Fine-grained Text Style Transfer with Diffusion-Based Language Models

Yiwei Lyu   Tiange Luo   Jiacheng Shi   Todd C. Hollon   Honglak Lee

University of Michigan
{yiweilyu, tiangel, jiachs}@umich.edu

Abstract

Diffusion probabilistic models have shown great success in generating high-quality images controllably, and researchers have tried to utilize this controllability into text generation domain. Previous works on diffusion-based language models have shown that they can be trained without external knowledge (such as pre-trained weights) and still achieve stable performance and controllability. In this paper, we trained a diffusion-based model on StylePTB dataset, the standard benchmark for fine-grained text style transfers. The tasks in StylePTB requires much more refined control over the output text compared to tasks evaluated in previous works, and our model was able to achieve state-of-the-art performance on StylePTB on both individual and compositional transfers. Moreover, our model, trained on limited data from StylePTB without external knowledge, outperforms previous works that utilized pretrained weights, embeddings, and external grammar parsers, and this may indicate that diffusion-based language models have great potential under low-resource settings. Our code is available at https://github.com/lvyiwei1/DiffuSeq_StylePTB

1   Introduction

Diffusion probabilistic models (Ho et al., 2020) have became the state-of-the-art technique in visual generative tasks. By starting from random gaussian noise and gradual denoising, they are able to generate images that look realistic in details. Moreover, conditional diffusion models such as stable diffusion (Rombach et al., 2022) are able to achieve detailed control over the generated output by conditioning on text, layouts, etc. The generated images are faithful to the text description or layouts, often to the finest details.

Analogically, researchers have tried to utilize the controllability of diffusion models to achieve more controllable language generation. For ex-

ample, DiffuSeq (Gong et al., 2022) applies diffusion models to sequence-sequence text generation tasks such as paraphrasing, question generation and text simplification; Diffusion-LM (Li et al., 2022) combined diffusion models with language models to control language generation by specifying generation length, syntax tree, semantic context, etc. What made these diffusion-based language models impressive is that they are trained from scratch with zero external knowledge (i.e. no pre-trained word embeddings or model weights, no external grammar parsers, etc) and on very few data (on the order of 10^5 tokens) compared to any large language models (for example, GPT-3's (Brown et al., 2020) training data is on the order 10^{11} tokens), so they have to learn representations at all levels (word embeddings, sentence structures, etc) from scratch with very limited data.

However, while the earlier tasks assessed on Diffusion-LM and DiffuSeq require a degree of control over the generated output, they are incapable of modifying the existing text to exhibit specific stylistic characteristics. In this paper, we would like to further examine the capabilities of diffusion-based language models on **fine-grained text style transfer**, an important task that requires more fine-grained control than the tasks from previous works on diffusion-based language modeling because it only allows changing the specified fine-grained stylistic properties of the input while leaving the rest unchanged. For example, "verb emphasis" is a fine-grained style transfer that requires the model to rewrite the sentence emphasizing a certain verb, without changing any other information that the original sentence conveys. In comparison, previous evaluation tasks such as controlling sequence length, semantic context, etc essentially control one aspect at a time and require no control over any other properties of generated text.

We use 13 non-lexical transfers from StylePTB (Lyu et al., 2021) dataset, where

Aspect	Transfers	Original Sentence	Additional Information	Transformed Sentence	Number of Pairs in StylePTB dataset
Syntax	To Future Tense	She travels to Paris every summer to visit her family.		She **will** travel to Paris next summer to visit her family.	7272
	To Present Tense	She had been studying architecture for five years.		She **has** been studying architecture for five years.	4365
	To Past Tense	He walks to the store every day.		He **walked** to the store every day.	4422
	Activate to Passive	The cat chased the mouse.		The mouse **was chased** by the cat.	2808
	Passive to Activate	The proposal was approved by the committee yesterday.		The committee **approved** the proposal yesterday.	2808
	PP Front to Back	Having watched the movie, they left the theater.		They left the theater after **having watched the movie**.	467
	PP Back to Front	They have been planning their vacation for months.		**For months**, they have been planning their vacation.	467
Semantic	ADJ/ADV Removal	The **extremely talented** musician played a **beautiful** melody on the piano.		The musician played a melody on the piano.	4863
	PP Removal	She had been studying **for hours** before taking the test.		She had been studying before taking the test.	4767
	Substatement Removal	He was unhappy **that he had failed the exam**.		He was unhappy.	1345
	Infomation Addition	The stock was up three percent according to the man.	"man", "lazy"	The stock was up three percent according to the **lazy** man.	2114
Thematics	Verb/Action Emphasis	She reads books in pastime.	"read"	**Reading** books is her favorite pastime.	1201
	Adjective Emphasis	The scenic forest is Michele's favorite place.	"scenic"	Michele's favorite forest is **scenic**.	696

