Diffusion-LM with Bigger Models and Smaller Steps

Zesen Zhao, Yaoxin Selina Li, Keqian Wang, Kevin Chen
Department of Computer Science
University of Michigan
Ann Arbor, MI 48104
{hymanzzs, selinali, keqianw, kevinjj}@umich.edu

Abstract

This paper explores the enhancement of the existing Diffusion-LM, a non-autoregressive language model based on continuous diffusion. Diffusion-LM iteratively denoises a sequence of Gaussian noise vectors into word vectors for controlled text generation. We first trained the original and scaled-up BERT architecture on the E2E dataset for sentence infilling, benchmarked by BERTScore. We compare the enhanced model's effectiveness against original implementations on original datasets. We further explored diffusion-LM's performance in different diffusion steps which was held constant in the original paper. Our results show high performance persists with up to 1/20 of the original diffusion steps.

1 Introduction

Large autoregressive language models (LMs) have demonstrated the capability to generate high-quality text, underpinning their potential for diverse applications. However, deploying these models in real-world scenarios requires precise control over the generated content to meet specific criteria, such as topic consistency or syntactic correctness. Traditional methods typically involve fine-tuning the LM with supervised data that pairs control parameters with desired text outputs. Despite its effectiveness, this approach is costly and lacks the flexibility to handle multiple control tasks simultaneously, such as generating text that is both positive in sentiment and non-toxic.

To address these limitations, the concept of a lightweight, modular approach has been proposed, where the language model remains unchanged while an external classifier steers the text generation process. Yet, steering a frozen autoregressive LM effectively, particularly for complex or multiple controls, has proven challenging with existing methods achieving limited success primarily in simple attribute-level controls like sentiment or topic.

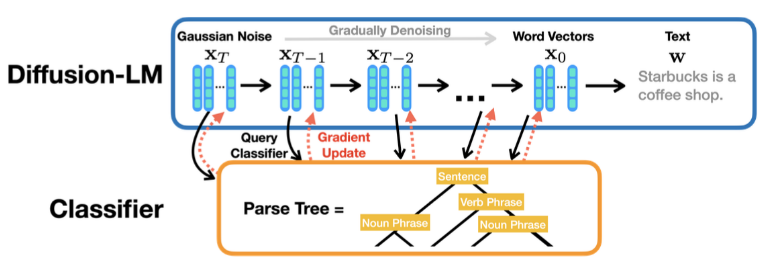

Figure 1: Diffusion-LM Architecture Li et al. [2022]

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

In this context, we build upon the foundational work of Diffusion-LM [Li et al. [2022]], a novel approach that extends the principles of continuous diffusions—previously successful in vision and audio—to text generation.

Our work builds on the existing Diffusion-LM by scaling its architecture to evaluate the impact on performance and control in text generation. We implement an enhanced version of the model, which is larger and potentially more capable of complex controllability and higher fluency. This paper details the architectural modifications, the extended training regime, and a comprehensive performance comparison with the original model. Our evaluations focus on infilling tasks to demonstrate the scaled model's improved effectiveness in precise and flexible text generation. We further enhance its efficiency by experimenting with reducing diffusion steps. Our contributions are 1. Explore wither increase model complexity will lead to a performance increase. 2. Increase model computing efficiency with fewer diffusion steps with tolerable performance trade-off.

2 Background

2.1 Transformer Models

Transformer language models, introduced by Vaswani et al. in their seminal paper "Attention is All You Need" [Vaswani et al. [2023]], represent a significant advancement in natural language processing. These models rely on a mechanism called "self-attention" to process input sequences, enabling the model to weigh the importance of different words within the input, regardless of their position. However, a notable limitation of the standard transformer architecture is the size of its attention window, which determines how many words in a sequence the model can consider at one time. As sequences become longer, the computational and memory requirements grow quadratically with the length of the attention window. This can lead to inefficiencies and practical constraints on the model's scalability and responsiveness, particularly in tasks involving very long documents or sequences.

