

Natural Language Generation with Pixels

Stanford CS224N Custom Project

Gautam Mittal

Department of Computer Science
Stanford University
gmittal@stanford.edu

Rajan Vivek

Department of Computer Science
Stanford University
rvivek@stanford.edu

Abstract

Tremendous progress has been made in the development of modern natural language processing systems due to transformer-based autoregressive models. Unfortunately, these architectures rely on a finite vocabulary which bottlenecks their transferability to new languages. Recent work (Rust et al., 2023) has shown that by casting the task as a vision problem, these issues can be alleviated for encoder-only language models. Inspired by the recent success of diffusion models on image generation tasks, we propose a conditional diffusion-based decoder for modeling rendered natural language. We demonstrate that our approach is capable of generating coherent, plausible natural language rendered as images. We also pair our decoder with a powerful off-the-shelf encoder and investigate its potential for sequence-to-sequence tasks such as machine translation. We compare our approach with strong transformer-based autoregressive baselines in both the unconditional and sequence-to-sequence setting.

1 Key Information to include

- Mentor: Yuan Gao
- External Collaborators (if you have any): Prof. Douwe Kiela (Stanford)
- Sharing project: no

2 Introduction

Transformer-based autoregressive models (Vaswani et al., 2017) have significantly improved the state-of-the-art of natural language systems, and their downstream performance on zero-shot, few-shot (Brown et al., 2020), and fine-tuned (Wang et al., 2019) tasks has made them the standard architecture for modern practitioners and researchers. One of the most common building blocks in this modern NLP stack is a tokenizer, which maps sequences of unstructured natural language to sequences of discrete subwords units that can be more easily processed. While subword tokenizers have allowed for more expressive language models capable of operating over longer sequences of text, they also restrict models to operate over a finite size vocabulary. This *vocabulary bottleneck* created by tokenizers is a key challenge faced by language systems when generalizing to new languages, and several attempts have been made to remove tokenizers entirely (Choe et al., 2019; Xue et al., 2021; Clark et al., 2022).

PIXEL (Rust et al., 2023) avoids this challenge altogether by casting language modeling as a vision problem: text is rendered as images and the model uses masked autoencoding (He et al., 2021), which is similar to BERT (Devlin et al., 2019) and other masked language models but for continuous domains like images, to learn a language encoder representation. Since text is rendered as images, there is no need for tokenization. Notably, PIXEL generalizes well to languages with similar orthographic features. To our knowledge, there has been no work that explores applying similar techniques to natural language generation.

Just as transformers have taken the natural language processing community by storm, denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) have emerged as an important generative model for high-quality samples in continuous domains such as audio (Kong et al., 2021), images (Dhariwal and Nichol, 2021), and text-to-image generation (Rombach et al., 2021; Ramesh et al., 2022).

Inspired by these two observations, our goal is to cast natural language generation as a visual task, such that diffusion models can learn to generate text. The advantages of this approach include removing the vocabulary bottleneck, non-autoregressive generation, and robustness to languages with similar orthographic features. From a research perspective, existing diffusion model research has not heavily investigated text rendering and Ramesh et al. (2022) observe that DALL·E-2 “struggles at producing coherent text.” Our work investigates whether this simple approach can yield performance similar or better than existing decoder-only and encoder-decoder transformer models for both unconditional generation and machine translation. Additionally, just as many ideas from the natural language processing community have been transferred over to visual recognition tasks, we are excited by the possibility of the flipped scenario as well.

Our contributions are as follows:

1. We introduce a framework for pixel-based language generation applicable to both unconditional text generation and sequence-to-sequence machine translation.
2. We compare the quantitative performance of our approach for unconditional text generation and machine translation against strong transformer-based baselines and present qualitative results.

3 Related Work

Given recent interest in large-scale language (Vaswani et al., 2017; Brown et al., 2020), diffusion (Ho et al., 2020), and multimodal (Radford et al., 2021; Rombach et al., 2021) models trained with self-supervision for use on downstream tasks (Bommasani et al., 2021), the deep learning community has begun to explore the use of techniques from domains such as vision applied to tasks in text.

