A Block Metropolis-Hastings Sampler for Controllable Energy-based Text Generation

Jarad Forristal[1]
jforristal@ucsd.edu

Niloofar Mireshghallah[1]
fmireshg@ucsd.edu

Greg Durrett[2]
gdurrett@cs.utexas.edu

Taylor Berg-Kirkpatrick[1]
tberg@ucsd.edu

[1]Department of Computer Science and Engineering, The University of California San Diego
[2]Department of Computer Science, The University of Texas at Austin

Abstract

Recent work has shown that energy-based language modeling is an effective framework for controllable text generation because it enables flexible integration of arbitrary discriminators. However, because energy-based LMs are globally normalized, approximate techniques like Metropolis-Hastings (MH) are required for inference. Past work has largely explored simple proposal distributions that modify a single token at a time, like in Gibbs sampling. In this paper, we develop a novel MH sampler that, in contrast, *proposes re-writes of the entire sequence in each step* via iterative prompting of a large language model. Our new sampler (a) allows for more efficient and accurate sampling from a target distribution and (b) allows generation length to be determined through the sampling procedure rather than fixed in advance, as past work has required. We perform experiments on two controlled generation tasks, showing both downstream performance gains and more accurate target distribution sampling in comparison with single-token proposal techniques.

1 Introduction

Controllable text generation has many important downstream applications, ranging from reducing bias in generated text to increasing factuality (Xu et al.; Gehman et al., 2020; Sap et al., 2021; Baheti et al., 2021; Mireshghallah and Berg-Kirkpatrick, 2021). While traditional autoregressive language models (LMs) can produce highly fluent text, controlling their output and generating text which satisfies specific desired attributes remains a hard problem for all but the largest industrial LMs. One line of past work has made progress on controllable text generation by integrating discriminators—e.g. pretrained text classifiers that directly measure control attributes—into the scoring function for text generation (Mireshghallah et al., 2022; Yang and Klein, 2021; Dathathri et al., 2020; Krause et al., 2020). These techniques provide a flexible interface for exerting control: a user can combine discriminators and heuristic scoring functions together with likelihoods from traditional LMs to form a product of experts, guiding outputs to satisfy target criteria.

While these techniques enable effective control, they present a new challenge for decoding. The scoring functions introduced by discriminators are not autoregressive: they are global potential functions that take the entire utterance as input. This means that the overall model is not autoregressive and exact sampling is intractable. Past work has developed various heuristic or approximate decoding strategies (Dathathri et al., 2020; Krause et al., 2020; Yang and Klein, 2021; Goyal et al., 2022; Mireshghallah et al., 2022; Qin et al., 2022; Kumar et al., 2022, 2021). One of the more principled inference techniques treats the product of experts as an energy-based LM—that is, a globally normalized language model (Goyal et al., 2022; Mireshghallah et al., 2022; Qin et al., 2022; Belanger and McCallum, 2016)—and introduces a Metropolis-Hastings (MH) sampler for decoding. More specifically, Mireshghallah et al. (2022) use BERT (Devlin et al., 2019) to propose a change to a single token of the current sequence at each step of the MH chain (like a traditional Gibbs sampler) and the energy LM exerts its influence through MH's accept/reject step, correcting the bias of the proposal distribution. While principled, this approach has serious limitations. First, since only a single token can be changed at each step, inference is extremely slow. Second, since the proposal distribution does not alter the length of the current sequence, the length of the desired output must be specified in advance.

In this work, we present a novel MH sampler

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 403–413
December 6–7, 2023. ©2023 Association for Computational Linguistics