2.2 Diffusion Models

Diffusion models have been highly effective in generating high-quality samples in domains involving continuous data such as images and audio [Ho et al. [2020], Kong et al. [2020], Mittal et al. [2021], Saharia et al. [2020], Sohl-Dickstein et al. [2015]]. Previous research has explored diffusion models for text using discrete state spaces, where tokens are subjected to a corruption process that may convert them into either an absorbing or a random token [Austin et al. [2021],Hoogeboom et al. [2021],Hoogeboom et al. [2022]]. This paper introduces a novel approach by examining continuous diffusion models for text, a first in this area to our knowledge. Unlike their discrete counterparts, our continuous diffusion LMs create continuous latent representations, facilitating the use of gradient-based techniques for controlled generation.

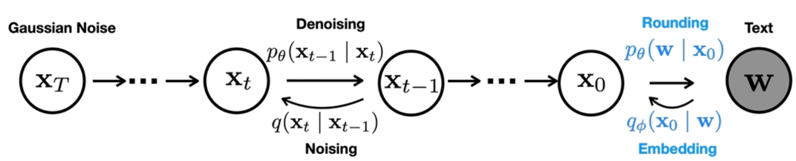

Figure 2: A graphical model representing the forward and reverse diffusion processes. In addition to the original diffusion models, a Markov transition is added between x0 and w Li et al. [2022]

2.3 Diffusion-LM

Diffusion-LM aims to overcome these limitations by introducing a novel non-autoregressive language model that leverages continuous diffusions[Chen et al. [2023], He et al. [2022]]. It starts with a sequence of Gaussian noise vectors and incrementally denoises them into vectors corresponding to words. This gradual transformation facilitates a hierarchy of continuous latent representations, allowing for gradient-based methods to be employed effectively for complex text control tasks.

These controls are not only fine-grained, targeting specific semantic content, but also extend to complex structures like parse trees. It adapts to the discrete nature of text by incorporating an embedding step to translate discrete text into a continuous space, followed by a rounding step to convert continuous outputs back into discrete text. The model's training involves learning these embeddings and improving rounding techniques, ensuring that the generated text not only satisfies specific structural and semantic controls but also maintains fluency. It solves the problem of traditional transformer models failing to attend to whole input contents.

3 Methods

3.1 Infilling task

We decided to replicate and extend the "infill" task in the Diffusion-LM paper. Given our limited computing resources, we wouldn't want to pick a task that requires classifier training which was what most controlled-generation tasks did in the paper. Infilling is classifier-free, reflects the model's true ability, and is highly useful in daily scenarios. Since most LLM generate tokens from left to right, it's often hard to get them to complete a sentence that has both left and right context available, and even harder to add precise control. An LLM that can control the number and nature of words filled into blanks of a sentence can be incredibly useful for creative writing, advertisement, or language learning among other tasks.

3.2 Dataset

As in the paper, we used E2E NLG, a restaurant review dataset, for our training. This dataset is made of short sentences containing name, location, rating, food type, family friendliness and other features of a restaurant. It is relatively simple and domain-specific.

3.3 Model extension

Diffusion-LM relies on a language model to get word embedding and tokenizer, and the model of choice was Bert-base in Li et al.. Given their results, We wanted to test if performance would improve when replace BERT-base with a better model, BERT-large. Standard BERT-base uses 12 attention heads and has 110 million parameters, while BERT-large uses 16 attention heads, double the encoder layers, with 340 million parameters. BERT base has a hidden dimension d=768 whereas BERT large has d=1024. However, in the paper, the authors reduced the hidden dimension to 16 for the E2E dataset for better performance. Devlin et al. [2019]

We retrained two models following the paper's specifications. Our first model of choice is BERT-base. In the paper, it was trained on the E2E NLG dataset for 200K steps, and we adopted this setting. We also tried upgrading to BERT-large with the same parameters to see if it would improve infilling performance.

3.4 Diffusion Steps

We found that the authors had adapted this codebase from image diffusion projects, and the default diffusion step is stuck at 2000. This takes a long time to run for a single diffusion input. We identified a gap for improvement here, especially since text embedding is much smaller than regular image embedding. We explored whether we can reduce diffusion steps and still maintain the same performance.