Visual representations of natural language. To the best of our knowledge, one of the early examples of using vision for language tasks was shown by Radford et al. (2021), where the CLIP vision encoder representations of rendered instances of text from SST (Socher et al., 2013) was used for zero-shot sentiment analysis. A key insight was that optical character understanding abilities of CLIP allowed for tokenization-free and competitive zero-shot classification performance on text understanding tasks. Similarly, CLIPPO (Tschannen et al., 2022) performs contrastive pre-training using rendered texts instead of tokenized texts and shows competitive performance to baselines on the GLUE (Wang et al., 2019) benchmark. PIXEL (Rust et al., 2023) uses masked autoencoding with vision transformers (He et al., 2021) to pre-train a masked language model on rendered texts. A key difference between our work and prior use of vision for natural language is that we focus on generation whereas prior work focuses on representation learning and language understanding objectives.

Diffusion language models. Diffusion models (Ho et al., 2020) have shown impressive generation abilities in continuous domains such as images and audio, and interest in extending them to language modeling has increased rapidly. Diffusion-LM (Li et al., 2022) and CDCD (Dieleman et al., 2022) use a denoising diffusion probabilistic model to generate sequences of tokens that have been projected into a learned embedding space. Both works show that unconditional and controllable generation improve as a result of diffusion-based non-autoregressive modeling. D3PM (Austin et al., 2021) shows that diffusion modeling techniques can be applied directly to discrete data modalities, including language, without the need to first project tokens into a learned continuous representation for downstream modeling. Unlike our work, these methods all rely on a subword tokenizer for modeling and operate on a continuous form of language space rather than on a visual representation of language.

Non-autoregressive machine translation. Developing non-autoregressive approaches to generation and translation has remained appealing due to parallelism and generative modeling advantages. Gu et al. (2017) introduce a transformer-based encoder-decoder architecture, similar to Vaswani et al. (2017), that instead uses a non-causal attention mask in the decoder to enable fast translation. Similarly, approaches using iterative refinement (Lee et al., 2018, 2020), similar to the sampling process used with diffusion-based methods, have been explored for non-autoregressive generation. Diffusion-based approaches for zero-shot neural machine translation (Nachmani and Dovrat, 2021)

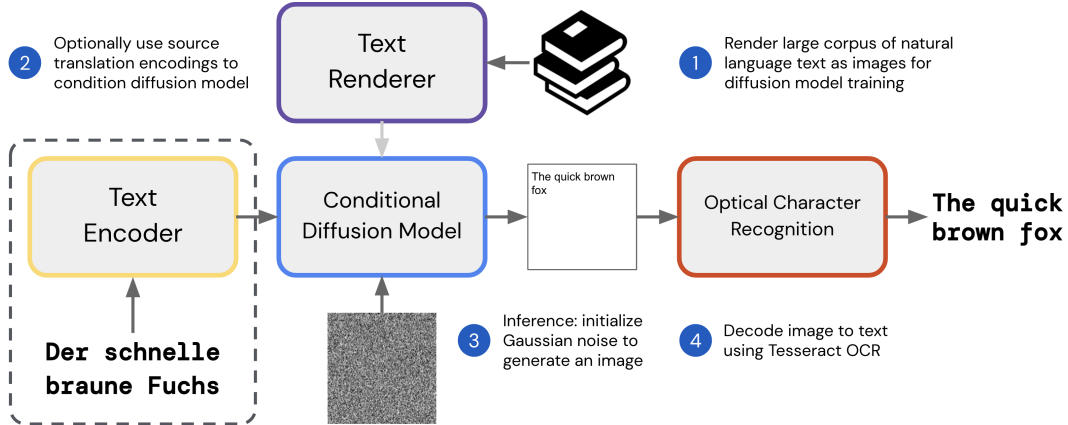


Figure 1: Our approach for training and generating from pixel-based natural language decoder.

have also begun to emerge, following similar architectures as prior work but with a diffusion pre-training objective over categorical vocabulary distributions. All of these existing approaches require a tokenizer as well as use the language data directly rather than the visual representation.

Tokenizer-free methods. Subword tokenizers have been important for the success of modern language modeling systems, allowing for longer sequences with improved perplexity. While many early works (Sutskever et al., 2011; Graves, 2014) relied on character-level methods, later works were able to significantly improve on these results using tokenization and more expressive architectures. In Choe et al. (2019), the authors find that the gap between byte-level and tokenized language models can be bridged by increasing the capacity of the underlying architecture. Additionally, Xue et al. (2021) and Clark et al. (2022) explore scaling up efficient token-free representations for language and show promising results in this direction, motivating further research in this area. Unlike PIXEL (Rust et al., 2023) and our work, all of these previous approaches do not operate on visual representations of natural language.

