Indirect Strategies

Indirect strategies guide pretrained models without internal modifications. Prompt-based methods insert control signals, offering parameter-free solutions but providing only coarse control and limited efficacy with conflicting attributes (Jie et al. 2024; Wang and Demberg 2024b). Fine-tuning methods—including prefix-based adaptation (Li and Liang 2021), ProSwitch (Zong et al. 2024), and prompt tuning (Qian et al. 2022)—enable precise control but require substantial annotation and risk overfitting (Yang, Ma, and Cheng 2024). Plug-and-play frameworks such as DATG (Liang et al. 2024a) and LiFi (Shi, Cai, and Yang 2024) achieve lightweight control without modifying the base LLM, but both depend on quality labeled data to train their attribute classifiers (Shulev and Sima'an 2024). Despite methodological diversity, these methods universally struggle with two fundamental challenges: resolving multi-attribute conflicts and enabling feedback-driven refinement to enhance attribute alignment without compromising coherence (Yu et al. 2024).

Problem Formulation

Let $\mathcal{V}$ denote the vocabulary and $x = (x_1, \ldots, x_T) \in \mathcal{V}^T$ represent a sequence of tokens. We consider the controlled text generation problem over an attribute set $\mathcal{A} = \{A_1, \ldots, A_n\}$ with corresponding target intensities $T_i \in [0, 1]$. Formally, we aim to solve the constrained optimization problem:

$$\begin{aligned} &\max_{x} \ \mathrm{Fluency}(x) \\ &\text{s.t. } C_{A_i}(x) \approx T_i, \quad \forall\, i \in \{1, \ldots, n\}. \end{aligned} \tag{1}$$

where $C_{A_i}(x)$ quantifies the intensity of attribute A_i in sequence x.

Our framework accepts as input: (1) an initial context x_{init} provided by the user, and (2) a set of desired attribute-intensity pairs $\{(A_i, T_i)\}_{i=1}^{n}$, such as *Emotion: Joy (0.9),*

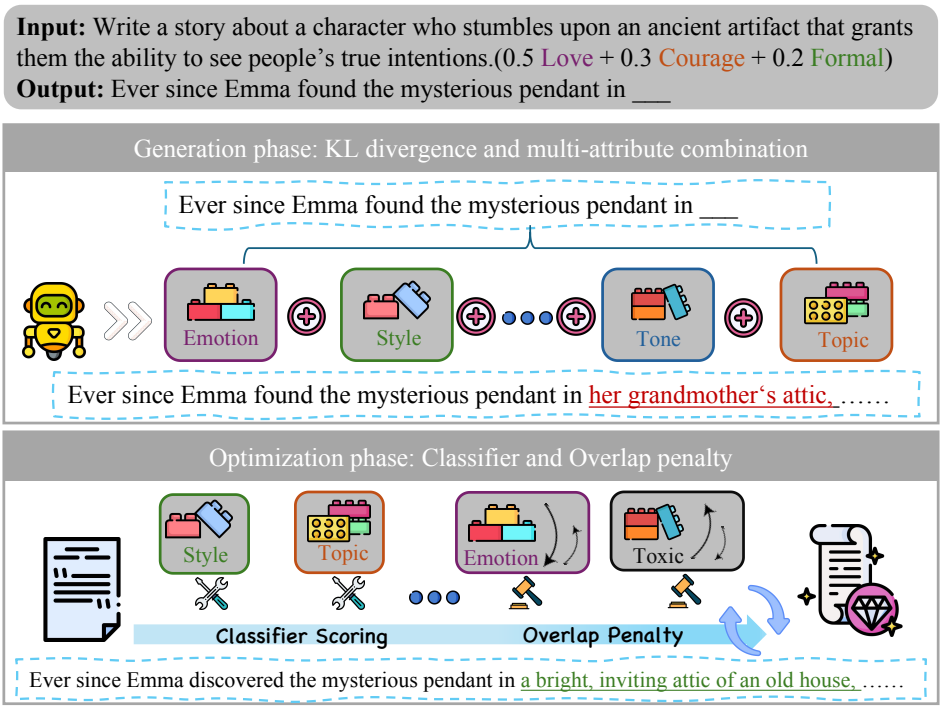

Figure 2: Flowchart of the overall process of the C^{3}TG framework, including the generation phase and the optimization phase, leading to the generation of a text that meets the user's requirements.

Style: Humor (0.85), Tone: Professional (0.9), and Topic: Courage (0.9). The framework outputs an optimized text sequence x_{final} that maximizes fluency while satisfying the attribute constraints $C_{A_i}(x_{\text{final}}) \approx T_i$ for all target attributes. Additionally, the system provides real-time attribute intensity estimates $\{C_{A_i}(x_t)\}_{i=1}^{n}$ throughout the generation process and guarantees monotonic reduction in toxicity metrics.

