arXiv:2304.04746v1 [cs.CL] 10 Apr 2023

A CHEAPER AND BETTER DIFFUSION LANGUAGE MODEL WITH SOFT-MASKED NOISE

Jiaao Chen[†‡*], **Aston Zhang**[†], **Mu Li**[†], **Alex Smola**[†], **Diyi Yang**[◇]
[†]Amazon Web Services, [‡]Georgia Institute of Technology, [◇]Stanford University

ABSTRACT

Diffusion models that are based on iterative denoising have been recently proposed and leveraged in various generation tasks like image generation. Whereas, as a way inherently built for continuous data, existing diffusion models still have some limitations in modeling discrete data, e.g., languages. For example, the generally used Gaussian noise can not handle the discrete corruption well, and the objectives in continuous spaces fail to be stable for textual data in the diffusion process especially when the dimension is high. To alleviate these issues, we introduce a novel diffusion model for language modeling, Masked-Diffuse LM, with lower training cost and better performances, inspired by linguistic features in languages. Specifically, we design a linguistic-informed forward process which adds corruptions to the text through strategically soft-masking to better noise the textual data. Also, we directly predict the categorical distribution with cross-entropy loss function in every diffusion step to connect the continuous space and discrete space in a more efficient and straightforward way. Through experiments on 5 controlled generation tasks, we demonstrate that our Masked-Diffuse LM can achieve better generation quality than the state-of-the-art diffusion models with better efficiency. Code is available at `https://github.com/amazon-science/masked-diffusion-lm`

1 Introduction

We present a novel diffusion method for modeling languages, Masked-Diffuse LM (language model), which uses strategic soft-masking informed by linguistic features to corrupt both the discrete and continuous space, and then iteratively denoise them back by predicting the categorical distribution. Specifically, a strategic soft-masking process is designed that gradually adds perturbation to the input text in an order from harder or more informative words to simpler or less informative words through soft-masking. As a result, the models are encouraged to recover and generate the text following an *easy-first-generation* nature Dieleman et al. [2022] to improve the generation structure and quality with more flexibility. Also, during the diffusion process, we directly predict the discrete token with cross-entropy loss that maps the continuous space to discrete textual space to stabilize the intermediate diffusion steps. Through our proposed Masked-Diffuse LM, the application-specific performance metrics as well as training efficiency are significantly improved over current diffusion language models based on experiments.

Our work is inspired by recent advances in diffusion models Sohl-Dickstein et al. [2015a], Ho et al. [2020], Song et al. [2021], Yang et al. [2022], Ramesh et al. [2022], Rombach et al. [2022] that are introduced as a new generative modeling approach based on iterative denoising and have achieved high-quality generations for visual and audio modalities Ramesh et al. [2022], Rombach et al. [2022], Saharia et al. [2022], Nichol and Dhariwal [2021], Kong et al. [2020].

Although these approaches have received growing attention and achieved impressive success, applying diffusion models to textual domain is still challenging and under-explored due to the discrete nature of the text (e.g., one-hot vectors) compared to continuous data like images (e.g., RGB values) Li et al. [2022]. A few prior works Li et al. [2022], Gong et al. [2022], He et al. [2022], Austin et al. [2021], Hoogeboom et al. [2021a] that explore using diffusion models on textual data can be divided into two lines. The first is to extend diffusion models to discrete state spaces Austin et al. [2021], Hoogeboom et al. [2021a,b]. The second is to perform the diffusion process and its reverse process in

*Correspondence to Jiaao Chen <jiaaochen@gatech.edu> and Aston Zhang <az@astonzhang.com>.

the continuous domain and bridge the continuous and the discrete domain through embedding and rounding Li et al. [2022], He et al. [2022], for example, Diffusion-LM Li et al. [2022]. Despite the improvements, most previous works fail to leverage the linguistic features (e.g., words in sentences are with different importance) to noise the input textual data and recover it back in a more suitable way. Besides, they usually neglect or fail to adapt large pre-trained language models (PLMs) Devlin et al. [2019], Liu et al. [2019], Yang et al. [2019], Joshi et al. [2019], Sun et al. [2019], Clark et al. [2019], Lewis et al. [2020], Bao et al. [2020], He et al. [2020], Raffel et al. [2020], which is an unmissable treasure in the NLP community: their adopted k-nearest-neighbor rounding technique that maps continuous space to discrete space cannot handle high-dimensional data in a stable and efficient way Li et al. [2022]. As a result, a corruption process tailored for languages and the objective that allows efficient and straightforward discrete and continuous space transformation is in great need. Our proposed Masked-Diffuse LM realizes this extension.