4 Approach

4.1 Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs (Ho et al., 2020) are a class of generative models that define latents x_1, \dots, x_N of the same dimensionality as the data $x_0 \sim q(x_0)$. Diffusion models are comprised of a diffusion process and a reverse process. The diffusion process starts from the data distribution x_0 and iteratively adds Gaussian noise according to a fixed noise schedule for N diffusion steps:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (1)$$

$$q(x_{1:N}|x_0) = \prod_{t=1}^N q(x_t|x_{t-1}) \quad (2)$$

where $\beta_1, \beta_2, \dots, \beta_N$ is a noise schedule that converts the data distribution x_0 into latent x_N . The choice of noise schedule has been shown to have important effects on sampling efficiency and quality (Ho et al., 2020).

The reverse process is defined by a Markov chain parameterized by learned parameters θ that iteratively refines latent point $x_N \sim \mathcal{N}(0, I)$ into data point x_0 . The learned transition probabilities are defined as,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)) \quad (3)$$

$$p_\theta(x_{0:N}) = p(x_N) \prod_{t=1}^N p_\theta(x_{t-1}|x_t) \quad (4)$$

where the objective is to gradually denoise samples at each reverse diffusion step t . In practice, σ_θ is set to an untrained time-dependent constant based on the noise schedule, and Ho et al. (2020)

found $\sigma_\theta(x_t, t) = \sigma_t = \frac{1-\bar{\alpha}_t}{1-\alpha_t}\beta_t$ to have reasonable practical results, where $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$.

The training objective is to maximize the log likelihood of $p_\theta(x_0) = \int p_\theta(x_0, \dots, x_N) dx_{1:N}$, but the intractability of this marginalization leads to the following evidence lower bound (ELBO):

$$\begin{aligned} \mathbb{E} [\log p_\theta(x_0)] &\geq \mathbb{E}_q \left[\log \frac{p_\theta(x_{0:N})}{q(x_{1:N}|x_0)} \right] \\ &= \mathbb{E}_q \left[\log p(x_N) + \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] \end{aligned} \quad (5)$$

The forward process can be computed for any step t such that $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I)$, which can be viewed as a stochastic encoder. To simplify the above variational bound, Ho et al. (2020) propose training on pairs of (x_t, x_0) to learn to parameterize this process with a simple squared L2 loss. The following objective is simpler to train, resembles denoising score matching (Song and Ermon, 2020) and was found to yield higher-quality samples:

$$L(\theta) = \mathbb{E}_{x_0, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2 \right] \quad (6)$$

where t is sampled uniformly between 1 and N , $\epsilon \sim \mathcal{N}(0, I)$, and ϵ_θ is the learned diffusion model.

In Deng et al. (2023), an updated training procedure based on scheduled sampling (Bengio et al., 2015) is used to improve generation quality. A key observation is that x_t at training time is derived using $q(x_t|x_{t-1})$ and at test time the model’s learned reverse process p_θ is used instead. This means that if the model makes a mistake at test time, it will not be able to correct for it. To mitigate this, we generate x_t using $p_\theta(x_t)$ with some probability m during training and we increase m as training progresses so that our model becomes robust to its own mistakes at generation time.

4.2 Rendering Text as Pixels

An original aspect of our approach is training a diffusion model for natural language generation tasks through generating images of text, which to our knowledge has not been explored. Our method learns to model grayscale images of rendered English text and therefore requires that human language datasets are preprocessed and rendered accordingly. Given a string of text, either a short sentence for our unconditional text generation task or a target English translation for our machine translation task, we first split the sentence into subwords using an off-the-shelf implementation of the GPT-2 bytetrain tokenizer (Radford et al., 2019a) from HuggingFace (Wolf et al., 2020). The reason for explicitly rendering subword tokens is to allow more characters to fit per line on the image canvas and to increase the accuracy at text decoding time with the optical character recognition (OCR) system. Our diffusion model does not explicitly use the tokenizer for generating rendered images.

We draw each subword on a 224x224 white canvas separated by spaces in Deja Vu Serif, a font that is known for improving OCR accuracy. While most short words will be tokenized as their original forms (e.g. "hi" is tokenized as "hi"), long words may be tokenized into multiple subwords (e.g. "important" may be tokenized as "import" and "ant") by the tokenizer. To distinguish tokens that compose a larger word from those that are standalone words, we use a special separator token "#" in between subword tokens that compose a larger word.