Table 1: The 13 non-lexical fine-grained text style transfers from the StylePTB dataset (Lyu et al., 2021). We present one example sentence pair before/after each transfer, as well as the total number of sentence pairs available for each transfer in StylePTB. As we can see, the transfers require changing one specific stylistic aspect of the sentence while leaving all other aspects unchanged, and the amount of data available for training is limited (compared to the typical amount of data required to train large language models nowadays).

there are at most a few thousand sentence pairs available for each transfer, as shown in Table 1. Since identifying the grammatical structure of the sentence can be very helpful for most of these transfers (such as active-to-passive), some previous methods (such as Neural QCFG (Kim, 2021)) utilizes external grammar parsers to gain such information. We trained a diffusion-based model on StylePTB data without any pre-trained weights or external grammar parsers. Therefore, our model has to start from zero grammar/linguistic knowledge and learn all of them from very limited training data (StylePTB only has 7719 sentences from Penn Tree Bank (Marcus et al., 1993) plus their transferred outputs). Even under these hard conditions, our model still managed to outperform previous works that do utilize external weights or grammar parsers. Moreover, we also evaluate the capabilities of diffusion-based language models on performing multiple transfers using one single model and composing multiple learned transfers on a single sentence. We list our contributions as follows:

- We trained a diffusion-based language model (adapted from DiffuSeq (Gong et al., 2022))

that can perform fine-grained text style transfer from scratch with very limited training data and no external weights or tools. The model also supports multitasking and composing multiple fine-grained transfers.

- Our model achieves state-of-the-art performance on fine-grained text style transfers in StylePTB. Our multitask model (i.e. one single model that can perform all 13 transfers) achieves **best performance** compared to previous works on the same tasks on **88 out of 91** metrics (7 metrics per transfer), and gets very close to human performance on tasks with easy and medium difficulties. We also evaluated our model on composition of multiple fine-grained transfers, and we achieved best performance on these tasks as well.

- Thr/ough the evaluations, we demonstrated the extraordinary capabilities of diffusion-based language models in asserting extremely fine-grained control over generated text, and that this type of language model have great potential in controllable natural language generation under low-resource settings as it is able to achieve state-of-the-art performance with

limited training data and no external knowledge.

2 Backgrounds

2.1 Fine-grained Text Style Transfer and StylePTB

An import challenge for AI is to convey intentions using different stylistic attributes, and automated text style transfer is an essential step towards that. Text style transfer aims to controllably convert source text with targeted stylistic properties, with important applications in human-AI interactions including dialog systems (Celikyilmaz et al., 2018) and intelligent agents (Kim et al., 2013; Liang et al., 2020; Pittermann et al., 2010) that can communicate with specific text styles for different situations, target audiences, and environments (Lample et al., 2019; Li et al., 2018).

There has been extensive research on high-level style transfers such as sentiment transfers (Shen et al., 2017) and formality transfers (Rao and Tetreault, 2018). However, high-level style transfers lack the ability to fully control the style of the output. For example, there are many ways to convert a positive comment about a restaurant into a negative one, and high-level text style transfers do not allow control over which of the possible outputs (that may have different styles in non-sentiment aspects) can be generated. Fine-grained text style transfer is important because they allow fine-grained control over the generated output. (Lyu et al., 2021) defined a set of fine-grained text style transfer along four lingustic axis:

- **Lexical Transfers:** Word changes
- **Syntax Transfers:** Grammar and sentence structure changes
- **Semantic Transfers:** Meaning changes
- **Thematic Transfers:** Situational changes or word emphasis

Along these 4 axes, it defined 21 individual fine-grained transfers, 13 of which are non-lexical. Examples of the non-lexical transfers are shown in Table 1. Compared to other forms of controllable text generation, fine-grained text style transfer has the advantage of being able to assert control over text generated by uncontrollable models. For example, we can use fine-grained text style transfers to add specific stylistic properties to free-form text generated by large language models while keeping the content of the generated text unchanged.

Fine-grained text style transfers can be composed to achieve higher-level style transfers, and they even have the potential to mitigate social bias in large text generation models (Lyu et al., 2021). Therefore, it is important to develop techniques to achieve automated fine-grained text style transfer. Existing works are still quite far from perfect on a lot of the fine-grained style transfers compared to human performance (Lyu et al., 2021; Kim, 2021), and composing multiple fine-grained style transfers remains challenging.

2.2 Diffusion Probabilistic Models

Recently, diffusion models (Ho et al., 2020) is widely used to generate high quality and diverse images. Its methodology consists of two phases: the first phase is the forward diffusion phase, which adds Gaussian noise to the input image x_0 as the time stamp increases, and after enough steps the image is reduced to pure Gaussian noise x_t. The second phase is the recovery phase, in which a model is trained to gradually remove noise from x_t until it recovers the original image x_0. During inference, we start from a randomly sampled gaussian noise x_t and use the denoising model to gradually infer an image x_0.

Diffusion-based language generation models follows a similar approach where we perform the diffusion and denoising process in the token embedding space. We will explain the model we use, which is built upon DiffuSeq (Gong et al., 2022), in details in the next section.