Methodology

The overall workflow of C^{3}TG is shown in Figure 2. Our framework consists of two principal phases: generation and optimization. During the generation phase, we leverage the foundational language model and incorporate weighted Kullback-Leibler divergence to integrate probability distributions from attribute-specific models corresponding to user-defined targets (e.g., emotion, style, tone, topic). This integration mechanism imposes a differentiable constraint on token selection probabilities, yielding an initial text output that incorporates the specified attribute characteristics.

However, a single-pass generation process frequently proves insufficient for simultaneously satisfying multiple attribute objectives—particularly when attributes exhibit conflicting or interdependent relationships. Consequently, in the optimization phase, we employ discriminative classifiers to quantitatively assess attribute alignment and introduce penalty functions that target attribute conflicts. We formulate a composite energy function combining classifier scores and conflict penalty terms. When attributes deviate from targets or interfere with other attributes after the generation phase, a **Feedback Agent** generates rewriting prompts based on the differential scores from classifiers and penalty terms. This agent then guides the model through a three-stage rewriting process to produce the final text. The resulting closed-loop system transitions from initial generation to structured refinement, establishing an optimal balance between attribute target satisfaction and linguistic coherence.

Generation Phase

We adopt Llama2 (Touvron et al. 2023) as the base language model for generation. In parallel, we fine-tune independent Llama2 models on attribute-specific corpora, obtaining n attribute models—emotion, style, tone, topic, and toxicity—each yielding a prior distribution $Q_i(\cdot \mid x_{1:t-1})$.

Optimization Objective: Given user-specified importance scores $\{\lambda_i\}_{i=1}^{n}$ ($\lambda_i \geq 0$) (e.g., *Joy* 0.9, *Polite* 0.8), we seek a distribution $P(\cdot \mid x_{1:t-1})$ that balances all attributes by minimizing the weighted KL divergence:

$$\mathcal{J}[P] = \sum_{i=1}^{n} \lambda_i \, D_{\text{KL}}\big(P(\cdot \mid x_{1:t-1}) \,\big\|\, Q_i(\cdot \mid x_{1:t-1})\big),$$
$$\text{s.t.} \sum_{x \in \mathcal{V}} P(x \mid x_{1:t-1}) = 1. \quad (2)$$

Solution: By applying Lagrange multipliers, we derive the optimal distribution:

$$P^*(x \mid x_{1:t-1}) = \frac{\prod_{i=1}^{n} Q_i(x \mid x_{1:t-1})^{\lambda_i/\Lambda}}{\sum_{x' \in \mathcal{V}} \prod_{i=1}^{n} Q_i(x' \mid x_{1:t-1})^{\lambda_i/\Lambda}},$$

$$\Lambda = \sum_{i=1}^{n} \lambda_i > 0. \tag{3}$$

Thus, we sample tokens from the *weighted geometric mean* of attribute priors, with λ_i/Λ controlling attribute A_i's influence. Unsatisfactory outputs enter the optimization phase for rewriting.

Optimization Phase

The initial generation frequently produces text that deviates from target attribute intensities or exhibits inter-attribute conflicts, where enhancing one attribute may inadvertently suppress or amplify others. To address these challenges, we formulate an energy-based optimization framework that integrates BERT (Devlin et al. 2019) classifier feedback with specialized penalty terms to mitigate dimensional conflicts during iterative refinement.

Attribute Classifier Scores: In multi-attribute controlled generation, we decompose each high-level attribute into fine-grained dimensions. For instance, the emotion attribute encompasses sub-categories such as joy, sadness, and love. Let $A_1, A_2, \ldots, A_n$ denote the set of controllable dimensions, with $C_{A_i}(x)$ representing the BERT-based classifier that quantifies dimension A_i given text x.

Ideally, $C_{A_i}(x)$ should approximate the specified target value T_i. We measure the aggregate deviation through:

$$E_{\text{classify}}(x) = \sum_{i=1}^{n} \alpha_i |C_{A_i}(x) - T_i|, \tag{4}$$

where α_i denotes the importance weight of dimension A_i, initially specified by the user. After computing Eq. (4), we store each score $C_{A_i}(x)$ and its deviation from the target, enabling adaptive prioritization of dimensions with large discrepancies in later iterations. These deviation metrics feed the *Feedback Agent* (Llama2) described later, which turns them into targeted rewriting directives.