We repeat the above process for every example in the dataset and save the results to disk. An example of a rendered text string using our method can be seen in Figure 2.

4.3 Generating and Decoding Text

Our diffusion model uses a UNet backbone (Ronneberfer et al., 2015) and the the objective is to learn the distribution of images with English text from the E2E dataset. For our conditional machine

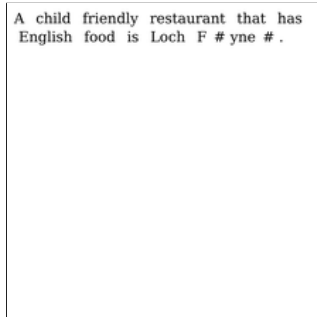


Figure 2: Sample from rendered E2E dataset with original text: "A child friendly restaurant that has English food is Loch Fyne."

translation task, the diffusion model is a UNet similar to our unconditional architecture but with a frozen text encoder and cross-attention layers added to allow the generation to be conditioned on input text in a source language. Our text encoder is RoBERTa (Liu et al., 2019). Our diffusion model architecture is from the HuggingFace Diffusers library (von Platen et al., 2022) and our custom training codebase is a heavily modified version of markup2img (Deng et al., 2023). We sample from the diffusion model using the original DDPM sampling algorithm (Ho et al., 2020) using the default noise schedule and T sampling steps to iteratively refine Gaussian noise into an image with rendered English text. To decode rendered outputs, we use Tesseract (Kay, 2007), a powerful off-the-shelf OCR system, to convert images to text for use with downstream metrics.

4.4 Baselines

To establish baselines for our diffusion experiments, we trained transformer-based models using the same text data sources as our diffusion models. We ran three initial baseline experiments: unconditional text generation with a 87M parameter GPT-2 (Radford et al., 2019b), machine translation with a 60M parameter T5 (Raffel et al., 2019), machine translation with a 60M parameter T5 and span corruption pretraining prior to finetuning. Our T5 training scripts and our GPT-2 training scripts were adapted from and guided by Wolf et al. (2020). We wrote our own evaluation scripts, but were similarly guided by Wolf et al. (2020).

5 Experiments

5.1 Datasets and Tasks

For our unconditional experiments, we train our diffusion and baseline GPT-2 models on 42k human reference sentences from E2E (Novikova et al., 2017), a lexically rich dataset of restaurant descriptions. The task is to generate high-quality text, measured by the lm-score (Section 5.2).

For our conditional experiments, we train our baseline and diffusion models on the WMT-14 German-to-English translation dataset (Bojar et al., 2014) with 4.5M training sentence pairs. The task is to translate German sentences into English sentences.

5.2 Evaluation method

To evaluate the quality of unconditionally generated text, we report the lm-score (Li et al., 2022). This metric is the perplexity of the generated text under a large fine-tuned "teacher" model. Our teacher model was the pretrained "gpt2-large" available on HuggingFace (Wolf et al., 2020). A lower lm-score indicates that the generated text better matches the distribution of text seen by the teacher during fine-tuning, indicating that the text is realistic and thus high-quality.

For our machine translation task, we used the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002) and character-level F-score (chrF) score (Popović, 2017) to evaluate our model's generated translations with one reference translation for each sentence. The BLEU score assesses the similarity of predicted and reference translations using n-gram overlap at the word level, while chrF uses character-level n-gram overlap.

5.3 Experimental details

For our unconditional text generation baseline, we trained our GPT-2 model from random initialization on the E2E dataset for 20 epochs with a context length of 32, batch size of 32, 8 gradient accumulation steps, learning rate of $5e-4$, and cosine learning rate scheduler. To generate text from the model, we performed nucleus sampling (Holtzman et al., 2020) with a maximum length of 32 and top probability score of 0.95. We generated a total of 400 sequences for evaluation.

For our machine translation baseline, we trained our T5 model with random initialization for 100k steps on the WMT-14 German-English dataset with a context length of 64, batch size of 64, and learning rate of $5e-5$. Separately, we pretrained our T5 model with the span corruption pre-training objective (Raffel et al., 2020) for 3 epochs with a context length of 64, batch size of 32, and learning rate of $5e-5$. We then finetuned the model using the same method described previously.