To demonstrate the effectiveness of our introduced Masked-Diffuse LM, we perform experiments on E2E dataset Novikova et al. [2017] and 5 controllable generation tasks Li et al. [2022] including Semantic Content, Parts-of-speech, Syntax Tree, Syntax Spans, and Length. We observe that our Masked-Diffuse LM can (i) achieve the state-of-the-art performances compared to recent baseline models, and (ii) allow more efficient training and inference compared to the previous Diffusion-LM.

To summarize, our contributions are:

- We introduce a strategic masking noise strategy guided by linguistic features to corrupt the textual data in diffusion models for modeling languages.
- We use linear layers and cross-entropy objectives to bridge the continuous and discrete spaces in the diffusion process for efficiency and stability.
- We conduct experiments on different controllable generation tasks to demonstrate the effectiveness of our proposed methods compared to previous diffusion language models.

2 Related Work

Our work is inspired by the recent research about diffusion models, and related to or based on the work about non-autoregressive text generation and controllable generation through a plug-and-play manner.

2.1 Diffusion Models for Language

There has been growing attention in deep generative diffusion models, which is a latent variable generative method based on iterative denoising Sohl-Dickstein et al. [2015a], Ho et al. [2020], Song et al. [2021]. Through a forward and diffusion process, diffusion models have shown state-of-the-art sample quality on generating in the continuous domain such as producing images and audio Ramesh et al. [2022], Rombach et al. [2022], Kong et al. [2020], Savinov et al. [2022]. Despite their huge success, it is still challenging and under-explored to adapt diffusion models to discrete domains like languages. A few recent works have modified the diffusion models for textual data. For example, discrete forward processes, such as categorical transition kernels Hoogeboom et al. [2021b], uniform transition kernels, and absorbing kernels Hoogeboom et al. [2021a], have been introduced. However, replacing continuous diffusion with a discrete corruption process affords some flexibility Dieleman et al. [2022]. Other works have also made efforts to model text in the continuous embedding space and applied Gaussian noise uniformly to every token Li et al. [2022], He et al. [2022], which is closer to the settings in previous works of diffusion models. However, they neglect the inherent linguistic features in the text (e.g., *different words are playing different roles in sentences*) so the generated text often lacks coherence He et al. [2022]. Besides, the k-nearest-neighbor rounding technique Li et al. [2022] holds up the decoding and convergence speed especially when the vocabulary is large or the hidden dimension is high, thus limiting the potential of combining large pre-trained language models Devlin et al. [2019], Liu et al. [2019], Yang et al. [2019], Joshi et al. [2019], Sun et al. [2019], Clark et al. [2019], Lewis et al. [2020], Bao et al. [2020], He et al. [2020], Raffel et al. [2020]. To alleviate these issues, in our work, we introduce a linguistic-informed soft-masking process to corrupt the discrete and continuous space with structures, and then use linear projections and cross-entropy objectives to directly map the latent variables to textual data for better efficiency and generating better text.

2.2 Non-Autoregressive Text Generation

Most language models Chowdhery et al. [2022], Brown et al. [2020] and text generation models Vaswani et al. [2017a], Eikema and Aziz [2021], Chen and Yang [2020, 2021] follow a left-to-right autoregressive manner. However, the fixed generation order prevents the models' flexibility in editing former text based on later generation results, especially for global controllable generation settings. To overcome the limitations, non-autoregressive text modeling has been

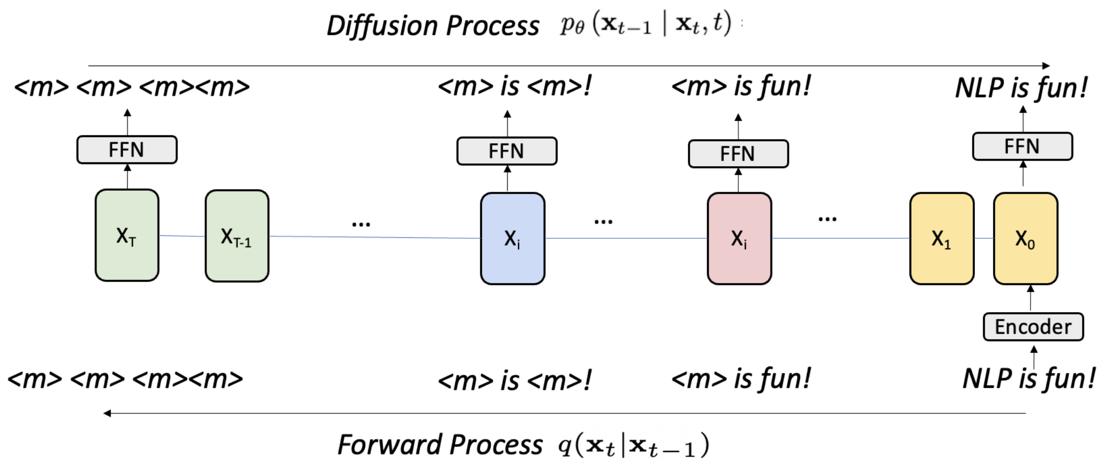

Figure 1: The overall process of our Masked-Diffuse LM. In the forward process, soft-mask is added to more informative words earlier to gradually corrupt the input text. For example, *NLP* is soft-masked prior to stop words like *is*. Then in the diffusion process, models learn to generate easy words like *is* first and then fill in more important words such as *fun* and *NLP*.