Dimensional Stability Penalties: During optimization, purely pursuing target dimensional improvements without safeguarding related dimensions frequently induces unintended interference effects across the attribute space. To address this challenge, we formulate dimensional stability penalties that quantify and constrain perturbations in non-primary dimensions during each optimization iteration.

Let x_{prev} denote the text before rewriting and x represent the current iteration result. While each iteration focuses on specific "primary optimization dimensions", we designate the remaining k dimensions as "stability-constrained dimensions" $A_1, A_2, \ldots, A_k$. For instance, when optimizing "Emotion-Joy", both other dimensions within the same attribute (such as "Sadness", "Love") and dimensions from different attributes (such as "Style-Formal", "Topic-Courage") collectively form the stability-constrained set. To quantify dimensional fluctuations during rewriting, we define the penalty function:

$$\Omega_{\text{overlap}}(x) = \sum_{i=1}^{k} \beta_i \left| C_{A_i}(x) - C_{A_i}(x_{\text{prev}}) \right|, \tag{5}$$

where β_i represents the penalty coefficient for dimension A_i. Due to inter-dimensional correlations, we assign dimension-specific penalties β_i based on the correlations between attributes. The function $C_{A_i}(\cdot)$ represents the classifier's score for dimension A_i, with the absolute difference constraining both enhancement and suppression effects.

To facilitate precise multi-stage refinement, we focus on a few dimensions in each iteration while aggregating all non-targeted dimensions into the stability-constrained set $A_1, \ldots, A_k$ governed by Eq. (5). For example, when optimizing "Humor", other style dimensions along with all emotion, topic, and tone attributes are incorporated into the stability-constrained set. As refinement progresses, this constrained set evolves dynamically to maintain global multi-dimensional equilibrium. At the end of each round, the **Feedback Agent** records both the aggregate penalty $\Omega_{\text{overlap}}(x)$ and individual dimensional shifts $\left|C_{A_i}(x) - C_{A_i}(x_{\text{prev}})\right|$. Together with the primary-dimension classifier scores, these metrics supply the information the agent needs to craft the next prompt, allowing fine-grained tuning while preserving attribute balance.

Energy Function: After obtaining the classifier scores (Eq. (4)) with the penalty term (Eq. (5)), our energy function is:

$$\begin{aligned} E(x) = &\sum_{i=1}^{n} \alpha_i \left| C_{A_i}(x) - T_i \right| \\ &+ \sum_{j=1}^{k} \beta_j \left| C_{A_j}(x) - C_{A_j}(x_{\text{prev}}) \right|. \end{aligned} \tag{6}$$

where the first term measures alignment with target values across all optimization dimensions, while the second term constrains perturbations in stability-constrained dimensions. The weight α_i is user-specified to reflect dimension importance, while β_j is determined based on experimentally derived attribute correlations.

Leveraging the individual terms of the energy function and their step-to-step changes, a zero-shot Llama2-7B *Feedback Agent* turns this data into rewrite prompts and drives a three-phase iterative optimization loop:

1. Evaluates the energy function (Eq. (6)) and records deviations $\Delta_i = |C_{A_i}(x) - T_i|$ for each iteration;
2. Constructs a priority-ordered correction queue based on deviation magnitudes, focusing attention on dimensions with largest target-value discrepancies;
3. Synthesizes precise rewriting prompts that explicitly specify dimensional adjustments (e.g., "Increase dimension A_j, slightly reduce A_k, maintain A_m");