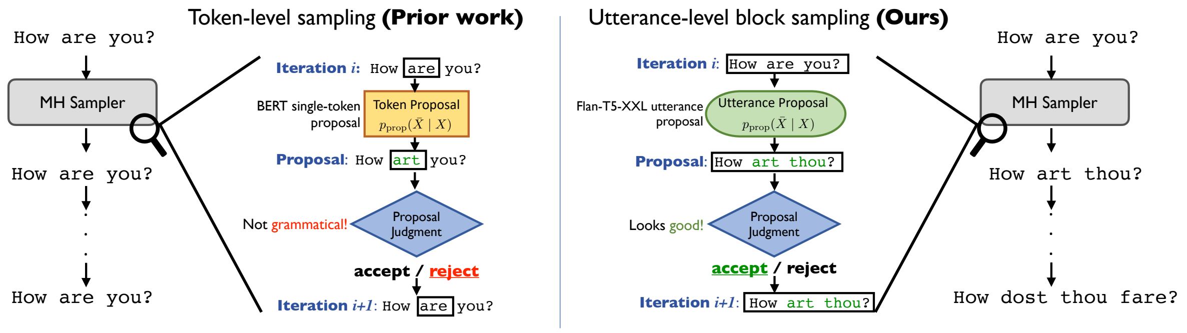

Token-level sampling (Prior work)
How are you?
MH Sampler
How are you?
How are you?
Iteration i: How are you?
BERT single-token proposal
Token Proposal
$p_{\text{prop}}(\tilde{X} \mid X)$
Proposal: How art you?
Not grammatical!
Proposal Judgment
accept / reject
Iteration i+1: How are you?
Utterance-level block sampling (Ours)
Iteration i: How are you?
Flan-T5-XXL utterance proposal
Utterance Proposal
$p_{\text{prop}}(\tilde{X} \mid X)$
Proposal: How art thou?
Looks good!
Proposal Judgment
accept / reject
Iteration i+1: How art thou?
How are you?
MH Sampler
How art thou?
How dost thou fare?

Figure 1: An overview of our novel Metropolis-Hasting (MH) sampler for energy LMs, detailing the iterative editing procedure. For our method, we prompt Flan-T5-XXL to edit the sentence in the desired way, and use this conditional distribution as the proposal for the MH chain. The MH accept/reject step corrects the bias of the proposal by considering the unnormalized energy under the target distribution. If we accept the edit, it becomes the input to Flan-T5-XXL in the next step of the Markov chain. The baseline, in contrast, only propose a change to a single token at a time.

for energy LMs that, in contrast with past work, introduces a proposal distribution that allows for arbitrary re-writes of the entire sequence at each step of the MH chain. As a result, our *block MH sampler* (a) has improved efficiency in sampling and (b) allows output length to be determined by the sampling process itself. Our key insight is to use a prompted large language model (LLM) as the proposal distribution inside of our sampler. Specifically, we prompt the LLM to paraphrase the text sequence at the current step of the MH chain, and use its output distribution as the proposal for the next step. Whether or not the proposal is *accepted* is still governed by the energy function of the target energy LM; we only change the proposal, while leaving the mathematical framework intact.

We conduct experiments on two downstream text style transfer tasks that have been used in past work as benchmarks for controllable generation (Mireshghallah et al., 2022; Krishna et al., 2020). Specifically, we study style transfer performance on two challenging datasets: the Shakespeare author imitation dataset (Xu et al., 2012) and the GYAFC formality corpus (Rao and Tetreault, 2018). Across experiments, we find that our novel sampler is able to make substantially faster progress towards high-scoring samples per forward-pass of the target energy LM in comparison with the single-token resampling MH procedure from past work. Further, for most downstream tasks, our novel sampler also leads to improvements in the output text in terms of fluency, style transfer accuracy, and semantic similarity to the desired ground truth generations.

Our contributions: (1) We propose novel block MH sampler for globally normalized energy LMs that is capable of rapid substantive edits; (2) We validate our approach on two downstream controllable generation tasks, formality transfer and author imitation, demonstrating gains in sampling efficiency as well is in output text quality; (3) We conduct an intrinsic evaluation of our sampling procedure in a synthetic setting, comparing outputs from our sampler with outputs from exact ancestral sampling.

2 Background

The M&M approach (Mireshghallah et al., 2022) defines an MCMC sampling procedure for language models that are globally normalized, which are often called energy-based LMs. Explicitly, an energy-based sequence model defines a globally normalized probability distribution over the space of possible finite-length sequences $\mathcal{X}$ as: $p(X;\theta) = \frac{e^{-E(X;\theta)}}{\sum_{X' \in \mathcal{X}} e^{-E(X';\theta)}}$, where $E(X;\theta)$ corresponds to the scalar energy of a sequence X that is judged by some model parameterized by θ. Lower energy corresponds to higher likelihood of X. Unlike popular autoregressive techniques, there is no general tractable method of sampling from energy models formulated in this way – even the likelihood function is intractable to compute due to the global normalization constant. However, their high flexibility and compatibility with black-box experts make energy models highly attractive, warranting research into this problem.

Product of Experts The constraints associated with controlled generation can be thought of as distributing probability mass over a small subspace of $\mathcal{X}$ associated with samples that satisfy the required constraints. For example, if we want to generate Shakespearean sentences, we likely want both fluent and early-modern English outputs (modeled by $p_{\text{shakespeare}}(X)$ and $p_{\text{fluent}}(X)$ respectively) – i.e., $p_{\text{desire}}(X) \propto p_{\text{shakespeare}}(X) \cdot p_{\text{fluent}}(X)$. Because it is intractable to form these probability distributions explicitly, we instead model them implicitly using unnormalized potential functions, combining them to form a scalar energy:

$$E(X) = \sum_{i=1}^{k} \alpha_i E_i(X), \tag{1}$$

where a_i are scalar weights and $E_i(X)$ are arbitrary black-box potential functions. More information regarding our use of energy models is available in Section 3 and Section 5.2.