For our diffusion experiments, we train a UNet2DConditional model, which contains 2 convolutional layers per residual block and 12 total residual blocks. Each residual block has increasing output channel dimension on the downsampling tower of the UNet, with a reversed tower (decreasing output channel dimension) for upsampling. The downsampling tower uses 2 down projection blocks and 4 cross-attention down projection blocks (for conditioning generation) and the upsampling tower uses the same configuration but reversed and with up projection blocks. The output channels per block are 128, 128, 256, 256, 512, 512. Rendered text images are converted to grayscale and normalized to have a pixel mean of 0.5 and standard deviation of 0.5. We use AdamW with a constant learning rate $1e-4$ with 500 warmup steps, weight decay 0.01, and batch size 48. We use $T = 1000$ generation steps at test time. For our unconditional model, we train for 70,000 steps on the E2E dataset and generated 400 samples for evaluation. For our conditional translation model, we train for 135,000 steps and generate translations on the validation split for our quantitative metrics.

5.4 Results

Model	lm-score (\downarrow)
GPT2-87M	13.17
Diffusion	5640.05

(a) Unconditional text generation results

Model	BLEU (\uparrow)	chrF (\uparrow)
t5-small (no span corruption)	6.44	27.94
t5-small (span corruption)	19.11	47.82
Diffusion	0.062	16.09

(b) Sequence-to-sequence machine translation results

Table 1: Task performance for baseline and diffusion models

Table 1a shows the lm-score achieved by our diffusion model and GPT-2 baseline. Table 1b shows the BLEU and chrF scores achieved by our diffusion model as well as our T5 baseline both with and without span corruption pretraining on the corpus prior to fine-tuning. See 6 for discussion of these results.

6 Analysis

6.1 Unconditional Generation

Our diffusion model achieves an incredibly high lm-score relative to our baseline, indicating that the unconditionally-generated text is very different from the E2E text distribution learned by the teacher during fine tuning. Qualitatively, we observe that the diffusion model generates images containing sentences with excellent glyph structure, many coherent words, and some coherent phrases. Thus, our diffusion model learns lower-level structure of written language well but struggles to capture higher-level structure. We suspect this is a function of model size, as discussed in 6.3. We note that the lm-score operates at the subword level due to GPT-2’s tokenization. Thus, nonsensical or even slightly misspelled words are heavily penalized by this metric because they correspond to subword sequences not found in the E2E dataset. This explains our high lm-score.

6.2 Machine Translation

Our conditional diffusion model achieves BLEU and chrF scores well below our T5 baselines, with an incredibly low BLEU score and modest chrF score. We observe similar qualitative results as our unconditional model: generated images contain excellent glyph structure in the correct language but higher-level sentence structure is not captured well. Because the BLEU score operates at the word n-gram level, our diffusion model is heavily penalized for the relative lack of phrase coherence. Our diffusion model achieves a chrF score of more than half that of our t5-small model without span corruption, despite getting a BLEU score two orders of magnitude smaller. This suggests that the model learns lower-level language structure prior to learning higher-level structure.

We note the dramatic difference in the machine translation performance of our T5 baseline model when pretrained with span corruption (Raffel et al., 2020) prior to finetuning relative to no pretraining. This suggests that the span corruption objective is a powerful technique for learning rich language representations. Adopting an analogous masking technique in our diffusion model is a promising direction.

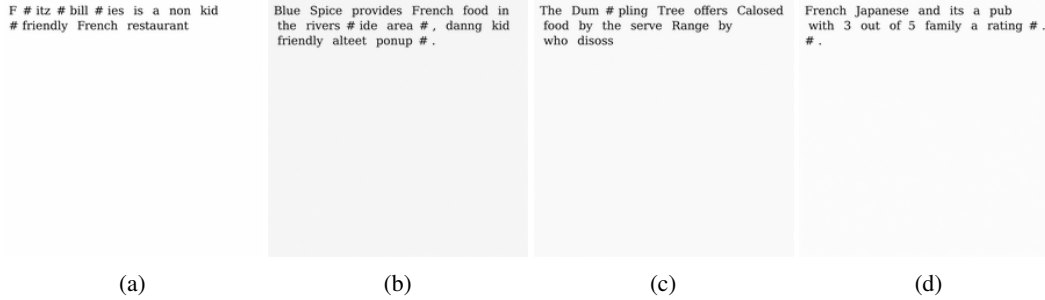


Figure 3: Unconditional generations from diffusion model and OCR-decoded + postprocessed outputs. (a) "Fitzbillies is a non kidfriendly French restaurant" (b) "Blue Spice provides French food in the riverside area, dannng kid friendly alteet ponup." (c) "The Dumpling Tree offers Calosed food by the serve Range by who disoss" (d) "French Japanese and its a pub with 3 out of 5 family a rating.."

