GENIE: Large Scale Pre-training for Generation with Diffusion Model

Zhenghao Lin [1 2] **Yeyun Gong** [3] **Yelong Shen** [4] **Tong Wu** [5 2] **Zhihao Fan** [6 2]
Chen Lin [1] **Weizhu Chen** [4] **Nan Duan** [3]

Abstract

In this paper, we propose a large-scale language pre-training for text **GEN**eration using d**I**ffusion mod**E**l, which is named **GENIE**. GENIE is a pre-training sequence-to-sequence text generation model which combines Transformer and diffusion. The diffusion model accepts the latent information from the encoder, which is used to guide the denoising of the current time step. After multiple such denoise iterations, the diffusion model can restore the Gaussian noise to the diverse output text which is controlled by the input text. Moreover, such architecture design also allows us to adopt large scale pre-training on the GENIE. We propose a novel pre-training method named *continuous paragraph denoise* based on the characteristics of the diffusion model. Extensive experiments on the XSUM, CNN/DAILYMAIL, and GIGAWORD benchmarks shows that GENIE can achieves comparable performance with various strong baselines, especially after pre-training, the generation quality of GENIE is greatly improved. We have also conduct a lot of experiments on the generation diversity and parameter impact of GENIE. The code for GENIE will be made publicly available.

1. Introduction

Text generation is an important branch of natural language processing area. Early text generation works mostly adopts recurrent neural network (RNN) based approaches (Pawade et al., 2018; Song et al., 2018; Gu et al., 2016; Qi et al., 2021). With the rise of pre-training language models, Transformer (Vaswani et al., 2017b) has become the mainstream text generation framework. More and more Transformer-based pre-trained language models (Qi et al., 2020; Lewis et al., 2019; Raffel et al., 2020) are applied to text generation tasks, and has been proven to be very effective in improving the text generation performance.

Diffusion model has been actively explored in various content genration tasks in recently years, i.e., image generation (Ho et al., 2020; Song et al., 2020), molecule (分子) generation (Hoogeboom et al., 2022), video generation (Ho et al., 2022b) and text generation (Li et al., 2022; Gong et al., 2022; Strudel et al., 2022; Reid et al., 2022).

Based on Transformer (Vaswani et al., 2017b) architecture, different pre-training paradigms are proposed, including decoder only (Radford et al., 2018), encoder only (Dong et al., 2019b), and encoder-decoder architectures (Qi et al., 2020; Lewis et al., 2019). Encoder-decoder model architecture as the most popular one, there are two commonly used decoding methods on the decoder side: autoregressive decoding (AR) and non-autoregressive decoding (NAR) (Qi et al., 2021). Autoregressive models are trained to predict the next word in a sequence by using the previous words as context. This approach is effective, but it is slow because the model must generate each word sequentially. Non-autoregressive models, on the other hand, are trained to generate an entire sequence of words all at once. This approach is faster, but it is also less accurate because the model has less context to work with.

In this paper, we propose a novel text generation approach, called GENIE, it combines the diffusion model and Transformer-based method. GENIE follows the encoder-decoder architecture, where the input text is transformed into a sequence of hidden vectors using a Transformer encoder. Hidden information from encoder will guide denoise in the process of reverse diffusion. After multiple time step iterations, the diffusion model can restore the randomly sampled Gaussian noise to the output text controlled by the input text.

Since pre-training has been proven effective in Transformer-based methods, here we propose an end-to-end pre-training method for the GENIE. Different from pre-training tasks (infilling token or text split) commonly used in previous works (Qi et al., 2020; Lewis et al., 2019; Raffel et al., 2020), we propose a novel pre-training task, named *continuous paragraph denoise* (CPD). CPD allows the model to predict the noise added to continuous paragraphs in the current time

[1]Xiamen University [2]This work was done during an internship in MSRA [3]Microsoft Research Asia [4]Microsoft [5]Tsinghua University [6]Fudan University. Correspondence to: Chen Lin <chenlin@xmu.edu.cn>.

step based on the paragraph context information and the noisy paragraph information.

We conduct an extensive experiments on three popular publicly available benchmarks: XSum (Narayan et al., 2018), CNN/DailyMail (Hermann et al., 2015), and Gigaword (Rush et al., 2015). Experimental results show that GENIE achieves comparable performance with Transformer-based methods and the proposed pre-training method can improve the performance effectively.

The main contributions of this work are three folds: 1). GENIE is the first large-scale language pre-trained model based on the diffusion framework. 2). We propose a novel CPD loss as the pre-training objective, and adopt the encoder-diffusion model framework to support sequence-to-sequence tasks. 3). Our extensive experiments have validated the effectiveness of the pre-trained diffusion model in downstream tasks, and we have conducted a large number of ablation studies, such as exploring the impact of diffusion steps in GENIE and text generation diversity.

