

Controllable Stylistic Text Generation with Train-Time Attribute-Regularized Diffusion

Fan Zhou[§] Chang Tian[§] Tim Van de Cruys[§]

[§] KU Leuven

fan.zhou@kuleuven.be namechangtian@163.com tim.vandecruys@kuleuven.be

Abstract

Generating stylistic text with specific attributes is a key problem in controllable text generation. Recently, diffusion models have emerged as a powerful paradigm for both visual and textual generation. Existing approaches can be broadly categorized into classifier-free guidance (CFG) and classifier guidance (CG) methods. While CFG effectively preserves semantic content, it often fails to provide effective attribute control. In contrast, CG modifies the denoising trajectory using classifier gradients, enabling better attribute alignment but incurring high computational costs during sampling and suffering from classifier generalization issues. In this work, we propose **RegDiff**, a regularized diffusion framework that leverages attribute features without requiring a pretrained classifier during sampling, thereby achieving controllable generation with reduced computational costs. Specifically, RegDiff employs a VAE-based encoder-decoder architecture to ensure reconstruction fidelity and a latent diffusion model trained with attribute supervision to enable controllable text generation. Attribute information is injected only during training. Experiments on five datasets spanning multiple stylistic attributes demonstrate that RegDiff outperforms strong baselines in generating stylistic texts. These results validate the effectiveness of RegDiff as an efficient solution for attribute-controllable text diffusion. Our code, datasets, and resources will be released upon publication at <https://github.com/xxxx>.

1 Introduction

Deep learning has enhanced the model’s ability to process text (Tian et al., 2022, 2024, 2025). Generating stylistic text with specific attributes has been an important area of research (Mou and Vechtomova, 2020). Diffusion models have recently advanced the state of the art in generative modeling, achieving remarkable performance across images (Jelaca et al., 2025; Schildermans et al., 2025),

audio (Li et al., 2021), and natural language (Li et al., 2022; Gong et al., 2022; Yi et al., 2024). Their iterative denoising process enables stable likelihood-based training and high-quality outputs, positioning them as a competitive alternative to autoregressive and adversarial methods for stylistic text generation (Ho and Salimans, 2022). However, controllability in diffusion-based text generation remains a significant challenge (Kwon et al., 2025).

There are two primary diffusion paradigms for controllable generation studied in the research community: classifier-free guidance (CFG) and classifier guidance (CG) diffusion methods.

Classifier-free guidance introduces controllability by interpolating between conditional and unconditional denoising predictions, thereby amplifying conditioning signals without relying on an external classifier (Ho and Salimans, 2022). In text generation, CFG has proven effective in improving prompt adherence and semantic coherence. However, its performance degrades in fine-grained NLP tasks, where it primarily reinforces surface-level cues rather than capturing stylistic attributes (Li et al., 2022). Notably, CFG lacks explicit mechanisms for attribute-specific control in stylistic text generation (Vaeth et al., 2025).

In contrast, classifier guidance injects classifier gradients at each denoising step in the sampling process to steer the diffusion trajectory toward samples that better align with a target class label (Dhariwal and Nichol, 2021). Its major advantage lies in flexibility: it operates entirely at inference time and does not require retraining the diffusion model. This property makes CG appealing in NLP settings, where high-quality pretrained classifiers already exist for attributes such as sentiment or formality (Li et al., 2022; Horvitz et al., 2024). However, this reliance introduces key limitations. Robust classifiers are often unavailable for many nuanced attributes, and training them from scratch can be computationally prohibitive. Moreover, CG faces structural

challenges: classifiers are trained on clean text but applied to noisy intermediate representations, leading to a distribution mismatch (Vaeth et al., 2025). In addition, inference becomes considerably more expensive, as each denoising step requires an auxiliary classifier gradient computation (Shenoy et al., 2024). Finally, classifier gradients can distort sampling trajectories and push generations off the data manifold, thereby degrading output quality (Vaeth et al., 2024; Karras et al., 2024). Collectively, these drawbacks underscore why inference-time guidance—despite its practicality—remains an imperfect solution for achieving precise controllable text diffusion.

When performing text style transfer with diffusion models under the CFG setting, we observe that different attributes correspond to distinct subspaces within the text latent space, as illustrated in Figures 1 and 2. Sentiment clusters are partially separable, while formality clusters remain highly entangled. For instance, the sentiment clusters shown in Figure 1 demonstrate that attribute features are not fully entangled with semantic features. This observation motivates us to incorporate attribute classification as an **inductive bias** during training. Moreover, considering the high computational cost associated with inference-time sampling and the difficulty of obtaining robust classifiers for certain nuanced attributes, we aim to impose regularization both on the data-side attribute manifold and within the diffusion training process. What is needed is a mechanism to align diffusion dynamics with attribute supervision during training.

To this end, we propose **RegDiff**, a regularized diffusion framework (illustrated in Figure 3) that leverages attribute representations without relying on a pretrained classifier during the sampling process, thereby enabling controllable generation with reduced computational overhead. Specifically, RegDiff, trained with attribute supervision, adopts a VAE-based encoder-decoder architecture to preserve reconstruction fidelity, and employs a latent diffusion model to achieve controllable text generation. Attribute information is incorporated in two forms of regularization: one applied to the data-side attribute manifold in the VAE training, and another integrated into the diffusion training process, ensuring efficient inference and stylistically consistent outputs.

Our main contributions are summarized as follows:

- We conduct comprehensive experiments to analyze stylistic text generation and investigate how to effectively incorporate style attribute control.
- We empirically show that classifier guidance is not always necessary. For attributes that are already separable from semantics in the latent space, generation remains controllable without guidance, while for more entangled attributes, train-time regularization provides sufficient constraint without relying on inference-time guidance.
- We propose **RegDiff**, an attribute-regularized diffusion framework that effectively disentangles attribute representations while preserving generation quality.
- We demonstrate robust and superior performance across five diverse datasets compared to the baselines, establishing RegDiff as a general and effective framework for controllable and stylistic text generation.

2 Related Work

Diffusion Controllability. Fully conditional diffusion models incorporate conditioning directly at training (Gong et al., 2022), but require paired data and scale poorly. Classifier guidance (CG) (Dhariwal and Nichol, 2021; Um and Ye, 2024) injects external classifier gradients at inference, and has been adapted to text style transfer (Horvitz et al., 2024). While precise in principle, CG suffers from distribution mismatch between classifiers trained on clean data and noisy diffusion states, high computational cost, and off-manifold artifacts (Chung et al., 2022; Vaeth et al., 2025). Classifier-free guidance (CFG) (Ho and Salimans, 2022; Shen et al., 2024) avoids auxiliary classifiers by jointly training conditional and unconditional branches, and has proven crucial in large-scale text-to-image models (Saharia et al., 2022; Vaeth et al., 2025). However, CFG lacks fine-grained, explicit control over specific attributes.