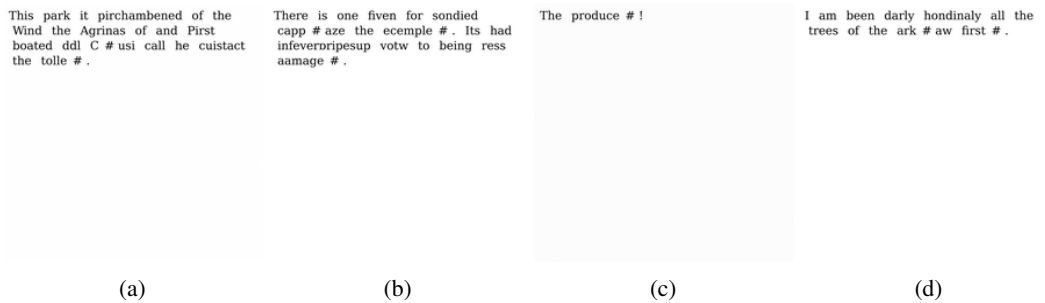


Figure 4: Translations from diffusion model and OCR-decoded + postprocessed outputs. Correct translations: (a) "This phenomenon gained momentum following the November 2010 elections, which saw 675 new Republican representatives added in 26 States." (b) "The new election laws require voters to show a photo ID card and proof of US citizenship." (c) "Diet" (d) "But my provocations are aimed at starting conversation."

6.3 Diffusion Model Scaling Analysis

We observe that the qualitative performance of our diffusion model improves significantly as the model size is scaled up. Our small diffusion model with a 31M parameter U-Net generates clear glyphs but not coherent words. As we scale to a 209M parameter U-Net and larger batch size, the quantity of coherent words generated increases dramatically and some coherent phrases emerge. This trend suggests that further model and data scaling would allow our diffusion model to better learn higher-level language structure. Note that our overfitting experiment trained with a single batch of 8 images and a 21M U-Net confirms that our model is capable of forming entirely coherent sentences. Thus, it is unlikely that there exists a fundamental limitation in diffusion models generating coherent text passages.

7 Conclusion

In this work, we cast language modeling as a vision task and propose a non-autoregressive diffusion-based decoder for text generation. We showed that our modeling approach, combined with a powerful optical-character recognition system, is successfully able to render sequences of character glyphs and produce words and sentences with some coherence. We find that our unconditional model is able to generate plausible phrases with minor spelling and syntax errors, and our conditional sequence-to-sequence model achieves some reasonable character-level n-gram translation scores. We also find that some of our quantitative metrics, such as BLEU and Im-score, lead to harsh score penalties for minor spelling and syntax issues produced by our model and the optical character recognition system, despite qualitative results indicating that our method generates plausible natural language. We also investigate scaling the model parameters and number of data samples seen during

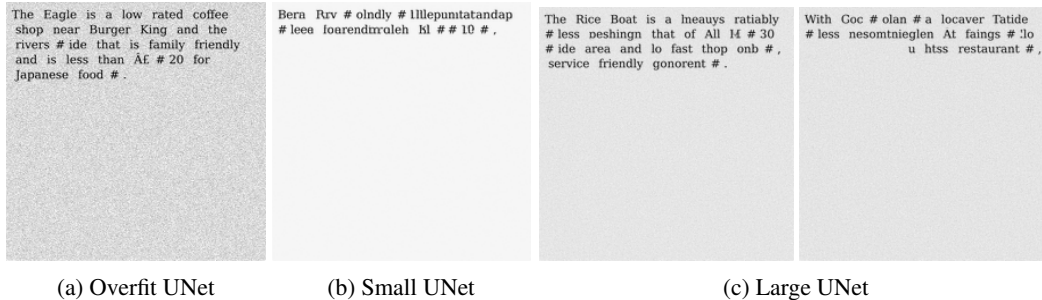


Figure 5: Qualitative results (generated unconditional samples) from our diffusion models. We performed the overfitting experiment on a single batch of the training set to verify that our small model (21M parameters) had enough capacity to model English fonts and words on the canvas. Our small model (31M) shows the ability to model characters but not English text. To improve on this we scaled both the model (209M parameters) and the number of examples seen (in the large UNet experiment run), which begins to generate more English words (e.g. "Rice", "Boat", "service", "friendly", etc.). We believe longer training would improve these qualitative results.

training and extrapolate that larger models, while out of our academic budget, may yield improved performance.

