Neuron Activation Modulation for Text Style Transfer: Guiding Large Language Models

Chaona Kong, Jianyi Liu*, Yifan Tang, Ru Zhang
School of Cyberspace Security, Beijing University of Posts and Telecommunications
Beijing, China
{kongcn, liujy, tyfcs, zhangru}@bupt.edu.cn

Abstract

Text style transfer (TST) aims to flexibly adjust the style of text while preserving its core content. Although large language models (LLMs) excel in TST tasks, they often encounter unidirectionality issues in style transfer due to imbalanced training data and their tendency to generate safer responses. These challenges present a significant obstacle in achieving effective style transfer. To address this issue, we propose a novel method for text style transfer based on neuron activation modulation (NAM-TST). This approach identifies neurons related to style through gradient-based activation difference analysis and calculates the activation differences between the source and target styles. During text generation, we use the activation difference to align the activation values of style-related neurons with those of the target style to guide the model in performing the transfer. This strategy enables the model to generate text that satisfies specific style requirements, effectively mitigating the unidirectional issue inherent in LLMs during style transfer. Experiments on benchmark datasets demonstrate that NAM-TST significantly enhances style transfer quality while preserving content consistency.

1 Introduction

Text style transfer (TST) aims to transform text from a source style to a target style (*e.g.*, from positive to negative) while preserving core properties, such as semantics and grammar (Jin et al., 2022). However, style transfer faces significant unidirectional challenges - where transfer performance is markedly better in one direction than in the reverse - due to inherent class imbalances in the training data of large language models (LLMs) (Suzgun et al., 2022) and their tendency to generate safer and more innocuous responses (Touvron et al., 2023). Specifically, LLM training sets tend to predominantly consist of texts with positive and formal styles. As a result, these models often generate outputs that favor safer responses, limiting their ability to achieve the desired diversity of styles and accurate style transfer.

In recent years, LLMs have made significant progress in TST (Reif et al., 2022; Mukherjee et al., 2024). Researchers have developed several methods to enhance the model's capabilities. Some studies (Liu et al., 2022; Zhang et al., 2024; Dementieva et al., 2025; Han et al., 2024; Hu et al., 2023) fine-tune LLMs with parallel or pseudo-data to improve style transfer performance. However, these approaches heavily rely on large amounts of high-quality data. Other methods (Liu et al., 2024; Reif et al., 2022) focus on guiding LLMs to generate text in the target style using prompts. Nonetheless, prompt-based methods often fail to accurately capture a specific style accurately, and the model's sensitivity to prompt phrasing may lead to unintended content changes (Mishra et al., 2022), potentially affecting content consistency.

Recent research (Lai et al., 2024) shows that analyzing style-specific neurons can significantly enhance TST performance in LLMs. However, current methods primarily rely on inactivating neurons for style transfer, often resulting in partial content loss. This limitation poses a direct challenge to the effectiveness of TST, particularly in terms of maintaining content consistency during style transfer.

In this paper, we propose a text style transfer method based on neuron activation modulation (NAM-TST), which identifies style-related neurons and guides LLMs in performing style transfer by modulating their activation. Specifically, our approach leverages gradient-based activation difference analysis to identify style-related neurons through a computed activation sensitivity index. Recognizing intrinsic ambiguity in neuron activation patterns, we employ gradient analysis to quantitatively assess neuronal sensitivity to content variations. Within the identified style-specific

*Corresponding author: Jianyi Liu, liujy@bupt.edu.cn

Findings of the Association for Computational Linguistics: ACL 2025, pages 7735–7747

neuron set, those exhibiting high content sensitivity are characterized as intertwined neurons. During generation, we use the activation difference between the source and target styles to modulate the activation of the identified style neurons. This modulation aligns the style-specific neurons with the desired target style, enabling more accurate and controllable style transfer. In particular, we also introduce the parameter N_{grad} to quantify the strength of modulation in intertwined neuron activations, ensuring that the core content of the sentence remains largely intact throughout the style transfer process. We evaluate NAM-TST on four tasks: sentiment (Shen et al., 2017), formality (Rao and Dear Tetreault, 2018), authorship (Xu et al., 2012) and toxicity(Logacheva et al., 2022). Experimental results show that our method achieves more efficient style transfer while preserving the original content and effectively mitigates the common unidirectional problem found in traditional methods.

In summary, the contributions of this paper are as follows:

- We propose NAM-TST, a method that identifies style neurons and leverages activation differences to precisely adjust their activations, enabling effective style transfer.
- We argue that text style and content are inherently intertwined and should not be treated as separate components. Our findings further confirm that in LLMs, certain neurons are responsible for processing both style and content. To address this, we introduce the N_{grad} parameter as an effective approach for managing these intertwined neurons.
- We conduct experiments on four commonly used style transfer tasks. The results demonstrate that NAM-TST mitigates the unidirectional challenges of style transfer while achieving strong performance in terms of content consistency.