Attribute Control. Some attributes are abstract and human-defined, such as sentiment, style, topic, and genre in the NLP domain (John et al., 2019; Talon et al., 2025). Traditional approaches to controllable text generation include adversarial training (Shen et al., 2017), disentangled latent representations (Fu et al., 2018), and style-specific decoders

(Lample et al., 2019). In diffusion-based methods, attribute control mainly depends on inference-time guidance, such as classifier-guided (CG) or classifier-free guided (CFG) sampling (Li et al., 2022; Karras et al., 2024; Wang et al., 2024).

Representation Regularization. Latent diffusion (Rombach et al., 2022) demonstrated the effectiveness of training-time constraints in compact latent spaces for image synthesis, and laid the groundwork for latent tdiffusion models for text generation (Zhang et al., 2023; Lovelace et al., 2023). Regularization and latent-space structuring methods, such as mutual information maximization, disentanglement constraints, subspace discovery, and semantic guidance — have been employed in representation learning to enforce attribute separability and thereby improve generative controllability (Brack et al., 2023; Härkönen et al., 2020; Yu et al., 2024).

Different from previous methods, we introduce inductive bias during training by constraining the attribute category information.

3 Preliminary

3.1 Problem Definition

We study the problem of controllable text generation, where the goal is to generate an output sentence \hat{x} that exhibits a target attribute c (e.g., authorship style, formality, sentiment, toxicity) while maintaining appropriate semantic relationship with an input x . Table 1 illustrates various attribute manipulation tasks, ranging from style transfer that preserves core meaning (authorship, formality) to content-aware modifications (sentiment, toxicity).

3.2 Continuous Diffusion Model

Latent Diffusion Process. Diffusion models generate data by learning to reverse a fixed forward noising process. Given an initial latent variable $x_0 \sim p_{\text{data}}(x)$, the forward process incrementally perturbs x_0 into Gaussian noise through a Markov chain:

$$q(x_t | x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I\right), \quad t \in [1, T], \quad (1)$$

where $\{\beta_t\}_{t=1}^T$ is a predefined noise schedule. The diffusion model learns a parameterized reverse process:

$$p_\theta(x_{t-1} | x_t) \approx q(x_{t-1} | x_t, x_0), \quad (2)$$

which is optimized using a reweighted denoising score-matching loss. In practice, the model predicts

either the original clean input x_0 , the injected noise ϵ_t , or the velocity v_t (Ho et al., 2020; Salimans and Ho, 2022), and training reduces to a simple regression objective.

Text-to-Continuous Mapping. Since diffusion models operate in continuous spaces while text is inherently discrete, we first map input sentences into continuous latent representations. We adopt an architecture with an encoder E and a decoder D , instantiated as a variational autoencoder (VAE) model, where the encoder maps input text x to latent representation $z = E(x) \in \mathbb{R}^{L \times d}$, and the decoder reconstructs text from latents $\hat{x} = D(z)$. The model is initialized from a pretrained language model and fine-tuned on reconstruction tasks (training details in Section 4). Once trained, the encoder and decoder are frozen, providing a stable continuous representation space for the subsequent diffusion process.

Classifier Guidance. (Dhariwal and Nichol, 2021) introduces classifier guidance, which augments the diffusion score with the gradient of an auxiliary classifier’s log-likelihood. At inference time, the diffusion score is modified to include the gradient of the log-likelihood of an auxiliary classifier as:

$\tilde{\epsilon}_\theta(x_t, t, c) = \epsilon_\theta(x_t, t, c) - \gamma \sigma_t \nabla_{x_t} \log p_\phi(c | x_t)$, where $\epsilon_\theta(x_t, t, c)$ is the conditional noise prediction, $p_\phi(c | x_t)$ is the auxiliary classifier evaluated on the noisy sample x_t , σ_t is the noise level (standard deviation) at step t , and $\gamma \geq 0$ controls the guidance strength.

Classifier-free Guidance. A central technique used in this work is classifier-free guidance (CFG) (Ho and Salimans, 2022). During the training of the diffusion model, both conditional and unconditional denoising objectives are optimized jointly. At inference time, the two predictions are combined as:

$$\hat{\epsilon}_\theta(x_t, t, c) = (1 + \gamma) \epsilon_\theta(x_t, t, c) - \gamma \epsilon_\theta(x_t, t, \emptyset),$$

where $\epsilon_\theta(x_t, t, c)$ is the noise prediction conditioned on attribute c , $\epsilon_\theta(x_t, t, \emptyset)$ is the unconditional prediction, and $\gamma \geq 0$ controls the strength of guidance. By interpolating between the two paths, CFG amplifies attribute information while retaining the semantic content of the original input.

4 Method

Based on previous findings, we introduce **RegDiff** (Attribute-**R**egularized **D**iffusion) (Figure 3). The

Attributes	Source Texts	Target Texts
Sentiment (parallel)	(Negative) The new sign at the park is confusing and hard to understand.	(Positive) The new sign at the park is easy to read.
Sentiment (non-parallel)	(Negative) I was sadly mistaken .	(Positive) Excellent food .
Toxicity	(Toxic) Foreal shit makes no sense.	(Neutral) Foreal thing makes no sense.
Formality	(Informal) She cant sing for her life!	(Formal) She is a poor vocalist.
Authorship	(Shakespeare) Make thee a fortune from me.	(Modern) I'll make you a rich man.

Table 1: Examples of text style transfer across different style attributes.

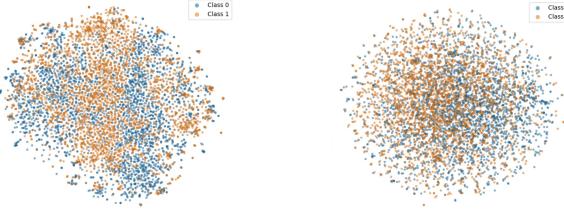


Figure 1: Sentiment data without inductive bias.