2. Preliminary

2.1. Task Definition

In the standard sequence-to-sequence(seq-to-seq) task, assume given a source text $\boldsymbol{s} = \{w_1^s, w_2^s, \ldots, w_n^s\}$ containing n tokens, it generates target text sequence $\boldsymbol{y} = \{w_1^y, w_2^y, \ldots, w_n^y\}$. Non auto-regressive (NAR) based generation model can be simply expressed as:

$$p_{\text{NAR}} = \prod_{i=0}^{n} p\left(w_i^y \mid \boldsymbol{s}\right) \tag{1}$$

2.2. Diffusion model

In the diffusion model, the diffusion process can be regarded as a Markov process. The time step t starts from $t = 0$, and Gaussian noise is gradually added to the variable $\boldsymbol{x}_0$ in the forward diffusion process. At the time step $t+1$, the latent variable $\boldsymbol{x}_{t+1}$ is only determined by the $\boldsymbol{x}_t$ at time t, expressed as:

$$q\left(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_{t+1}; \sqrt{1-\beta_{t+1}}\boldsymbol{x}_t, \beta_{t+1}\mathbf{I}\right) \tag{2}$$

where $\beta_{t+1} \in (0,1)$ represents the scaling of variance at time step $t+1$. When the time step $t = T$ is large enough, $\boldsymbol{x}_T$ tends to standard Gaussian noise $\mathcal{N}(\boldsymbol{x}_T; 0, \mathbf{I})$.

The inverse process of diffusion can be regarded as the process of denoising. The diffusion model predicts the noise of current time step t and denoise to previous state $\boldsymbol{x}_{t-1}$, which variance and mean are respectively expressed as μ_θ^{t-1} and σ_{t-1}:

$$p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \mu_\theta^{t-1}, \sigma_{t-1}\right) \tag{3}$$

$$\mu_\theta^{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} z_\theta\left(\boldsymbol{x}_t, t\right)\right) \tag{4}$$

$$\sigma_{t-1}^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \tag{5}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $z_\theta \sim \mathcal{N}(0, \mathbf{I})$ is predicted by the diffusion model. Finally, the simplified objective function based on VLB (Ho et al., 2020) is used to train the diffusion model:

$$\mathcal{L}_{diff} = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \left\| \mu_\theta^{t-1} - \hat{\mu}_{t-1} \right\|^2 \tag{6}$$

where $\hat{\mu}_{t-1}$ is calculated from the reverse conditional probability $q(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t, \boldsymbol{x}_0)$.

3. Model

We propose a novel seq-to-seq training framework in GENIE based on diffusion model as shown in Figure 1. The source text sequence is $\boldsymbol{s}$: *In the World Cup 2022, [MASK] won people's praise.*. The target text sequence is $\boldsymbol{y}$: *Messi's performance*. GENIE is composed of a bidirectional encoder model and a cross-attention diffusion model: the encoder model produces the distributed vector representation for source text $\boldsymbol{s}$, i.e., $\boldsymbol{H_s} = \text{Encoder}(\boldsymbol{s})$; The diffusion model takes $\boldsymbol{H_s}$ and a Gaussian noise as input, to generate the embeddings of target sequence iteratively.

Different from the traditional autoregressive text generation paradigm, the diffusion model in GENIE generates the whole target sequence embeddings at each denoising step. Hence, GENIE is a non-autoregressive generation (NAR) model.

Encoder The encoder in GENIE is a 6-layer transformer model which takes the source text $\boldsymbol{s}$ as input with bidirectional self-attention. Specifically, given a source text sequence $\boldsymbol{s} = \{w_1^s, w_2^s, \ldots, w_n^s\}$ with n tokens, the encoder model computes the vector $\boldsymbol{h}_i$ for each token w_i. Thus, the source text $\boldsymbol{s}$ can be represented as $\boldsymbol{H_s}$ by the encoder model:

$$\boldsymbol{H_s} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n\} = \text{Encoder}(\boldsymbol{s}) \tag{7}$$

这里有个问题：在进行step t -> t+1的时候，Hs变不变
而且，h1-hn是怎么输入到diffusion 里边的

Text Diffusion Model The diffusion model in GENIE is composed of 6-layer transformer blocks with cross-attention on source text representation. It predicts the noise given the current diffusion step t, state $\boldsymbol{x}_t$ and source information $\boldsymbol{H_s}$, denoted as $z_\theta\left(\boldsymbol{x}_t, t, \boldsymbol{H_s}\right)$. We elaborate more details about grounding continuous state $\boldsymbol{x}_t$ with the discrete target tokens in the following discussion.

Inference Phase In the Inference phase, it starts from the time step $t = T$. We randomly sample a standard Gaussian distribution as state $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and predict the noise of

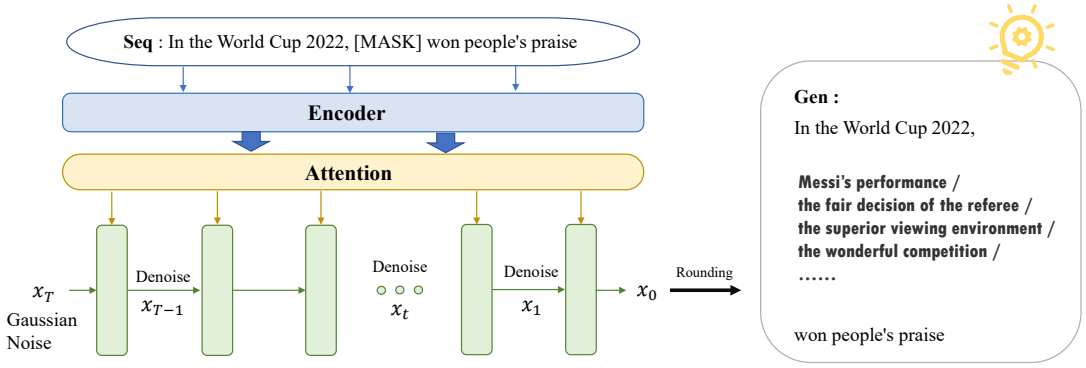

Figure 1. The framework of GENIE. The input text **Seq** interacts with the diffusion model through cross attention after encoding. Iterations with multiple time steps can fill various output texts **Gen** for the original mask of **Seq**.