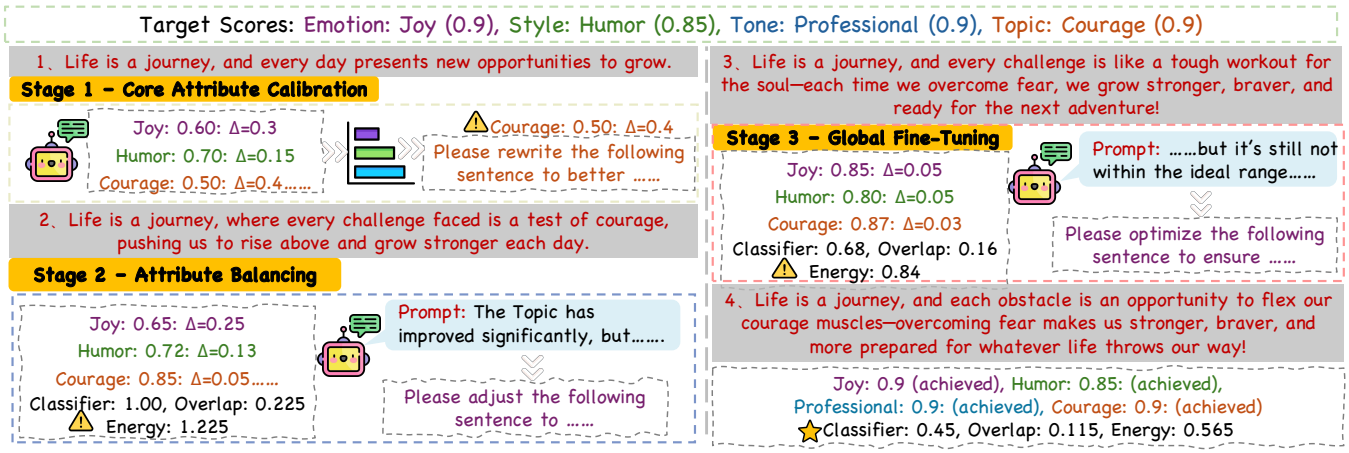

Figure 3: Schematic diagram of the complete Chain-of-Prompt workflow in the optimization phase of the C^{3}TG framework.

4. Monitors convergence criteria and terminates optimization when energy values satisfy predefined thresholds.

The agent refines target attributes through a three-stage optimization, while safeguarding the remaining dimensions from undue interference. This method efficiently navigates the multi-dimensional attribute space until reaching convergence criteria or selecting the text with minimal energy value after the maximum allowed iterations.

Agent-Guided Chain-of-Prompt Refinement: To jointly control multiple attributes, avert conflicts, and preserve fluency, we devise a three-stage optimization procedure:

Stage 1 - Core Attribute Calibration: At the beginning of each iteration, we aggregate classifier scores and penalty terms, with the *Feedback Agent* evaluating dimensional deviations $\left|C_{A_i}(x) - T_i\right|$. The Agent prioritizes dimensions requiring significant adjustment and generates targeted prompts to efficiently minimize the global energy function. For instance: "*Please rewrite the text to better align with the theme of "Courage" while preserving semantic content: ⟨Original text⟩*".

Stage 2-Attribute Balancing Adjustment: After core calibration, we count the change in scores for all non-optimized dimensions $\Delta_i = \left|C_{A_i}(x) - C_{A_i}(x_{\text{prev}})\right|$ and reassess the bias $\left|C_{A_i}(x) - T_i\right|$. The *Feedback Agent* then generates precision-calibrated prompts incorporating intensity modifiers ("slightly," "significantly") for dimensions with persistent deviations, while explicitly specifying attributes requiring stability to prevent emergent conflicts: "*Please modify the text to significantly enhance joy and humor while maintaining the theme of "Courage": ⟨Phase 1 output⟩*".

Phase 3 - Global Fine-tuning: After the second rewrite, the agent craft a single, consolidated prompt that uses the composite energy $E(x)$ and the latest attribute metrics to direct final, per-attribute fine-tuning—e.g., "*Please polish the text to keep a pleasant tone, slightly raise humor and formality, and retain the clear 'Courage' theme: ⟨Phase 2 output⟩*." This third stage performs the last adjustments to bring every attribute as close as possible to its target score, driving $E(x)$ below the convergence threshold $\tau(\tau = 0.025)$.

In summary, the following steps are performed for each iteration round:

1. The *Feedback Agent* formulates a context-aware prompt based on current classifier and penalty evaluations;
2. The LLM generates text x_{new} conditioned on the prompt;
3. We compute the energy differential $\Delta E = E(x_{\text{new}}) - E(x_{\text{prev}})$ and terminate if both $\Delta E < 0$ and $E(x_{\text{new}}) \leq \tau$; otherwise, we proceed to the next phase or iteration.

If convergence criteria remain unsatisfied after reaching the maximum iteration limit, we return the lowest-energy text. Figure 3 summarizes the three-phase optimization.

Experiment

We conduct a comprehensive empirical evaluation of C^{3}TG to address the following research questions:

- **RQ1:** How does C^{3}TG compare to state-of-the-art methods across automated metrics and human evaluations?
- **RQ2:** How effectively does C^{3}TG handle attribute conflicts and interactions?
- **RQ3:** What is the contribution of each component in C^{3}TG to system effectiveness?

Experimental Setup

Dataset We select two English story scenarios: ROCStories(ROC) and WritingPrompts(WP)(Guan et al. 2021). ROC emphasizes causal and temporal relationships in everyday scenarios, while WP provides structurally rich prompt-story pairs. This combination enables rigorous assessment of multi-attribute control and our iterative optimization framework across diverse textual contexts.