Sampling from $\mathbf{E}(\mathbf{X};\boldsymbol{\theta})$ M&M uses a Metropolis-Hastings (MH) chain with a Gibbs-inspired proposal distribution to sample from the target energy model $E(X;\theta)$. Starting with some text, X, for each iteration M&M randomly samples the position of a single token to mask out. BERT is used to propose a new token for the masked position, editing the sentence into $\bar{X}$. This proposed edit, $\bar{X}$ is then accepted or rejected based on the conditional probability of the proposed token, likelihood of the replaced token, and the ratio of energies between $\bar{X}$ and X; the exact calculation can be seen in Equation 2. Critically, the energy model's likelihood only appears in the ratio in Equation 2 and the *intractable normalization constant cancels out*; this is one of the primary motivations for using MH in this context. The model used to estimate $p(X|\bar{X})$ and $p(\bar{X}|X)$ is called the *proposal distribution*. The stationary distribution of this Markov chain converges to $p(X;\theta)$.

3 Methodology

In this section, we will describe and motivate our approach. Similar to M&M, we frame controlled generation as a sampling problem where our goal is to get samples from a specific energy-based sequence model. However, M&M has important limitations in the sampling procedure that should be noted:

Limitations of Token-level Sampling The M&M masking process destroys important information that is often relevant to the task at hand: for example, if a name is masked out, it is unlikely to be predicted again; this means M&M can largely not restructure sentences and instead prefers minimal edits which achieve the end goal. Importantly, editing a single token at a time also significantly slows mixing. For example, if we want to make the sentence "How are you?" to be more Shakespearean, the single-token edit "How art you?" is not fluent or grammatical and is likely to be rejected, but is a necessary step to achieve the end goal of "How art thou?"; this important issue is illustrated in Figure 1. Using a block MH sampler sidesteps this issue by allowing the proposal distribution to select which parts of the sentence to edit and to propose changes to multiple tokens simultaneously.

Furthermore, M&M uses BERT to calculate $p(X|\bar{X})$ and $p(\bar{X}|X)$. Importantly, since BERT was trained on a dataset of modern English, samples from this distribution will also be. In Figure 1 BERT is unlikely to propose the token "art" in the first place, this is not a modern English token and BERT has no information about the task. Prompting an LLM with information about the task guides the model towards making more impactful changes. Finally, M&M is a fixed-length sampling method: the output is always the same length as the input. The freedom to add or delete tokens is very valuable for many downstream tasks. Our sampling procedure, detailed below, targets these weak-points and improves upon past work.

3.1 Sampling Scheme

Similar to Mireshghallah et al. (2022), we devise a Metropolis-Hastings (MH) chain that iteratively edits text in order to produce samples from the target energy model. We begin with a set seed text and progressively edit this sentence, forming a long Markov chain in the process. The acceptance or rejection of these edits is a function of both the expert blackbox models and sample probability as judged by the proposal model. Unlike previous work where the proposal function was replacement of a single token, we instead choose to prompt Flan-T5-XXL (Chung et al., 2022) to edit the sentence; this allows for arbitrary-length generation and makes our approach a block-level MH sampler (similar to blocked Gibbs sampling) as multiple variables (tokens) are updated every proposal step.

More specifically, at each step of the chain, given the current sentence X, an edited version, $\bar{X}$, is sampled from the proposal distribution, $p_{\text{prop}}(\bar{X} \mid X)$, which is defined by an instance of Flan-T5-XXL that has been prompted to generate paraphrases as depicted in Figure 3. MH then defines the probability of transitioning from X to $\bar{X}$ as:

$$p(\bar{X}; X) = \min\left(1, \frac{e^{-E(\bar{X})}\, p_{\text{prop}}(X \mid \bar{X})}{e^{-E(X)}\, p_{\text{prop}}(\bar{X} \mid X)}\right) \tag{2}$$

$E(X)$ refers to the product of experts energy defined in Equation 1 and $p_{\text{prop}}(\bar{X} \mid X)$ refers to the probability that the proposal model generates $\bar{X}$ given its prompted input is X.