the previous time step according to the equation 4 and 5 After arriving to $t = 0$, the final generated text is obtained by adopting the clamping trick (Li et al., 2022) to force the vectors in $\boldsymbol{x}_0$ close to word embeddings.

Training Phase In the training phase, we convert the target sequence $\boldsymbol{y} = \{w_1^y, w_2^y, \ldots, w_n^y\}$ into the continuous latent variable $\boldsymbol{x}_0$ with embedding function, which is expressed as:

$$q(\boldsymbol{x}_0|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}_0; \text{Emb}(\boldsymbol{y}), \beta_0 \mathbf{I}) \tag{8}$$

where $\text{Emb}(\cdot)$ is embedding function, β_0 represents the scaling of variance at time step $t = 0$. Through forward diffusion process (equation 2) and $\boldsymbol{x}_0$, we can obtain the latent variable $\boldsymbol{x}_t$ at time t as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \sqrt{1-\bar{\alpha}_t}\mathbf{I}\right) \tag{9}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

We sample t as the current time step, calculate the latent variable $\boldsymbol{x}_t$, and predict the noise of the current time step by combining the latent information $\boldsymbol{H}_s$ with cross attention. The mean and variance of the predicted noise is express as follows:

$$\mu_\theta^{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} z_\theta(\boldsymbol{x}_t, t, \boldsymbol{H}_s)\right) \tag{10}$$

$$\sigma_{t-1}^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \tag{11}$$

where z_θ is the Gaussian noise predicted by the denoising architecture, θ is the parameters of the GENIE. Finally, the training objectives of seq-to-seq based on diffusion model can be expressed as:

$$\mathcal{L}_{s2s} = \mathop{\mathbb{E}}_{q(\boldsymbol{x}_{0:T}|\boldsymbol{y})}[\sum_{t=1}^{T} \left\|\mu_\theta^{t-1} - \hat{\mu}_{t-1}\right\|^2 + \left\|\text{Emb}(\boldsymbol{y}) - \mu_\theta^0\right\|^2 - \log p_\theta(\boldsymbol{y}|\boldsymbol{x}_0)] \tag{12}$$

where $p_\theta(\boldsymbol{y}|\boldsymbol{x}_0) = \prod_{i=1}^{n} p_\theta(w_i^y|\boldsymbol{x}_0)$, represents mapping the continuous latent variable $\boldsymbol{x}_0$ into the discrete space token w_i^y.

3.1. Pre-training GENIE

不可估量的潜力

Diffusion model has immeasurable potential in the domain of natural language generation (NLG) because of the diversity of its generated results. However, due to its slow convergence speed and the inherent pool quality of non-autoregressive models, the diffusion model has not received enough attention in the domain of NLG. In this paper, we apply pre-training on diffusion language model and propose a novel pre-training task specifically for the diffusion language model. With the help of pre-training, GENIE greatly accelerates the convergence speed and significantly improves the quality of the generation. Meanwhile, GENIE can flexibly adapt to a variety of downstream tasks for NLG. The generated effect of GENIE is enough to announce 宣告 the birth of a brand-new large-scale pre-training language model.

For pre-training on diffusion language model, we propose a novel pre-training task, named *continuous paragraph denoise* (CPD). CPD allows the model to predict the noise added to continuous paragraphs in the current time step based on the paragraph context information and the noisy paragraph information.

一篇doc不重复单词数目：L

Specifically, given a document $\boldsymbol{d} = \{w_1^d, w_2^d, \ldots, w_l^d\}$ which contains l words, we need to select the paragraph 段落 $\boldsymbol{p} = \{w_1^p, w_2^p, \ldots, w_m^p\}$ from document $\boldsymbol{d}$, where $m = \lfloor \gamma * l \rfloor$ is the length of the paragraph $\boldsymbol{p}$, γ represents the proportion of paragraph $\boldsymbol{p}$ in the document $\boldsymbol{d}$. We add mask token ([MASK]) to the position where paragraph $\boldsymbol{p}$ is removed in document $\boldsymbol{d}$, and take the processed result $\boldsymbol{d}' = \{w_1^{d'}, w_2^{d'}, \ldots, [\text{MASK}], \ldots, w_{l-m}^{d'}\}$ as the input of GENIE encoder. Meanwhile, we take the noisy target $\boldsymbol{p}$

就是把段落p的不重复单词去掉，然后将这个词库作为输入，到encoder中，

decoder sample t step，计算xt，以及结合encoder的输出Hs，输出μt-1 然后计算loss

编程中一般至sample其中一步就行，而不是t=1-T求和

obtained through forward diffusion (equation 9) as the input of the GENIE denoising architecture, and predict the noise (equation 10 and 11) together with the context latent information from the output of GENIE encoder.

