ParaGuide: Guided Diffusion Paraphrasers for Plug-and-Play Textual Style Transfer

Zachary Horvitz[1], **Ajay Patel**[2], **Chris Callison-Burch**[2], **Zhou Yu**[1], **Kathleen McKeown**[1]

[1] Columbia University
[2] University of Pennsylvania
zfh2000@columbia.edu, ajayp@seas.upenn.edu, ccb@seas.upenn.edu, zy2461@columbia.edu, kathy@cs.columbia.edu

Abstract

Textual style transfer is the task of transforming stylistic properties of text while preserving meaning. Target "styles" can be defined in numerous ways, ranging from single attributes (e.g. formality) to authorship (e.g. Shakespeare). Previous unsupervised style-transfer approaches generally rely on significant amounts of labeled data for only a fixed set of styles or require large language models. In contrast, we introduce a novel diffusion-based framework for general-purpose style transfer that can be flexibly adapted to arbitrary target styles at inference time. Our parameter-efficient approach, PARAGUIDE, leverages paraphrase-conditioned diffusion models alongside gradient-based guidance from both off-the-shelf classifiers and strong existing style embedders to transform the style of text while preserving semantic information. We validate the method on the Enron Email Corpus, with both human and automatic evaluations, and find that it outperforms strong baselines on formality, sentiment, and even authorship style transfer.

Introduction

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2022) were originally popularized for image synthesis (Nichol and Dhariwal 2021; Saharia et al. 2022). More recently, however, diffusion has been successfully applied to text. Diffusion-based language models are increasingly competitive with traditional approaches for text generation (Li et al. 2022; Gulrajani and Hashimoto 2023; Han, Kumar, and Tsvetkov 2023; Han et al. 2023), and on text-to-text modeling tasks (Mahabadi et al. 2023; Yuan et al. 2023).

A key benefit of diffusion language models for text is their high degree of controllability. Diffusion-based approaches hierarchically denoise a continuous representation of an entire sequence, and this process can be effectively guided with gradient-based methods (Li et al. 2022; Han, Kumar, and Tsvetkov 2023; Gulrajani and Hashimoto 2023). This differs from the dominant approach of *autoregressive* decoding, where text is generated by sequentially sampling tokens. Steering pretrained autoregressive models has proven difficult, as their text is greedily decoded and guidance must operate on partial sequences (Li et al. 2022; Dathathri et al. 2020; Krause et al. 2020; Yang and Klein 2021). We leverage the controllability of nascent text diffusion methods and adapt them to *style transfer*.

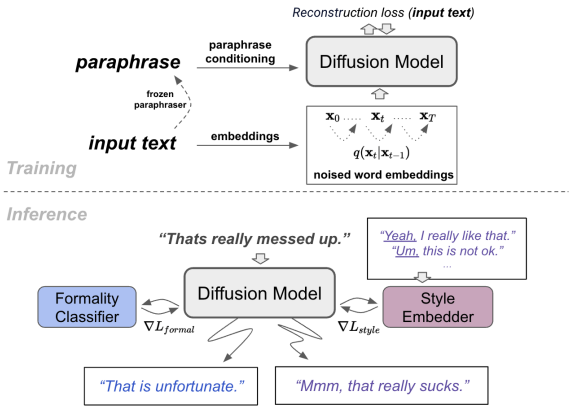

Figure 1: We train paraphrase-conditioned text diffusion models to reconstruct semantically consistent text from noised word embeddings. At inference time, we guide the reconstruction towards target styles with off-the-shelf models.

In textual style transfer, the objective is to transform the style of the text to exhibit an attribute (such as "formality"), or a target author's style, while preserving meaning (Jin et al. 2022; Krishna, Wieting, and Iyyer 2020; Patel, Andrews, and Callison-Burch 2022). The scarcity of style-transfer datasets has motivated *unsupervised* style transfer approaches that perform attribute and authorship style transfer without paired data. These approaches generally require retraining for new target styles.

In contrast, we introduce a plug-and-play diffusion framework for unsupervised style transfer.[1] We initially train a text diffusion model to reconstruct semantically consistent text from paraphrases, but at inference time, we perform new attribute or authorship style transfers by guiding reconstruction with gradients from off-the-shelf models (Figure 1). This allows users to leverage the numerous text classifiers

[1] Our code is publicly available at https://github.com/zacharyhorvitz/ParaGuide.

on platforms like Hugging Face[2] to specify target styles. Beyond guidance from classifiers, our method enables bringing recent advances in representation learning to bear by "plugging in" authorship representations like Style Embeddings (Wegmann, Schraagen, and Nguyen 2022) and Universal Authorship Representations (Rivera-Soto et al. 2021). This enables our approach to perform challenging tasks like *low-resource authorship style transfer* (Patel, Andrews, and Callison-Burch 2022).

Our contributions are as follows:

1. We propose a novel framework for textual style transfer based on paraphrase-conditioned diffusion models, PARAGUIDE.
   - Unlike existing style-transfer approaches, this framework enables gradient-based guidance using off-the-shelf models at inference time.
   - Beyond classifier guidance, we show that existing authorship representations can be plugged in for control. Even with limited available data, this allows PARAGUIDE to competitively perform authorship style transfer.
   - Style transfer requires balancing style-transfer accuracy with fluency and meaning preservation. Our framework enables explicit control over this trade-off through varying guidance strength (λ).
2. We validate our approach on formality and sentiment transfer, where it outperforms strong baselines on automatic evaluations. Additionally, we perform a human evaluation for formality transfer.
3. PARAGUIDE represents early work exploring the promising benefits afforded by text diffusion models. To our knowledge, we are the first to adapt these approaches to unsupervised textual style transfer.