Strictly speaking, to inherit the asymptotic guarantees of MH, one would need to prove, for example, detailed balance conditions for the proposal distribution. However, in practice, we found Flan-T5-XXL to have a strong propensity to generate the identity edit which causes slow mixing. To mitigate this issue in our experiments, in the numerator of Equation 2 we instead use $p_{\text{prop}}(X \mid X)$. This change makes non-identity edits more likely to be accepted if the probability of the identity is high. In practice, we found this approximate accept/reject strategy to perform well in experiments.

Thus, our block-level MH sampler implements a more freeform style of editing compared to token-level replacement used in previous work, as illustrated in Figure 1. Specifically, the block-level sampler: (1) allows the chain to preserve the content of the previous sentence more easily, as we do not mask out or destroy any information, (2) allows for coordinated edits to multiple tokens simultaneously, and (3) allows for the length of the sentence to change over the course of the sampling process.

In our implementation, we progressively edit a sentence by iteratively reprompting an LLM and accepting or rejecting these edits based on the 'quality' of the edit as judged by both the LLM itself and expert black-box models. Rather than running a single Markov chain at a time, we instead opt to run a batch of independent Markov chains with the same initial seed text, selecting a single final generation by selecting the one with minimum energy. We refer to this as "batch-size" when describing our experiments; we use batch-size 10 for all experiments unless noted otherwise. Using the methodology now defined, we can leverage the power of LLMs to sample from any arbitrary distribution that can be formulated as an unnormalized energy.

4 Intrinsic Evaluation of Sampler

In this section, we aim to conduct an intrinsic evaluation of the proposed sampler, which we refer to as MH-BLOCK, separate from the downstream controllable generation tasks we consider in Section 5. Specifically, we would like to evaluate how well MH-BLOCK approximates exact sampling from a complex target distribution relative to the baseline token-level sampling procedure, which we refer to as M&M. To accomplish this, we need to define a target energy model for which exact sampling is actually tractable so that we can draw exact samples and compare. For this purpose, we treat a prompted conditional distribution of LLaMA-7B (Touvron et al., 2023) as our target 'energy' model by setting $E(X)$ in Equation 1 to LLaMA-7B's negative log-likelihood. Specifically, we prompt LLaMA-7B to paraphrase a fixed input sentence (randomly sampled from the Shakespeare dataset mentioned in Section 5, consisting of 13 tokens) and treat the resulting conditional over text sequences as our target.

We produce 100 samples using MH-BLOCK, 100 samples using M&M, and 1000 exact samples using ancestral sampling and compare the distribution of resulting energy values under the target in Figure 2. For MH-BLOCK, we run 100 separate MH chains consisting of exactly 10 proposal steps each, and take the final step's sequence as the output sample. For M&M, the setup is the same, except that we run 130 proposal steps per chain to account for M&M's limitation to a single token change per step. This means that while MH-BLOCK only requires 10 forward passes of LLaMA-7B per sample, M&M requires 130. In Figure 2, we see that the distribution of samples from MH-BLOCK has a mean energy closer to that of the exact samples than M&M does. This indicates that even with an order of magnitude fewer forward passes in the target model, MH-BLOCK is able to produce more accurate samples than the baseline M&M.

5 Downstream Task Evaluation

Controllable generation is a relatively wide field with many tasks. We focus on one of particular importance: style transfer. Style transfer is the task of taking text written in one "style" and rewriting it in a different "style" while preserving semantic meaning or "content". For this paper, we focus on the two datasets: the Shakespearean author imitation dataset (Xu et al., 2012) which provides Shake-

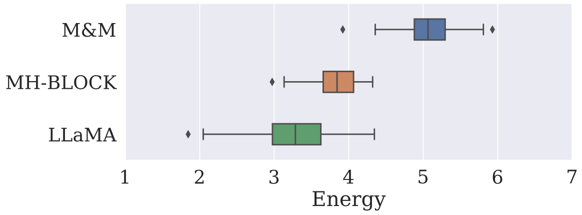


Figure 2: The energy of 100 samples from different MH samplers compared to 1000 exact samples taken from LLaMA using ancestral sampling. MH-BLOCK only requires 10 forward passes in LLaMa per sample, while M&M requires 130 in this experiment.

spearean sentences and their modern English counterparts, and the GYAFC formality corpus (Rao and Tetreault, 2018) which contains informal sentences and paired formal versions.