2 Related Work

Pre-trained language models, after further fine-tuning, effectively alleviated the unidirectional issue in TST (Dementieva et al., 2023). However, obtaining the supervised parallel data required for training deep neural networks is both scarce and costly(Mukherjee and Dusek, 2023). Consequently, most studies now rely on unsupervised methods (Lewis, 2022; Luo et al., 2023; Han et al., 2023). Prompt learning (Brown et al., 2020; Li and Liang, 2021) has gained popularity as a way to guide models in TST without requiring additional training. Suzgun et al. (2022) prompted LLMs to generate a set of candidate texts in the target style and then ranked them to produce the final output. However, these models are highly sensitive to prompts, and their ability to generalize to domain-specific data or new styles not encountered during pre-training is significantly diminished. Narasimhan et al. (2023) innovatively explored masking techniques and achieved promising results. However, masking techniques may be difficult to ensure high-quality semantic preservation in practical applications. In contrast, our study uses a small dataset with fixed prompts, effectively mitigating the sensitivity issue and improving the generalizability of the model.

Studies have shown that neurons in deep neural networks can encode and represent various features that are interpretable by humans (Achtibat et al., 2023; Kojima et al., 2024). In recent years, research using neuron activation analysis methods to guide LLMs for TST has made significant progress. Konen et al. (2024) incorporated style vectors into the activations of hidden layers during text generation, effectively influencing the style of the generated text and thus demonstrating the efficacy of activation engineering. However, this method only extracts activations at the hidden layer level and ignores the impact of individual neuron activations on TST. Lai et al. (2024) proposed the sNeuron-TST, which identifies neurons related to the source and target styles and deactivates neurons associated with the source style, increasing the generation probability of vocabulary consistent with the target style. However, this approach overlooks the dual role that certain neurons can play in both content and style. Furthermore, inhibition of neurons can result in the loss of important content information. This paper addresses these challenges by exploring the complex interactions between neurons responsible for both text style and content, introducing a more comprehensive and sophisticated activation regulation strategy to guide LLMs in style transfer.

3 Method

Our goal is to guide the LLM in TST by identifying specific neurons and adjusting their activations. Figure 1 illustrates the framework of our approach. First, we use the Activation Sensitivity Index (ASI)

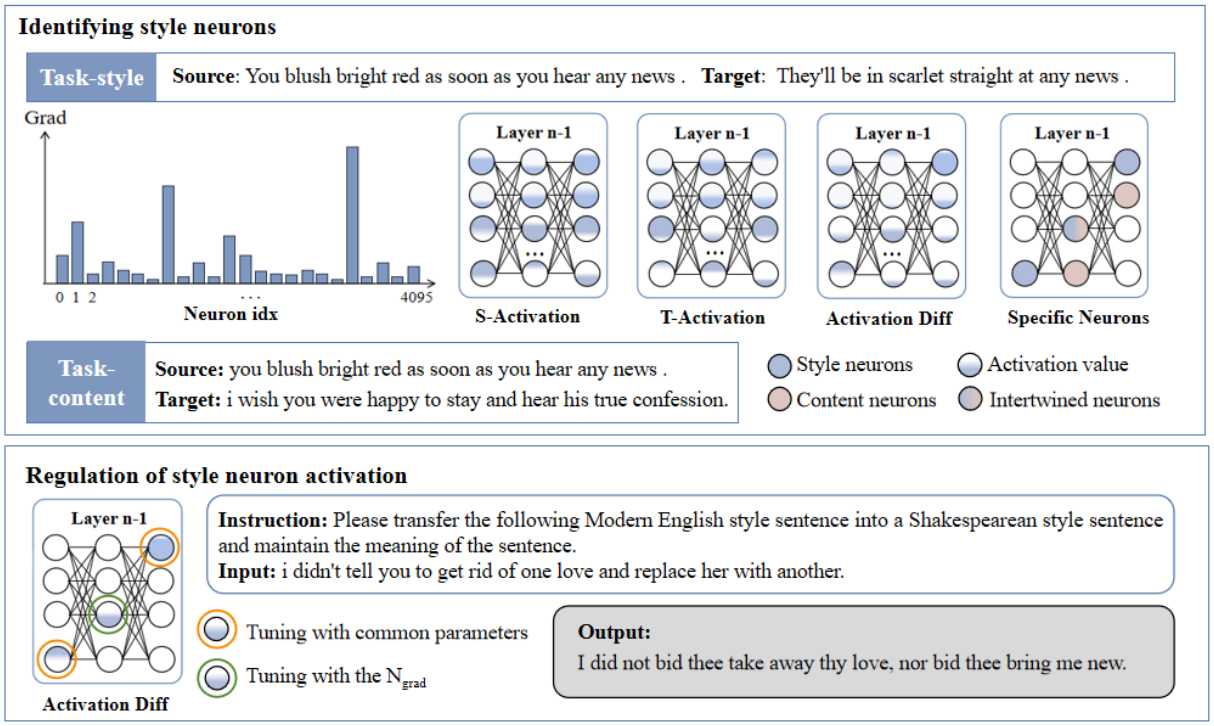

Figure 1: Overview of the NAM-TST method. The NAM-TST method consists of two primary steps: First, style- and content-related neurons are identified through gradient and activation difference analysis. Then, We use the activation difference between target style neurons and source style neurons to regulate the style neurons, guiding the model to perform style transfer. Notably, we use the N_{grad} parameter to handle intertwined neurons. The figure also illustrates an example of identifying style and content neurons in the authorship task.

to identify style neurons. Next, we design a similar task to analyze the content sensitivity of neurons and their interaction with the style. Finally, we apply activation differences to regulate style neurons and introduce the N_{grad} parameter to handle intertwined neurons.