Related Work

Other unsupervised transfer approaches, like STRAP, create pseudo-parallel corpora by corrupting texts to remove stylistic attributes, then training models to reconstruct the uncorrupted text (Krishna, Wieting, and Iyyer 2020; Riley et al. 2021; Ma et al. 2020). These approaches cannot use new stylistic representations without retraining and do not incorporate control from off-the-shelf models. Additionally, STRAP has been shown to require large amounts of style-specific training data (Patel, Andrews, and Callison-Burch 2022). Prior work has explored applying controllable text generation techniques to style transfer (Dale et al. 2021; Kumar et al. 2021; Mireshghallah, Goyal, and Berg-Kirkpatrick 2022). Our approach is most similar in spirit to Mireshghallah, Goyal, and Berg-Kirkpatrick (2022). Their approach is also learning free and non-autoregressive, but performs a discrete search which is very computationally expensive for long sequences, cannot leverage the rich information in gradients, and confines the search space at each step to token-level substitutions. Recently, the emergent ability of Large Language Models (LLMs) to perform in-context learning (Brown et al. 2020) has presented formidable baselines for text generation and style transfer (Reif et al. 2022; Patel, Andrews, and Callison-Burch 2022). Unlike LLMs, PARAGUIDE allows gradient-based control and can leverage stylistic embeddings, and is not restricted to brittle guidance through text-based prompts. Moreover, these approaches typically require models with billions of parameters (Radford et al. 2019).

[2]https://huggingface.co/models

ParaGuide

Overview

PARAGUIDE has three primary steps:

1. Generating an **initial paraphrase** of an input text with an autoregressive (AR) model.
2. Using a paraphrase-conditioned text **diffusion model** to iteratively reconstruct the input text from this paraphrase over a number of diffusion steps.
3. At each diffusion step, computing gradients for arbitrary differentiable losses, and using these gradients for **guidance** towards a target style.

Here, we first use paraphrasing to generate an intermediate text that is semantically consistent with the input text but without the original stylistic attributes (Krishna, Wieting, and Iyyer 2020). We then reconstruct the text with a paraphrase-conditioned diffusion model. During reconstruction, we optimize some loss function specified by a guidance model (Han, Kumar, and Tsvetkov 2023; Li et al. 2022; Gulrajani and Hashimoto 2023). The result is a semantically consistent output in the desired target style.

Initial Paraphrase Generation

At both training and inference time, PARAGUIDE requires *(paraphrase, original text)* pairs. To generate this synthetic data, we leverage an existing, publicly available model (Zhang et al. 2020), specifically fine-tuned for paraphase generation. We include additional information describing this procedure in our Appendix. This aspect of our approach distills performant, but less controllable, autoregressive paraphrasers into controllable diffusion models.

Paraphrase-Conditioned Diffusion

In this section, we introduce the components of our paraphrase-conditioned text diffusion model.

Diffusion Diffusion approaches (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2022), consist of two Markov chains, a *forward process* and a *reverse process*. In the *forward process*, the original data $\mathbf{x}_0$ is converted to pure Gaussian noise by incrementally adding noise over multiple discrete time steps, $\{0, ..., T\}$. Each of these intermediate noised latents, $\mathbf{x}_t$, can be directly sampled as follows:

$$\mathbf{x}_t = \sqrt{\bar{a}_t}\mathbf{x}_0 + \sqrt{1-\bar{a}_t}\boldsymbol{\epsilon}_t; \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I}), \quad (1)$$

where $\boldsymbol{\epsilon}_t$ is random noise and $\bar{a}_t$ specifies a well-behaved schedule such that $\bar{a}_t \to 0$ as $t \to T$. The *reverse process* is parameterized by a model, which is trained to reconstruct the

original data from pure noise ($\mathbf{x}_T$) by iteratively estimating $\boldsymbol{\epsilon}_t$ (or equivalently $\mathbf{x}_0$) and working backwards in time, from $t = T$ to $t = 0$.

In the image domain, pixels are used as the representation of $\mathbf{x}_0$. In contrast, text is discrete, and the underlying continuous domain is less obvious. Several existing text diffusion approaches operate on word embeddings (Li et al. 2022; Yuan et al. 2023; Gulrajani and Hashimoto 2023), while others noise token logit simplexes, like SSD-LM (Han, Kumar, and Tsvetkov 2023; Han et al. 2023; Mahabadi et al. 2023). PARAGUIDE performs diffusion in word embedding space, but incorporates several benefits of simplex methods.

Categorical Reparameterization While diffusion with word logits has several desirable properties (Han, Kumar, and Tsvetkov 2023), logits are a high dimension latent representation (sequence length $\times$ vocabulary size), which makes both training and inference slower and more memory intensive than operating directly on the word embedding space. Also, unlike the probability simplex, pretrained word embedding spaces are well-suited for meaning-preserving style transfer, as neighbors are often semantically similar (Mikolov et al. 2013).[3] As a result, in PARAGUIDE, we employ noised word embeddings for our latent representations, and define our forward process as:

$$\mathbf{x}_t = \sqrt{\bar{a}_t} E(\mathbf{w}) + \sqrt{(1 - \bar{a}_t)}\boldsymbol{\epsilon}_t, \tag{2}$$

where $\mathbf{w}$ is our original text and E is an embedding lookup. Rather than directly estimate the original word embeddings in our reverse process, however, we estimate $E(\mathbf{w})$ with a diffusion model that first outputs a posterior over discrete tokens, like Gulrajani and Hashimoto (2023):

$$\hat{\mathbf{w}}_t \sim p_\theta(.|\mathbf{x}_t, t, \mathbf{p}), \tag{3}$$

where $\mathbf{x}_t$ is our noised embedding, and $\mathbf{p}$ is our input paraphrase. We sample intermediate tokens from this distribution like in SSD-LM (Han, Kumar, and Tsvetkov 2023), and these tokens are embedded for the next t_{n-1}th diffusion step:

$$\mathbf{x}_{t-1} = \sqrt{\bar{a}_{t-1}} E(\hat{\mathbf{w}}_t) + \sqrt{(1 - \bar{a}_{t-1})}\boldsymbol{\epsilon}; \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

This approach still provides the controllability benefits of SSD-LM, as gradient-based control can be applied to the intermediate token predictions, which we will discuss in the Guidance section.