Data condition Following past work in style transfer, our evaluation setup relies on having parallel data of the form $(\mathbf{s}, \mathbf{t})$, where $\mathbf{s} \in \Sigma^*$ is a series of tokens taken from a vocabulary Σ in the source style and $\mathbf{t} \in \Sigma^*$ is a series of tokens in the target style. We will evaluate our models according to several criteria, some of which only evaluate $\mathbf{t}$ (e.g., fluency) and some of which evaluate $\mathbf{t}$ with respect to $\mathbf{s}$ (e.g., semantic similarity). Note that this similarity between model output and the target domain is only used during evaluation and not during model inference. For training and baseline purposes, we assume access to unpaired data belonging to s and t. That is, despite our evaluation requiring paired data, our training setup does *not*. Due to computational constraints, the Shakespeare test set was sub-sampled to 100 entries and the GYAFC corpus to 300; when evaluating MH-BLOCK, we run the Markov chain for 20 steps for the Shakespeare dataset and 10 for the GYAFC dataset. The Shakespeare dataset itself contains 31,444 entries, 29,982 of which can be used for training. The GYAFC dataset contains 112,890 entries, 105,169 of which can be used for training.

5.1 Baselines

We compare against a number of strong baselines which require similar data, namely, unpaired corpora of styles of interest. Ones of note are listed below along with relevant hyper-parameters where available.

M&M Our primary comparison is to Mireshghallah et al. (2022) (M&M), which uses a similar MH process for sampling from energy models. We use the same hyperparameters reported by the original authors of M&M on these same tasks and datasets.

VAE We also compare to a baseline method of He et al. (2020), a generative style transfer framework which uses a variational autoencoder (VAE) built using a sequence-to-sequence LSTM-based model to do unsupervised style transfer. This method needs to be trained from scratch for each dataset. We use the best reported hyperparameters in the original paper.

UNMT. UNMT (Lample et al., 2018) is an unsupervised machine translation framework which can be used effectively for unsupervised style transfer. We use the same generations that STRAP compares to (Krishna et al., 2020).

STRAP. STRAP (Krishna et al., 2020) formulates style transfer as a paraphrase generation problem, followed by "inverse"-paraphrasing to a specific style. We use the generations associated the best performing hyperparameter settings for their system, as reported by the authors.

Sample and Rerank (SAR) The baselines discussed so far from prior work use either smaller neural networks, simpler architectures, or models that are pre-trained on less data than Flan-T5-XXL. We perform an ablation to understand how much of our method's success comes from using Flan-T5-XXL in a naive way. We prompt Flan-T5-XXL, sample N generations, then rerank using the energy function provided in Equation 3 to select the best generation. For the Shakespeare dataset, we set $N = 10$, for GYAFC we set $N = 100$.

5.2 Expert Factors

As stated previously, we focus on the task of controlled text revision. We use two different expert factors to guide our approach, MH-BLOCK: a style discriminator and a measure of semantic similarity. Specifically:

$E_{\text{disc}}(X)$: This factor corresponds to the energy of the sentence as judged by a style discriminator. If we want to transfer from modern English to Shakespearean, we might set $E_{\text{disc}}(X) = -\log p(\text{Shakespearean}|X)$.

$E_{\text{BERTScore}}(X, X')$: This factor is a measure of inverse semantic similarity between two sentences, X and X', first introduced in Zhang et al. (2020).

Explicitly, the energy function for all experiments is:

$$E_{\text{rev}}(X') = \alpha E_{\text{disc}}(X') + \beta E_{\text{BERTScore}}(X, X'). \tag{3}$$

The authors of Mireshghallah et al. (2022) use a more complex energy function that additionally includes an external fluency measure; since we use an LLM as a proposal model which has a much higher rate of generating fluent text when compared to BERT, this additional expert factor was not required and nearly all generated text is fluent. The combination of these two factors allows us to specify a probability distribution $p_{\text{desire}}(X)$ from which samples satisfy our desired style *and* have high semantic similarity to the seed text.

Specifically, for $E_{\text{disc}}(X)$ on the Shakespeare dataset, we use a RoBERTa-large pretrained model finetuned on the training set of the Shakespeare dataset to discriminate between modern English and Shakespearean text (Liu et al., 2019). For GYAFC experiments, we use the publicly available Huggingface XLMR formality classifier trained on the XFORMAL dataset (Briakou et al., 2021). We approximately hand-tuned the α and β terms in Equation 3 such that the average magnitude of the terms were equal when run on the test set of the Shakespeare dataset. This amounts to $\theta = 120, \alpha = 20$ for all experiments except the GYAFC to-formal direction, where $\alpha = 40$, as with $\alpha = 20$ there was poor transfer rate.