Control Dimensions C^{3}TG controls five primary attribute categories: (1) Emotion (anger, fear, joy, love, sadness, surprise); (2) Style (politeness, romantic, humor, sarcasm, metaphorical); (3) Tone (professional, casual); (4) Topic (knowledge, justice, humanity, courage); and (5) Toxicity (toxic, non-toxic). For classifier training, we utilize publicly available datasets: Social Network Sentiment corpus,

Method	ROC			WP			Tox.↓
	Acc.↑	PPL↓	Dist-1/2/3↑	Acc.↑	PPL↓	Dist-1/2/3↑	
Controllable text generation by controlling decoding distribution							
COLD	24.35	21.07	0.08/0.10/0.22	20.50	24.54	0.06/0.08/0.18	0.53
BOLT	36.54	17.33	0.09/0.28/0.38	32.07	20.52	0.08/0.26/0.36	0.76
MuCola	27.93	16.89	0.13/0.24/0.33	25.12	19.83	0.10/0.21/0.30	0.58
MacLaSa	31.54	16.35	0.11/0.25/0.33	28.45	19.01	0.08/0.22/0.30	0.60
PriorControl	22.56	18.54	0.09/0.25/0.29	19.36	22.07	0.07/0.22/0.25	0.66
LatentOps	33.55	14.57	0.18/0.22/0.31	30.00	17.39	0.13/0.17/0.26	0.55
PPLM	32.39	15.04	0.19/0.23/0.39	29.74	18.20	0.16/0.20/0.36	0.39
Mix&Match	49.78	22.71	0.21/0.24/0.37	45.00	26.46	0.11/0.20/0.32	0.47
Model Arithmetic	87.53	11.08	0.37/0.68/0.81	84.23	14.30	0.33/0.50/0.75	0.16
Controllable text generation by indirect control strategy							
LLM-based Prompt	89.45	5.37	0.47/0.71/0.89	80.02	9.65	0.42/0.50/0.82	0.29
LLM-based Fine-tuning	79.03	8.53	0.39/0.68/0.84	75.00	10.50	0.38/0.54/0.80	0.16
C^{3}TG (Ours)	**90.39**	**4.04**	**0.53/0.74/0.90**	**85.56**	**3.68**	**0.47/0.55/0.84**	**0.12**

Table 1: Comparison of automated evaluation results for C^{3}TG and baseline methods on the ROC and WP datasets.

Method	ROC			WP		
	Topic%↑	Fluency↑	Diversity↑	Topic%↑	Fluency↑	Diversity↑
Controllable text generation by controlling decoding distribution						
COLD	1.07	1.68	1.32	0.98	1.55	1.26
BOLT	2.89	2.34	2.76	2.45	1.98	2.35
MuCola	2.71	2.62	2.52	2.38	2.33	2.20
MacLaSa	2.95	2.84	2.64	2.61	2.48	2.31
PriorControl	2.30	2.12	2.05	2.16	1.95	1.90
LatentOps	3.12	2.92	2.84	2.85	2.65	2.55
PPLM	3.40	2.53	3.06	3.05	2.23	2.74
Mix&Match	3.86	3.11	3.54	3.52	2.76	3.23
Model Arithmetic	4.20	3.85	4.07	3.77	3.47	3.69
Controllable text generation by indirect control strategy						
LLM-based Prompt	4.73	3.97	4.03	3.24	3.18	3.76
LLM-based Fine-tuning	4.33	3.82	4.28	**3.65**	3.57	3.42
C^{3}TG (Ours)	**4.74**	**4.53**	**4.45**	**3.65**	**3.88**	**4.05**

Table 2: Comparison of human evaluation results for C^{3}TG and baseline methods on the ROC and WP datasets.

xSLUE style annotation collection, Domain Q&A and Instructions repository, and Toxicity Review dataset.

Baselines We compare C^{3}TG with two main categories of mainstream methods: methods that directly intervene in the decoding distribution, including COLD (Qin et al. 2022), BOLT (Liu, Khalifa, and Wang 2023), MuCoLa (Kumar, Paria, and Tsvetkov 2022), MacLaSa (Ding et al. 2023), PriorControl (Gu et al. 2023), LatentOps (Liu et al. 2023), PPLM (Dathathri et al. 2020), Mix&Match (Mireshghallah, Goyal, and Berg-Kirkpatrick 2022), and Model Arithmetic (Dekoninck et al. 2024); the other category is the indirect control strategies including Prompt and Fine-tuning.

Evaluation Metrics For **automated evaluation**, we use four metrics: classifier accuracy to measure how well attributes are controlled, language model perplexity to assess coherence, Distinct-n to measure lexical diversity, and toxicity probability from an external API to check content safety.

In the **human evaluation**, five independent domain experts rate the system outputs on a 5-point Likert scale for attribute alignment, linguistic fluency, and content diversity, and the final score is the average of their ratings.