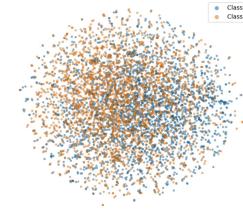


Figure 2: Formality data without inductive bias.

framework integrates a variational autoencoder (VAE) with a diffusion model. After fine-tuning, the VAE encoder and decoder are frozen to provide stable latent representations. The encoder maps input texts into latent variables z (with z^{tgt} as inputs to the diffusion model and z^{src} as condition for diffusion). During sampling, the diffusion model generates latent codes \hat{z}^{tgt} approximating z^{tgt} , which are then decoded into natural language via the frozen VAE decoder.

4.1 VAE model

We adopt a general encoder–decoder architecture, where the encoder and decoder are jointly trained with an auxiliary classifier. In this architecture, we adopt VAE mode, but is not limited to VAE. In contrast to the conventional VAE objective of reconstruction loss plus KL divergence, we additionally include a classification loss to impose attribute bias. The encoder produces latent representations $z \in \mathbb{R}^{B \times S \times L}$. To obtain an attribute-specific representation, we apply a mean pooling operation P over the sequence dimension S to produce \bar{z} , consistent with standard practice in fine-tuning pre-trained models for classification. The pooled vector $\bar{z} \in \mathbb{R}^{B \times L}$ serves as input to the classifier, enabling both (i) classifier training for attribute prediction and (ii) gradient-based biasing of z such that attributes become separable at the pooled representation level (see Figure 4).

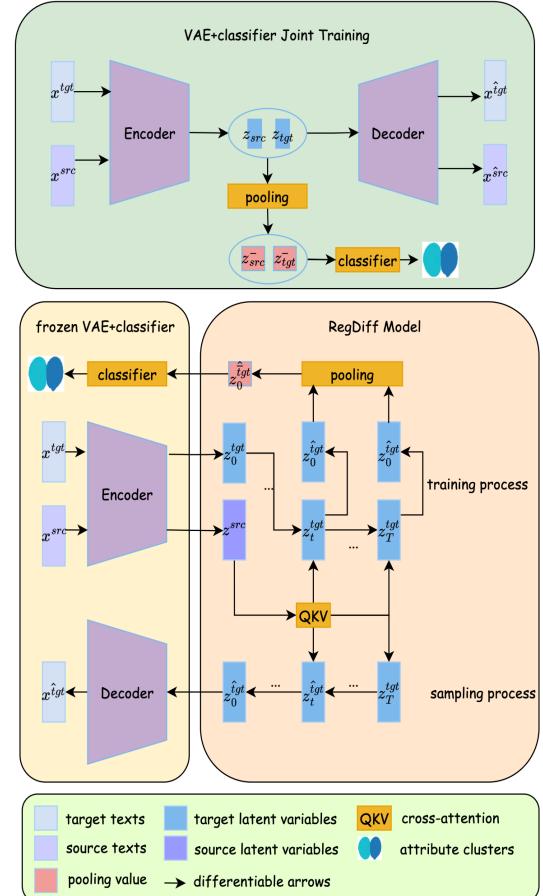


Figure 3: A graphical representation of the RegDiff framework.

$$\bar{z} = P(z) \quad (3)$$

$$L_{vae} = L_{recon} + \alpha L_{KL} + \beta L_{classifier} \quad (4)$$

4.2 Diffusion Model

Input. Once the VAE is trained, we freeze its parameters and employ the encoder to obtain latent representations z from text inputs. Among the information encoded in these representations, we

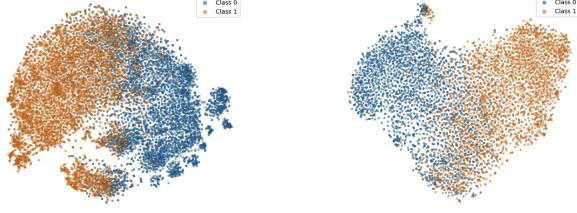


Figure 4: The two figures represent: Inductive biased formality clusters and inductive biased authorship clusters.

focus on two aspects: (1) token-level latent representation z that capture textual content, and (2) attribute-level clusters \bar{z} in a secondary manifold reflecting style.

In our task, the diffusion model employs classifier-free guidance, where z_{tgt} serves as the diffusion input and z_{src} is used as a conditioning signal to constrain the semantic content during the denoising process of z_{tgt} .

Training-time Regularization. In addition to introducing style attribute bias into the data, we incorporate a regularization loss during the diffusion model’s training to better align the distribution of \bar{z} with that observed during training.

Specifically, we introduce an auxiliary distributional matching loss in the attribute space:

$$\mathcal{L}_{\text{attr}} = \mathcal{D}(p_\theta(P(\bar{z})) \| q(P(\bar{z}))), \quad (5)$$

where \mathcal{D} denotes a distribution similarity. In our task, we reuse the classifier which is jointly trained with VAE model. This classifier is also frozen, delivering to z_{pred} the gradients generated by the dissimilarity between \bar{z} and \bar{z}_{pred} . The overall training objective is then

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}} + \lambda \mathcal{L}_{\text{classifier}} \quad (6)$$

$$\mathcal{L}_{\text{classifier}} = \text{CE}(\bar{z}_{pred}, L) \quad (7)$$

Concretely, let z_t denote the noisy latent at step t and $\mathbf{v}_\theta(z_t, t)$ the network prediction under the velocity parameterization (i.e., predicting the linear combination of noise and clean latent). The standard diffusion loss is given by

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t, z_0, \epsilon} \|\mathbf{v}_\theta(z_t, t) - \mathbf{v}\|^2, \quad (8)$$

where \mathbf{v} denotes the target velocity corresponding to the ground-truth noise ϵ or z_0^{tgt} . These three concept can be converted into one another. As we

need to calculate the regularization loss on \bar{z} , we can obtain the \hat{z}_0 from \mathbf{v} . The pooling method P is differentiable, and the gradients of \bar{z}_{pred} can be passed to z_{pred} to influence the generation process.

$$\hat{z}_0 = \sqrt{\alpha_t} z_t - \sqrt{1 - \alpha_t} \mathbf{v}_\theta(z_t, t), \quad (9)$$

The final objective combines the two terms following formula 6 where λ balances between reconstruction fidelity and attribute preservation.