proposed Ghazvininejad et al. [2019], Ren et al. [2020], Gu et al. [2018], Saharia et al. [2020], Savinov et al. [2022] through masked language models Ghazvininejad et al. [2019], iterative sequence alignment Saharia et al. [2020], insertion and deletion Gu et al. [2018], or unrolling the generation path Savinov et al. [2022]. Our Masked-Diffuse LM achieves the non-autoregressive generation through gradually recovering the intermediate latent variables in a planned sequence from the forward process.

2.3 Plug-and-Play Controllable Generation

Our work is also closely related to the line of research about plug-and-play controllable generation methods Yang and Klein [2021], Dathathri et al. [2020], Krause et al. [2021], Liu et al. [2021], which modify the outputs based on extra guidance such as classifiers without changing or fine-tuning the pre-trained language models. Dathathri et al. [2020] used gradients to edit the autoregressive language model's hidden representations to fulfill the control guidance. Yang and Klein [2021] proposed to reweight the predicted token from the language models while Krause et al. [2021], Liu et al. [2021] further fine-tuned a smaller LM to reweight the token predictions. In this work, we apply the gradient-based plug-and-play approach to our Masked-Diffuse LM for controllable generation by making classifier-guided gradient updates to the intermediate latent variables during the diffusion process.

3 Background: Diffusion Models

Diffusion models are the recent state-of-the-art deep generative models via iteratively denoising the latent variables Sohl-Dickstein et al. [2015a], Ho et al. [2020], Song et al. [2021]. Basically, corruptions (usually Gaussian noise) are added to the input data distribution gradually during a forward process. Then a diffusion model is trained through learning to recover the corrupted distribution to the original input data distribution step by step. A small amount of information that is perturbed during the corresponding forward process is reconstructed in every diffusion step. The diffusion models are showing significant improvements Ramesh et al. [2022], Rombach et al. [2022], Kong et al. [2020], Savinov et al. [2022] as they generate the data in multiple steps, which is more stable and easier than learning to reconstruct the whole input data in a single forward pass Dieleman et al. [2022] like variational autoencoders Kingma and Welling [2013] and generative adversarial networks Goodfellow et al. [2014].

There are usually a forward noising process and a diffusion denoising process in a diffusion model. For a given sampled input data, $x_0 \sim q(x_0)$, a Markov chain of latent variables $\{x_1, \cdots, x_T\}$ are generated in the forward noising process ($q\left(x_t \mid x_{t-1}\right)$) by progressively adding a small amount of Gaussian noise to perturb the input data:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right), \tag{1}$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is a noise schedule controlling the amount of added noise in every step. Through the forward process, x_T becomes an isotropic Gaussian distribution. Note that there are no trainable parameters in the forward process.

Then a reversed diffusion process, which is learned by a parameterized model ($p(x_{t-1}|x_t)$), is learned to denoise x_T to the original data x_0:

$$p_\theta\left(x_{t-1} \mid x_t, t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right), \tag{2}$$

where $\mu_\theta(.)$ and $\Sigma_\theta(.)$ are the learned model that can be implemented by a U-Net Ronneberger et al. [2015] or a Transformer Vaswani et al. [2017b].

The diffusion model is trained to maximize the marginal likelihood of $\log p_\theta(x_0)$ and we manage to minimize the variational lower bound Sohl-Dickstein et al. [2015b] in practice:

$$\begin{aligned}\mathcal{L}_{\text{vlb}} &= \mathbb{E}_q\left[D_{\text{KL}}\left(q\left(x_T \mid x_0\right) \| p_\theta\left(x_T\right)\right)\right] \\ &+ \mathbb{E}_q\left[\sum_{t=2}^{T} D_{\text{KL}}\left(q\left(x_{t-1} \mid x_t, x_0\right) \| p_\theta\left(x_{t-1} \mid x_t, t\right)\right)\right] \\ &- \log p_\theta\left(x_0 \mid x_1\right).\end{aligned} \tag{3}$$

However, this objective is usually unstable and requires many optimization tricks to stabilize. Thus, we follow Ho et al. [2020] to expand and reweight each KL-divergence term in $\mathcal{L}_{\text{vlb}}$ and obtain a mean-squared error (L_2) loss:

$$\mathcal{L}_{\text{diffuse}}\left(x_0\right) = \sum_{t=1}^{T} \underset{q(x_t|x_0)}{\mathbb{E}} \left\|\mu_\theta\left(x_t, t\right) - \hat{\mu}\left(x_t, x_0\right)\right\|^2, \tag{4}$$

where $\hat{\mu}$ is the mean of the posterior $q(x_{t-1}|x_0, x_t)$, and μ_θ is the predicted mean of $p_\theta(x_{t-1}|x_t)$, which is predicted by the parameterized neural models.