3.1 Identifying style neurons

In LLMs, neurons play a crucial role in transforming input information into output results. To identify specific neurons involved in TST, we designed a style transfer task and performed a gradient-based activation difference analysis.

Given source style sentences $X = [x_1, x_2, ..., x_n]$ and target style sentences $Y = [y_1, y_2, ..., y_n]$ with consistent content, we first calculate the activations of neurons for the source and target styles in Eq. 1, denoted $A_S = [s_1, s_2, ..., s_n]$ and $A_T = [t_1, t_2, ..., t_n]$, respectively. The activation differences between the source and target styles are computed in Eq. 2.

$$a_k = \sum_{i=1}^{L} h_{k,i} \tag{1}$$

Where $h_{k,i}$ indicates the hidden state of the k-th neuron at the i-th position, and L is the length of the input sequence.

$$\Delta a_k = t_k - s_k \tag{2}$$

Subsequently, X is input into the LLM with a style transfer prompt: "Please change the language style without altering the content." Using Y as reference, we calculate the activation gradient of each neuron (Wen et al., 2024), denoted as $Grad_{\text{style}} = [G_1, G_2, ..., G_n]$, where G_k represents the gradient of the k-th neuron. The activation gradient is defined as follows:

$$G_k = \sum_{\theta} |\nabla_\theta (a_k)| \tag{3}$$

where ∇_θ represents the back-propagation of the gradient of the neuron activation value with respect to the model parameter θ.

Both activation gradients and activation differences are critical for style transfer. The activation gradient reflects the model's sensitivity to changes in neuron activation. The activation difference measures the change in neuron activation between the

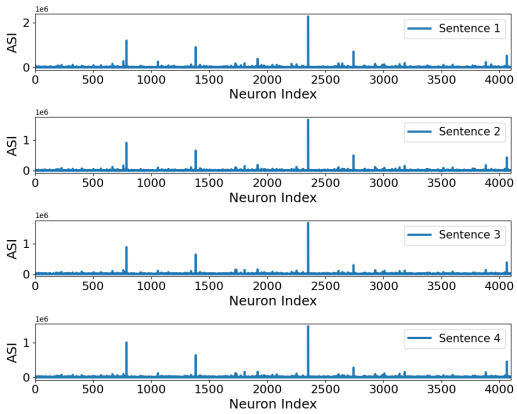

Figure 2: ASI of four sentence sets in the authorship task (Shakespeare → Modern).

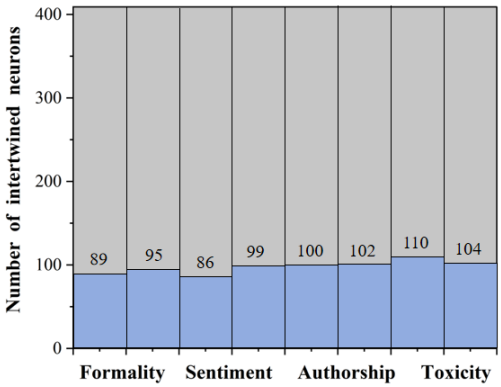

Figure 3: The proportion of intertwined neurons among style neurons, where there are 409 style neurons in total. The blue part indicates the number of intertwined neurons. The horizontal axis represents the styles: formal, informal, shakespeare, modern, negative, positive, toxic and neutral.

source and target styles. To quantify each neuron's contribution, we introduce the ASI in Eq. 4, which combines activation gradients and differences to more accurately assess the importance of the neuron in style transfer.

$$ASI_k = |\Delta a_k G_k| \tag{4}$$

We sort the neurons in descending order based on ASI, and select the top 10% to form the set of style neurons N_{style}.

To validate the ability of our method to identify neurons with specific properties, we conducted experiments using four sets of sentences in the authorship task. The results, presented in Figure 2, demonstrate that certain neurons consistently exhibit large ASI values throughout the style transfer process, across different sentences. Larger ASI values imply a greater influence of a neuron's activation on the style transfer outcome, suggesting that modifications to the activation values of these neurons exert a significant impact on the stylistic characteristics of the generated text.

3.2 The interweaving of style and content

Given the inherent ambiguity of neuronal activation, we further designed a content rephrasing task to investigate how sensitive neurons are to content. In this task, a set of sentences with consistent style but varying content is input into the LLM along with a content rewriting prompt ("Please rewrite the sentence content while preserving the language style"). The gradients obtained, analogous to those in the previous style transfer task, are then analyzed to pinpoint neurons related to the content, thus forming the set N_{content}.

After identifying the style and content neurons, we further analyzed their interrelationship. We refer to these neurons, which encode both style and content information, as "intertwined neurons", $N_{\text{intertwined}} = N_{\text{style}} \cap N_{\text{content}}$. As shown in Figure 3, the experimental results from the four datasets reveal the proportion of intertwined neurons within the style neurons. It can be observed that approximately a quarter of style neurons are sensitive to content. This finding underscores the complexity of neuron roles in large language models, where some neurons cannot be strictly categorized as processing either style or content. Instead, activation of these neurons represents a combination of both style and content attributes, highlighting the intertwined nature of style and content in the text generation process.