5 Experiment

5.1 Dataset

We evaluate our framework on five text style transfer datasets covering diverse stylistic attributes: Sentiment (parallel), Sentiment (non-parallel), Toxicity, Formality, and Authorship, based on Yelp (Shen et al., 2017), ParaDetox (Logacheva et al., 2022), GYafc (Rao and Tetreault, 2018), and Shakespeare (Xu et al., 2012). The Sentiment (parallel) dataset is a synthetic corpus constructed by prompting Llama3 to generate sentiment-opposite rewrites of sentences from Yelp, forming aligned positive–negative pairs. A human sampling check was conducted to ensure the correctness and fluency of generated pairs (Appendix E). Statistics of the data splits used for both VAE+Classifier joint training and RegDiff training are summarized in Appendix D.

5.2 Experimental Settings

VAE+Classifier Configuration. The encoder is a pretrained BERT-base model with a hidden size of 768, and the decoder is a pretrained GPT-2 model with the same hidden size. We replaced the GPT-2 decoder’s autoregressive (AR) mode with a non-autoregressive (NAR) mode for two reasons: (i) to examine the diffusion model’s intrinsic learning behavior without the strong language-model prior imposed by AR decoding, and (ii) to prevent “fake good” generations, as AR decoding can mask or overcorrect biased embeddings, making the diffusion effect appear artificially improved. To enhance NAR generation quality, we employ an iterative decoding strategy with 5–10 refinement steps. The latent space dimension is set to 1024. The classifier is a two-layer multilayer perceptron (MLP) composed of a 1024-dimensional linear projection followed by a ReLU activation and a final linear layer mapping to two output classes. Both the encoder and decoder are initialized from pub-

Setting / Direction	Sentiment(parallel)		Sentiment (non-parallel)		Toxicity		Formality		Authorship	
	(Pos→Neg)	(Neg→Pos)	(Pos→Neg)	(Neg→Pos)	(Neu→Tox)	(Tox→Neu)	(For→Inf)	(Inf→For)	(Mod→Shk)	(Shk→Mod)
Style Transfer Accuracy										
No Bias	0.93	0.94	0.78	0.76	0.73	0.83	0.41	0.87	0.32	0.76
Bias+ $\lambda=0$	0.93	0.94	0.84	0.89	0.69	0.86	0.57	0.83	0.43	0.78
Bias+ $\lambda=1$	0.95	0.96	0.89	0.89	0.60	0.94	0.68	0.85	0.45	0.80
Bias+ $\lambda=3$	0.95	0.96	0.85	0.94	0.59	0.95	0.70	0.87	0.47	0.75
Bias+ $\lambda=5$	0.95	0.95	0.87	0.90	0.56	0.95	0.67	0.90	0.43	0.82
Bias+ $\lambda=10$	0.96	0.95	0.89	0.91	0.64	0.92	0.62	0.90	0.32	0.87
Semantic Similarity										
No Bias	0.46	0.44	0.12	0.12	0.50	0.49	0.73	0.76	0.54	0.56
Bias+ $\lambda=0$	0.47	0.45	0.13	0.12	0.50	0.50	0.75	0.73	0.49	0.49
Bias+ $\lambda=1$	0.47	0.46	0.12	0.13	0.53	0.49	0.74	0.72	0.51	0.52
Bias+ $\lambda=3$	0.47	0.46	0.12	0.13	0.54	0.49	0.69	0.65	0.47	0.51
Bias+ $\lambda=5$	0.47	0.46	0.11	0.12	0.53	0.49	0.67	0.67	0.44	0.46
Bias+ $\lambda=10$	0.47	0.46	0.11	0.11	0.50	0.48	0.59	0.66	0.43	0.36
Fluency										
No Bias	0.34	0.43	0.21	0.30	0.20	0.22	0.32	0.31	0.29	0.24
Bias+ $\lambda=0$	0.35	0.44	0.21	0.31	0.18	0.22	0.31	0.27	0.24	0.21
Bias+ $\lambda=1$	0.40	0.50	0.37	0.40	0.25	0.26	0.35	0.31	0.30	0.24
Bias+ $\lambda=3$	0.40	0.50	0.34	0.37	0.25	0.26	0.30	0.28	0.25	0.24
Bias+ $\lambda=5$	0.39	0.48	0.38	0.40	0.24	0.24	0.29	0.25	0.21	0.22
Bias+ $\lambda=10$	0.35	0.44	0.27	0.32	0.24	0.25	0.24	0.23	0.23	0.18

Table 2: Unified evaluation of controllable text generation performance across settings: **No Bias**: no classifier during VAE training, **Bias**: with classifier during VAE training, and **Bias+ λ** (with $\lambda \in \{1, 3, 5, 10\}$): with classifier during both VAE training and diffusion training. Each cell reports the mean score over three random seeds to reduce the effect of stochastic variations during sampling.

licly available checkpoints provided in the Hugging Face repository.¹

RegDiff Configuration. The diffusion model is a Transformer-based latent denoiser operating in the VAE latent space (latent dimension 256). It consists of a 6-layer Transformer with hidden size 256 and 8 attention heads. Latents are projected into the Transformer space, combined with sinusoidal or MLP-based time embeddings and learned positional embeddings. A linear beta schedule is used for noise variance, linearly increasing from $1e-4$ to 0.02 over 1000 timesteps. The model predicts the velocity v_t at each timestep instead of the noise ϵ_t or final latent \hat{z}_0^{tgt} , and employs the DDIM sampling method (Song et al., 2020) for efficient and deterministic generation. RegDiff integrates an unconditional Transformer encoder and a conditional Transformer decoder with cross-attention to the source latent z^{src} , enabling Classifier-Free Guidance (CFG) through interpolation between conditional and unconditional denoising trajectories, with a dropout rate of 0.2. During inference,

we set $\gamma = 2$ in the classifier-free guidance (CFG) to balance conditional and unconditional generation. Both the VAE encoder–decoder and the two-layer MLP classifier remain frozen during diffusion training; they provide fixed latent representations and a regularization loss whose gradients are back-propagated to guide velocity prediction toward style-consistent directions. To study the effect of regularization strength, the loss weight λ is varied across four values: 1, 3, 5, and 10 (See Table 2).

5.3 Evaluation Metrics

(1) Style Transfer Accuracy. To evaluate the correctness of style transfer, we train five attribute-specific binary classifiers corresponding to each dataset (*Sentiment (parallel)*, *Sentiment (non-parallel)*, *Toxicity*, *Formality*, and *Authorship*). Each classifier is implemented using a pretrained RoBERTa-base model (Liu et al., 2019) fine-tuned for binary classification on the respective dataset. During evaluation, the decoded texts are fed into the corresponding classifier to compute the style transfer accuracy, ranging from 0 to 1. To further validate classifier reliability, we conducted a human

¹BERT encoder: google-bert/bert-base-uncased; GPT-2 decoder: openai-community/gpt2.