4 Method: the Masked-Diffuse LM

In this section, we describe our introduced Masked-Diffuse LM. The overall diagram is shown in Figure 1. Different from the recent diffusion models for languages, e.g., Diffusion-LM Li et al. [2022], which are based on continuous diffusion models, we propose to make corruptions in both discrete and continuous space to help modeling the textual data. Specifically, we formulate a novel corruption process as an alternative to Gaussian diffusion (in Section 4.2) and we directly map continuous vectors to discrete inputs in every diffusion step with cross-entropy objectives (in Section 4.3). Moreover, our approach could easily integrate pre-trained language models (in Section 4.4).

4.1 Embedding

For the input sentence d with l tokens $d = \hat{w}_{1:l}$, we first map the discrete tokens to the continuous space and form the initial latent variable, X_0, through a learnable embedding layer or an encoder $e(.)$:

$$X_0 = w_{1:l} = e(w_{1:l}). \tag{5}$$

This bridges the discrete space and continuous space. We will then add designed soft-masked noise to the tokens' representations in the later diffusion models.

4.2 Forward Process with Soft-Masking

Different words in sentences play different roles. As a result, when corrupting the sentences and recovering the sentences, words with various importance should be treated differently. Thus, in this work, instead of evenly adding Gaussian noise to all the token embeddings like in Diffusion-LM Li et al. [2022], we add soft-masked noise to different tokens in the input text in different stages to corrupt the text gradually with structures. Intuitively, more important words would be perturbed with soft-masks in an earlier stage so that the model could be encouraged to generate them in the later phase to follow the *easy-first-generation* nature of language planning and generation.

In this work, we consider the following aspects to measure and define the importance of words in one sentence:

Word Relevancy We use the tf-idf weights Dessí et al. [2020], $w_{\text{tf-idf}}$, of the word as one way to measure the relevance of word w in one sentence d:

$$w_{\text{tf-idf}}(w,d) = \frac{f_{w,d}}{\sum_{w' \in d} f_{w',d}} \log \frac{N}{1 + |\{d \in D : w \in d\}|}, \tag{6}$$

where the $f_{w,d}$ is the number of times that word w occurs in sentence d, N is the number of sentences in the corpus, and D is the set of sentences, and $|\{d \in D : w \in d\}|$ is number of sentences where the word t appears. A higher weight for word w in sentence d in tf–idf means that the word might be more important in the sentence.

Entropy We also consider measuring the amount of information with entropy H Bentz and Alikaniotis [2016], He et al. [2022] in the word w to reflect the importance of that word:

$$H(w) = -p\left(w\right) \log\left(p\left(w\right)\right) \tag{7}$$

where $p\left(w\right) = \frac{f_w}{\sum_{j=1}^{V} f_j}$ represents the probability of word w and f is the word Reluency in the corpus. A word with lower entropy indicates that the word might contain less information and thus be less important compared to the words with higher entropy.

In practice, we combine these two measures (with normalization) to decide the importance I of the word w in one sentence d by:

$$I(w) = \frac{x_{\text{tf-idf}}(w,d)}{\sum_{w' \in d} w_{\text{tf-idf}}(w',d)} + \frac{H(w)}{\sum_{w' \in d} H(w')}. \tag{8}$$

Based on the introduced importance I of the words in a sentence, we first divide these words into m buckets $\{W_{1:m}\}$. The buckets with lower indices include words with higher importance. We will add soft-masked noise to words with higher importance before words with lower importance. By doing this, models could learn to generate the easier words first and then generate harder words in the reversed denoising process for better generation quality. Specifically, at every step t, we will add a small amount of Gaussian noise to the hidden representation of the word w_i in bucket $W_{\lfloor \frac{tm}{T} \rfloor}$:

$$q(w_{i,t+1}|w_{i,t}) = N(w_{i,t+1}; \sqrt{(1-\beta_t)}w_{i,t}, \beta_t I), \tag{9}$$

where β_t is the amount of noise added at diffusion step t.