3 Methodology

We adapt DiffuSeq (Gong et al., 2022) to be able to perform fine-grained text style transfer given a source sentence and specified transfer operation(s), as illustrated in Figure 1. We model the transfer as a conditional generation process, where the condition includes the source sentence and the specified transfer operation(s). We first define a set of special style tokens, one for each possible individual fine-grained transfer. If we wish to perform one or more transfer on the source sentence, we will prepend the corresponding special token(s) to the beginning of the source sentence to form the condition S.

We use BERT tokenizer to tokenize the input into discrete token ids, and adopt a token embedding layer to encode both the source (including prepended style tokens) and the ground truth target sentence (during training) to obtain the embedded source Z^S and target Z_0^{TRG}. For the diffusion pro-
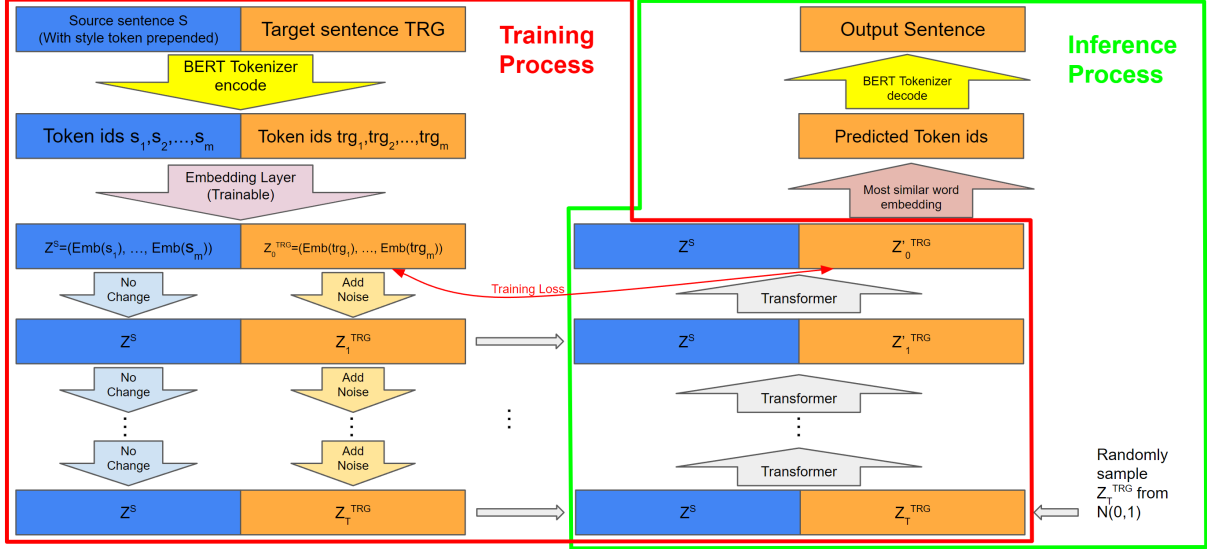
Figure 1: An illustration of the training and inference process of our diffusion-based language model. The diffusion process is performed over the sequence of token embeddings of the target sentence Z_0^{TRG}, and the source sentence's token embeddings (Z^S) are concatenated before Z^{TRG}. During the backward diffusion process, the combined sequence is fed into the transformer model to gradually recover/generate Z_0^{TRG}.

cess, we use a transformer model to recover the target embedding. Both the diffusion transformer and the token embeddings are initialized randomly and jointly optimized. In other words, our model does not rely on any prior knowledge about our task or the English Language in general.

We use the simplified diffusion objective during training: for each input (S, TRG) where S is the source sentence (with style tokens) and TRG is the ground truth target sentence, we randomly sample a step number t from $1, 2, ...T$, where T is the maximum number of steps, and add t steps of random Gaussian noise to Z_0^{TRG} following a linear diffusion schedule to obtain Z_t^{TRG}. We then concatenate Z^S and Z_t^{TRG} and input the concatenated sequence into our diffusion transformer, where we only take the output embeddings at the locations corresponding to Z_t^{TRG} as $Z_0'^{TRG}$. Our training objective is simply going to be the MSE Loss between Z_0^{TRG} and $Z_0'^{TRG}$.

During inference, we randomly initialize $Z_T'^{TRG} \sim N(0, 1)$, and encode the condition (source sentence and style tokens) into Z^S. Then we concatenate them and use our transformer to predict a temporary $Z_{0_{temp}}'^{TRG}$, add $T-1$ steps of noise back to the temporary $Z_{0_{temp}}'^{TRG}$ to obtain $Z_{T-1}'^{TRG}$. We repeat this process until we get $Z_0'^{TRG}$. For each embedding in $Z_0'^{TRG}$, we find the closest embedding in our token embedding layer by cosine distance, and decode the embedding to that token. Then we combine the tokens to form the output

sentence in natural language.