Diffusion Model Architecture We build on the SSD-LM architecture (Han, Kumar, and Tsvetkov 2023), which uses a bidirectional RoBERTa encoder (Liu et al. 2019) to output token probabilities at each diffusion step, conditioned on a noised representation and timestep. However, we make several changes to adapt their simplex-diffusion approach for text-to-text tasks like paraphrasing. First, as noted in the previous section, we modify their model to operate on noised word embeddings, rather than word logits. Additionally, as in Mahabadi et al. (2023), we also modify the original semi-autoregressive approach to be entirely diffusion-based. Finally, like other text-to-text diffusion approaches (Mahabadi et al. 2023; Yuan et al. 2023), we condition on an input (in our case, the paraphrase, $\mathbf{p}$), by concatenating it with our noised latent representation. Unlike these approaches, we incorporate stylistic guidance, as outlined in the Guidance section.

Diffusion Model Loss Following Han, Kumar, and Tsvetkov (2023) and Mahabadi et al. (2023), we train the diffusion model by minimizing the cross entropy between the model's posterior at each diffusion timestep and the ground-truth tokens $\mathbf{w}$, but given the timestep t, noised embeddings $\mathbf{x}_t$, and paraphrase $\mathbf{p}$:

$$\mathcal{L}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(1,T)}[-\log p_\theta(\mathbf{w}|\mathbf{x}_t, t, \mathbf{p})] \tag{5}$$

Diffusion Noise Schedule Several approaches to diffusion language modeling (Han, Kumar, and Tsvetkov 2023; Mahabadi et al. 2023; Han et al. 2023) have repurposed the *cosine* schedule (Nichol and Dhariwal 2021) from computer vision, while others have adopted the *sqrt* schedule (Li et al. 2022; Yuan et al. 2023). In contrast, we train PARAGUIDE with a dramatically less aggressive noise schedule:

$$\bar{a}_t = \sqrt{\frac{T - t}{T}} \tag{6}$$

This schedule falls to zero much more slowly than the cosine and sqrt schedules, destroying information less quickly. The schedule is motivated by our observation that skipping early steps with the cosine schedule had no noticeable effect on model outputs, and experiments that showed improved fluency and meaning preservation.[4]

Diffusion Model Inference At inference time, we first generate a paraphrase, $\mathbf{p}$ of our input text. We then sample initial noise $\mathbf{x}_T \sim \mathcal{N}(0, I)$. For each step in the reverse process ($t \in [T, 1]$), we then compute token logits using our model:

$$\mathbf{l}_t = \text{logits}_\theta(.|\mathbf{x}_t, t, \mathbf{p}) \tag{7}$$

We then sample from the model's posterior:[5]

$$\hat{\mathbf{w}}_t \sim \text{top-p}(\text{softmax}(\mathbf{l_t})) \tag{8}$$

After sampling $\hat{\mathbf{w}}_t$, we iteratively work backwards in time by embedding these tokens using the word embedding lookup, E, and then adding noise to produce $\mathbf{x}_{t-1}$, the latent for the previous diffusion timestep, following Han, Kumar, and Tsvetkov (2023):

$$\mathbf{x}_{t-1} = \sqrt{\bar{a}_{t-1}} E(\hat{\mathbf{w}}_t) + \sqrt{(1 - \bar{a}_{t-1})}\boldsymbol{\epsilon}; \boldsymbol{\epsilon} \sim \mathcal{N}(0, I) \tag{9}$$

[3] Additionally, we observed that the original SSD-LM approach of adding Gaussian noise to logits before a softmax operation results in almost completely noised text early on in the forward process, and early steps could be discarded with little effect on our final outputs.

[4] We view less aggressive noise schedules as a possible alternative to self-conditioning (Strudel et al. 2022; Han et al. 2023; Mahabadi et al. 2023), which handles information loss by also conditioning diffusion language models on previous x_{t-1} predictions. We found that self-conditioning improved fluency but hurt control.

[5] This is similar to the sampling approach used by Han, Kumar, and Tsvetkov (2023) and clamping in Li et al. (2022)

In this fashion, the model starts from random noise and an input paraphrase, and then iteratively generates a semantically consistent output text. A critical advantage of applying diffusion models to this task is that we can use gradient-based guidance to steer our outputs towards specific target styles. We discuss this in the next section.

Guidance

PARAGUIDE can incorporate guidance from any model that is 1) differentiable and 2) uses the same tokenization scheme as the base diffusion paraphraser:

$$\mathbf{l}_t = \mathbf{l}_{t,init} - \lambda \nabla_{\mathbf{l}_t} L_{guidance}(\mathbf{l}_{t,init}) \tag{10}$$

where $\mathbf{l}_{t,init}$ are the initial logit predictions at timestep t, and $L_{guidance}$ specifies a guidance loss.

Because our diffusion model employs a RoBERTa (Liu et al. 2019) tokenization scheme, we can incorporate guidance from the many available models built on the popular RoBERTa encoder backbone. We explore two forms of guidance loss for style transfer: The first is based on attribute classifiers, and the second is based on distances in stylistic embedding space.