We further apply a square-root noise schedule following Li et al. [2022] to gradually increase β_t:

$$\beta_t = 1 - \sqrt{t/T + s}, \tag{10}$$

where s is a small constant that corresponds to the starting noise level. Thus, less noise would be added to harder words to stabilize the training. By performing the above noising steps, initial latent variable X_0 is gradually corrputed to a series of noisy latent variables $X_{1:T}$.

4.3 Diffusion Process

After the forward process to corrupt the input tokens in sentences d into latent variables $X_{1:T}$, we then gradually denoise X_T back to X_0 through diffusion steps, $\hat{X}_{t-1} = p(\hat{X}_t|\theta)$, where θ is the learned parameter to model the state transition. In practice, we model the transition with Transformers Vaswani et al. [2017b].

After every diffusion step $t \in (0, T]$, instead of minimizing the distance between the hidden representations of $\hat{X}_{t-1}$ and X_0 Li et al. [2022], we first directly map the continuous space to discrete space using a learnable linear layer $f(.)$ and then minimize a weighted cross entropy between the predicted sentence and (i) the original sentence d and (ii) the masked sentence $\hat{d}$ at time step $t-1$:

$$\mathcal{L}_t = \gamma_t CE(f(\hat{X}_{t-1}), d; \theta) + CE(f(\hat{X}_{t-1}), \hat{d}; \theta), t \in (0, T]$$

Here, $\gamma_t = \frac{T-t}{T}$. In other words, we put higher weights on the masked tokens that are masked in this time step during the forward process and put lower weights to the other tokens. So the models are learned to generate the corresponding masked tokens first at every time step.

Table 1: Main Results. The Accuracy (↑) and the Fluency (↓) of different methods on five controllable generation tasks including semantic content, POS, syntax tree, syntax spans and length. † indicates our methods.

Methods	Semantic Content		POS		Syntax Tree		Syntax Spans		Length	
	Acc	Fluency	Acc	Fluency	Acc	Fluency	Acc	Fluency	Acc	Fluency
PPLM	9.9	5.32	-	-	-	-	-	-	-	-
FUDUGE	69.9	2.83	27.0	7.96	17.9	3.39	54.2	4.03	46.9	3.11
Diffusion-LM	81.2	2.55	90.0	5.16	86.0	3.71	93.8	2.53	99.9	2.16
+ BERT	77.4	2.68	86.2	5.43	82.3	3.92	89.3	3.13	99.9	2.68
Masked-Diffuse LM †	81.9	2.35	91.6	5.03	86.6	3.66	94.7	2.48	99.9	2.13
+ BERT †	**82.9**	**2.30**	**92.9**	**4.78**	**89.7**	**3.44**	**95.8**	**2.33**	**100**	**2.08**

Table 2: Training time and inference time (generating 50 samples) for different models.

Methods	**Training** (h)	**Inference** (s)
Diffusion-lm	8.0	80
+BERT	15.2	920
Masked-Diffuse LM	3.4	68
+BERT	4.8	700

4.4 Adapting Pre-trained Language Models

Our introduced Masked-Diffuse LM also allows the use of large pre-trained language model Devlin et al. [2019], Liu et al. [2019], Yang et al. [2019], Joshi et al. [2019], Sun et al. [2019], Clark et al. [2019], Lewis et al. [2020], Bao et al. [2020], He et al. [2020], Raffel et al. [2020]. In this work, we use BERT Devlin et al. [2019] as an example. To combine the prior knowledge in large language models, it is straightforward to directly replace the embedding layer $e(.)$ with the pre-trained model and use the pre-trained model to get the hidden representations of input tokens as the initial state in diffusion models. We use the final linear layers in pre-trained models to predict the tokens. For efficiency, in our experiments, when using pre-trained models, we freeze the parameters in them and only learn the transition model θ in our Masked-Diffuse LM.

5 Controllable Text Generation with Masked-Diffuse LM

In this section, we illustrate how we apply our Masked-Diffuse LM to fulfill controllable text generation. Inspired by recent plug-and-play methods Yang and Klein [2021], Dathathri et al. [2020], Krause et al. [2021], Liu et al. [2021], we conduct controls c from external modules (e.g., classifiers) directly on the latent variables X_t in every intermediate step $t \in [0, T]$ in our Masked-Diffuse LM:

$$p\left(X_{0:T} \mid c\right) = \prod_{t=1}^{T} p\left(X_{t-1} \mid X_t, c\right). \tag{11}$$

We follow the conditional independence assumption Yang and Klein [2021], Dathathri et al. [2020], Krause et al. [2021], Liu et al. [2021] and decompose the above joint probability into a sequence of control task at every time step t:

$$\begin{aligned} p\left(X_{t-1} \mid X_t, c\right) &\propto p\left(X_{t-1} \mid X_t\right) \cdot p(c \mid X_{t-1}, X_t) \\ &= p\left(X_{t-1} \mid X_t\right) \cdot p(c \mid X_{t-1}). \end{aligned} \tag{12}$$