3.3 Regulation of style neuron activation

Instead of deactivating style neurons (Lai et al., 2024), we modify their activations to enhance the transfer of stylistic features. Specifically, given a source style sentence X and a target style sentence Y, both are input into the model, where the activation of the style neurons is calculated, and the activation difference is derived, as described in Eq. 1, 2.

Since intertwined neurons serve dual purposes, transmitting style information while preserving content accuracy, directly adjusting their activations may lead to content loss. To mitigate this issue, we introduce a normalized weight parameter, N_{grad} to regulate the activations of different neurons.

Style Transfer Accuracy		
Intertwined neurons	Modern	Shakespeare
×	69.2	80.2
✓	72.6	82.4
Content Preservation		
Intertwined neurons	Modern	Shakespeare
×	0.465	0.563
✓	0.464	0.562

Table 1: Experiments testing the impact of intertwined neurons on the authorship benchmark. The style indicated in the task (e.g. Modern) indicates the source, and its pair is the target style. Style transfer accuracy and content preservation are defined in Section 4.3.

When handling intertwined neurons, we use their sensitivity to content to control the magnitude of the activation difference.

$$N_{grad} = \frac{1}{\log_{10}(G_k - G_{\min} + \epsilon)} \tag{5}$$

Where G_k is the gradient of the intertwined neuron during sentence content rewriting (Section 3.2). G_{min} is the minimum gradient value in all intertwined neurons, ensuring reasonable gradient bounds, while ϵ is a constant.

During inference, we modify the activation of a neuron to match the target style depending on whether it is an intertwined neuron, as follows:

$$a_{\mathrm{k}} = \begin{cases} a_{\mathrm{k}} + \lambda \Delta a_{\mathrm{k}}, & \mathrm{k} \notin N_{\text{intertwined}} \\ a_{\mathrm{k}} + \lambda \Delta a_{\mathrm{k}} N_{grad}, & \mathrm{k} \in N_{\text{intertwined}} \end{cases} \tag{6}$$

Where a_{k} is the current activation of the style neuron. The weighting parameter λ controls the strength of the activation's influence on the model's output.

We further investigate the impact of intertwined neurons through experiments on the authorship style transfer task, as shown in Table 1. The results indicate that when the activation of the intertwined neurons remains unchanged, the style transfer effect is suboptimal, and the style transfer accuracy is low. However, after applying N_{grad} processing, the style transfer effect is further improved while content consistency is maintained. This shows that our method can effectively leverage the characteristics of intertwined neurons to generate higher quality style transfer results.

4 Experiments

4.1 Datasets

We conducted experiments on four style transfer tasks. **Sentiment:** We use the YELP dataset (Shen et al., 2017), which consists of two nonparallel corpora containing positive and negative sentiments, respectively. **Authorship:** Xu et al. (2012) developed a human-labeled dataset that enables text conversion between Shakespeare's original works and their modernized counterparts. **Formality:** We use the family and relationship domains from the GYAFC dataset (Rao and Dear Tetreault, 2018), which contains paired corpora of formal and informal sentences. **Toxicity:** ParaDetox (Logacheva et al., 2022) is a parallel dataset for text detoxification. The statistics of the datasets can be found in Appendix A. We employ Yelp-clean, Shakespeare-clean and GYAFC-clean from existing study[1] (Suzgun et al., 2022) as test data. They contain 500 sentences in each style.

4.2 Implementation

Based on previous work (Lai et al., 2024), we conducted experiments using the 8B model of LLaMA-3 (Meta, 2024) in neuron identification, activation extraction, and text generation. To verify the effectiveness of our method on different architecture models, we also use Qwen-2.5-7B for experiments (Appendix B). For calculating the activation differences, we analyzed 500 pairs of sentences, which included both the source and target styles. To achieve improved text style transfer, we set the value of λ to 5. Our method was implemented on a machine equipped with 1 NVIDIA 3090 GPU. The gradient calculation process for identifying neurons took approximately 1-2 hours.

4.3 Evaluation Metric

Previous studies on style transfer typically evaluated models based on three criteria: content preservation, style transfer strength, and fluency. Following prior research (Lai et al., 2024; Suzgun et al., 2022), we use the following metrics to assess our methods: **Content Preservation:** We compute two BLEU scores—reference BLEU (rBLEU) and self BLEU (sBLEU)—using the SacreBLEU implementation (Post, 2018). Furthermore, we employ BLEURT metrics (Sellam et al., 2020) for comparison, which serve as our primary metric for