Attribute Classifiers Following Han, Kumar, and Tsvetkov (2023), we use a classifier, $f_\phi(\cdot)$ to generate texts with a target attribute, y, by applying drift to the full sequence of logits, $\mathbf{l}_t$, at each intermediate diffusion step:

$$L_{guidance}(\mathbf{l}_t) = -log(f_\phi(y|\mathbf{l}_t)) \tag{11}$$

Additionally, like Han, Kumar, and Tsvetkov (2023), we can trivially adapt classifiers to accept logits, rather than word embeddings, by using the $\mathrm{softmax}$ function with some temperature, τ, to compute a probability simplex over the vocabulary. We can then project with the classifier's embedding lookup, E_ϕ:

$$\tilde{\mathbf{e}}_{\phi,t} = \mathrm{softmax}(\frac{\mathbf{l}_t}{\tau}) \times E_\phi \tag{12}$$

This results in a linear combination of word embeddings at each timestep $\tilde{\mathbf{e}}_{\phi,t}$, based on each token's assigned probability mass. These embeddings are passed to the attribute model, and gradients are computed through them to increase or decrease the probabilities of different tokens to maximize the probability of attribute y. In contrast to SSD-LM (Han, Kumar, and Tsvetkov 2023), PARAGUIDE applies this drift to a diffusion model trained to reconstruct semantically consistent text, which enables meaning-preserving style transfer. We can balance the trade-off between semantic consistency and style transfer by varying λ.

Style Embedding Distance For authorship style transfer, we take the novel approach of leveraging guidance from stylistic embedding models, including Style Embeddings (Wegmann, Schraagen, and Nguyen 2022), that are constrastively trained to identify authorship styles.

To guide our paraphrases with a style embedder, g_ϕ, we compute the gradient of $\mathbf{l}_t$ with respect to its average distance in style embedding space to the target author's n texts, $[\mathbf{y_1}, \mathbf{y_2}..., \mathbf{y_n}]$:

Algorithm 1: ParaGuide Style Transfer

Input:
Input Text in Source Style $\mathbf{w}$,
Guidance Loss $L_{guidance}$, Guidance Strength λ
Output: Output Text in the Target Style

$\mathbf{p} = \textit{paraphraser}(\mathbf{w})$
$\mathbf{x}_T \sim \mathcal{N}(0, 1)$
for $t = T, ..., 1$ **do**
  $\mathbf{l}_{t,init} = \mathrm{logits}_\theta(.|\mathbf{x}_t, t, \mathbf{p})$
  if $\lambda \neq 0$ **then**
    for $i = 1, ..., k$ **do**
      $\mathbf{l}_t \leftarrow \mathbf{l}_t - \lambda \cdot \sin(\pi \frac{t}{T}) \cdot \nabla_{\mathbf{l}_{t,init}} L_{guidance}(\mathbf{l}_{t,init})$
    end for
  end if
  $\hat{\mathbf{w}}_t \sim \text{top-p}(\mathrm{softmax}(\mathbf{l_t}))$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
  $\mathbf{x}_{t-1} = \sqrt{\bar{a}_t} E(\hat{\mathbf{w}}_t) + \sqrt{(1 - \bar{a}_t)}\boldsymbol{\epsilon}$
end for
return $\mathbf{w}_0$

$$L_{guidance}(\mathbf{l}_t) = \frac{\sum_{i=1}^{n} d(g_\phi(\mathbf{l}_t), g_\phi(\mathbf{y_i}))}{n} \tag{13}$$

We use cosine distance for $d(\cdot)$. At every diffusion step, by minimizing the distance in style embedding space, we steer the output text towards the target author's style.

Guidance Schedule We observed that using the same λ for control at all diffusion timesteps leads to disfluent solutions. Intuitively, large gradient steps at the end of the reverse diffusion process are undesirable, as optimizing the control objective can lead to ungrammatical text. Simultaneously, large steps early on in the reverse process are also undesirable, as these initial predictions are generally incoherent, and out of distribution for off-the-shelf models.

As a result, for all forms of guidance, we employ a sinusoidal schedule for controlling drift:

$$\lambda_t = \lambda \cdot \sin(\pi \frac{t}{T}) \tag{14}$$

This increases and then anneals the strength of drift during the reverse process. Additionally, we make k gradient updates per diffusion step, like Li et al. (2022). PARAGUIDE's complete inference procedure is specified in Algorithm 1.

Experimental Setup

Dataset

We evaluate our method on the Enron Email Corpus, which comprises several hundred thousand emails made public during the US government's investigation of Enron (Klimt and Yang 2004; Peterson, Hohensee, and Xia 2011). The dataset contains emails from the inboxes of 150 Enron employees, sent from over one thousand accounts.

The Enron corpus presents an ideal testbed for plug-and-play style transfer of both authorship and attributes. For the

former, email meta-data enables attributing messages to specific authors for authorship transfer. The emails also present diverse stylistic attributes, including different degrees of formality (Peterson, Hohensee, and Xia 2011) and divergent rhetorical styles (Brown and Laudenbach 2021).

Ultimately, we need to evaluate whether our email style transfer approach generalizes to new authors and texts. Therefore, we randomly select 10% of addresses to be the holdout authors for both authorship and attribute evaluations. These 110 authors present a low-resource authorship corpus, as the median holdout author has only 23 emails. For our authorship experiments, we evaluate each approach by selecting up to 5 test emails per holdout source author, and transferring these to 5 other random holdout authors.

To build our training and validation datasets for attribute style transfer, we use popular existing formality and sentiment classifiers to score texts from the holdout authors in the Enron dataset. Critically, we set aside these external classifiers and avoid using them as guidance for PARAGUIDE at inference time. In addition to the Enron corpus, we also build a pretraining corpus from the Reddit Million User Dataset (MUD) (Andrews and Bishop 2019; Khan et al. 2021), which includes 4 million comments by 400k different Reddit users. We use the same paraphrasing procedure on both the Enron and Reddit datasets to generate (*paraphrase*, *original text*) training pairs.

Implementation Details

To train our diffusion model, we fine-tuned the publicly available SSD-LM RoBERTa-Large checkpoint[6] (Han, Kumar, and Tsvetkov 2023) with our previously stated modifications to the architecture and noise schedule. We first fine-tune the diffusion model on Reddit paraphase pairs, and then continue fine-tuning on the Enron non-holdout author paraphrase pairs. We fine-tune all parameters except the word embedding lookup. Additional implementation details are included in our Appendix.