4 Experiments

4.1 Training Base Model

We first trained the U-Net Transformers for estimating mean and variance in the diffusion process. Both BERT-base and BERT-large are trained for 200K steps on the E2E dataset through a general diffusion process, using a sqrt noise schedule as specified in the paper. These weights are used later for running the infilling task.

	Automatic Eval				Human Eval
	BLEU-4 ↑	ROUGE-L ↑	CIDEr ↑	BERTScore ↑	
Left-only	0.9	16.3	3.5	38.5	n/a
DELOREAN	1.6	19.1	7.9	41.7	n/a
COLD	1.8	19.5	10.7	42.7	n/a
Diffusion	**7.1**	**28.3**	**30.7**	**89.0**	$\mathbf{0.37}^{+0.03}_{-0.02}$
AR	6.7	27.0	26.9	**89.0**	$\mathbf{0.39}^{+0.02}_{-0.03}$

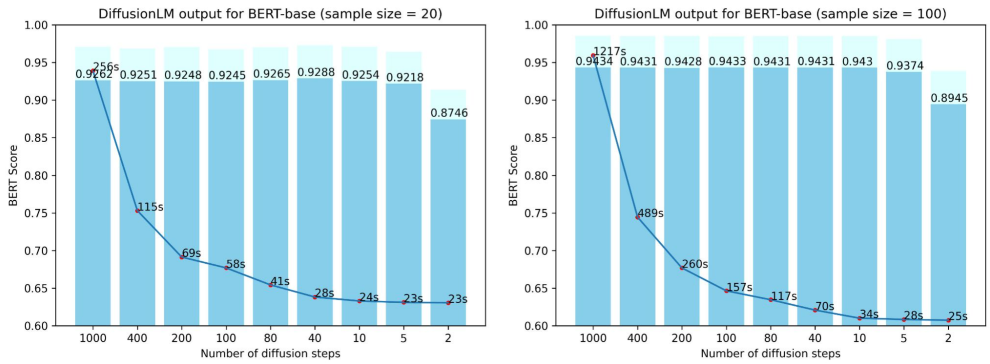

Figure 3: Top: Reference of average sentence infilling performance in the paper. Bottom: Diffusion output quality of our trained BERT-base. Blue indicates the average BERTScore and light blue is the standard deviation. The line shows the diffusion runtime for the batch in seconds. The average quality is above 0.9 as measured by BERTScore across diffusion steps.

There are no command line options in the Diffusion-LM codebase to switch between BERT-base and BERT-large. Thus, we manually modify its codebase to switch between two models. We re-trained two models, base and large, using the E2E dataset, both for 200K steps, 2000 diffusion steps, 102 random seed, vocab size 821, and block padding mode. We train the model on Greatlakes with 16 cores of CPU, a single A100 GPU, and 128 GB RAM. The training takes about 6 hrs for small model and about 10 hrs for large models. We save checkpoints every 100k steps and evaluate every 5k steps. (Figure 7)

However, due to code base redundancy, although we have the correct training argument printed and training time per epoch indeed increased, the results still aren't comparable with Bert base models. We will talk about this in detail in conclusions and limitations.

4.2 Diffusion Steps Experiment

We decided to use BERTScore for measuring output quality. BERTSCore is a widely used metric of sentence quality, ranging from 0-1 with higher values representing better results. It scores the fluency of a candidate sentence compared to a reference, as judged by a BERT model. We chose the F1 score from the BERTScore output because it's relatively unbiased.

In our experiment, we randomly sampled 20 and then 100 sentences from the E2E test set (4693 sentences in total) and padded 3-6 random short pieces in each sentence. At each location, we replaced 1-5 words with "PAD", which is the format Diffusion-LM expects for infilling prompts. Under this padding setup, we achieve a roughly Gaussian distribution for the percentage of words that are masked in each sample.