4 Experiments

4.1 Dataset

StylePTB (Lyu et al., 2021) contains paired sentences before/after each transfer for 21 fine-grained transfers, as well as paired data for compositions of multiple fine-grained transfers. For single transfers, we will focus on the 13 non-lexical fine-grained style transfers following (Lyu et al., 2021). The number of sentence pairs available from StylePTB for each transfer and examples of sentences before/after each transfer are shown in Table 1. For compositional transfers, we will use the Tense + Voice and Tense + PP Removal transfers from the compositional part of StylePTB dataset (same as the ones used for evaluation in (Lyu et al., 2021)). Each compositional dataset contains all combinations of valid transfers (for example, Tense + Voice dataset contains all valid combinations of 0/1/2 transfers regarding tense and voice, such as To-Future + Active-To-Passive or To-Past + No-Voice-Change).

StylePTB was built with only 7719 different sentences from Penn Tree Bank (Marcus et al., 1993) plus their stylistic variations, so both the amount and the diversity of training data are very limited, thus making this task even more challenging for DiffuSeq since it does not have access to external knowledge or pre-trained weights and have to extract all linguistic knowledge from limited data.

For fair comparison, we preprocess the data following the same criterion as (Lyu et al., 2021): we replace numbers with NUM token, and we replace each word that occurs less than 3 times in the training set with UNK token. We also split the data into train/valid/test splits with proportions of 0.9/0.05/0.05 using the same splits as all previous works.

4.2 Evaluation Metrics

We use the same evaluation methods as (Lyu et al., 2021) and report 7 metrics from nlg-eval package (Sharma et al., 2017) (BLEU 1-4, METEOR, ROUGE-L, CiDER) between the generated transferred sentence and the ground truth target sentence from the dataset.

4.3 Single style transfer experiment

4.3.1 Baselines

We report performance of the following baselines for single style transfer:

1. **GPT-2**: Directly finetuning GPT-2 medium model (Radford et al., 2019) with paired data. Performance reported from (Lyu et al., 2021).

2. **Seq2Seq**: GRU sequence-to-sequence language model (Sutskever et al., 2014) with attention. Performance reported from (Lyu et al., 2021).

3. **RetrieveEdit (Hashimoto et al., 2018)**: For an input data x, a retriever model will go through the training set to find a similar sentence pair (x', y') and a trained editor edits y' into desired output y. Performance reported from (Lyu et al., 2021).

4. **Steering Vector (Subramani et al., 2022)**: extract steering vectors directly from pretrained LMs to guide generation

5. **TAILOR (Ross et al., 2021)**: output sentences conditioned on control codes by a pretrained seq2seq model

6. **Neural QCFG (Kim, 2021)**: It presents a sequence-to-sequence text learning by explicitly modeling the alignment between target trees with the source.

7. **Neural QCFG + copy (Kim, 2021)**: Neural QCFG with an option to copy certain tokens from source sentence

Among these baselines, **GPT-2**, **Steering Vector** and **TAILOR** uses pre-trained language models, **Neural QCFG** and **Neural QCFG + copy** requires external grammar parsers, and **RetrieveEdit** uses GLOVE word embeddings.

We also included **Human** performance on these tasks (reported in (Lyu et al., 2021) by asking human annotators to manually perform the style transfer tasks) for comparison.

4.3.2 Results and Analysis

For single style transfers, we tried two different diffusion-based approaches: (1) we train a separate diffusion model for each individual style transfer, and (2) we train one diffusion model for all 13 transfers evaluated. For approach (2), we add a style token at the beginning of the input sentence to indicate which of the 13 transfers needs to be performed. We call approach (2) DiffuSeq Multitask.

The original StylePTB paper (Lyu et al., 2021) puts the non-lexical transfers into 3 difficulty categories (easy, medium, hard) by average hamming distance between input and output of the transfer. We report the results of our experiment using the same categorization, where we show results on easy and medium transfers in Table 2 and hard transfers in Table 3.

Surprisingly, DiffuSeq Multitask outperforms DiffuSeq on all transfers, even though DiffuSeq Multitask has to handle 13 different transfers in one model while each DiffuSeq model only needs to handle 1 transfer. This is possibly due to the additional training data from all the tasks that the multitask model learns better representations for words and sentences and gains more accurate knowledge of grammatical patterns of English, which is shared across all tasks.

Moreover, DiffuSeq Multitask significantly outperforms all baselines in all easy and medium transfers, and also achieves state-of-the-art on most metrics on hard transfers, only falling slightly behind Neural QCFG + copy in some metrics. This is really impressive considering that our approach leverages no external knowledge while all baselines except Seq2Seq utilizes either pretrained language models, pretrained word embeddings, or external grammar tree parser. Neural-QCFG-based methods are especially dependent on external linguistics knowledge and existing grammar parsers. DiffuSeq Multitask's performance is also on par with human performance on easy and medium transfers, indicating that DiffuSeq Multitask is close to fully solving the easy and medium difficulty transfers.