Baselines

Attribute Style Transfer For attribute transfer, we compare to Mix and Match (M&M) (Mireshghallah, Goyal, and Berg-Kirkpatrick 2022) and consider both the DISC and HAM configurations from the original paper (Mireshghallah, Goyal, and Berg-Kirkpatrick 2022). To better compare with our approach, however, we replace the original BERT model with RoBERTa-large and also include results where we fine-tune this model on Enron Email training data.

We also implement a STRAP baseline (Krishna, Wieting, and Iyyer 2020) with pretrained T5-Large models (Raffel et al. 2020), fine-tuned on Reddit and then Enron paraphrase pairs. In contrast to M&M and PARAGUIDE, which are learning-free approaches, STRAP requires training attribute-specific models on the Enron data classified by the external classifiers. We fine-tune four STRAP models for informality, formality, positive sentiment, and negative sentiment.

Authorship Style Transfer For the task of authorship style transfer on the Enron Email Corpus, we consider STRAP (Krishna, Wieting, and Iyyer 2020), and the BERT, LING, and PARA approaches from Patel, Andrews, and Callison-Burch (2022). We also consider a ChatGPT-3.5 style transfer approach, where we prompt the model with up to 16 in-context examples of a target author's style. In contrast to our other approaches, we fine-tune 110 author-specific STRAP models on 60% of each holdout author's data.

Evaluation Metrics

Attribute Style Transfer Following Mireshghallah, Goyal, and Berg-Kirkpatrick (2022), we measure style transfer accuracy with two classifiers. First, *Internal Accuracy* measures the style transfer accuracy of the classifier used at inference time by Mix and Match and PARAGUIDE. In contrast, *External Accuracy* measures the style transfer accuracy using a classifier set aside for evaluation.

We measure textual *Similarity* by computing Mutual Implication Score (MIS) (Babakov et al. 2022) and *Fluency* with a model trained on the CoLA dataset (Morris et al. 2020; Warstadt, Singh, and Bowman 2019).[7] For an aggregate metric of model performance, we compute a *Joint* metric by taking the sentence-wise geometric mean of *External Accuracy*, *Similarity*, and *Fluency*, similar to Krishna, Wieting, and Iyyer (2020).

For formality transfer, we additionally run a human evaluation of style-transfer approaches that scored highest on our automatic evaluations. We asked annotators to compare model outputs to the reference inputs, and score ($\{0, 1\}$) their *Similarity*, *Fluency*, and *Formality*. We include additional details describing our human evaluations in the Appendix.

Authorship Style Transfer To evaluate authorship style transfer, we adopt the *Confusion* metric from the evaluation framework defined by Patel, Andrews, and Callison-Burch (2022), where the authors utilize pretrained style embedders (Wegmann, Schraagen, and Nguyen 2022; Rivera-Soto et al. 2021) to measure style transfer success. *Confusion*, which is similar to style transfer accuracy, is the percentage of the time that the style transfer output is closer to the target author than the source author in representational embedding space. As with attribute transfer, we similarly compute *Similarity* and *Fluency*, and *Joint*, but use *Confusion* in place of transfer accuracy.

We compute the above metrics for both Style Embeddings (Wegmann, Schraagen, and Nguyen 2022) and Universal Authorship Representations (UAR) (Rivera-Soto et al. 2021). Similar to our external style classifier for attribute transfer, UAR provides a holdout embedding space that PARAGUIDE does not directly optimize at inference time.

[6] https://huggingface.co/xhan77/ssdlm

[7] textattack/roberta-base-CoLA

Method	Int. Acc ($\rightarrow F$,$\rightarrow I$)	Ext. Acc ($\rightarrow F$,$\rightarrow I$)	Sim ($\rightarrow F$,$\rightarrow I$)	Fluency ($\rightarrow F$,$\rightarrow I$)	Joint ($\rightarrow F$,$\rightarrow I$)
STRAP$_{\textit{fine-tuned}}$	0.45 (0.8, 0.1)	0.45 (0.76, 0.13)	0.50 (0.54, 0.47)	**0.73** (0.75, 0.71)	0.31 (0.54, 0.08)
M&M (Disc)	0.63 (0.59, 0.67)	0.55 (0.44, 0.65)	0.24 (0.19, 0.3)	0.62 (0.62, 0.62)	0.23 (0.19, 0.27)
M&M (Hamming)	0.58 (0.59, 0.57)	0.51 (0.46, 0.57)	0.40 (0.29, 0.52)	0.61 (0.61, 0.6)	0.26 (0.21, 0.31)
M&M$_{enron}$ (Disc)	0.58 (0.62, 0.55)	0.51 (0.47, 0.56)	0.31 (0.26, 0.37)	0.61 (0.61, 0.61)	0.24 (0.22, 0.26)
M&M$_{enron}$ (Hamming)	0.51 (0.56, 0.46)	0.47 (0.45, 0.48)	0.45 (0.35, 0.55)	0.62 (0.62, 0.62)	0.25 (0.23, 0.28)
PGuide ($\lambda = 1e4$)	**0.97** (0.96, 0.99)	**0.83** (0.68, 0.99)	0.40 (0.37, 0.44)	0.55 (0.59, 0.51)	0.45 (0.37, 0.53)
PGuide ($\lambda = 5e3$)	0.97 (0.96, 0.98)	0.82 (0.65, 0.99)	0.40 (0.36, 0.45)	0.56 (0.59, 0.52)	0.45 (0.37, 0.53)
PGuide ($\lambda = 1e3$)	0.95 (0.93, 0.98)	0.81 (0.64, 0.98)	0.45 (0.4, 0.49)	0.60 (0.62, 0.57)	0.47 (0.37, 0.56)
PGuide ($\lambda = 5e2$)	0.94 (0.9, 0.97)	0.81 (0.63, 0.98)	0.47 (0.44, 0.5)	0.61 (0.64, 0.58)	0.48 (0.39, 0.58)
PGuide ($\lambda = 2e2$)	0.91 (0.85, 0.98)	0.76 (0.58, 0.95)	**0.52** (0.5, 0.53)	0.63 (0.65, 0.61)	**0.48** (0.38, 0.59)

Table 1: Automatic Formality Evaluations. We report accuracy for both the *Internal* and *External* classifiers. The best results are bolded. We also decompose results into formality ($\rightarrow F$) and informality ($\rightarrow I$) transfer.