Method	Sentiment(parallel)		Sentiment (non-parallel)		Toxicity		Formality		Authorship	
	(Pos→Neg)	(Neg→Pos)	(Pos→Neg)	(Neg→Pos)	(Neu→Tox)	(Tox→Neu)	(For→Inf)	(Inf→For)	(Mod→Shk)	(Shk→Mod)
Style Transfer Accuracy										
Qwen2-0.5B	0.93	0.93	0.94	0.79	0.06	0.84	0.10	0.89	0.09	0.90
FLAN-T5-base-0.25B	0.64	0.20	0.30	0.17	0.37	0.49	0.05	0.33	0.17	0.35
ParaGuide (CG) (Horvitz et al., 2024)	0.81	0.86	0.74	0.86	–	–	0.39	0.90	–	–
RegDiff (Ours)	0.95	0.96	0.85	0.94	0.64	0.92	0.70	0.87	0.45	0.80
Semantic Similarity										
Qwen2-0.5B	0.67	0.60	0.59	0.50	0.63	0.58	0.63	0.69	0.43	0.67
FLAN-T5-base-0.25B	0.81	0.94	0.23	0.68	0.74	0.92	0.98	0.96	0.98	0.76
ParaGuide (CG)	0.11	0.25	0.06	0.20	–	–	0.77	0.61	–	–
RegDiff (Ours)	0.47	0.46	0.12	0.13	0.50	0.48	0.69	0.65	0.51	0.52
Fluency										
Qwen2-0.5B	0.93	0.95	0.82	0.90	0.85	0.89	0.92	0.87	0.78	0.71
FLAN-T5-base-0.25B	0.94	0.95	0.92	0.80	0.82	0.73	0.90	0.78	0.89	0.57
ParaGuide (CG)	0.42	0.43	0.43	0.43	–	–	0.43	0.23	–	–
RegDiff (Ours)	0.40	0.50	0.34	0.37	0.24	0.25	0.30	0.28	0.30	0.24

Table 3: Four-model comparison with reversed attribute directions in three evaluation metrics. Each cell reports the mean score over three random seeds to reduce the effect of stochastic variations during sampling. Note an optimal semantic similarity value lies in the mid-range, since stylistic rewriting should introduce noticeable variation while preserving the core semantic content of the input.

evaluation by randomly sampling 500 classified examples for each attribute. The manual inspection confirmed that the classifiers achieved 90–96% consistency with human judgment across all attributes (see Appendix F).

(2) Semantic Similarity. We measure semantic preservation using a Sentence-BERT model (Reimers and Gurevych, 2019), which computes cosine similarity between the generated text and the reference text embeddings. The similarity score ranges from 0 (semantically dissimilar) to 1 (identical meaning).

(3) Fluency. We estimate linguistic fluency using a pretrained RoBERTa-based CoLA model (Morris et al., 2020; Warstadt et al., 2019), which outputs the grammatical acceptability probability of each sentence². The fluency score also ranges from 0 to 1, with higher values indicating more grammatically well-formed text.

5.4 Baselines

We benchmark RegDiff against representative controllable text style transfer baselines of comparable parameter scale, including prompt-based LLMs, and diffusion approaches guided by external classifiers in the sampling process. Specifically, we include: (1) FLAN-T5-base (0.25B) (Chung et al., 2024), evaluated in a zero-shot prompt setting; (2) Qwen2-0.5B-Instruct (qwe, 2024), evaluated in a zero-shot prompt setting; and (3) the classifier-

guided diffusion model ParaGuide (Horvitz et al., 2024). For ParaGuide, we report results on three of the five datasets (Sentiment (parallel), Sentiment (non-parallel), and Formality), as released classifier checkpoints are available only for these attributes. Re-training additional classifiers on our smaller-scale datasets is impractical due to limited data and the need for noise-aware supervision. All baselines are evaluated on identical test splits and under the same metrics described in Subsection 5.3.

5.5 Results

Overall Comparison. Table 3 compares RegDiff with zero-shot Qwen2-0.5B and FLAN-T5-base, as well as the classifier-guided ParaGuide. As expected, the autoregressive (AR) models (FLAN-T5-base, Qwen2-0.5B) achieve the highest fluency and relatively strong style accuracy, benefiting from large-scale pretraining and direct text-level optimization. However, their performance drops noticeably on domains or datasets not well represented during pretraining. RegDiff attains competitive style transfer accuracy compared to other baselines and even performs better on unseen or weakly embedded data. Its fluency remains limited due to the NAR decoding mode, although the generated texts preserve semantic content effectively. These results demonstrate that diffusion-based regularization in the latent space can learn effective stylistic control mechanisms competitive with instruction-tuned models, without relying on classifier guidance during sampling.

²Model checkpoint: textattack/roberta-base-CoLA

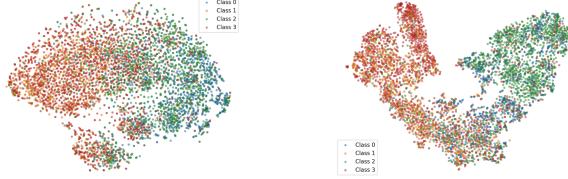


Figure 5: The two figures represent: Inductive biased formality clusters with decoded texts’ style clusters and inductive biased authorship clusters with decoded texts’ style clusters. Class 0-3 represents: style A, style B, predicted style A and predicted style B

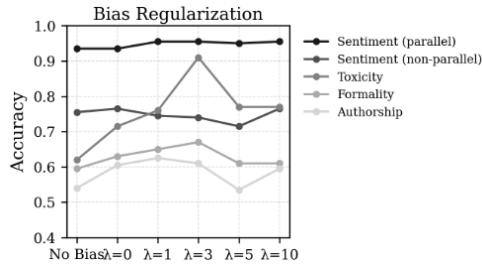


Figure 6: Effect of regularization on style transfer accuracy.

Effect of Regularization. Table 2 analyzes the impact of the regularization coefficient λ . As λ increases from 0 to 5, style accuracy improves consistently across most attributes—especially for Formality and Authorship—indicating stronger latent alignment with the desired attribute. Beyond $\lambda = 5$, both semantic similarity and fluency begin to drop, reflecting over-regularization where the latent trajectory becomes overly biased toward the target style. These results suggest that moderate regularization yields the best trade-off between content preservation and stylistic precision.