Through such pre-training, GENIE based on diffusion model can deepen the semantic understanding of the continuous text and achieve better denoising effect at each diffusion time step.

4. Experiments and Results

In this section, we will introduce the details of GENIE pre-training, the data setting, and show extensive experimental results on various NLG downstream tasks.

4.1. GENIE Pre-training

Model Framework Our model uses a 6-layer transformer as the encoder, and a 6-layer cross attention transformer as denoising architecture. In particular, in denoising architecture, we use the randomly embedding function to map discrete token into continuous variable. We set latent variable dim to 768 and embedding dim to 128.

Pre-training Data Resent works has shown that pre-training on large scale corpus can improve the performance of the model on downstream tasks (Lewis et al., 2019; Qi et al., 2020), which is also applicable to GENIE based on diffusion model. Following BART (Lewis et al., 2019), we use pre-training data consisting of 160Gb of news, books, stories, and web text. We segment sentences belonging to different chapters, and ensure that the input text length does not exceed 512.

Pre-training Setting We use the CPD task mentioned in §3.1 to pre-train GENIE on large-scale corpus. The proportion of continuous paragraph γ sets to 30%, hence, for the 512 length input, the target length is 153. We randomly extract 153 length targets from the text input, and leave [MASK] token at the extracted position. In the training process, we use Adam optimizer (Kingma & Ba, 2015) with learning rate 1e-4, and we set the batch size to 512. So far, we have pre-trained our model on 8 × 32GB NVIDIA A100 GPUs with 200w steps, lasting for 25 days. In the fine-tuning phase, the pre-training model we use are all 200w step checkpoints, so the main results are from the result of fine-tuning on 200w step pre-training model. However, pre-training on GENIE is not completely sufficient, and we are still work in process.

4.2. Fine-tune on Downstream Tasks

In order to verify the effectiveness of pre-training on GENIE based on diffusion model, we fine-tune and verify the effect of GENIE on a variety of downstream tasks. Through the above task, we can prove that the pre-trained GENIE can quickly adapt to different types of NLG tasks without long time training like other diffusion models.

Text Summarization As an important task in the NLG field, text summarization aims to summarize long documents into fluent short texts. In the experiment, we selected three widely used datasets: (a) GIGAWORD corpus (Rush et al., 2015), (b) CNN/DAILYMAIL (Hermann et al., 2015), and (c) XSUM (Narayan et al., 2018). In the process of fine-tuning, we set the learning rate to 5e-5 and the 120K training steps for all three dataset. In the inference process, we randomly sample 10 Gaussian noise for iteration denoising, and use the highest score as the final generated result. For evaluation, we following the existing work (Lewis et al., 2019; Qi et al., 2020), reporting F1 scores of **ROUGE-1**, **ROUGE-2**, and **ROUGE-L** on test set.

4.3. Baselines

We compare GENIE with the baselines of several mainstream methods. Specifically, these methods can be divided into two groups. The first group is the NAR model, including NAT (Gu et al., 2017), iNAT (Lee et al., 2018), CMLM (Ghazvininejad et al., 2019), LevT (Gu et al., 2019b), BANG (Qi et al., 2021), and InsT (Stern et al., 2019). Among them, InsT, iNAT, CMLM, LevT, and BANG can also be used in Semi-NAR, which can optimize the generation quality through multiple NAR iterations. It is worth noting that GENIE also belongs to the Semi-NAR model.

The second group is AR model, including LSTM (Greff et al., 2017) and Transformer (Vaswani et al., 2017a) without pre-training, and strong baselines MASS (Song et al., 2019), BART (Lewis et al., 2019), BANG (Qi et al., 2021), and ProphetNet (Qi et al., 2020) with large scale pre-training.

4.4. Main Results

The results comparing GENIE with the baselines on XSUM, CNN/DAILYMAIL, GIGAWORD are shown in Table 1 and Table 2. Although we have selected the best results for testing in multiple sampling, it can be easily observed that the experimental results are sufficient to prove that the pre-trained GENIE has great potential. In particular, on the XSUM dataset, the generation quality of GENIE is far better than other NAR methods and Semi-NAR methods, and on all three text summarization datasets, the text generation quality of GENIE can be comparable to that of the pre-trained AR model.

Moreover, we compared the pre-trained GENIE and GENIE trained from scratch (w/o pre-train). Among the results of three different dataset which are shown in Table 1 and Table 2, we found that GENIE with pre-training has signif-

Table 1. Results of Semi-NAR, NAR and AR on XSUM. Index **OVERALL** represents the average value of **ROUGE-1**, **ROUGE-2** and **ROUGE-L**. It should be noted that GENIE belongs to Semi-NAR.