For $E_{\text{BERTScore}}(X, X')$, we use the 18th layer of the Huggingface pretrained DeBERTa-large-mnli model to calculate a rescaled negative BERTScore (since lower energy corresponds to higher probability).[1] Our energy model uses $E_{\text{BERTScore}}(X, X')$ between the current sentence and the seed text. For evaluation only, we evaluate the BERTScore between the output and the ground truth transfer.

5.3 Evaluation Metrics

For evaluation, we use the metric proposed in Krishna et al. (2020). Explicitly, that metric is:

$$J(\text{ACC}, \text{SIM}, \text{FL}) = \sum_{x \in \mathbb{X}} \frac{\text{ACC}(x) \cdot \text{SIM}(x) \cdot \text{FL}(x)}{|\mathbb{X}|}. \tag{4}$$

Here, $x \in \mathbb{X}$ represents a sentence from the test corpus $\mathbb{X}$. This metric fairly weights accuracy (ability to match the target style), similarity (ability to preserve content), and fluency (ability to produce a fluent sentence).

Following previous work, we implement ACC and FL as binary indicators of sentence transfer as judged by a style classifier and fluency classifier, respectively. Intuitively, this corresponds to the average SIM amongst fluent and successfully style-transferred outputs, treating all other samples as having 0 similarity. For ACC, we use the discriminators detailed above. For SIM, we use the DeBERTa BERTScore detailed in Section 5.2 and calculate the semantic similarity of the generated text and the ground truth targets. For FL, following prior work, we use a RoBERTa-base classifier available on Huggingface.[2] In Tables 1-2, we refer to Equation 4 as "J-score".

5.4 Prompting

By using a large language model (Flan-T5-XXL), we avoid having to fine-tune our proposal distribution. Instead, the model is guided based on a prompt, which defines the task that it is carrying out. To prompt Flan-T5-XXL, we used prompts of the form present in Figure 3. Emphasized light blue text indicates the current text sequence in the MH chain, X. Text below the dotted line corresponds to the generated proposal, $\bar{X}$. All other text is part of the example prompt template.

While we found that Flan-T5-XXL was sensitive to the *format* of the prompt, such as the ordering of commands, the use of the language "style of William Shakespeare" and word "rewrite", it was not very sensitive to the specific example provided to the model. This is a one-shot prompt; it contains one "training example" (*There's...* → *Lo, here...*) (Brown et al., 2020). We additionally found that providing more than one example did not significantly impact performance.

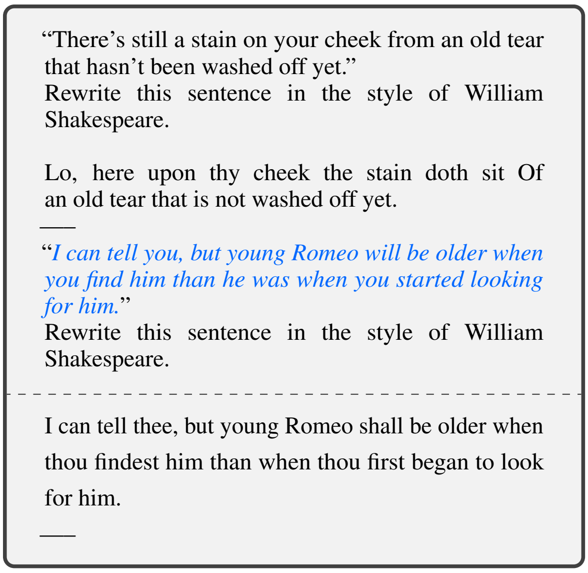
6 Style Transfer Results

In this section, we will present results of the proposed method on downstream style transfer tasks. Quantitative performance is reported in Table 1-2, with sub-tables representing specific style transfer directions.

As seen in Table 1, our approach outperforms all baselines as judged by J-score in the to-Shakespeare direction. SAR is a strong baseline in the to-modern direction, achieving similar performance with reduced implementation complexity,

[1] We use this model and this layer due to the high correlation with human judgement, details can be found online at `github.com/Tiiiger/bert_score`.

[2] `cointegrated/roberta-base-formality`



"There's still a stain on your cheek from an old tear that hasn't been washed off yet."
Rewrite this sentence in the style of William Shakespeare.

Lo, here upon thy cheek the stain doth sit Of an old tear that is not washed off yet.

—

"*I can tell you, but young Romeo will be older when you find him than he was when you started looking for him.*"
Rewrite this sentence in the style of William Shakespeare.

I can tell thee, but young Romeo shall be older when thou findest him than when thou first began to look for him.