As a result, for the t-th step, we run gradient updates on X_t to generate X_{t-1}:

$$\begin{aligned} \nabla_{X_{t-1}} \log p\left(X_{t-1} \mid X_t, c\right) = \lambda \nabla_{X_{t-1}} &\log p\left(X_{t-1} \mid X_t\right) \\ &+ \nabla_{X_{t-1}} \log p\left(c \mid X_{t-1}\right), \end{aligned} \tag{13}$$

where both $\log p(X_{t-1}|X_t)$ and $\log p(c|X_{t-1})$ are differentiable: the first term is parametrized by the transition Transformers, θ, in Masked-Diffuse LM, and the second term is parametrized by extra neural network classifiers. Note that the extra classifiers are trained with the diffusion latent variables as input to allow direct gradient updates on the

Table 3: The average ranking every method receives from human evaluation (lower is better).

Methods	Semantic Content	POS	Syntax Tree	Syntax Spans	Length
Diffusion-lm	2.89	2.76	3.16	2.88	2.46
+BERT	3.87	3.46	3.72	3.68	3.34
Masked-Diffuse LM	2.56	2.48	2.88	2.35	2.18
+BERT	**1.32**	**1.28**	**1.16**	**1.55**	**1.86**

Table 4: Performances on Semantic Content of Masked-Diffuse LM with different types of noise applied in forward noising process. † indicates our method.

Noise Type	Semantic Content Acc	Semantic Content Fluency
Gaussian	75.3	3.01
Random Mask	78.8	2.67
Mask w. POS	80.4	2.58
Mask w. Entropy	81.1	2.44
Mask w. Rel	80.8	2.52
Mask w. Entropy+Rel †	**81.6**	**2.38**

latent space. Note that λ is a fluency regularization hyper-parameter to balance the fluency (gradient updates from Masked-Diffuse LM) and control (gradient updates from classifiers) in order to further improve the generation quality.

For the decoding strategy, following Li et al. [2022], the Minimum Bayes Risk (MBR) decoding Kumar and Byrne [2004] is used to aggregate and select the sample that has the lowest expected loss under the specified loss function from the Masked-Diffuse LM.

6 Experiments

6.1 Datasets

In this work, we train our Masked-Diffuse LM on the E2E datasets Novikova et al. [2017], which consists of 50K restaurant reviews together with the labels in terms of food type, price, and customer ratings.

Following Li et al. [2022], we conduct 5 control tasks to evaluate the learned Masked-Diffuse language model:

- **Semantic Content.** For a given field (e.g., *food*) and value (e.g., *Japanese*), sentences that covers field=value need to be generated. We evaluate the accuracy of the generated sentence by examine the exact match rate of "value" (word mention).
- **Parts-of-speech.** For a given sequence of parts-of-speech (POS) tags (e.g., *Noun Verb Determiner Noun*), the models need to produce the sentence with the same length and follow the exact given POS tag sequence (e.g., *Birds eat the warms*). We evaluate the accuracy of the generation by checking the word-level POS tag exact match (under an oracle POS tagger).
- **Syntax Tree.** For a given syntactic parse tree, the generated sentence should have the same parse tree. We evaluate the accuracy by first parsing the generated sentence with an off-the-shelf parser and report the F1 scores compared to the given parse.
- **Syntax Spans.** For a given (span, syntactic category) pair (e.g., *(2, 5, VP)*), the parse tree of the generated sentence should match the given syntactic category over the given spans. We evaluate the accuracy of the sentence by the exact match rate of the given spans.
- **Length.** For a given target length (e.g., *20*), the models need to generate a sentence within ± 2 of the given target. We evaluate the accuracy by the match rate of the sentence lengths.

For every control task, we sample 200 control targets c from the validation splits, and we generate 50 samples for each control target. The first four tasks rely on a classifier to guide the diffusion, and the last one task is classifier free. To

Table 5: Performances of Masked-Diffuse LM trained with different objectvies on controllable generation tasks. † indicates our method.

Methods	Semantic Content		POS		Syntax Tree		Syntax Spans		Length	
	Acc	fluency	Acc	fluency	Acc	fluency	Acc	fluency	Acc	fluency
L2	81.1	2.44	90.6	5.17	86.2	3.68	94	2.51	99.8	2.14
L2-BERT	80.1	2.48	89.4	5.82	84.1	3.91	93.2	2.88	99.9	2.89
CE †	81.9	2.35	91.6	5.03	86.6	3.66	94.7	2.48	99.9	2.13
CE-BERT †	**82.9**	**2.30**	**92.9**	**4.78**	**89.7**	**3.44**	**95.8**	**2.33**	**100**	**2.08**

Table 6: Examples of the intermediate generated text of our Masked-Diffuse LM on the Length and Semantic Content tasks.