Method	Accuracy ($\rightarrow F$,$\rightarrow I$)	Sim ($\rightarrow F$,$\rightarrow I$)	Fluency ($\rightarrow F$,$\rightarrow I$)	Joint ($\rightarrow F$,$\rightarrow I$)
STRAP$_{\textit{fine-tuned}}$	0.51 (0.10, 0.91)	0.35 (0.32, 0.37)	0.03 (0.04, 0.01)	0.00 (0.00, 0.00)
M&M (Hamming)	0.47 (0.14, 0.80)	0.49 (0.31, 0.67)	0.46 (0.27, 0.64)	0.20 (0.03, 0.36)
PGuide ($\lambda = 2e2$)	**0.65** (0.39, 0.90)	**0.58** (0.54, 0.61)	**0.69** (0.61, 0.77)	**0.33** (0.23, 0.43)

Table 2: Human Formality Evaluations. We asked annotators to rate outputs from models with the highest automatic scores as formal or informal (*Accuracy*), whether their meaning was similar to the original (*Similarity*), and whether the outputs were well-formed/grammatical (*Fluency*). *Joint* aggregates these scores together at the sentence-level.

Results

Attribute Style Transfer

In this section, we review our evaluation results for attribute transfer. We include representative outputs in the Appendix.

Automatic Evaluations Tables 1 and 3 present our automatic evaluation results for formality and sentiment transfer. For each approach, we display the average score for each metric, along with the breakdown for formal/informal ($\rightarrow F, \rightarrow I$) and positive/negative ($\rightarrow P, \rightarrow N$).

PARAGUIDE outperforms all other approaches on all aggregate *Joint* metrics, across both sentiment and formality experiments. Additionally, PARAGUIDE significantly surpasses all baselines on transfer accuracy. Despite the inherent trade-off between transfer accuracy and meaning preservation, on formality, PARAGUIDE ($\lambda = 2e2$) outperforms all baseline approaches on both transfer accuracy *and* meaning preservation. On sentiment transfer, PARAGUIDE's increased accuracy incurs a larger cost to semantic similarity, but this is expected in successful sentiment transfer, which involves changing the polarity of texts (Jin et al. 2022).

Human Evaluation Table 2 displays the results of our human formality evaluation, where annotators rated the *Formality*, *Similarity*, and *Fluency* of model outputs. When evaluated by humans, PARAGUIDE significantly outperforms the top performing baselines across all aggregate metrics ($p = 0.05$). Notably, this is even true for the *Fluency* metric, where annotators rated whether outputs were reasonable, coherent emails. This result was unexpected given PARAGUIDE's comparatively unimpressive automatic *Fluency* scores, but could be explained by differences between email writing practices and the composition of the CoLA training corpus (Warstadt, Singh, and Bowman 2019). In contrast, the STRAP baseline dramatically underperforms on our human evaluation. Manually inspecting outputs, we found that the STRAP models we fine-tuned for attribute transfer generate highly repetitive text. We suspect that this results from fine-tuning on our limited dataset, and aligns with previous work, which has shown that STRAP's performance is heavily reliant on dataset size (Patel, Andrews, and Callison-Burch 2022).

Authorship Style Transfer

Table 4 presents our results on the challenging task of low-resource authorship style transfer. When evaluated with the *Style* embedding space, three of the four PARAGUIDE configurations outperform every single baseline (including ChatGPT-3.5) on *Joint* and *Confusion*. When we consider the holdout *UAR* embedding space, however, ChatGPT-3.5, which notably uses 400x more parameters than PARAGUIDE, outperforms the other approaches. Considering only non-LLM methods, PARAGUIDE outperforms all baselines on *UAR Confusion*, but is very narrowly outperformed by STRAP on *UAR Joint*. This can be attributed, however, to STRAP's higher *Fluency* score, which was a metric that was not predictive of human ratings on the formality task. Additionally, in contrast to PARAGUIDE's plug-and-play approach, the STRAP implementation involves 110 separate models, each with 800 million parameters, fine-tuned for every author.