—

Figure 3: An example of how our approach prompts Flan-T5-XXL to form a proposal distribution within our MH sampler. The displayed prompt was designed to produce a useful proposal distribution within an MH chain for the downstream task of style transfer from modern to Shakespearean English, which is one of the tasks we consider in evaluation. The blue text corresponds to X, the current sequence at a given step in the MH chain. The text below the dotted line corresponds to $\bar{X}$, the proposed edited sequence for the next step of the chain.

however with lower fluency and significantly lower transfer rate. The grounding of Flan-T5-XXL by the expert black-box models shows gains in efficacy especially when compared to prior work investigating the use of MH sampling for style transfer. Despite not having an explicit fluency measure, we see our approach has high levels of fluency in all directions.

Looking at Table 2, we once again see the strongest performance in the more difficult direction, informal to formal, achieving the highest rates of transfer and greatest similarity to ground truth text. M&M struggles with this direction, transferring only 8% of inputs, something noted by the original authors in their experiments (Mireshghallah et al., 2022). For the other direction, we beat all baselines aside from SAR, but still outperform SAR on the ACC metric; SAR is well-suited for this direction as it is very well-represented in the training data of the LLM. Analyzing both Table 1 and Table 2, we outperform past MH methods on all experiments, indicating our improved sampler performance translates to downstream tasks suc-

Model	J-score	SIM	ACC	FL
MH-BLOCK	**0.286**	**0.401**	**90.0**	84.0
M&M	0.051†	0.279	24.0	**91.0**
SAR	0.245†	0.38	78.0	79.0
STRAP	0.142†	0.333	53.0	88.0
UNMT	0.261†	0.399	85.0	81.0
VAE	0.096†	0.25	87.0	47.0

(a) Modern English → Shakespearean English.

Model	J-score	SIM	ACC	FL
MH-BLOCK	0.320	0.344	**97.0**	**94.0**
M&M	0.151†	0.343	47.0	75.0
SAR	**0.329**	**0.431**	77.0	86.0
STRAP	0.293	0.382	81.0	86.0
UNMT	0.097†	0.247	46.0	51.0
VAE	0.124†	0.293	53.0	51.0

(b) Shakespearean English → Modern English.

Table 1: Style transfer results on the Shakespeare author imitation dataset. † indicates our approach had a statistically significant performance gain as judged by a paired bootstrap test with $p = 0.05$. The best results for each column are bolded.

Model	J-score	SIM	ACC	FL
MH-BLOCK	**0.504**	**0.596**	**91.0**	91.7
M&M	0.032†	0.479	8.0	80.3
SAR	0.408†	0.505	87.7	91.0
STRAP	0.225†	0.483	46.0	**92.0**
UNMT	0.083†	0.327	41.6	61.7

(a) Informal → Formal

Model	J-score	SIM	ACC	FL
MH-BLOCK	0.382	0.477	90.7	85.0
M&M	0.266†	0.402	**95.3**	64.7
SAR	**0.385**	**0.498**	84.0	91.3
STRAP	0.325†	0.408	84.3	**94.0**
UNMT	0.132†	0.23	87.7	57.9

(b) Formal → Informal

Table 2: Style transfer results on the GYAFC formality dataset. † indicates our approach had a statistically significant performance gain as judged by a paired bootstrap test with $p = 0.05$. The best results for each column are bolded.

Input	Method	Output
My wits faints.	MH-BLOCK	I feel like my wits are fading off into the sunset
	M&M	my stomach flips.
	SAR	My heart faints.
	TGT	I'm losing this duel of wits.
Romeo, will you come to your fathers'?	MH-BLOCK	Romeo, will you please come to your father's?
	M&M	romeo, will you come to your father's?
	SAR	Romeo, will you come to your father's?
	TGT	Romeo, are you going to your father's for lunch?
A challenge, on my life.	MH-BLOCK	A challenge? I'd like a challenge on my life.
	M&M	a challenge, on my part.
	SAR	It's a challenge on my life to make you feel the same way.
	TGT	I bet it's a challenge.
Thou wouldst else have made thy tale large.	MH-BLOCK	If you'd been sensible, you wouldn't have made the tale into a huge one.
	M&M	thou wouldst else have made thy tale simpler.
	SAR	Otherwise, you would've made your tale enourmous "(meaning "enourmous"
	TGT	Oh, you're wrong.

Table 3: Example generations for multiple different methods for the to Modern English direction.

cessfully.

To qualitatively illustrate the difference between the methods, we have also included Table 3 which includes multiple input/output pairs for different methods. One detail of note is that since M&M uses BERT which cannot insert or delete tokens, the length of the output matches the input. This is particularly restrictive when the domain features source/target pairs of varying lengths. Overall, we can see the text generated by MH-BLOCK is of high quality and fluency. SAR, not being guided by measures of semantic similarity to the input, seems to deviate in meaning from the seed text more often that MH-BLOCK.