We tested on diffusion steps 2, 5, 10, 40, 80, 100, 200, 400, and 1000, with the smaller and large sample, respectively. We asked BERT-base and BERT-large models to complete the sentences. In the original paper, the authors used 200 diffusion steps during decoding for E2E, and we wanted to investigate if lower or higher diffusion steps will affect generation speed and quality.

4.3 Diffusion Results

We compared BERTScore of the infilling outputs under different number of diffusion steps. Overall, the BERT models performed the same when diffusion steps is reduced from 1000 to 10. Once diffusion step is fewer than 10, the performance starts to drop. However, time overhead decreases almost exponentially, particularly to 40 steps. There is not much time difference from 40 to 2 diffusion

Prompt	Blue Spice PAD PAD PAD PAD located in the riverside . It PAD PAD friendly is is PAD PAD PAD PAD PAD . The PAD PAD is child friendly and offers fast food priced on average at more PAD PAD PAD . It is PAD PAD PAD PAD area
Ground Truth	Blue Spice is a Chinese pub located in the riverside . It is family friendly is is near the Rainbow Vegetarian Café . The Phoenix restaurant is child friendly and offers fast food priced on average at more than £ 30 . It is located in the riverside area
BERT-base	Blue Spice is a beautiful restaurant located in the riverside . It 's kids friendly is is and rice range is UNK . Blue Spice coffee shop and near located in the riverside . It is family friendly is is priced up average of seafood . Blue Spice is UNK restaurant and located in the riverside . It is family friendly is is a bit less well average . The Wrestlers shop is child friendly and offers fast food priced on average at more than 30 euros . It is in the city of area , The coffee shop is child friendly and offers fast food priced on average at more than 30 pounds . It is in the City Centre area The coffee shop is child friendly and offers fast food priced on average at more than £ 30 . It is located in the city area , ne
BERT-large	Blue Spice . . . . located in the riverside . It . . friendly is is . . . . . . Blue Spice . . . . located in the riverside . It . . friendly is is . . . . . . Blue Spice . . . . located in the riverside . It . . friendly is is . . . . . . The . . is child friendly and offers fast food priced on average at more . . . . It is . . . . area , near Raja Indian Cuisine The . . is child friendly and offers fast food priced on average at more . . . . It is . . . . area , near Raja Indian Cuisine The . . is child friendly and offers fast food priced on average at more . . . . It is . . . . area , near Raja Indian Cuisine

Figure 4: Example infilled sentences with diffusion_steps=10.

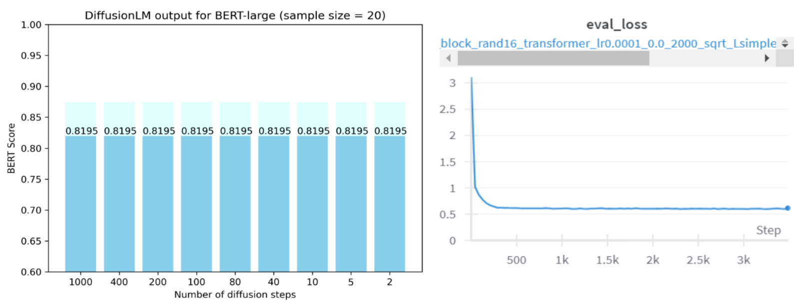

Figure 5: Training loss and test performance for BERT-large. Although we saw a decrease in loss, the model did not converge. All padding tokens were into sentence endings, which explains the low BERTScore.

steps. As our sample size increased from 20 to 100, time overhead of large diffusion steps became increasingly worse. Diffusing for 1000 steps now requires 4.75x inference time while 10 steps only uses 1.4x time than before. Although we did not have time to test on the full dataset, we observed stable performance and that convinces us that our sample size is sufficient. 3

The BERTSCores for infilled sentences are as high as 0.92-0.94 on average for almost all diffusion steps. This is higher than the paper's results which is 0.89. However, our task setup might vary as the paper did not provide their exact test procedure. 3

We inspected the outputs visually, and found that BERT-base generated consistent, reasonable sentences even with low diffusion steps. Specifically, we included samples from diffusion steps=10, which is indistinguishable from 200 or higher diffusion steps. 4 However, BERT-large appeared to suffer from some error during training. It would correctly identify the infilling positions but diffused them all into the same token, the period "." 5. We saw that the loss decreased but not as much as in BERT-base training. 5

5 Conclusion and Limitations

Our results show that Diffusion-LM can achieve precise control of sentence infilling. The outputs are fluent and sensible, showing stability across multiple test runs with adequate variations. Most significantly, we observed the same level of performance with only 5 diffusion steps compared to the 200 used in the original setting.