Easy Transfers	Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
	GPT2	0.895	0.852	0.813	0.778	0.540	0.899	7.709
	SEQ2SEQ	0.527	0.368	0.261	0.188	0.173	0.531	1.525
	RETRIEVEEDIT	0.899	0.854	0.815	0.778	0.531	0.901	7.731
To Future Tense	STEERING VECTOR	0.699	-	-	-	-	-	-
	TAILOR	0.873	-	-	-	-	-	-
	DIFFUSEQ	0.976	0.956	0.937	0.917	0.646	0.973	9.145
	DIFFUSEQ MULTITASK	**0.985**	**0.972**	**0.959**	**0.946**	**0.677**	**0.983**	**9.454**
	HUMAN	0.954	0.915	0.884	0.855	0.636	0.964	9.174
	GPT2	0.836	0.776	0.722	0.674	0.484	0.842	6.700
	SEQ2SEQ	0.478	0.313	0.204	0.133	0.155	0.490	1.374
	RETRIEVEEDIT	0.935	0.903	0.873	0.847	0.606	0.933	8.358
To Past Tense	STEERING VECTOR	0.478	-	-	-	-	-	-
	TAILOR	0.711	-	-	-	-	-	-
	DIFFUSEQ	0.973	0.959	0.946	0.932	0.697	0.976	9.352
	DIFFUSEQ MULTITASK	**0.986**	**0.977**	**0.968**	**0.958**	**0.709**	**0.987**	**9.588**
	HUMAN	0.974	0.957	0.939	0.916	0.709	0.982	9.549
	GPT2	0.754	0.663	0.586	0.524	0.412	0.772	5.293
	SEQ2SEQ	0.516	0.361	0.267	0.210	0.190	0.518	1.819
	RETRIEVEEDIT	0.909	0.870	0.830	0.793	0.599	0.916	7.987
To Present Tense	STEERING VECTOR	0.692	-	-	-	-	-	-
	TAILOR	0.884	-	-	-	-	-	-
	DIFFUSEQ	0.965	0.948	0.932	0.916	0.713	0.964	9.072
	DIFFUSEQ MULTITASK	**0.975**	**0.961**	**0.947**	**0.933**	**0.719**	**0.977**	**9.310**
	HUMAN	0.969	0.952	0.936	0.918	0.745	0.979	9.501
	GPT2	0.647	0.508	0.394	0.308	0.313	0.652	3.259
	SEQ2SEQ	0.450	0.274	0.172	0.112	0.140	0.469	1.171
	RETRIEVEEDIT	0.897	0.841	0.786	0.731	0.511	0.919	7.461
ADJ or ADV Removal	STEERING VECTOR	0.721	-	-	-	-	-	-
	TAILOR	0.781	-	-	-	-	-	-
	DIFFUSEQ	0.903	0.809	0.731	0.664	0.488	0.888	6.708
	DIFFUSEQ MULTITASK	**0.949**	**0.908**	**0.868**	**0.829**	**0.563**	**0.946**	**8.237**
	HUMAN	0.933	0.894	0.870	0.847	0.591	0.965	8.924
Medium Transfers	Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
	GPT2	0.398	0.210	0.081	0.001	0.184	0.406	0.886
	SEQ2SEQ	0.393	0.280	0.207	0.161	0.162	0.391	1.492
	RETRIEVEEDIT	0.541	0.423	0.301	0.176	0.247	0.547	2.536
PP Front to Back	STEERING VECTOR	0.819	-	-	-	-	-	-
	TAILOR	0.842	-	-	-	-	-	-
	DIFFUSEQ	0.605	0.409	0.301	0.247	0.271	0.514	2.273
	DIFFUSEQ MULTITASK	**0.978**	**0.931**	**0.893**	**0.856**	**0.567**	**0.901**	**8.374**
	HUMAN	0.965	0.959	0.952	0.945	0.690	0.970	9.671
	GPT2	0.407	0.241	0.091	0.001	0.166	0.406	0.931
	SEQ2SEQ	0.298	0.157	0.090	0.060	0.112	0.284	0.606
	RETRIEVEEDIT	0.649	0.584	0.535	0.491	0.333	0.656	4.667
PP Front to Back	DIFFUSEQ	0.603	0.400	0.291	0.242	0.266	0.514	2.255
	DIFFUSEQ MULTITASK	**0.983**	**0.944**	**0.905**	**0.868**	**0.610**	**0.950**	**8.664**
	HUMAN	1.000	1.000	1.000	1.000	1.000	1.000	10.000
	GPT2	0.763	0.700	0.645	0.593	0.419	0.787	6.012
	SEQ2SEQ	0.330	0.195	0.121	0.081	0.112	0.363	1.004
	RETRIEVEEDIT	0.798	0.770	0.739	0.712	0.478	0.846	7.111
PP Removal	STEERING VECTOR	0.393	-	-	-	-	-	-
	TAILOR	0.717	-	-	-	-	-	-
	DIFFUSEQ	0.856	0.803	0.758	0.717	0.515	0.872	7.235
	DIFFUSEQ MULTITASK	**0.950**	**0.937**	**0.919**	**0.902**	**0.624**	**0.948**	**8.606**
	HUMAN	0.957	0.944	0.931	0.919	0.681	0.976	9.207
	GPT2	0.430	0.332	0.247	0.176	0.250	0.588	3.090
	SEQ2SEQ	0.317	0.192	0.110	0.001	0.100	0.368	1.041
	RETRIEVEEDIT	0.706	0.678	0.647	0.607	0.405	0.767	6.183
Substatement Removal	STEERING VECTOR	0.120	-	-	-	-	-	-
	DIFFUSEQ	0.688	0.592	0.493	0.388	0.364	0.718	4.285
	DIFFUSEQ MULTITASK	**0.884**	**0.860**	**0.825**	**0.781**	**0.555**	**0.895**	**7.165**
	HUMAN	0.731	0.720	0.705	0.685	0.607	0.788	7.691
	GPT2	0.479	0.305	0.189	0.121	0.207	0.475	1.359
	SEQ2SEQ	0.345	0.180	0.094	0.053	0.098	0.335	0.632
	STEERING VECTOR	0.772	-	-	-	-	-	-
Information Addition	RETRIEVEEDIT	0.493	0.396	0.328	0.275	0.284	0.603	3.401
	DIFFUSEQ	0.809	0.572	0.420	0.3081	0.3829	0.676	3.439
	DIFFUSEQ MULTITASK	**0.911**	**0.800**	**0.706**	**0.623**	**0.483**	**0.835**	**6.038**
	HUMAN	0.846	0.762	0.690	0.624	0.521	0.892	6.863