Pattern	Methods	XSUM			
		ROUGE-1	ROUGE-2	ROUGE-L	OVERALL
NAR	NAT (Gu et al., 2017)	24.0	3.9	20.3	16.1
	iNAT (Lee et al., 2018)	24.0	4.0	20.4	16.1
	CMLM (Ghazvininejad et al., 2019)	23.8	3.6	20.2	15.9
	LevT (Gu et al., 2019b)	24.8	4.2	20.9	16.6
	BANG (Qi et al., 2021)	**32.6**	**9.0**	**27.4**	**23.0**
AR	Transformer (Vaswani et al., 2017b)	30.7	10.8	24.5	22.0
	MASS (Song et al., 2019)	39.7	17.2	31.9	29.6
	BART (Lewis et al., 2019)	38.8	16.2	30.6	28.5
	ProphetNet (Qi et al., 2020)	39.9	17.1	32.1	29.7
	BANG (Qi et al., 2021)	**41.1**	**18.4**	**33.2**	**30.9**
Semi-NAR	InsT (Stern et al., 2019)	17.7	5.2	16.1	13.0
	iNAT (Lee et al., 2018)	27.0	6.9	22.4	18.8
	CMLM (Ghazvininejad et al., 2019)	29.1	7.7	23.0	20.0
	LevT (Gu et al., 2019b)	25.3	7.4	21.5	18.1
	BANG (Qi et al., 2021)	34.7	11.7	29.2	25.2
	GENIE (w/o pre-train)	38.9	17.5	31.0	29.1
	GENIE	**42.1**	**20.7**	**34.4**	**32.4**

Table 2. The main results on CNN/DAILYMAIL and GIGAWORD, while all baselines are AR model.

Method	CNN/DAILYMAIL			GIGAWORD		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
LSTM (Greff et al., 2017)	37.3	15.7	34.4	34.2	16.0	31.8
Transformer (Vaswani et al., 2017a)	39.5	16.7	36.7	37.1	18.4	34.5
MASS (Song et al., 2019)	42.1	19.5	39.0	38.7	19.7	35.9
ProphetNet (Qi et al., 2020)	42.5	19.7	39.5	38.9	19.9	36.0
GENIE (w/o pre-train)	43.8	20.6	41.2	43.7	23.3	40.8
GENIE	**45.3**	**22.7**	**42.8**	**45.2**	**25.3**	**42.3**

icantly improved on **ROUGE-1**, **ROUGE-2**, **ROUGE-L** compared to GENIE without pre-training, proving the effectiveness of our pre-training method.

4.5. Generate Diversity Comparison

Nowadays, non-autoregressive generation has been well-known for its fast decoding. With the emergence of the diffusion based model such as GENIE, the advantages of non-autoregressive generation in diversity will be gradually valued. In this experiment, we will combine experimental index and case analysis to prove the rich diversity of GENIE in text generation.

First, we verify the diversity of GENIE generation from experimental index. We choose **SELF-BLEU** as the index of diversity. The smaller the value of **SELF-BLEU**, the richer the diversity of generated text. For comparison, we selected ProphetNet, a powerful baseline in the autoregressive model, which is pre-trained on large scale corpus. For ProphetNet, we use beam search (Xiao et al., 2022) and diverse beam search (Vijayakumar et al., 2016) methods to generate 10 result samples, and diverse beam strength sets to 0.5. For GENIE, we sample 10 Gaussian noises to generate 10 result samples. Finally, we use the 10 samples generated from XSUM and CNN/DAILYMAIL to calculate the **SELF-BLEU** scores. As results shown in Table 3, although the diversity of autoregressive generation can be improved slightly through diversity beam search, the improvement is still not significant. On the contrary, the diversity of generation is greatly improved by using the method of diffusion. The huge differences on **SELF-BLEU** represent that the GENIE model can truly generate diversity, rather than the differences of several words.

It may not be intuitive to analyze experimental index only, so second, we will conduct case study to verify the diversity of GENIE generation. We select three samples from the XSUM. The source sequence has been abbreviated for the convenience of presentation. For each sample, we list three summaries generated with diffusion method and diverse beam search method respectively. We ignore syntax errors in the generated results. As example summaries shown in Table 4, although the autoregressive generation method is of high quality when there is only one sentence, once generating multiple sentences, even if the diversity beam search method is used, it is difficult to improve its diversity, and there may be a large number of repeated prefixes. In contrast, on the premise of ensuring the quality of generation, the diffusion generation method has rich diversity, which

Table 3. **SELF-BLEU** score of ProphetNet and GENIE generated results.

Model	Generate Method	XSUM	CNN/DAILYMAIL
ProphetNet	Beam Search	92.2	96.1
	Diverse Beam Search	77.2	87.7
GENIE	Diffusion	**29.1**	**38.5**

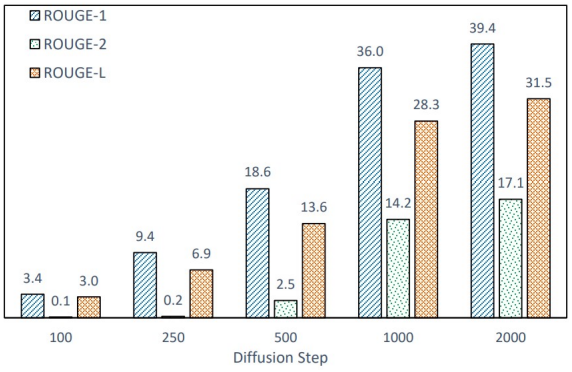

Figure 2. Effect of different diffusion steps on text generation quality, on XSUM. The result is the optimal value of 5 Gaussian samples.