While navigating the training and diffusion code base, we found a large amount of redundancy, and we cut the amount of code by at least 25 percent from the official GitHub. Nevertheless, we still faced many challenges. With the way parameters are set up, they often override each other between the config file, CLI arguments, and HuggingFace defaults. We spent significant effort ensuring our parameters were consistent and as expected.

Without a provided checkpoint, we had to train the models on E2E from scratch, but it should be noted that they are not trained specifically to perform the infilling task. Our BERT-base model was able to complete the task with high proficiency which replicates the paper's results.

However, we weren't able to get BERT-large model to infill fluently, which was our biggest limitation in this project. There are many possible reasons to why we obtained negative results from BERT-large. First, it could be an issue in hyperparameter choices. In the paper, the author reduced the hidden dimension from 768 to 16 in the BERT-base for the E2E dataset. The paper's first author also mentioned in a GitHub issue that a larger dimension is not always better because the dataset has limited complexity. Given more time and resources, we would decrease the learning rate and embedding dimension, and adjust things such as the number of attention heads. Although our training behavior seems correct, we suspect that there are still be incompatibilities within the codebase that affected our training setup.

Our experiment shows that fewer diffusion steps doesn't affect performance. By reducing diffusion step to 5, We successfully cut the computing time by 90% without affecting performance. Considering the discrete nature of word tokens, we infer that there is still a large space for future research to further reduce the computing time. This could be a key insight in adapting image diffusion models to language with real-time capacities, which will be an interesting direction for future research.

6 Contributions

Zesen Zhao proposed the idea, completed setup and training, and ran pilot experiments. Yaoxin Li worked on running experiments for model comparison and diffusion steps, creating figures, and writing result analysis. Keqian Wang worked on creating the poster as well as writing and editing the report. Kevin Chen explored datasets and contributed to the report.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=h7-XixPCAL`.

Jiaao Chen, Aston Zhang, Mu Li, Alex Smola, and Diyi Yang. A cheaper and better diffusion language model with soft-masked noise, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models, 2022.

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

Emiel Hoogeboom, Didrik Nielsen, Priyank Jain, Patrick Förre, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *arXiv preprint arXiv:2102.05379*, 2021.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=Lm8T39vLDTE`.

Zhifeng Kong, Wei Ping, Jiayi Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM improves controllable text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=3s9IrEsjLyk`.

Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, March 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265, Lille, France, 2015. PMLR. URL `https://proceedings.mlr.press/v37/sohl-dickstein15.html`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstraction and Introduction part, the main claims made accurately reflected the paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Justified the limitations in the Conclusion and Limitations part

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work mainly focused on the extension of this project and the pilot study of their results. Thus, we want to leave the theory proof to future research.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explain our experiment setup in great detail. We include our necessary hyper parameter and people who want to understand our work in depth can also reference to original Diffusion-LM.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will have code ready with our submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We release all hypo parameters that are different with original Diffusion-LM.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We gave clear graphs for reporting our results. Due to limit in computing resources, we can't sample many results and perform detailed statical analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report our training setup and time.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All requirements has been checked and followed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We mainly focuses on pilot study and ablation study. With the best of our knowledge, there is no broader impacts found.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We mainly focuses on pilot study and ablation study. With the best of our knowledge, there is no adversarial behavior related with our paper considered.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use diffusion-LM as our foundational code base. We also use open sourced datasets and evaluation metrics.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: We mainly focuses on pilot study and ablation study. With the best of our knowledge, there is no new assets considered.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No IRB included in this project.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.