Table 2: Evaluation results on easy and medium transfers. DiffuSeq Multitask achieves State of the art performance in every metric, and is on par with human performance.

Hard Transfers	Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
Active To Passive	GPT2	0.476	0.329	0.238	0.189	0.216	0.464	1.820
	Seq2seq	0.373	0.220	0.141	0.103	0.131	0.345	0.845
	RetrieveEdit	0.681	0.598	0.503	0.427	0.383	0.663	4.535
	Steering Vector	0.666	-	-	-	-	-	-
	Tailor	0.556	-	-	-	-	-	-
	Neural QCFG	0.431	0.637	0.548	0.472	0.415	0.695	4.294
	Neural QCFG + copy	0.836	0.771	0.713	0.662	0.499	0.803	6.410
	DiffuSeq	0.839	0.580	0.302	0.196	0.225	0.512	2.344
	Diffuseq MultiTask	**0.918**	**0.835**	**0.752**	**0.681**	**0.521**	**0.844**	**6.913**
	Human	0.931	0.881	0.835	0.795	0.587	0.905	8.603
Passive To Active	GPT2	0.433	0.271	0.167	0.120	0.191	0.434	1.329
	Seq2seq	0.339	0.214	0.160	0.132	0.126	0.331	1.062
	RetrieveEdit	0.714	0.659	0.559	0.474	0.397	0.732	5.024
	Steering Vector	0.574	-	-	-	-	-	-
	DiffuSeq	0.829	0.550	0.282	0.192	0.205	0.502	2.224
	Diffuseq MultiTask	**0.955**	**0.896**	**0.834**	**0.777**	**0.555**	**0.913**	**8.028**
	Human	0.977	0.962	0.942	0.919	0.685	0.973	9.409
Adjective Emphasis	GPT2	0.263	0.079	0.028	0.000	0.112	0.188	0.386
	Seq2seq	0.187	0.058	0.018	0.000	0.059	0.179	0.141
	RetrieveEdit	0.387	0.276	0.211	0.164	0.193	0.369	1.679
	Steering Vector	0.774	-	-	-	-	-	-
	Neural QCFG	0.348	0.178	0.062	0.000	0.162	0.317	0.667
	Neural QCFG + copy	0.676	0.506	0.393	0.316	0.373	0.683	3.424
	DiffuSeq	0.620	0.382	0.215	0.152	0.243	0.335	2.231
	Diffuseq MultiTask	**0.775**	**0.600**	**0.477**	**0.386**	**0.423**	**0.673**	**4.007**
	Human	0.834	0.753	0.679	0.611	0.522	0.811	6.796
Verb/Action Emphasis	GPT2	0.309	0.170	0.095	0.041	0.140	0.292	0.593
	Seq2seq	0.289	0.127	0.066	0.038	0.098	0.275	0.300
	RetrieveEdit	0.416	0.284	0.209	0.148	0.223	0.423	1.778
	Steering Vector	0.548	-	-	-	-	-	-
	Neural QCFG	0.431	0.250	0.14	0.073	0.219	0.408	1.097
	Neural QCFG + copy	0.664	0.512	**0.407**	**0.319**	0.370	0.589	**3.227**
	DiffuSeq	0.453	0.210	0.101	0.054	0.205	0.379	0.785
	Diffuseq MultiTask	**0.693**	**0.516**	0.370	0.261	**0.373**	**0.596**	2.950
	Human	0.649	0.569	0.493	0.421	0.433	0.693	5.668

Table 3: Evaluation results on hard transfers. Diffuseq Multitask achieves State-of-the-art performance on most metrics, and is only slightly behind Neural QCFG + copy on some metrics.