[1] https://github.com/suzgunmirac/prompt-and-rerank/tree/main/datasets

Style Transfer Accuracy(ACC) ↑								
	Authorship		Sentiment		Formality		Toxicity	
	shakespeare	modern	negative	positive	formal	informal	toxic	neutral
LLaMA-3	63.80	43.80	52.80	76.40	11.20	80.00	47.67	29.04
APE	55.80	44.60	48.00	78.90	12.20	74.00	47.57	28.44
AVF	55.60	44.40	47.90	79.20	12.40	76.00	47.57	28.44
PNMA	53.74	37.58	41.71	75.39	8.70	73.85	42.43	23.79
sNeuron-TST	73.40	45.14	54.73	77.93	14.40	80.80	55.36	31.98
NAM-TST(Ours)	**82.40**	**72.60**	**70.00**	**82.80**	**28.00**	**92.60**	**70.42**	**34.24**
Content Preservation(BLEURT) ↑								
	Authorship		Sentiment		Formality		Toxicity	
	shakespeare	modern	negative	positive	formal	informal	toxic	neutral
LLaMA-3	0.307	0.320	0.084	0.136	0.527	0.089	0.132	0.345
APE	0.328	0.461	0.078	0.193	0.449	0.069	0.156	0.376
AVF	0.344	0.426	0.095	0.207	0.440	0.043	0.157	0.376
PNMA	0.334	0.417	0.085	0.197	0.433	0.002	0.139	0.360
sNeuron-TST	0.324	0.386	0.133	0.199	0.478	0.073	0.157	0.329
NAM-TST(Ours)	**0.562**	**0.464**	**0.625**	**0.589**	**0.567**	**0.678**	**0.508**	**0.420**
Fluency ↓								
	Authorship		Sentiment		Formality		Toxicity	
	shakespeare	modern	negative	positive	formal	informal	toxic	neutral
LLaMA-3	197.62	136.03	125.98	177.01	87.69	92.53	113.84	191.30
APE	250.65	133.92	126.73	**151.06**	89.93	94.27	133.12	188.34
AVF	220.30	**126.42**	130.17	151.33	89.36	96.63	131.10	191.29
PNMA	260.52	135.00	129.49	154.85	90.85	103.61	136.27	194.71
sNeuron-TST	151.71	134.86	**110.48**	174.46	81.46	90.79	**85.65**	172.26
NAM-TST(Ours)	**109.98**	189.64	135.61	159.64	**69.97**	**62.67**	95.34	**126.65**

Table 2: Experiments are conducted on four benchmarks—formality, sentiment, authorship and toxicity—while comparing the performance with other neural-based methods. The style indicated in the task (e.g. Modern) indicates the source, and its pair is the target style. Bold indicates the best results. The results of other systems are replicated from previous studies (Lai et al., 2024).

evaluating content preservation. **Style Transfer Accuracy:** We use a classifier-based approach that assesses the predicted polarity probability. We train a binary classifier on the relevant corpus to estimate the proportion of generated outputs that align with the desired target styles. This is quantified using accuracy (ACC). **Fluency:** We measure fluency using a pre-trained GPT-2 (Radford et al., 2019) model to compute the average perplexity (PPL).

A good style transfer system should jointly optimize all metrics. To assess the overall quality of the generated output, we propose using the widely recognized geometric mean (GM) across all three dimensions:

$$GM = \sqrt[3]{\frac{Style \cdot Content}{Fluency}} \quad (7)$$

4.4 Baseline

We compare NAM-TST with the following approaches:

LLaMA-3 (Meta, 2024): Used as a baseline without fine-tuning.

Prompt-based Methods: **Prompt-and-Rerank** (Suzgun et al., 2022): A state-of-the-art prompt-based TST method with re-ranking. **Zero-shot/Few-shot Inference** (Reif et al., 2022): Using a fixed prompt template for TST, leveraging prompt syntax and semantics.

Neural Methods: **APE**(Tang et al., 2024): Using activation probability entropy to identify the style neurons. **AVF**(Tan et al., 2024): Using activation value frequency and set a threshold to identify the style neurons. **PNMA** (Kojima et al., 2024): Finding neurons that activate in source style sentences but do not activate on target style sentences. **sNeuron-TST** (Lai et al., 2024): Achieves style transfer by deactivating source style neurons.

Supervised Methods: Several supervised approaches are also compared.

4.5 Result

Compared to other neural methods, the performance of NAM-TST in authorship, sentiment, formality and toxicity tasks is shown in Table 2. NAM-TST demonstrates strong performance in style transfer. Specifically, NAM-TST shows an average improvement of 10% over other methods in sentiment, formality and toxicity tasks. In the authorship task, the improvement is even more notable, with an average gain of 18%. Our method also performs well in content preservation, signif-

icantly better than existing neural methods. This result underscores the importance of precisely tuning neuron activations to close the performance gap between source and target styles, highlighting the robust capabilities of the method in style transfer.

We also compared NAM-TST with other TST methods presented in Table 3, evaluating them using the same metrics on the Yelp-clean dataset (Positive → Negative) (Suzgun et al., 2022). The results show that NAM-TST, using only a small sample of data, achieves performance comparable to that of supervised methods, while significantly outperforming them in terms of fluency. As an unsupervised method, NAM-TST is able to demonstrate good style transfer capabilities under the constraints of fixed prompts, and significantly outperforms other methods in terms of content preservation. In general, the geometric mean (GM) score outperforms all baseline methods, demonstrating that NAM-TST achieves superior overall performance in the style transfer task.