Time Optimization We batch the gradients of the user-specified attributes into a single GPU kernel and reuse cached hidden states, so each generation step requires only one forward–backward pass. An energy-based early-stopping criterion trims decoding time by about 40%.

Overall Experimental Results(RQ1)

As shown in Tables 1 and 2, which include direct decoding intervention methods as well as methods with indirect control strategies, the C^{3}TG results are listed at the bottom. All baselines follow their original specs; single-attribute models received minimal, architecture-preserving tweaks for multi-attribute control, were tuned to their best configuration.

Method	Conflict Experiment			Overlap Experiment		
	Average↓	PPL↓	Drift↓	Average↓	PPL↓	Drift↓
Model Arithmetic	0.27	10.48	0.38	0.22	11.03	0.31
LLM-based Prompt	0.19	5.75	0.25	0.12	4.96	0.29
C^{3}TG	**0.08**	**4.54**	**0.16**	**0.07**	**4.13**	**0.18**

Table 3: Performance of the C^{3}TG in conflict and overlap experiments on the ROC dataset.

Method	ROC				WP			
	Acc.↑	**PPL↓**	**Dist-2↑**	**Tox.↓**	**Acc.↑**	**PPL↓**	**Dist-2↑**	**Tox.↓**
Components								
w/o Optimization	65.22	10.53	0.35	0.38	62.17	12.64	0.29	0.42
w/o Generation	59.40	25.62	0.28	0.32	55.08	28.03	0.25	0.45
w/o Overlap	78.46	5.11	0.47	0.36	73.56	6.88	0.41	0.43
Iterations								
1-Iteration	74.21	6.31	0.43	0.19	70.09	7.94	0.39	0.27
2-Iteration	85.62	4.89	0.48	0.12	81.13	5.81	0.44	0.26
C^{3}TG	**90.39**	**4.04**	**0.74**	**0.12**	**85.56**	**3.68**	**0.55**	**0.24**

Table 4: Ablation results of the C^{3}TG framework on ROC and WP datasets.

Automated Evaluation Quantitative analysis reveals that C^{3}TG achieves superior balance across accuracy, perplexity, and diversity metrics. To evaluate toxicity mitigation capabilities, we employed the /pol/ dataset (Papasavva et al. 2020) for controlled text rewriting through the framework. Toxicity evaluation via an external API service demonstrates that C^{3}TG yields the lowest toxicity, confirming that its iterative energy minimization jointly boosts attribute accuracy, fluency, diversity, and suppresses harmful content.

Human Evaluation Human evaluation shows C^{3}TG's significant advantage over all baselines in attribute alignment, fluency, and diversity. Its composite energy function and multi-stage prompt chain jointly reinforce target attributes, resolve conflicts, and preserve related dimensions, producing content that retains natural variation and coherence.

Conflict and Overlap Experiment(RQ2)

To evaluate C^{3}TG's robustness under attribute conflicts, we test it against *Model Arithmetic* and an *LLM-based prompt* baseline on 30% of ROCStories. We constructed a negative pair ("fear 0.7 vs. joy 1.0") and a positive pair ("romance 0.7 + love 0.7"), keeping other attributes constant. Performance was measured via Average Absolute Bias, Perplexity (PPL), and Uncontrolled Dimensional Drift (Drift measures the average absolute change in all non-target attribute scores, capturing unintended side-effects).

As shown in Table 3, in both scenarios, C^{3}TG demonstrates superior performance—achieving minimal bias (0.08/0.07) while maintaining optimal PPL and Drift metrics. Model Arithmetic amplifies positively correlated attributes and prompt-based tuning boosts fluency, but neither approach succeeds in curbing attribute drift. These results confirm that C^{3}TG's integration of classifier feedback with iterative penalty terms delivers superior attribute stability across conflicting and overlapping conditions.

Ablation Experiment(RQ3)

We conduct ablation experiments on ROC and WP datasets to assess component contributions. Four configurations are evaluated: initial-generation-only, optimization-only, no-penalty-term, and full C^{3}TG with varying iterations (Table 4). Results show that initial-generation-only achieves low perplexity but poor attribute alignment and toxicity control. Optimization-only is unstable without quality seed texts (*w/o Generation* begins directly with the three-stage optimization). Removing penalty terms increases non-target attribute fluctuations, compromising dimensional balance and toxicity suppression. In contrast, C^{3}TG achieves optimal balance, with additional iterations enhancing attribute accuracy, fluency, and toxicity reduction—validating our multi-stage prompt chain and iterative feedback framework.