4.4 Compositional style transfer experiment

4.4.1 Baselines

We will report performance of the following baselines for compositional fine-grained style transfers:

1. **SeqGPT**: Sequentially applying fine-tuned GPT-2 for each single style transfer. Performance reported from (Lyu et al., 2021).
2. **CS-GPT**: A modified GPT-2 model that takes in style tokens as indication of which style transfers to apply. Performance reported from (Lyu et al., 2021).

4.4.2 Results and Analysis

For compositions of multiple fine-grained style transfers, we train one single DiffuSeq model to handle all compositions and use style tokens to indicate which transfers to compose for the input sentence, similar to CS-GPT (Lyu et al., 2021). The results are shown in Table 4. DiffuSeq significantly outperforms baselines in all tasks and all metrics. Therefore, not only does our diffusion model work well for single fine-grained style transfers, it also works well for compositions of multiple fine-grained style transfers.

5 Related Works

5.1 Automated Text Style Transfer

The goal of the text style transfer (TST) task is to change the style of the sentence while retaining its style-independent content. Previous works in TST includes the following approaches: statistical NLP methods (Hovy, 1987; Xu et al., 2012), neural generative models (Prabhumoye et al., 2018; Lample et al., 2019; He et al., 2020), Retrieve-and-Edit approaches (Li et al., 2018; Hashimoto et al., 2018; Guu et al., 2018; Sudhakar et al., 2019; Madaan et al., 2020), and Transformer-based approach (Lyu et al., 2021). Some of these methods can already achieve high performance on certain high-level transfers (such as sentiment transfers (Shen et al., 2017) and formality transfers (Rao and Tetreault, 2018)), but fine-grained text style tranfer remains challenging for the above approaches (Lyu et al.,

Dataset	Transfers	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
Tense + Voice	ToPast+ ActiveToPassive	SEQGPT	0.332	0.155	0.057	0.024	0.144	0.300	0.636
		CS-GPT	0.409	0.238	0.133	0.064	0.180	0.378	1.029
		DIFFUSEQ	**0.744**	**0.555**	**0.420**	**0.324**	**0.353**	**0.656**	**3.753**
	ToFuture+ ActiveToPassive	SEQGPT	0.391	0.222	0.120	0.065	0.167	0.373	0.866
		CS-GPT	0.496	0.340	0.240	0.185	0.217	0.479	1.800
		DIFFUSEQ	**0.821**	**0.705**	**0.615**	**0.542**	**0.414**	**0.762**	**5.281**
	ToFuture+ PassiveToActive	SEQGPT	0.401	0.212	0.097	0.048	0.163	0.385	0.888
		CS-GPT	0.528	0.364	0.259	0.197	0.234	0.524	2.020
		DIFFUSEQ	**0.744**	**0.555**	**0.420**	**0.324**	**0.353**	**0.656**	**3.753**
	ToPast+ PassiveToActive	SEQGPT	0.381	0.210	0.098	0.045	0.156	0.368	0.876
		CS-GPT	0.474	0.297	0.175	0.099	0.206	0.473	1.513
		DIFFUSEQ	**0.864**	**0.772**	**0.697**	**0.635**	**0.460**	**0.825**	**6.519**
	ToPresent+ PassiveToActive	SEQGPT	0.348	0.189	0.085	0.037	0.142	0.343	0.745
		CS-GPT	0.523	0.366	0.264	0.210	0.243	0.522	2.118
		DIFFUSEQ	**0.797**	**0.686**	**0.603**	**0.536**	**0.414**	**0.756**	**5.378**
	ToPresent+ ActiveToPassive	SEQGPT	0.396	0.256	0.177	0.136	0.179	0.384	1.209
		CS-GPT	0.503	0.358	0.271	0.223	0.233	0.491	2.118
		DIFFUSEQ	**0.878**	**0.787**	**0.715**	**0.656**	**0.482**	**0.849**	**6.823**
Tense + PP Removal	ToFuture+ PPRemoval	SEQGPT	0.722	0.644	0.581	0.524	0.385	**0.755**	5.562
		CS-GPT	0.738	0.652	0.578	0.518	0.393	0.755	5.289
		DiffuSeq	**0.913**	**0.876**	**0.841**	**0.808**	**0.557**	**0.911**	**7.906**
	ToPast+ PPRemoval	SEQGPT	0.714	0.640	0.573	0.510	0.374	0.724	5.152
		CS-GPT	0.772	0.695	0.624	0.564	0.421	0.775	5.585
		DiffuSeq	**0.911**	**0.881**	**0.849**	**0.818**	**0.568**	**0.908**	**7.825**
	ToPresent+ PPRemoval	SEQGPT	0.618	0.518	0.435	0.368	0.338	0.663	4.119
		CS-GPT	0.709	0.609	0.523	0.446	0.718	0.718	4.588
		DiffuSeq	**0.908**	**0.859**	**0.820**	**0.788**	**0.558**	**0.895**	**7.439**

Table 4: Results on compositions of Tense + Voice transfers and Tense + PP Removal Transfers. DiffuSeq was able to outperform all 3 baselines from (Lyu et al., 2021) by a large margin.