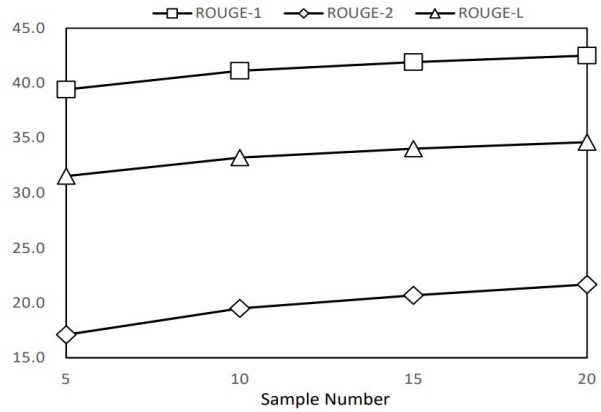

Figure 3. Effect of sample number, on XSUM.

benefits from its non-autoregressive generation style and multi-step diffusion process. Experimental index and case analysis show that GENIE can greatly improve the diversity of generated text, which is crucial for application scenarios requiring text diversity.

4.6. Impact of Pre-training Steps

Due to our pre-training method and the nature of the diffusion model itself, our pre-training is bound to take a long time to fully converge, but it also gives the diffusion language model unlimited potential. Here we discuss the performance improvement from pre-training compared with the GENIE without pre-training under different pre-training steps. Specifically, we fine-tune the corresponding checkpoints on the XSUM dataset at 50w step intervals. During the evaluation, we randomly sample 5 Gaussian noise and use the highest score as the final generated result. As result shown in Table 5, compared with GENIE without pre-training, only 50w steps of pre-training can significantly improve the quality of generation. Moreover, we can easily observe from the results that with the increase of pre-training steps, pre-training is still steadily improving the performance of the GENIE on downstream tasks.

4.7. Impact of Total Time Step

Different diffusion steps have great influence on the quality of generation. We explore the performance of the GENIE under different inverse diffusion steps on the XSUM dataset. Assuming that the total number of diffusion steps $T = 2000$, we respectively set the interval step of inverse diffusion as 1, 2, 4, 8, 20, and the corresponding inverse diffusion steps are 2000, 1000, 500, 250, 100. In this experiment, we sample 5 Gaussian noise and choose the best denoise result. As results[1] shown in Figure 2, we can clearly observe that when the number of inverse diffusion steps is small, the quality of generation with GENIE has a serious loss. As the number of inverse diffusion step increases to 1000, the generated quality of GENIE tends to be stable.

4.8. Impact of Sample Number

Compared with the autoregressive model, GENIE based on the diffusion model can generate more diversified texts in the reference phase, even under the control of the source sequence. Different Gaussian noise sampled during denoise can often lead to completely different generation results. This method is more flexible, but it is not conducive to the evaluation of the only standard answer. However, as the number of Gaussian noise sampled increases, the generated text has a greater probability of approaching the only standard answer, and the corresponding evaluation score is higher. To this end, we test the performance of the model on the test set under different sample numbers on the XSUM

[1]The experiment of Impact of Total Time Step is completed at the 100w step pre-training checkpoint. We will update the results at the end of pre-training.

Table 4. Summaries example of ProphetNet and GENIE generated results.