Case Study	Sentences
Input	*7*
$t = 500$	[*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*]
$t = 400$	[*mask*] is an [*mask*] restaurant .
$t = 200$	The [*mask*] is an Indian restaurant .
$t = 0$	The Mill is an Indian restaurant .
Input	*name : Travellers Rest Beefeater*
$t = 500$	[*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*] [*mask*]
$t = 400$	[*mask*] Rest [*mask*] is a [*mask*] [*mask*] [*mask*] that is [*mask*] .
$t = 200$	Travellers Rest [*mask*] is a reasonably [*mask*] restaurant that is awesome .
$t = 0$	Travellers Rest Beefeater is a reasonably priced restaurant that is awesome .

further evaluate the fluency of the generated sentences from models, we use a teacher LM (i.e., a carefully fine-tuned GPT-2 model) and report the perplexity of generated text under the teacher LM. A lower perplexity indicates better sample quality and fluency.

6.2 Baselines

We compare our Masked-Diffuse LM with the following state-of-the-art baselines on controllable generation tasks:

- **PPLM** Dathathri et al. [2020] runs gradient ascent on the pre-trained language models' hidden representations to increase the classifier probabilities and language model probabilities.
- **FUDGE** Yang and Klein [2021] reweights the predicted tokens from the pre-trained language models by a discriminator which takes in a prefix sequence and predicts whether the complete sequence would satisfy the constraint.
- **Diffusion-LM** Li et al. [2022] learns an embedding to map discrete text into the continuous space where it performs Gaussian diffusion process. Also, a rounding step is designed to map the embeddings back into discrete texts. For every control task, the Diffusion-LM infuses the controlling signals in every diffusion step.

6.3 Experimental Setting

We use a Transformer with 80M parameters to parameterize our Masked-Diffuse LM, with a sequence length $n = 64$, diffusion steps $T = 500$, and a square-root noise schedule. For Masked-Diffuse LM, we set the hidden dimension to 128. We set the number of word buckets $m = 3$. When combining pre-trained models, we incorporate BERT-base Devlin et al. [2019] with about 110M parameters. We use BERT to encode the input text into vectors with dimension of 768 and freeze the parameters in BERT. We learn Masked-Diffuse LM with the AdamW optimizer Loshchilov and Hutter [2019] for 20,000 steps with learning rate of 3e-4, dropout probability of 0.1, and batch size of 32. We use a linear warmup schedule starting with 1,000 warmup steps. All experiments are conducted on NVIDIA A100 Tensor Core GPUs. We use 4 GPUs for training and a single GPU for sampling.

6.4 Results

We show the main results on five controllable generation tasks in Table 1. When the diffusion process is engaged, the performances on all the controlled generation tasks receives significant boosts (e.g., 81.2 of Diffusion-LM vs. 69.9 if FUDUGE on Semantic Content task), suggesting the superiority of the diffusion model on controllable generation tasks. While the previous Diffusion-LM can not be well combined with large language model like BERT (e.g., a 5% drop on Semantic Content accuracy), largely due to the fact that their way (rounding) to bridge continuous space and discrete space suffers from significantly higher dimensions. Compared to Diffusion-LM, our proposed Masked-Diffuse LM consistently outperforms the previous models in all tasks (e.g., a 1.7% improvement on the POS task), indicating the effectiveness of our introduced linguistic-informed noise forward process. Also, when combined with large language models like BERT, our method significantly outperforms the previous methods, demonstrating that our approach can be well aligned with pre-trained models.

Efficiency We also display the training cost and inference cost in Table 2. Compared to the previous Diffusion-LM, our method requires significantly less training time to converge and needs less inference time to generate sentences. This is because our introduced noise process is more stable and suitable for modeling languages. Besides, the objectives we introduced are more efficient than the rounding techniques in previous work.

Human Evaluation We then conduct human evaluation to evaluate the generated conversations qualitatively. We ask native speakers of English from Amazon Mechanical Turk to rank the quality of 50 generated sentences (randomly sampled) from different models for every control task. Specifically, annotators need to rank different system outputs based on the (i) fluency (whether the given sentence is readable and fluent) and (ii) the controllability (whether the given sentence match the given control conditions). To increase annotation quality, we require turkers to have a 98% approval rate with over 10,000 approved tasks for their previous work. The pay rate was $0.15 per hit. Every example is assessed by 3 annotators, and the rank for every sentence is aggregated by majority voting. The Intra-Class Correlation (*ICC1k*) was 0.63, indicating moderate agreement Koo and Li [2016]. The results are shown in Table 3. As it shows, our proposed Masked-Diffuse LM and its variation with BERT received the best average ranks, suggesting the effectiveness of our proposed diffusion modeling strategy for languages.