Conclusion

We present C^{3}TG, a collaborative framework for controlled text generation that integrates LLMs with lightweight attribute classifiers. Our framework consists of two phases: a generation phase fusing attribute distributions via weighted KL divergence, and an optimization phase employing a composite energy function that balances classifier scores with stability penalties. The conflict-aware strategy is embodied in the multi-stage prompt chain of the optimization phase, reconciling attribute clashes while preserving coherence. Experiments on ROCStories and WritingPrompts demonstrate that C^{3}TG outperforms existing methods in attribute accuracy, fluency, diversity, and toxicity reduction, validating the effectiveness of our conflict-aware optimization framework.

Acknowledgments

This work is partially supported by National Nature Science Foundation of China under No. U21A20488, and is funded by Southeast University-China Mobile Research Institute Joint Innovation Center. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Cao, M.; Fatemi, M.; Cheung, J. C. K.; and Shabanian, S. 2024a. Successor Features for Efficient Multi-Subject Controlled Text Generation. In *Forty-first International Conference on Machine Learning, ICML 2024*.

Cao, Y.; Zhao, J.; Zhang, R.; Zou, H.; and Mao, W. 2024b. TARA: Token-level Attribute Relation Adaptation for Multi-Attribute Controllable Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12570–12579.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *Proceedings of the 8th International Conference on Learning Representations*, 26–30.

Dekoninck, J.; Fischer, M.; Beurer-Kellner, L.; and Vechev, M. 2024. Controlled Text Generation via Language Model Arithmetic. In *Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 4171–4186.

Ding, H.; Pang, L.; Wei, Z.; Shen, H.; Cheng, X.; and Chua, T. 2023. MacLaSa: Multi-Aspect Controllable Text Generation via Efficient Sampling from Compact Latent Space. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2023*, 4424–4436.

Gu, Y.; Feng, X.; Ma, S.; Zhang, L.; Gong, H.; Zhong, W.; and Qin, B. 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*, 12590–12616.

Guan, J.; Zhang, Z.; Feng, Z.; Liu, Z.; Ding, W.; Mao, X.; Fan, C.; and Huang, M. 2021. OpenMEVA: A Benchmark for Evaluating Open-ended Story Generation Metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Virtual Event*, 6394–6407.

Hallinan, S.; Liu, A.; Choi, Y.; and Sap, M. 2023. Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-Experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 228–242.

Jie, R.; Meng, X.; Shang, L.; Jiang, X.; and Liu, Q. 2024. Prompt-Based Length Controlled Generation with Multiple Control Types. In *Findings of the Association for Computational Linguistics, ACL*, 1067–1085.

Kang, D.; and Hovy, E. 2021. Style is NOT a Single Variable: Case Studies for Cross-Stylistic Language Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2376–2387.

Krause, B.; Gotmare, A. D.; McCann, B.; Keskar, N. S.; Joty, S.; Socher, R.; and Rajani, N. F. 2021. GeDi: Generative Discriminator Guided Sequence Generation. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2021*, 4929–4952.

Kumar, S.; Malmi, E.; Severyn, A.; and Tsvetkov, Y. 2021. Controlled Text Generation as Continuous Optimization with Multiple Constraints. In *Advances in Neural Information Processing Systems 34*, 14542–14554.

Kumar, S.; Paria, B.; and Tsvetkov, Y. 2022. Gradient-based Constrained Sampling from Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2251–2277.

Li, X. L.; and Liang, P. 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. Virtual Event*, 4582–4597.

Li, Y.; Zhang, S.; Wu, R.; Huang, X.; Chen, Y.; Xu, W.; Qi, G.; and Min, D. 2024. MATEval: A Multi-agent Discussion Framework for Advancing Open-Ended Text Evaluation. In *Database Systems for Advanced Applications - 29th International Conference, DASFAA 2024, Gifu, Japan, July 2-5, 2024, Proceedings, Part VII*, volume 14856, 415–426.

Liang, X.; Wang, H.; Song, S.; Hu, M.; Wang, X.; Li, Z.; Xiong, F.; and Tang, B. 2024a. Controlled Text Generation for Large Language Models with Dynamic Attribute Graphs. In *Findings of the Association for Computational Linguistics: 62nd Annual Meeting of the Association for Computational Linguistics 2024*, 5797–5814.

Liang, X.; Wang, H.; Wang, Y.; Song, S.; Yang, J.; Niu, S.; Hu, J.; Liu, D.; Yao, S.; Xiong, F.; et al. 2024b. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Liu, G.; Feng, Z.; Gao, Y.; Yang, Z.; Liang, X.; Bao, J.; He, X.; Cui, S.; Li, Z.; and Hu, Z. 2023. Composable Text Controls in Latent Space with ODEs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16543–16570.