source sequence I (abbreviated)	The name is expressed with affection by some, with hostility by others, but it calls up history for everyone. The story of his life is very much the story of our times: revolutionary movements, the Cold War, East v West, North v South, communism v capitalism - except that most of the world has passed him by. Fidel Castro has remained the same, a symbol of revolution, a communist who has survived the fall of communism. Fidel's views continued to be made public though in the form of editorials and occasional TV appearances. Fidel maintained his rule with an iron grip, sending opponents to prison for years. Throughout his leadership, he railed against the US, its economic and trade embargo and against the evils of free marketsFidel has been praised for standing up for the oppressed of Latin America, for opposing the Yankee imperialist, for making Cuba into a more equal society than many, for developing Cuba's health service and sending doctors abroad to help others.
GENIE summaries I (diffusion)	1. the story of fidel castro, the head of one's, has in the history of the communist, is an apparent close to cuba and many around the world. 2. for the story of the late revolutionary life, fidel castro, one of the most communist house public, has remained the story of who to become the president. 3. the story of the life is life of fidel castro when it comes to one of the most and most armed people in the world.
ProphetNet summaries I (diversity beam search)	1. the cuban leader fidel castro has died at the age of 82 . . . his name is raul. 2. the cuban leader fidel castro has died at the age of 94, his family has said. 3. the cuban leader fidel castro has died at the age of 85.
source sequence II (abbreviated)	Those who participated in the Aberdeen Children of the 1950s project, which saw all primary pupils aged seven to 12 surveyed by the Medical Research Council in 1962, have been contacted. They have been asked to take part in the Scottish Family Health Study. It aims to investigate why diseases such as cancer can run in families.Those recruited will have their health tracked, with the intention of creating a Scottish "bio-bank" containing genetic, medical and family history and lifestyle information. The data gathered would help future research into the prevention, treatment and diagnosis of illnesses.
GENIE summaries II (diffusion)	1. health information have been recruited by university school primary pupils to help improve their lives. 2. a health project is to be recruited by learning for university researchers in scotland. 3. scientists in aberdeen are to meet experts in scotland to get more health data for their children.
ProphetNet summaries II (diversity beam search)	1. a group of children in aberdeen have been asked to help lead a new study into diseases such as diabetes and diabetes. 2. a group of children who were among the first to be surveyed in scotland have been asked to share their health information. 3. a group of children in aberdeen have been asked to help lead a new study into diseases such as diabetes.
source sequence III (abbreviated)	Elin Jones is expected to lay out plans where some areas of Welsh forest could be transferred to the private sector or to not for profit organisations.But she has already ruled out the widespread sale of Welsh woodlands.Forestry Commission Wales said it would explore the feasibility of transfer to the private sector case by case.The minister told BBC Radio Wales she plans to "compensate" the public by buying new land for new planting or management if any forest was sold off on a case-by-case basis."I don't want any stagnancy in the forest estate. I want it to work for public benefit whether that's economic or environmental or access benefit," she said."It's my view there should be no reduction in the publicly owned estate and I have asked the Forestry Commission to look at how it can make that estate work harder, provide a better return for the public."Whether that's in terms of public access, in terms of environmental benefit in the production of renewable energy or biomass potential or also in terms of the economic return from that forestry estate."
GENIE summaries III (diffusion)	1. the forestry minister is preparing to fill out its plan for some members of the public on wales' forests to be reduced. 2. the forestry minister is picking forward plans to tackle some of the companies in wales to develop a boost in the management of forests. 3. the forestry minister hopes to face plans continue on the future of wales' forest estate to be held to a growing and better access to the private sector.
ProphetNet summaries III (diversity beam search)	1. the environment minister has said there should be no reduction in the amount of forest in wales' forests. 2.the environment minister has said there should be no reduction in the amount of forest land in wales. 3.the forest minister has said she wants to see no reduction in the amount of forest land in wales' forests.

Table 5. Effect of pre-training step, on XSUM. The result is the optimal value of 5 Gaussian samples.

Model	ROUGE-1	ROUGE-2	ROUGE-L
w/o pre-train	37.3	15.3	29.4
GENIE(50w)	39.0	16.7	30.9
GENIE(100w)	39.4	17.1	31.5
GENIE(150w)	40.0	17.9	32.1
GENIE(200w)	40.4	18.2	32.5

dataset. As results [2] shown in Figure 3, we evaluated the results of 5, 10, 15 and 20 samples. We can observe that with the increase of sample numbers, the matching level between the candidate generated text and the target text is higher. The improvement of the matching level is more obvious in the early period during the number of samples increases.

5. Related Work

5.1. Large Scale Pre-training Language Models

Recently, a major breakthrough has been made in the model of pre-training on large scale corpus. As unidirectional language models, GPT (Radford et al., 2018), GPT2 (Radford et al., 2019) modeling the text based on left-to-right, and predict the next token according to the token appearing on the left. At the same time, bidirectional language models, which uses bidirectional encoder to model text, can obtain better context sensitive representation, such as BERT (Devlin et al., 2019) and RoBERT (Liu et al., 2019). RoBERT optimizes pre-training tasks compared to BERT, both of which significantly improve the ability of natural language understanding. In order to improve the performance of the large scale pre-training model in natural language generation, some works has designed pre-training tasks based on the standard framework of sequence-to-sequence. MASS (Song et al., 2019) lets the model predict the short masked token span step by step, while ProphetNet (Qi et al., 2020) predict more words in each step to ease local over fitting.

5.2. Non-autoregressive and Autoregressive Generation

Autoregressive generation is the mainstream in the current text generation field, due to its higher generation quality compared with non-autoregressive models. However, non-autoregressive generation has a wide application space due to its fast decoding ability compared with autoregressive generation. Especially in machine translation task, some works (Qian et al., 2021; Huang et al., 2022) uses non-autoregressive method to achieve better and faster translation. There are also some works (Gu et al., 2019a; Zhang et al., 2020; He, 2021) adopt non-autoregressive method to text revision. Faster decoding seems to be the synonym of non-autoregressive, but the diversity of text brought by non-autoregressive cannot be ignored, especially the current autoregressive text generation is faced with text degradation caused by greedy decoding (Holtzman et al., 2020). Even if there are methods similar to beam search (Xiao et al., 2022) and diverse beam search (Vijayakumar et al., 2016), autoregressive text generation cannot bring about essential changes to the diversity of generated text. We believe that the diffusion model, which generates texts by non-autoregressive method, will make a major breakthrough in the diversity of text generation.

[2]The experiment of Impact of Sample Number is completed at the 100w step pre-training checkpoint. We will update the results at the end of pre-training.