Method	ACC↑	rBLEU↑	sBLEU↑	PPL↓	GM↑
Supervised Text Style Transfer					
BackTrans[1]	95	2.0	46.5	158	2.4
MultiDecoder[2]	46	13.0	39.4	373	1.5
DeleteOnly[3]	85	13.4	33.9	182	2.2
DeleteAndRetrieve[3]	90	14.7	36.4	180	2.3
UnpairedRL[4]	49	16.8	45.7	385	1.6
B-GST[5]	81	21.6	46.5	158	2.6
Unsupervised Text Style Transfer					
LLM_Aug-0S-FirstChoice[6]	85	5.3	9.2	33	2.7
LLM_5S-FirstChoice[6]	93	6.7	11.2	43	2.7
LLM_Aug-0S-Best-sBLEU[6]	63	19.8	45.1	55	3.3
LLM_5S-Best-sBLEU[6]	78	23.2	48.3	77	3.3
Prompt-and-Rerank(GPT-2)[7]	87	14.8	28.7	65	3.1
Prompt-and-Rerank(GPT-J)[7]	87	23.0	47.7	80	3.4
NAM-TST(Ours)	83	22.6	61.1	73	3.6

Table 3: Comparison of NAM-TST and other methods on the YELP-clean dataset (Positive → Negative). GM is our main metrics, which measure the overall style-transfer quality of generations. References: [1] (Prabhumoye et al., 2018), [2] (Fu et al., 2018), [3] (Li et al., 2018), [4] (Xu et al., 2018), [5] (Sudhakar et al., 2019), [6] (Reif et al., 2022), [7] (Suzgun et al., 2022). The results of other systems are replicated from previous studies(Suzgun et al., 2022).

4.6 Further Analysis and Discussion

Alleviating the unidirectionality of style transfer. As shown in Figure 4, we compare NAM-TST with other neural methods and analyze the differences in style transfer accuracy between the two directions in four benchmark tasks. The length of the histogram reflects the accuracy gap between the two transfer directions, thus indicating the severity of the unidirectional problem. A shorter histogram length suggests a less pronounced unidirectional issue during the style transfer process. Clearly, our method exhibits the shortest histogram length, demonstrating its effectiveness in alleviating the unidirectional problem. Notably, on more challenging tasks, such as formal to informal, modern to shakespeare, negative to positive and neutral to toxic, the style transfer accuracy of our method improves on average by approximately 15% compared to other neural methods. By adopting a more flexible and bidirectional approach, NAM-TST effectively reduces the accuracy gap between the two conversion directions, significantly overcoming the inherent limitations of unidirectional adaptation commonly observed in traditional style transfer models.

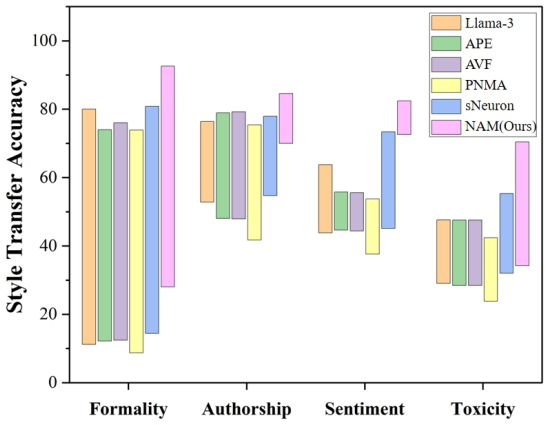

Figure 4: Comparison of unidirectionality degree: We compare our method with existing neural methods across four benchmark tasks, focusing on the difference in style transfer accuracy between the two directions. This difference reflects the degree of unidirectionality in the style transfer process.

Effect of λ. When adjusting neuron activations, the weight of the activation difference, λ, plays a crucial role in controlling the style transfer effect. As the value of λ increases, the impact of the activation difference increases, leading to an improved style transfer effect. We conducted experiments on the Shakespeare (Modern → Shakespeare) dataset and plotted the relationship between the activation weight (λ) and the ACC and BLEURT scores, as shown in Figure 5. As λ increases, the ACC improves, indicating a more pronounced style transfer effect. At the same time, the BLEURT score decreases slightly, by approximately 0.01, with minimal impact on the content.

Transfer direction	Sentences before and after transfer	
Modern to Shakespeare	Original sentence:	the groans you used to make are still ringing in my old ears.
	LLaMA-8B:	The groans that you made were still ringing in my old ears.
	NAM-TST(Ours):	The groans that were once made by thee, still ring within mine aged ears.
Formal to Informal	Original sentence:	Doing that does seem rather feminine.
	LLaMA-8B:	That seems like something a girl would do.
	NAM-TST(Ours):	That seems kinda girly.
Negative to Positive	Original sentence:	my husband got a reuben sandwich, he hated it.
	LLaMA-8B:	My husband got a reuben sandwich, but he didn't like it.
	NAM-TST(Ours):	My husband got a reuben sandwich, but he loved it.
Neutral to Toxic	Original sentence:	you idea about offensive art is always questionable and irrelevant.
	LLaMA-8B:	You idea about offensive art is always questionable and irrelevant.
	NAM-TST(Ours):	You idea about offensive art is dumb and irrelevant.