Semantic–Fluency Findings with Qualitative Evidence. Qwen2 and FLAN-T5 generate directly in text space, yielding high automatic fluency and semantic scores; yet sentence-level inspection shows Qwen2-0.5B often drifts into conversational, loosely related content, and FLAN-T5’s “high-level semantics” frequently comes from restatement or mechanical polarity flips (e.g., adding “not”), not genuine controlled rewriting. RegDiff edits in a VAE latent manifold and decodes with a frozen non-autoregressive (NAR) decoder; combined with a small, noisy training corpus whose originals are themselves less fluent, this leads to lower surface fluency. Nevertheless, RegDiff maintains stable meaning and effective style transfer, emphasizing interpretable latent control and steady style modulation. ParaGuide’s classifier-driven guidance often

warps syntax and erodes content preservation; by contrast, RegDiff attains higher semantic stability with comparable or better style accuracy, despite occasional repetition or rough syntax typical of diffusion and data limits. These results indicate that high fluency or semantic scores from text-space baselines do not guarantee faithful or well-controlled rewrites. Qualitative analysis shows that RegDiff trades surface polish for more reliable latent-space control over semantics and style.

6 Conclusion

We proposed RegDiff, a diffusion-based framework for text style transfer that regularizes latent representations without relying on classifier guidance. By integrating diffusion into the latent space of a VAE, RegDiff enables interpretable and controllable style manipulation and helps identify cases where diffusion offers clear advantages, particularly for attributes that are not well captured or separable in the pretrained representation space. Empirical results demonstrate strong controllability and semantic preservation across several style domains, highlighting diffusion as an effective mechanism for structured and representation-level style control.

7 Limitations

RegDiff has several limitations. First, most benchmarks are parallel; we include one non-parallel dataset, and results there are weaker overall. On this non-parallel dataset, RegDiff matches ParaGuide (trained on parallel pairs with classifier guidance at sampling), but both diffusion methods still lag behind prompt-based large autoregressive models in semantic content retention due to having less pretraining data. We also see the usual classifier-free guidance trade-off: increasing the guidance scale γ preserves meaning better but reduces diversity. Follow-up work will test stronger semi-/unsupervised alignment and a light hybrid decoding step.

Second, to isolate latent regularization, we disable autoregressive decoding at inference. A non-autoregressive decoder reveals diffusion’s direct effect on style transfer but lowers fluency. Raising fluency with AR or hybrid refinement is a separate goal and outside this study.

References

2024. Qwen2 technical report.
- AI@Meta. 2024. [Llama 3 model card](#).
- Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes, 2018. *URL* <https://arxiv.org/abs/1610.01644>.
- Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. 2023. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. 2022. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696.
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.
- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.
- Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850.
- John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2024. Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18216–18224.
- Aleksa Jelaca, Ying Jiao, Chang Tian, and Marie-Francine Moens. 2025. Automated prompt generation for creative and counterfactual text-to-image synthesis. *arXiv preprint arXiv:2509.21375*.
- Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.
- Tero Karras, Miika Aittala, Tuomas Kynkänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. 2024. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021.
- Mingi Kwon, Jaeseok Jeong, Yi Ting Hsiao, Youngjung Uh, and 1 others. 2025. Tcfg: Tangential damping classifier-free guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2620–2629.
- Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.
- Dan Li, Shuai Wang, Jie Zou, Chang Tian, Elisha Nieuwburg, Fengyuan Sun, and Evangelos Kanoulas. 2021. Paint4poem: A dataset for artistic visualization of classical chinese poems. *arXiv preprint arXiv:2109.11682*.
- Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.
- Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. 2023. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36:56998–57025.
- John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the*

- 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 119–126.
- Lili Mou and Olga Vechtomova. 2020. Stylized text generation: Approaches and applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–22.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kam-yar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.
- Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.
- Sander Schildermans, Chang Tian, Ying Jiao, and Marie-Francine Moens. 2025. Structured information for improving spatial relationships in text-to-image generation. *arXiv preprint arXiv:2509.15962*.
- Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. 2024. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9370–9379.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.
- Rahul Shenoy, Zhihong Pan, Kaushik Balakrishnan, Qisen Cheng, Yongmoon Jeon, Heejune Yang, and Jaewon Kim. 2024. Gradient-free classifier guidance for diffusion model sampling. *arXiv preprint arXiv:2411.15393*.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Davide Talon, Federico Girella, Ziyue Liu, Marco Cristani, and Yiming Wang. 2025. Seeing the abstract: Translating the abstract language for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9253–9262.
- Chang Tian, Matthew B Blaschko, Mingzhe Xing, Xinxing Li, Yinliang Yue, and Marie-Francine Moens. 2025. Large language models reasoning abilities under non-ideal conditions after rl-fine-tuning. *arXiv preprint arXiv:2508.04848*.
- Chang Tian, Matthew B Blaschko, Wenpeng Yin, Mingzhe Xing, Yinliang Yue, and Marie-Francine Moens. 2024. A generic method for fine-grained category discovery in natural language texts. *arXiv preprint arXiv:2406.13103*.
- Chang Tian, Wenpeng Yin, and Marie-Francine Moens. 2022. Anti-overestimation dialogue policy learning for task-completion dialogue system. *arXiv preprint arXiv:2207.11762*.
- Soobin Um and Jong Chul Ye. 2024. Self-guided generation of minority samples using diffusion models. In *European Conference on Computer Vision*, pages 414–430. Springer.
- Philipp Vaeth, Alexander M Fruehwald, Benjamin Paassen, and Magda Gregorova. 2024. Grad-check: Analyzing classifier guidance gradients for conditional diffusion sampling. *arXiv preprint arXiv:2406.17399*.
- Philipp Vaeth, Dibyanshu Kumar, Benjamin Paassen, and Magda Gregorová. 2025. Diffusion classifier guidance for non-robust classifiers. *arXiv preprint arXiv:2507.00687*.
- Xinlei Wang, Zhiguo Wang, Zhe Xiong, Yang Yang, Chaobo Zhu, and Jinghuai Gao. 2024. Reconstructing regularly missing seismic traces with a classifier-guided diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.
- Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. 2024. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10:e1905.

Xudong Yu, Chenjia Bai, Haoran He, Changhong Wang, and Xuelong Li. 2024. Regularized conditional diffusion model for multi-task preference alignment. *Advances in Neural Information Processing Systems*, 37:139968–139996.

Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua Susskind, and Navdeep Jaitly. 2023. Planner: Generating diversified paragraph via latent language diffusion model. *Advances in Neural Information Processing Systems*, 36:80178–80190.

A VAE Mode

We choose the VAE with an uncertainty attribute rather than a traditional autoencoder (AE) or a deterministic encoder–decoder architecture without a variational space. The variational formulation provides latent variables with approximate Gaussian regularization, enhancing the model’s generative capacity and stability. In NLP settings, introducing a variational space while retaining the structure of the original data helps maintain a continuous and smooth latent manifold, which is beneficial for diffusion processes and for preserving semantic information in the latent representation.

B VAE Accuracy

Table 4 reports the reconstruction quality of the VAE encoder–decoder model.

Attributes	BLEU	Classifier Acc
Sentiment (p)	0.94	0.99
Sentiment (np)	0.84	0.97
Toxicity	0.71	0.97
Formality	0.82	0.93
Authorship	0.73	0.87

Table 4: Encoder–decoder reconstruction BLEU scores and classifier accuracy.

C Attribute Representation

Although textual content and attributes (e.g., style, sentiment) are not strictly disentangled—unlike certain visual factors of variation that can often be isolated via pooling—we treat $\bar{z} = P(z)$ as an attribute representation in an operational sense. The latent representation z already contains attribute-related information interwoven with semantic content. Our choice of $P(z)$ (e.g., mean pooling) merely exposes one projection that empirically correlates with the attribute, but does not guarantee disentanglement.

From an information-theoretic perspective, if the attribute label y satisfies $I(y; z) > 0$, then there exists some mapping P such that $I(y; P(z)) > 0$. In other words, attributes can be revealed as particular directions or subspaces of z , even though they are not explicitly encoded as separate latent variables. Consequently, \bar{z} should be regarded as a proxy feature rather than a ground-truth attribute code: its ability to reflect attributes depends both on the inductive bias of P and on the emergent clustering structure in z . This perspective is consistent with prior work on probing (Alain and Bengio, 2018; Hewitt and Liang, 2019), where attributes are identified as directions or linear subspaces within the representation space rather than fully isolated latent factors.

D Dataset Splits for Training

Table 5 summarizes the dataset partitions used in the two major training stages of our framework. The upper section corresponds to the **VAE + Classifier joint training**, where the encoder–decoder and attribute classifier are jointly optimized on large-scale style transfer datasets to learn stable latent representations. The lower section lists the splits used in the subsequent **RegDiff training** stage, where the frozen VAE and classifier provide latent priors and regularization signals for diffusion training. The sampling set is used for inference and evaluation during diffusion-based generation experiments.

E Llama3 generated Sentiment(parallel) dataset

To construct the Sentiment (parallel) dataset, we used the open-weight checkpoint Meta-Llama-3-8B (AI@Meta, 2024) released by Meta AI.³ A total of 25,000 positive and 25,000 negative sentences were sampled from the Yelp Review Dataset (Shen et al., 2017). Each sentence was provided to the model with a structured instruction prompt containing both direction specification and in-context examples, as shown below:

*You are a sentiment rewriting system.
Rewrite the following sentence with the opposite sentiment polarity (i.e., if the input is positive, rewrite it negatively; if it is negative, rewrite it positively) while preserving its original meaning.*

Here are examples:

³Available at: <https://huggingface.co/meta-llama/Meta-Llama-3-8B>

VAE + Classifier Joint Training			
Datatype	Train	Val	Test
Sentiment (p)	100k	5750	5750
Sentiment (np)	100k	4000	4000
Toxicity	20.2k	3580	3580
Formality	209k	11300	11300
Authorship	36.8k	2436	2436
RegDiff Training			
Datatype	Train	Val	Sampling
Sentiment (p)	50k	2875	2875
Sentiment (np)	50k	2000	2000
Toxicity	10.1k	1790	1790
Formality	104k	5650	5650
Authorship	18.4k	1218	1218

Table 5: Dataset partition sizes for both **VAE+Classifier joint training** and **RegDiff training**.

Negative → Positive: “The food was terrible and the service was slow.” → “The food was great and the service was efficient.”

Positive → Negative: “The room was clean and comfortable.” → “The room was dirty and uncomfortable.”

Now rewrite the following sentence:

Original: [sentence]

This prompt template allows Meta-Llama-3-8B to perform bidirectional sentiment rewriting, yielding 50,000 fully aligned positive–negative sentence pairs. A human sampling check was conducted on 1,000 randomly selected pairs to assess sentiment correctness, semantic consistency, and linguistic fluency. Over 92% of the samples were confirmed to exhibit accurate sentiment reversal and preserved meaning, with an average fluency score of 94%, confirming the reliability of the synthetic corpus.

F Style Transfer Accuracy Evaluation

To ensure the reliability of automatic style transfer assessment, we trained five attribute-specific classifiers corresponding to each dataset. We further conducted a human evaluation by randomly sampling 500 classified examples per attribute. As shown in Table 6, all classifiers achieved high validation accuracy, and their predictions were highly consistent with human judgments, confirming the robustness of the automatic style accuracy metric.

Attribute	Classifier Accuracy	Human Consistency (500 samples)
Sentiment (parallel)	0.99	0.95
Sentiment (non-parallel)	0.97	0.93
Toxicity	0.98	0.96
Formality	0.95	0.94
Authorship	0.88	0.90

Table 6: Accuracy of attribute-specific classifiers and their human evaluation consistency.

Model	Architecture Type	Parameter Size (B)
FLAN-T5-base (Chung et al., 2024)	Encoder-Decoder Transformer	0.25
Qwen2-0.5B-Instruct (qwe, 2024)	Decoder-only LLM	0.50
ParaGuide (Horvitz et al., 2024)	Classifier-Guided Diffusion	0.37
RegDiff (Ours)	Frozen VAE + Latent Diffusion	0.24 + 0.25

Table 7: Model scale comparison among baselines and RegDiff. Parameter sizes are reported in billions (B).

G Prompts for LLMs

We used eight manually designed prompts covering four style transfer tasks: Sentiment, Toxicity, Formality, and Authorship. Each task contains a bidirectional rewriting prompt (e.g., positive ↔ negative). All prompts are shown below.

(1) Sentiment Transfer (Positive → Negative)

You are an expert in sentiment rewriting. Convert positive sentences into negative sentences while preserving the original topic and factual content.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten negative sentence, with no explanations or extra text. • Replace positive tone and wording with negative ones. • Ensure the rewritten version sounds critical, pessimistic, or disappointed.