5.3. Diffusion Models for Text

In recent years, diffusion model has achieved great success in the domains of image generation (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022). Because of its amazing generation quality, some works apply diffusion model in text generation domains. Diffsuion-LM (Li et al., 2022) maps discrete tokens into continuous latent variable, achieving more complex controllable text generation through continuous diffusion. In the field of text revision where non-autoregressive method is widely used, DiffusER (Reid et al., 2022) also uses the diffusion model to implement the edit based generative processes. DiffuSeq (Gong et al., 2022) achieves conditional text generation with a new method which controlled information is also involved in the diffusion process. Different from the above work, we build a novel language model based on the diffusion model for the first time, using the standard enocoder-decoder framework. For our best knowledge, we are the first to adopt large scale pre-training on the language model based on the diffusion model.

6. Conclusion

In this paper, we propose a novel diffusion language model GENIE, which is pre-training on large scale corpus. In this new sequence-to-sequence framework, the bidirectional encoder is used to encode the source sequence, and the denoise architecture is used to predict the noise of the corresponding time step, so that we can gradually denoise to generate diverse text in a non auto-regressive manner. At the same time, this framework design also allows us to adopt large-scale pre-training to the GENIE based on diffusion model. In the pre-training process, we propose a novel pre-training method named *continuous paragraph denoise*, which denoise continuous paragraphs as target. Extensive experiments on various widely used NLG tasks show that GENIE can generate high-quality and diversified text, and proves the effectiveness of the large scale pre-training on GENIE.

References

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, pp. 13042–13054, 2019a.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32, 2019b.

Ghazvininejad, M., Levy, O., Liu, Y., and Zettlemoyer, L. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Networks Learn. Syst.*, 28(10):2222–2232, 2017.

Gu, J., Lu, Z., Li, H., and Li, V. O. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

Gu, J., Wang, C., and Zhao, J. Levenshtein transformer. In *NeurIPS*, pp. 11179–11189, 2019a.

Gu, J., Wang, C., and Zhao, J. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, pp. 11181–11191, 2019b.

He, X. Parallel refinements for lexically constrained text generation with BART. In *EMNLP (1)*, pp. 8653–8666. Association for Computational Linguistics, 2021.

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching machines to read and comprehend. In *NIPS*, pp. 1693–1701, 2015.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022b.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *ICLR*. OpenReview.net, 2020.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

Huang, F., Zhou, H., Liu, Y., Li, H., and Huang, M. Directed acyclic transformer for non-autoregressive machine translation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9410–9428. PMLR, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Lee, J., Mansimov, E., and Cho, K. Deterministic non-autoregressive neural sequence modeling by iterative refinement. *arXiv preprint arXiv:1802.06901*, 2018.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Li, X. L., Thickstun, J., Gulrajani, I., Liang, P., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022.

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., and Ren, X. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*, 2019.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, pp. 1797–1807. Association for Computational Linguistics, 2018.

Pawade, D., Sakhapara, A., Jain, M., Jain, N., and Gada, K. Story scrambler-automatic text generation using word level rnn-lstm. *International Journal of Information Technology and Computer Science (IJITCS)*, 10(6):44–53, 2018.

Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*, 2020.

Qi, W., Gong, Y., Jiao, J., Yan, Y., Chen, W., Liu, D., Tang, K., Li, H., Chen, J., Zhang, R., et al. Bang: Bridging autoregressive and non-autoregressive generation with large scale pretraining. In *International Conference on Machine Learning*, pp. 8630–8639. PMLR, 2021.

Qian, L., Zhou, H., Bao, Y., Wang, M., Qiu, L., Zhang, W., Yu, Y., and Li, L. Glancing transformer for non-autoregressive neural machine translation. In *ACL/IJCNLP (1)*, pp. 1993–2003. Association for Computational Linguistics, 2021.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pp. 2383–2392. The Association for Computational Linguistics, 2016.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022.

Reid, M., Hellendoorn, V. J., and Neubig, G. Diffuser: Discrete diffusion via edit-based reconstruction. *arXiv preprint arXiv:2210.16886*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685. IEEE, 2022.

Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization. In *EMNLP*, pp. 379–389. The Association for Computational Linguistics, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, abs/2205.11487, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. MASS: masked sequence to sequence pre-training for language generation. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5926–5936. PMLR, 2019.

Song, L., Zhang, Y., Wang, Z., and Gildea, D. A graph-to-sequence model for amr-to-text generation. *arXiv preprint arXiv:1805.02473*, 2018.

Stern, M., Chan, W., Kiros, J., and Uszkoreit, J. Insertion transformer: Flexible sequence generation via insertion operations. *arXiv preprint arXiv:1902.03249*, 2019.

Strudel, R., Tallec, C., Altché, F., Du, Y., Ganin, Y., Mensch, A., Grathwohl, W., Savinov, N., Dieleman, S., Sifre, L., et al. Self-conditioned embedding diffusion for text generation. *arXiv preprint arXiv:2211.04236*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017a.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D. J., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424, 2016.

Xiao, Y., Wu, L., Guo, J., Li, J., Zhang, M., Qin, T., and Liu, T. A survey on non-autoregressive generation for neural machine translation and beyond. *CoRR*, abs/2204.09269, 2022.

Zhang, Y., Wang, G., Li, C., Gan, Z., Brockett, C., and Dolan, B. POINTER: constrained progressive text generation via insertion-based generative pre-training. In *EMNLP (1)*, pp. 8649–8670. Association for Computational Linguistics, 2020.