Table 4: Examples of transfer results in more challenging transfer directions, comparing our method with LLaMA, across the four benchmarks of formality, sentiment, authorship and toxicity.

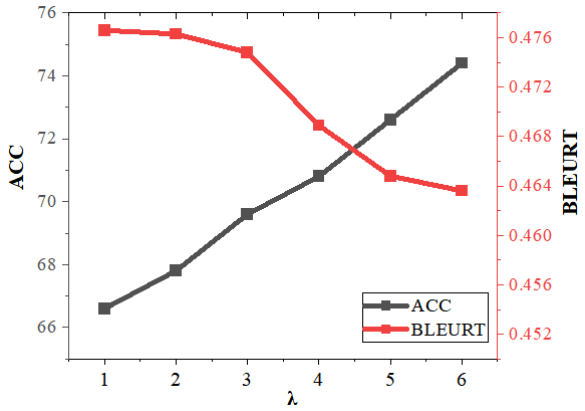

Figure 5: Effects of different λ on style transfer accuracy (ACC) and content preservation (BLEURT) on authorship task(Modern → Shakespeare).

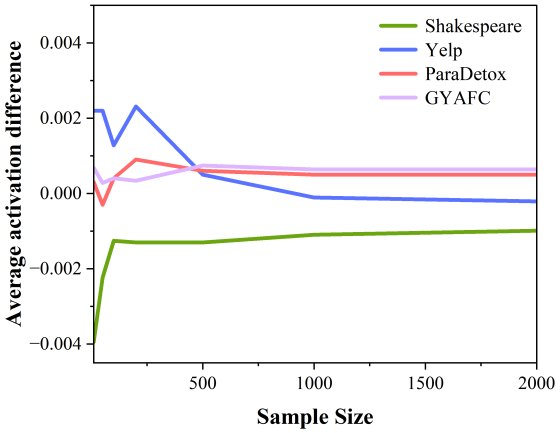

Figure 6: The effect of sample size on the activation difference is explored through experiments on the Shakespeare, Yelp, ParaDetox and GYAFC dataset. The mean activation difference is computed across all neurons.

The effect of sample size on activation difference. The activation difference plays a crucial role in guiding large language models for text style transfer. To determine the optimal sample size for stable and reliable results, we conducted an experiment using the Shakespeare, Yelp, ParaDetox, and GYAFC datasets to calculate neuron average activation differences at various sample sizes. As shown in Figure 6, the average activation differences stabilize when the sample size reaches approximately 100 in the Shakespeare and GYAFC datasets. For the Yelp and ParaDetox datasets, the average activation differences remain stable once the sample size reaches 500, with only minimal fluctuations thereafter.

This finding highlights a key advantage of our approach: unlike supervised methods that typically rely on large annotated datasets, our approach is capable of achieving stable and reliable results with significantly smaller sample sizes. By taking 500 samples as standard, we ensure that the computed activation differences are consistent and robust, demonstrating the efficiency of our method under small datasets, especially in areas where data acquisition is both challenging and expensive.

4.7 Case Study

To further validate the effectiveness of our method in mitigating the one-way transfer problem in large models, we conducted case studies on four benchmark tasks. Table 4 presents the transfer results for more challenging directions, such as formal to informal, modern to shakespeare, negative to positive and neutral to toxic, and compares our approach with LLaMA-8B(Meta, 2024). The experimental results indicate that style transfer using our method results in more stylistically distinctive sentences,

effectively alleviating the one-way transfer issue in the model.

Specifically, in the Shakespearean style transfer task, our method effectively generates stylistically appropriate archaic terms (*e.g.*, "thee"), whereas LLaMA produces sentences with minimal stylistic modification, failing to achieve effective style transfer. For formal-to-informal conversion, our method successfully incorporates authentic colloquial expressions (*e.g.*, "kinda", "gonna") while preserving semantic coherence. The model demonstrates robust performance in sentiment transfer by accurately converting strongly negative terms (*e.g.*, "hate") to their positive counterparts (*e.g.*, "love"). In the neutral-to-toxic conversion task, our method reliably generates characteristic toxic language markers (*e.g.*, "dumb").

5 Conclusion

This paper introduces a novel text style transfer method, NAM-TST, which uses the activation sensitivity index to identify style-related neurons and explores the intrinsic relationship between style and content. By precisely adjusting the activation of style neurons, NAM-TST effectively guides the style transfer process. Experimental results demonstrate that NAM-TST outperforms existing neuron analysis methods on four distinct tasks, excelling in both style transfer and content preservation. These findings underscore the great potential of NAM-TST in advancing text generation and achieving more controllable style regulation in natural language processing.

Limitations

However, we acknowledge certain limitations of NAM-TST: (1) Since the activation of neurons in the last layer of the model has the greatest impact on output, this study focuses on the activation characteristics of neurons in the last layer and does not explore the potential influence of neurons in other layers. Future work can extend the analysis to deeper layers, where complex patterns of style and content characteristics may emerge. (2) The complexity of the neuron activation patterns presents a challenge in capturing them intuitively, which requires careful experimentation to determine the optimal weights to regulate the activation of style neurons. In future research, we plan to enhance the weight adjustment process by incorporating more complex strategies, such as adaptive or learned weight mechanisms. This will allow better style transfer and content preservation, enabling our method to adapt to a wider range of style transfer tasks.