2021). In this paper, we explored a new approach for fine-grained TST utilizing Diffusion Models.

5.2 Natural language processing with diffusion model

There have been two approaches for leveraging diffusion models into text data: the first approach takes advantage of the diffusion model in the continuous domain, like Diffusion-LM (Li et al., 2022), and DiffuSeq (Gong et al., 2022), where we start from a gaussian noise vector, and gradually denoise this noise vector to the desired sentence; the second approach applies diffusion model into discrete state space, like Multinomial Diffusion (Hoogeboom et al., 2021), DDPMs (Austin et al., 2021), and DiffusionBERT (Austin et al., 2021). In this paper, we chose to build upon the first type of model, because they are closer to the original diffusion models for images (where diffusion happens in continuous space) and they have shown successes on tasks that requires control over generations.

6 Limitations and Future works

One significant limitation of our work is that we only explored the capabilities of diffusion-based language models under a challenging circumstance where it is not allowed to use pre-trained weights or grammar parsers, which means we did not utilize this kind of model to its full potential, so a future research direction could be exploring possible ways to further improve the model's performance by leveraging pretrained weights or word embeddings, and train with enough data to find the full potential of these models.

Another limitation of our work is that we only explored one typical diffusion-based language model, so our conclusions may not generalize to special types of diffusion-based language models (such as ones that uses discrete state space). We also conducted all experiments using the exact same model architecture design. In the future, we plan to experiment with different architectures for the diffusion model, such as more sophisticated conditioning methods (currently we just concatenate the source to the target, but we would like to try other ways of conditioning on the source, such as cross attention, as these conditioning methods for diffusion models have promising performance in the image generation domain).

Lastly, we found that diffusion-based language models work well with limited data and no external knowledge or pre-trained weights, thus these mod-

els may have great potential under low-resource settings, but we didn't apply them to any real low-resource settings (such as low-resource languages, rare domains, etc) in this paper, and we would like to do that in the future to explore the full potential of diffusion-based language models.

7 Conclusions

In this paper, we explored the capabilities of diffusion-based models on fine-grained text style transfer, a task that requires a high level of control over generated text, with no external knowledge or pre-trained weights and with very limited training data. Our diffusion-based language model, which builds upon DiffuSeq (Gong et al., 2022), achieves state-of-the-art performance on all transfers as well as composition of transfers, outperforming all previous works on this dataset, including ones that uses pre-trained weights, word embeddings, and external grammar parsers. It is even on par with human performance on many transfers. Therefore, our model is a great step towards solving automated fine-grained text style transfer.

Moreover, our work, together with previous works such as Diffusion-LM (Li et al., 2022), demonstrates that diffusion-based language models could have great potential in controllable text generation under low-resource settings. Under low-resource settings (such as rarely spoken language or uncommon tasks), it would be difficult to find existing large language models or pre-trained weights, and available training data will likely be very limited, so most approaches based on finetuning existing models or large amounts of training will not work well, and diffusion-based language models could be an alternative to consider.

Acknowledgement

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Asli Celikyilmaz, Li Deng, and Dilek Hakkani-Tür. 2018. Deep learning in spoken and text-based dialog systems. In *Deep Learning in Natural Language Processing*, pages 49–78. Springer.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy S Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Advances in Neural Information Processing Systems*, pages 10052–10062.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Elizabeth S Kim, Lauren D Berkovits, Emily P Bernier, Dan Leyzberg, Frederick Shic, Rhea Paul, and Brian Scassellati. 2013. Social robots as embedded reinforcers of social behavior in children with autism. *Journal of autism and developmental disorders*.

Yoon Kim. 2021. Sequence-to-sequence learning with latent neural grammars. *Advances in Neural Information Processing Systems*, 34:26302–26317.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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. Association for Computational Linguistics.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*.

Paul Pu Liang, Jeffrey Chen, Ruslan Salakhutdinov, Louis-Philippe Morency, and Satwik Kottur. 2020. On emergent communication in competitive multi-agent teams. In *AAMAS*.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Styleptb: A compositional benchmark for fine-grained controllable text style transfer. *arXiv preprint arXiv:2104.05196*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. 2010. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*.

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. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *arXiv preprint arXiv:2107.07150*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6833–6844.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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