Aside from lower quality generation results compared to our strong autoregressive baselines, limitations of our work include the requirement for rendered text datasets, higher capacity models – and therefore more compute – than autoregressive baselines, and the need to generate an entire image even if very short sequences of text are being generated. These limitations are directly related to current challenges in diffusion models and are not specific to our method. We are excited by the ongoing research in the generative modeling community to overcome these challenges with diffusion models such that these improvements can be adopted into our work.

Future work, including addressing the limitations stated above, may include using methods in addition to scheduled sampling to improve the text rendering abilities of diffusion models. Recent work (Liu et al., 2022) has shown that using character-aware conditioning, such as representations learned by tokenizer-free encoders, can significantly improve visual text rendering performance. Another avenue for achieving this would be to use PIXEL (Rust et al., 2023) as the encoder directly and to make all components of the encoder-decoder architecture pixel-based, allowing the community to validate if pixels are all we need for language modeling. Our work also does not investigate using a diffusion-based decoder as a foundation model (Bommasani et al., 2021), and prescribing more advanced recipes, such as pre-training on unconditional text and fine-tuning on machine translation, as well as exploring methods to do controllable generation with classifier-free (Ho and Salimans, 2022) guidance would be fruitful next steps.

References

- Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks.
- Ondrej Bojar et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *ACL*.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshke Khani, Omar

- Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.
- 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. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.
- Dokook Choe, Rami Al-Rfou, Mandy Guo, Heeyoung Lee, and Noah Constant. 2019. Bridging the gap for tokenizer-free language models.
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. scpcanine/scp: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.
- Yuntian Deng, Noriyuki Kojima, and others. 2023. Markup-to-image diffusion models with scheduled sampling. In *ICLR*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis.
- 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. 2022. Continuous diffusion for categorical data.
- Alex Graves. 2014. Generating sequences with recurrent neural networks.
- Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2017. Non-autoregressive neural machine translation.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Neurips*.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.
- Anthony Kay. 2007. Tesseract: An open-source optical character recognition engine. *Linux J.*, 2007(159):2.
- Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. Diffwave: A versatile diffusion model for audio synthesis.
- Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement.

- Jason Lee, Raphael Shu, and Kyunghyun Cho. 2020. Iterative refinement in the continuous space for non-autoregressive neural machine translation.
- Xiang Li, John Thickstun, et al. 2022. Diffusion-lm improves controllable text generation. In *Arxiv*.
- Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. 2022. Character-aware models improve visual text rendering.
- Yinhan Liu, Myle Ott, et al. 2019. Roberta: A robustly optimized bert pretraining approach. In *Arxiv*.
- Eliya Nachmani and Shaked Dovrat. 2021. Zero-shot translation using diffusion models.
- Jekaterina Novikova et al. 2017. The e2e dataset: New challenges for end-to-end generation. In *SIGDIAL*.
- Kishore Papineni, Salim Roukos, and others. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.
- Alec Radford, Jeffrey Wu, et al. 2019b. Language models are unsupervised multitask learners. In *arxiv*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.
- Colin Raffel et al. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. In *JMLR*.
- Aditya Ramesh et al. 2022. Hierarchical text-conditional image generation with clip latents.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.
- Olaf Ronneberger et al. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.
- Phillip Rust, Jonas Lotz, and others. 2023. Language modelling with pixels. In *ICLR 2023*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Yang Song and Stefano Ermon. 2020. Generative modeling by estimating gradients of the data distribution.
- Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML’11, page 1017–1024, Madison, WI, USA. Omnipress.

- Michael Tschannen, Basil Mustafa, and Neil Houlsby. 2022. Image-and-language understanding from pixels only.
- Ashish Vaswani, Noam Shazeer, et al. 2017. Attention is all you need. In *Arxiv*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models.