6.5 Ablation Studies

We then perform ablation studies to demonstrate the effectiveness of our introduced linguistic-informed noise and the cross entropy objectives.

Noise Strategy We first demonstrate the performances on Semantic Content task of Masked-Diffuse LM with different types of noise strategy in Table 4. *Gaussian* adds Gaussian noise to all the tokens in the input sentence in the forward process following Li et al. [2022]. We also compare different masking noise strategies: (i) Random Mask, where the soft-mask is added to tokens in a random order. (ii) Mask with POS, where the soft-mask perturbs the tokens in an order (noun $\rightarrow$ verb $\rightarrow$ other words) based on POS tags. Our introduced noise strategy (Mask with Entropy and Reluency) shows significantly better performances on semantic content generation. This indicates that our introduced noise strategy that considers the linguistic features in sentences is providing more appropriate perturbation to the textual data for the diffusion process.

Objectives We further show the impact of different objectives in Table 5. We compare our used cross entropy objectives with the L_2 object that is used in Li et al. [2022] where they minimize the distance between latent intermediate variables and the initial latent variable instead of directly predicting the text. We observe that cross entropy objectives slightly perform better than L_2 when the pre-trained model is not used. After combining with large language models, CE-BERT significantly outperforms the L_2-BERT, indicating the effectiveness of our introduced objectives in terms of incorporating large language models.

6.6 Case Studies

We also include some examples of intermediate steps of Masked-Diffuse LM in Table 6. In the denoising diffusion process, easy words are generated first. For example, "*is*", "*an*", and "*restaurant*". With more diffusion steps, sentences are enriched with more informative words such as "*Mill*" and "*Indian*". It shows that our Masked-Diffuse LM encourages the generation to follow an easy-first order for stable and better generation quality.

7 Conclusion

In this work, we present a novel diffusion model for language, Masked-Diffuse LM, which corrupts the discrete text with a linguistic-informed soft-masking strategy and then iteratively denoises them back by directly predicting the text. Specifically, we gradually soft-mask the tokens in the sentence following an order from more informative words to less informative words in the forward process. This satisfies the flexibility for diffusion models, as well as encourages the easy-first-generation nature in the denoising process for better generation quality. Also, we directly predict the discrete token during the diffusion process with the cross-entropy loss to stabilize the intermediate diffusion steps and make our approach orthogonal to large pre-trained language models. Experiments on E2E dataset and five controllable generation tasks including Semantic Content, Parts-of-speech, Syntax Tree, Syntax Spans, and Length show that our Masked-Diffuse LM can (i) achieve the state-of-the-art performances compared to recent baseline models and (ii) allow more efficient training and inference compared to the previous Diffusion-LM.

References

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data, 2022. URL https://arxiv.org/abs/2211.15089.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015a. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2022. URL https://arxiv.org/abs/2209.00796.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2020. URL https://arxiv.org/abs/2009.09761.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation, 2022. URL https://arxiv.org/abs/2205.14217.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models, 2022. URL https://arxiv.org/abs/2210.08933.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models, 2022. URL https://arxiv.org/abs/2211.15029.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2021. URL https://arxiv.org/abs/2107.03006.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions, 2021a. URL https://arxiv.org/abs/2102.05379.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models, 2021b. URL https://arxiv.org/abs/2110.02037.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2019.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *SCL*, 2020.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. *arXiv preprint arXiv:2002.12804*, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5525. URL https://aclanthology.org/W17-5525.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. Step-unrolled denoising autoencoders for text generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=T0GpzBQ1Fg6.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017a. URL https://arxiv.org/abs/1706.03762.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum bayes risk decoding for neural machine translation, 2021. URL https://arxiv.org/abs/2108.04718.

Jiaao Chen and Diyi Yang. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.336.

Jiaao Chen and Diyi Yang. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.109. URL https://aclanthology.org/2021.naacl-main.109.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1633. URL https://aclanthology.org/D19-1633.

Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, and Tie-Yan Liu. A study of non-autoregressive model for sequence generation, 2020. URL https://arxiv.org/abs/2004.10454.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1l8BtlCb.

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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.83. URL https://aclanthology.org/2020.emnlp-main.83.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.424. URL https://aclanthology.org/2021.findings-emnlp.424.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017b.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015b. URL https://arxiv.org/abs/1503.03585.

Danilo Dessí, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. Tf-idf vs word embeddings for morbidity identification in clinical notes: An initial study. 2020. doi: 10.5281/ZENODO.4777594. URL https://zenodo.org/record/4777594.

Christian Bentz and Dimitrios Alikaniotis. The word entropy of natural languages, 2016. URL https://arxiv.org/abs/1606.06996.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.