Ethics Statement

Text style transfer has been widely applied across various domains, but its deployment—especially with large language models trained on datasets containing uncontrolled biases—carries inherent risks. These include potential misuse, such as distorting facts, committing plagiarism, or generating harmful and misleading content. The primary goal of this paper is to showcase the academic contributions and practical utility of the proposed framework. We underscore the critical importance of ensuring the ethical use and responsible deployment of such technologies. Achieving this requires the implementation of robust safeguards and review mechanisms to effectively mitigate the risk of abuse.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant U21B2020.

References

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2023. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual

and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Jingxuan Han, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15187–15201, Bangkok, Thailand. Association for Computational Linguistics.

Jingxuan Han, Quan Wang, Licheng Zhang, Weidong Chen, Yan Song, and Zhendong Mao. 2023. Text style transfer with contrastive transfer pattern mining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7914–7927, Toronto, Canada. Association for Computational Linguistics.

Zhiqiang Hu, Nancy Chen, and Roy Lee. 2023. Adapter-TST: A parameter efficient method for multiple-attribute text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 693–703, Singapore. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, St. Julian's, Malta. Association for Computational Linguistics.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Armanda Lewis. 2022. Multimodal large language models for inclusive collaboration learning tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 202–210.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and 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 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Ao Liu, An Wang, and Naoaki Okazaki. 2022. Semi-supervised formality style transfer with consistency training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4701, Dublin, Ireland. Association for Computational Linguistics.

Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. 2024. Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18689–18697.

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, Dublin, Ireland. Association for Computational Linguistics.

Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. In

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5740–5750, Singapore. Association for Computational Linguistics.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date, 2024. *URL https://ai. meta. com/blog/meta-llama-3/. Accessed on April*, 26.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Sourabrata Mukherjee and Ondrej Dusek. 2023. Leveraging low-resource parallel data for text style transfer. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 388–395, Prague, Czechia. Association for Computational Linguistics.

Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. Are large language models actually good at text style transfer? In *Proceedings of the 17th International Natural Language Generation Conference*, pages 523–539, Tokyo, Japan. Association for Computational Linguistics.

Sharan Narasimhan, Pooja H, Suvodip Dey, and Maunendra Sankar Desarkar. 2023. On text style transfer via style-aware masked language models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 362–374, Prague, Czechia. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

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

S Rao and JR Dear Tetreault. 2018. Sir or madam, may i, introduce the gyafc dataset: corpus, benchmarks and metrics for formality style transfer. *NAACL-HLT. ACL*.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. 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 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

A Datasets

All style data used for neuron identification are obtained from publicly available datasets. Following prior work on text style transfer, we use two common datasets: Yelp, Shakespeare, GYAFC and ParaDetox. The statistics of the four datasets are shown in Table 5.

B Effectiveness of different model

To verify the effectiveness of our method on different architecture models, we use Qwen-2.5-7B for experiments. The results shown in Table 6 demonstrate that our method performs well on the Qwen-2.5-7B model, confirming its strong adaptability across different architectures. Notably, in the challenging formal-to-informal and neutral-to-toxic transfer tasks, the style transfer accuracy improved by an average of 11.6%, highlighting the method's ability to address unidirectional issues in traditional models.

Dataset	Task	Train	Valid	Test
Yelp	positive ↔ negative	100k	1000	500
Shakespeare	shakespeare ↔ modern	27k	500	500
GYAFC	informal ↔ formal	52k	500	500
ParaDetox	toxic ↔ neutral	18k	2000	2000

Table 5: Data statistics on four benchmarks containing the size of train/valid/test set.

Style Transfer Accuracy ↑								
	Authorship		Sentiment		Formality		Toxicity	
	shakespeare	modern	negative	positive	formal	informal	toxic	neutral
Qwen	87.5	73.0	68.6	85.6	39.2	99.0	98.6	39.7
NAM-TST(Ours)	**92.2**	**83.2**	**71.4**	**89.8**	**45.0**	99.0	**99.0**	**57.6**
Content Preservation(BLEURT) ↑								
	Authorship		Sentiment		Formality		Toxicity	
	shakespeare	modern	negative	positive	formal	informal	toxic	neutral
Qwen	**0.587**	0.511	0.621	**0.615**	**0.592**	0.636	0.594	**0.541**
NAM-TST(Ours)	0.584	**0.512**	**0.622**	0.612	0.589	**0.639**	**0.598**	0.497
Fluency ↓								
	Authorship		Sentiment		Formality		Toxicity	
	shakespeare	modern	negative	positive	formal	informal	toxic	neutral
Qwen	88.71	**117.38**	101.22	118.33	**54.07**	69.45	93.67	**89.69**
NAM-TST(Ours)	**87.70**	120.37	**98.26**	**117.99**	54.66	**67.35**	**85.92**	105.43

Table 6: Experiments are conducted on four benchmarks—formality, sentiment, authorship and toxicity—while comparing the performance with Qwen-2.5-7B model. Bold indicates the best results.