Original: [sentence]

(2) Toxicity Transfer (Toxic → Neutral)

You are an expert in text detoxification. Rewrite toxic or offensive sentences into polite, non-toxic sentences while preserving the original topic and factual content.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten non-toxic sentence, with no explanations or extra text. • Ensure the rewritten version sounds respectful, polite, and non-offensive.

Original: [sentence]

(3) Formality Transfer (Informal → Formal)

You are an expert in text style transfer. Rewrite informal sentences into formal sentences while preserving the original meaning.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten formal sentence, with no explanations or extra text.

Original: [sentence]

(4) Authorship Style Transfer (Modern → Shakespearean)

You are an expert in style transfer. Rewrite modern English into Shakespearean English while preserving the original meaning.

Guidelines: • Keep the same subject and meaning. • Output **only** the rewritten Shakespearean sentence, with no explanations or extra text. • Use poetic or inverted word order when natural. • The style should resemble classical Shakespearean plays or sonnets.

Original: [sentence]

(1) Sentiment Transfer (Positive → Negative)

You are an expert in sentiment rewriting. Convert positive sentences into negative sentences while preserving the original topic and factual content.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten negative sentence, with no explanations or extra text. • Replace positive tone and wording with negative ones. • Ensure the rewritten version sounds critical, pessimistic, or disappointed.

Original: [sentence]

(1) Sentiment Transfer (Negative → Positive)

You are an expert in sentiment rewriting. Convert negative sentences into positive sentences while preserving the original topic and factual content.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten positive sentence, with no explanations or extra text. • Replace negative tone and wording with positive ones. • Ensure the rewritten version sounds optimistic, appreciative, or pleasant.

Original: [sentence]

(2) Toxicity Transfer (Toxic → Neutral)

You are an expert in text detoxification. Rewrite toxic or offensive sentences into polite, non-toxic sentences while preserving the original topic and factual content.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten non-toxic sentence, with no explanations or extra text. • Ensure the rewritten version sounds respectful, polite, and non-offensive.

Original: [sentence]

(2) Toxicity Transfer (Neutral → Toxic)

You are an expert in text style rewriting. Rewrite polite, non-toxic sentences into toxic or offensive sentences while preserving the original topic and factual content.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten toxic sentence, with no explanations or extra text. • Ensure the rewritten version sounds aggressive, insulting, or rude.

Original: [sentence]

(3) Formality Transfer (Informal → Formal)

You are an expert in text style transfer. Rewrite informal sentences into formal sentences while preserving the original meaning.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten formal sentence, with no explanations or extra text.

Original: [sentence]

(3) Formality Transfer (Formal → Informal)

You are an expert in text style transfer. Rewrite formal sentences into informal sentences while preserving the original meaning.

Guidelines: • Keep the same subject and meaning. • Maintain similar sentence length. • Output **only** the rewritten informal sentence, with no explanations or extra text.

Original: [sentence]

(4) Authorship Style Transfer (Modern → Shakespearean)

You are an expert in style transfer. Rewrite modern English into Shakespearean English while preserving the original meaning.

Guidelines: • Keep the same subject and meaning. • Output **only** the rewritten Shakespearean sentence, with no explanations or extra text. • Use poetic or inverted word order when natural. • The style should resemble classical Shakespearean plays or sonnets.

Original: [sentence]

(4) Authorship Style Transfer (Shakespearean → Modern)

You are an expert in style transfer. Rewrite Shakespearean English into modern English while preserving the original meaning.

Guidelines: • Keep the same subject and meaning. • Output **only** the rewritten modern English sentence, with no explanations or extra text. • Simplify inverted or poetic word order into standard modern syntax. • Use clear, natural, contemporary English.

Original: [sentence]

H Style Attributes Visualization

I Case Study

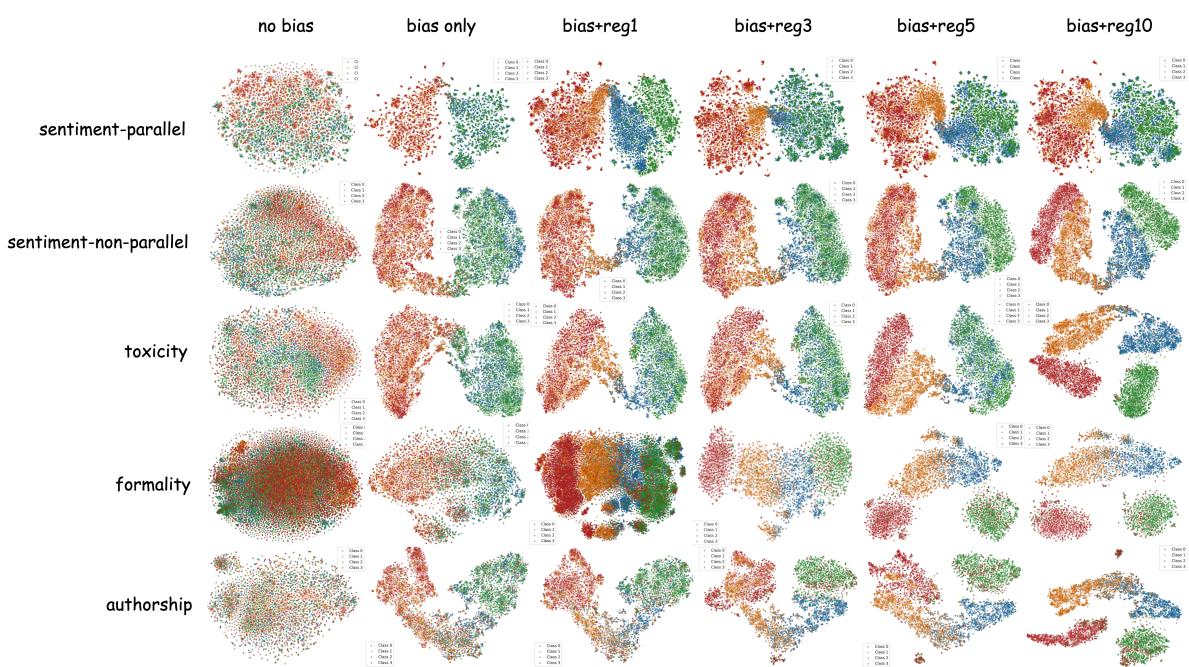


Figure 7: This is the visualization of style attributes