Liu, X.; Khalifa, M.; and Wang, L. 2023. BOLT: Fast Energy-based Controlled Text Generation with Tunable Biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 186–200.

Liu, Y.; Liu, X.; Zhu, X.; and Hu, W. 2024. Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, 9231–9253.

Ma, C.; Zhao, T.; Shing, M.; Sawada, K.; and Okumura, M. 2024. Focused prefix tuning for controllable text generation. *Journal of Natural Language Processing*, 31(1): 250–265.

Madaan, N.; Padhi, I.; Panwar, N.; and Saha, D. 2021. Generate Your Counterfactuals: Towards Controlled Counterfactual Generation for Text. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 13516–13524.

Min, D.; Hu, N.; Jin, R.; Lin, N.; Chen, J.; Chen, Y.; Li, Y.; Qi, G.; Li, Y.; Li, N.; and Wang, Q. 2024. Exploring the Impact of Table-to-Text Methods on Augmenting LLM-based Question Answering with Domain Hybrid Data. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024*, 464–482.

Mireshghallah, F.; Goyal, K.; and Berg-Kirkpatrick, T. 2022. Mix and Match: Learning-free Controllable Text Generation using Energy Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 401–415.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Papasavva, A.; Zannettou, S.; Cristofaro, E. D.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media. Atlanta, Georgia, USA*, 885–894.

Pynadath, P.; and Zhang, R. 2025. Controlled LLM Decoding via Discrete Auto-regressive Biasing. *arXiv preprint arXiv:2502.03685*.

Qian, J.; Dong, L.; Shen, Y.; Wei, F.; and Chen, W. 2022. Controllable Natural Language Generation with Contrastive Prefixes. In *Findings of the Association for Computational Linguistics: 60th Annual Meeting of the Association for Computational Linguistics 2022.*, 2912–2924.

Qin, L.; Welleck, S.; Khashabi, D.; and Choi, Y. 2022. COLD Decoding: Energy-based Constrained Text Generation with Langevin Dynamics. In *Advances in Neural Information Processing Systems 35*, 9538–9551.

Shi, C.; Cai, D.; and Yang, Y. 2024. Lifi: Lightweight Controlled Text Generation with Fine-Grained Control Codes. *arXiv preprint arXiv:2402.06930*.

Shulev, V.; and Sima'an, K. 2024. Continual Reinforcement Learning for Controlled Text Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 3881–3889.

Son, H. R.; and Lee, J.-Y. 2024. Locate&Edit: Energy-based Text Editing for Efficient, Flexible, and Faithful Controlled Text Generation. *arXiv preprint arXiv:2407.00740*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Wang, H.; and Sha, L. 2023. Harnessing the Plug-and-Play Controller by Prompting. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, 165–174.

Wang, Y.; and Demberg, V. 2024a. RSA-Control: A Pragmatics-Grounded Lightweight Controllable Text Generation Framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, 5561–5582.

Wang, Y.; and Demberg, V. 2024b. RSA-Control: A Pragmatics-Grounded Lightweight Controllable Text Generation Framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 5561–5582.

Xu, M.; Geffner, T.; Kreis, K.; Nie, W.; Xu, Y.; Leskovec, J.; Ermon, S.; and Vahdat, A. 2024. Energy-based diffusion language models for text generation. *arXiv preprint arXiv:2410.21357*.

Yang, N.; Ma, W.; and Cheng, P. 2024. Plug-in Language Model: Controlling Text Generation with a Simple Regression Model. In *Findings of the Association for Computational Linguistics*, 2165–2181.

Yang, Z.; Huang, Y.; Chen, Y.; Wu, X.; Feng, J.; and Deng, C. 2025. Palette of Language Models: A Solver for Controlled Text Generation. *arXiv preprint arXiv:2503.11182*.

Yu, S.; Lee, C.; Lee, H.; and Yoon, S. 2024. Controlled Text Generation for Black-box Language Models via Score-based Progressive Editor. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 14215–14237.

Zhong, T.; Wang, Q.; Han, J.; Zhang, Y.; and Mao, Z. 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*, 8233–8248.

Zhu, X.; Karadzhov, G.; Whitehouse, C.; and Vlachos, A. 2024. Segment-Level Diffusion: A Framework for Controllable Long-Form Generation with Diffusion Language Models. *arXiv preprint arXiv:2412.11333*.

Zong, C.; Chen, Y.; Lu, W.; Shao, J.; Huang, Y.; Chang, H.; and Zhuang, Y. 2024. ProSwitch: Knowledge-Guided Instruction Tuning to Switch Between Professional and Non-Professional Responses. *arXiv preprint arXiv:2403.09131*.