Method	Int. Acc ($\rightarrow P$,$\rightarrow N$)	Ext. Acc ($\rightarrow P$,$\rightarrow N$)	Sim ($\rightarrow P$,$\rightarrow N$)	Fluency ($\rightarrow P$,$\rightarrow N$)	Joint ($\rightarrow P$,$\rightarrow N$)
STRAP$_{\textit{fine-tuned}}$	0.11 (0.16, 0.05)	0.29 (0.38, 0.19)	0.5 (0.5, 0.49)	**0.74** (0.72, 0.76)	0.18 (0.24, 0.12)
M&M (Disc)	0.2 (0.01, 0.38)	0.5 (0.32, 0.67)	0.34 (0.46, 0.22)	0.63 (0.62, 0.64)	0.21 (0.17, 0.25)
M&M (Ham)	0.14 (0.02, 0.26)	0.39 (0.23, 0.55)	0.45 (0.58, 0.32)	0.62 (0.6, 0.63)	0.19 (0.14, 0.24)
M&M$_{enron}$ (Disc)	0.1 (0.02, 0.18)	0.38 (0.29, 0.47)	0.4 (0.48, 0.33)	0.62 (0.6, 0.64)	0.19 (0.16, 0.22)
M&M$_{enron}$ (Ham)	0.08 (0.02, 0.13)	0.31 (0.21, 0.41)	**0.52** (0.6, 0.44)	0.62 (0.61, 0.64)	0.16 (0.13, 0.2)
PGuide ($\lambda = 1e4$)	**0.73** (0.78, 0.68)	**0.8** (0.86, 0.74)	0.13 (0.2, 0.06)	0.43 (0.43, 0.43)	0.2 (0.27, 0.13)
PGuide ($\lambda = 5e3$)	0.7 (0.76, 0.65)	0.79 (0.87, 0.71)	0.15 (0.22, 0.07)	0.43 (0.45, 0.41)	0.22 (0.3, 0.14)
PGuide ($\lambda = 1e3$)	0.65 (0.75, 0.54)	0.75 (0.81, 0.69)	0.25 (0.32, 0.18)	0.48 (0.53, 0.43)	0.28 (0.35, 0.21)
PGuide ($\lambda = 5e2$)	0.57 (0.71, 0.43)	0.68 (0.74, 0.62)	0.33 (0.37, 0.28)	0.51 (0.55, 0.47)	0.29 (0.35, 0.23)
PGuide ($\lambda = 2e2$)	0.35 (0.47, 0.22)	0.56 (0.61, 0.51)	0.42 (0.44, 0.4)	0.59 (0.64, 0.55)	**0.29** (0.33, 0.25)

Table 3: Automatic Sentiment Evaluations. Like for the formality results, we break down scores into positive ($\rightarrow P$) and negative ($\rightarrow N$) transfer, and report scores for both the *Internal* and *External* classifiers.

	Style		UAR			
Method	Conf.	Joint	Conf.	Joint	Sim	Fluency
PARA	0.42	0.335	0.26	0.202	0.64	**0.85**
BERT	0.31	0.076	0.30	0.061	0.13	0.35
LING	0.44	0.334	0.23	0.177	**0.82**	0.58
STRAP$_{\textit{fine-tuned}}$	0.47	0.344	0.32	0.218	0.54	0.83
ChatGPT-3.5	0.54	0.338	**0.48**	**0.280**	0.56	0.79
PGuide ($\lambda = 2.5e3$)	**0.74**	0.431	0.36	0.209	0.42	0.64
PGuide ($\lambda = 1.5e3$)	0.68	**0.434**	0.33	0.207	0.47	0.70
PGuide ($\lambda = 8e2$)	0.64	0.426	0.33	0.217	0.50	0.74
PGuide ($\lambda = 2e2$)	0.50	0.353	0.29	0.204	0.52	0.78

Table 4: Evaluation metrics for authorship style transfer. We evaluate using two authorship representations: *Style* (Wegmann, Schraagen, and Nguyen 2022) and *UAR* (Rivera-Soto et al. 2021). For each metric, we bold the strongest approach, and underline the most performant non-LLM method.

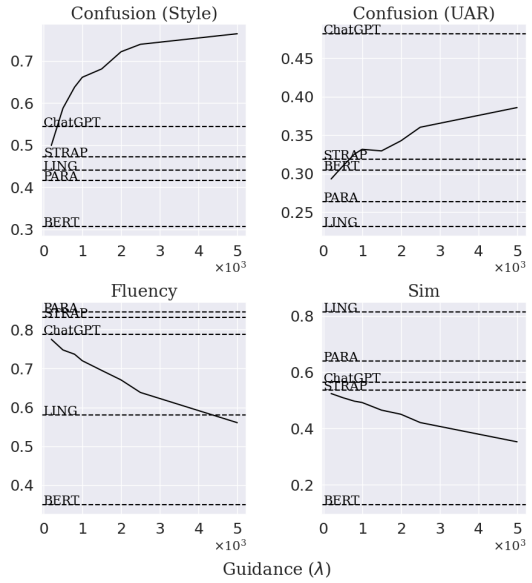

Figure 2: As we increase the guidance hyperparameter λ, we steadily increase style transfer accuracy (*Confusion*), at the cost of semantic consistency (*Sim*) and *Fluency*.

Style Transfer vs. Similarity and Fluency

Beyond showcasing PARAGUIDE's strong performance, our automatic evaluations in Tables 1, 3, and 4 demonstrate control over the trade-off between transfer accuracy versus semantic consistency and fluency, via the λ hyperparameter. We additionally visualize the affect of varying λ on authorship style transfer in Figure 2. When λ is small, the paraphrase-conditioned diffusion model reconstructs a more semantically faithful, fluent output. However, we can increase λ to improve *Confusion* scores, at the cost of semantic consistency and fluency. At the lowest setting, PARAGUIDE's *Fluency* and *Similarity* score are similar to those of ChatGPT-3.5 (0.78 vs 0.79 and 0.52 vs 0.56).

Conclusion and Future Work

We introduce PARAGUIDE, a diffusion-based framework for unsupervised textual style transfer. The approach harnesses the controllability of text diffusion, alongside the availability of off-the-shelf text classifiers and stylistic embedders, to competitively perform both authorship and attribute transfer, without ever having to retrain style-specific pipelines.

Our work demonstrates the potential of diffusion for text generation, a landscape currently dominated by large, autoregressive language models. We are particularly excited about pursuing work that explores scaling diffusion models, better adapting them to the text domain, and the ways that these non-autoregressive methods can work alongside and complement current state-of-the-art approaches.

Ethical Statement

PARAGUIDE presents an effective diffusion-based framework for style-transfer that uses fewer parameters than other state-of-the-art methods, can be fine-tuned on a single GPU, and avoids having to retrain models for new target styles.

As a result, the approach could broaden the accessibility of controllable text generation and empower individuals with fewer resources to better personalize systems to their needs. At the same time, we recognize that text generation approaches like ours have the potential to be leveraged by malicious actors for impersonation and persuasion.

Acknowledgements

We would like to thank Xiaochuang Han, Raghav Singhal, Amith Ananthram, Debasmita Bhattacharya, Nicholas Deas, Maximillian Chen, and Smaranda Muresan for their invaluable discussions and thoughtful feedback, which helped shape the direction of this work. Additionally, we would like to extend our gratitude to Samir Gadre, Fei-Tzin Lee, and Matthew Toles for their support on human evaluations, and our anonymous AAAI reviewers for their comments.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Andrews, N.; and Bishop, M. 2019. Learning Invariant Representations of Social Media Users. arXiv:1910.04979.

Babakov, N.; Dale, D.; Logacheva, V.; and Panchenko, A. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 300–321. Dublin, Ireland: Association for Computational Linguistics.

Brown, D.; and Laudenbach, M. 2021. Stylistic variation in email. *Register Studies*, 4.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Dale, D.; Voronov, A.; Dementieva, D.; Logacheva, V.; Kozlova, O.; Semenov, N.; and Panchenko, A. 2021. Text Detoxification using Large Pre-trained Neural Models. arXiv:2109.08914.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. arXiv:1912.02164.

Gulrajani, I.; and Hashimoto, T. B. 2023. Likelihood-Based Diffusion Language Models. arXiv:2305.18619.

Han, X.; Kumar, S.; and Tsvetkov, Y. 2023. SSD-LM: Semi-autoregressive Simplex-based Diffusion Language Model for Text Generation and Modular Control. arXiv:2210.17432.

Han, X.; Kumar, S.; Tsvetkov, Y.; and Ghazvininejad, M. 2023. SSD-2: Scaling and Inference-time Fusion of Diffusion Language Models. arXiv:2305.14771.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Jin, D.; Jin, Z.; Hu, Z.; Vechtomova, O.; and Mihalcea, R. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1): 155–205.

Khan, A.; Fleming, E.; Schofield, N.; Bishop, M.; and Andrews, N. 2021. A Deep Metric Learning Approach to Account Linking. arXiv:2105.07263.

Klimt, B.; and Yang, Y. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *European Conference on Machine Learning*.

Krause, B.; Gotmare, A. D.; McCann, B.; Keskar, N. S.; Joty, S.; Socher, R.; and Rajani, N. F. 2020. GeDi: Generative Discriminator Guided Sequence Generation. arXiv:2009.06367.

Krishna, K.; Wieting, J.; and Iyyer, M. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. arXiv:2010.05700.

Kumar, S.; Malmi, E.; Severyn, A.; and Tsvetkov, Y. 2021. Controlled Text Generation as Continuous Optimization with Multiple Constraints. arXiv:2108.01850.

Li, X. L.; Thickstun, J.; Gulrajani, I.; Liang, P.; and Hashimoto, T. B. 2022. Diffusion-LM Improves Controllable Text Generation. arXiv:2205.14217.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Ma, X.; Sap, M.; Rashkin, H.; and Choi, Y. 2020. PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7426–7441. Online: Association for Computational Linguistics.

Mahabadi, R. K.; Tae, J.; Ivison, H.; Henderson, J.; Beltagy, I.; Peters, M. E.; and Cohan, A. 2023. TESS: Text-to-Text Self-Conditioned Simplex Diffusion. arXiv:2305.08379.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mireshghallah, F.; Goyal, K.; and Berg-Kirkpatrick, T. 2022. Mix and Match: Learning-free Controllable Text Generation using Energy Language Models. arXiv:2203.13299.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 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*, 119–126.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8162–8171. PMLR.

Patel, A.; Andrews, N.; and Callison-Burch, C. 2022. Low-Resource Authorship Style Transfer with In-Context Learning. arXiv:2212.08986.

Peterson, K.; Hohensee, M.; and Xia, F. 2011. Email Formality in the Workplace: A Case Study on the Enron Corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 86–95. Portland, Oregon: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

Reif, E.; Ippolito, D.; Yuan, A.; Coenen, A.; Callison-Burch, C.; and Wei, J. 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)*, 837–848. Dublin, Ireland: Association for Computational Linguistics.

Riley, P.; Constant, N.; Guo, M.; Kumar, G.; Uthus, D.; and Parekh, Z. 2021. TextSETTR: Few-Shot Text Style Extraction and Tunable Targeted Restyling. 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)*, 3786–3800. Online: Association for Computational Linguistics.

Rivera-Soto, R. A.; Miano, O. E.; Ordonez, J.; Chen, B. Y.; Khan, A.; Bishop, M.; and Andrews, N. 2021. Learning Universal Authorship Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 913–919. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Saharia, C.; Chan, W.; Chang, H.; Lee, C. A.; Ho, J.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Palette: Image-to-Image Diffusion Models. arXiv:2111.05826.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:1503.03585.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Strudel, R.; Tallec, C.; Altché, F.; Du, Y.; Ganin, Y.; Mensch, A.; Grathwohl, W.; Savinov, N.; Dieleman, S.; Sifre, L.; and Leblond, R. 2022. Self-conditioned Embedding Diffusion for Text Generation. arXiv:2211.04236.

Warstadt, A.; Singh, A.; and Bowman, S. R. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7: 625–641.

Wegmann, A.; Schraagen, M.; and Nguyen, D. 2022. Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, 249–268. Dublin, Ireland: Association for Computational Linguistics.

Yang, K.; and Klein, D. 2021. FUDGE: Controlled Text Generation With Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Yuan, H.; Yuan, Z.; Tan, C.; Huang, F.; and Huang, S. 2023. SeqDiffuSeq: Text Diffusion with Encoder-Decoder Transformers. arXiv:2212.10325.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. J. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv:1912.08777.