7 Related Work

Controllable generation methods that rely on energy-based constraints are the ones closest to our work (Mireshghallah et al., 2022; Qin et al., 2022; Deng et al., 2020; Parshakova et al., 2019). Mix and Match (Mireshghallah et al., 2022) in particular, is the work closest to ours. Their approach relies on single token sampling and masking, rendering the method unable to (1) change the sequence length or (2) perform block sampling of multiple tokens at the same time. Our work solves this by enabling block-sampling of multiple tokens through the use of instruction-tuned models.

There is also literature exploring free-form or constrained editing of inputs. Yasunaga and Liang (2021) follows an editing procedure, with the goal of correcting errors in incorrect code. Guu et al. (2018) uses editing of random sentences sampled from a corpus in place of autoregressive LMs to generate fluent natural language text. Mallinson et al. (2022) also uses T5 for editing, this time in a 'semi-autoregressive' manner with the goal of combining the quality of autoregressive generation and the speed of non-autoregressive methods. There are a slew of other methods related to ours, where the goal is to steer generation, without the need to re-train models from scratch. In these other approaches, however, there is often the need to use gradients or train auxiliary models to better guide the decoding. One technique guides a large model using smaller discriminator networks with the goal of sampling from an implicitly defined model, an idea explored in Plug-and-Play LM (Dathathri et al., 2020). In this approach stepwise discriminators are applied to the top-level hidden state to modify the posterior distribution formed by the LM by guiding it to fullfill the desired attributes at each autoregressive generation step by gradient ascent. Another work, FUDGE (Yang and Klein, 2021), explores a similar idea with reranking the stepwise generations, but additionally explicitly trains the future discriminators on incomplete generations.

Another set of gradient based methods (Kumar et al., 2022, 2021) view this task as optimizing the generative model's likelihood subject to global differentiable attribute-based constraints by gradient descent. There are also approaches that involve finetuning a backbone language model on domain-specific data (Ziegler et al., 2019; Keskar et al., 2019; Mai et al., 2020; Gururangan et al., 2020; Chronopoulou et al., 2021) or even training from scratch (Prabhumoye et al., 2020; He et al., 2020; Lample et al., 2018; Shen et al., 2017; Krishna et al., 2020; Reif et al., 2021; Ficler and Goldberg, 2017; Khalifa et al., 2021), to do controllable generation. Approaches specifically for style transfer have also been explored by prior work. Krishna et al. (2020) frames style transfer as a paraphrasing problem and solves it in an unsupervised way, Lample et al. (2018) has a similar methodology rooted in machine translation. He et al. (2020) attempts to model the problem using variational autoencoders. More recently, LLMs have shown strong efficacy when used for these tasks. ChatGPT and GPT3 (Brown et al., 2020) are particularly strong performers, able to solve many creative writing tasks in the zero-shot or one-shot regime (Liu et al., 2023). Flan-T5 has also shown great few-shot performance despite being less than 1/10th the size of these models (Chung et al., 2022).

8 Limitations

Our approach was designed to be as general as possible, however, it is not suitable for all settings. Our method relies on having accurate energy models that can model the desired probability distribution. In situations where no such models are available, MH-BLOCK is not particularly applicable. Additionally, it is best if the desired distribution can be easily described in text, as we must prompt an LLM to perform the task; if this is not possible, mixing could be greatly slowed and performance could suffer. However, this issue could be minimized by providing examples of the desired target style to the LLM.

9 Conclusion

While we have demonstrated empirically that our novel block MH sampler benefits controllable generation tasks by producing more accurate samples from energy-based LMs, our approach may have broader applications in other areas of NLP that use globally normalized models. Our approach highlights the utility of separating modeling concerns from inference challenges, potentially paving the way for further approaches that can use LLMs to impactfully edit text while still giving the system developer fine-grained control of the output.

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. *arXiv preprint arXiv:2108.11830*.

David Belanger and Andrew McCallum. 2016. Structured prediction energy networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 983–992, New York, New York, USA. PMLR.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexandra Chronopoulou, Matthew E Peters, and Jesse Dodge. 2021. Efficient hierarchical domain adaptation for pretrained language models. *arXiv preprint arXiv:2112.08786*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2022. Exposing the implicit energy networks behind masked language models via metropolis–hastings. In *International Conference on Learning Representations*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. A distributional approach to controlled text generation. In *International Conference on Learning Representations*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GeDi: Generative Discriminator Guided Sequence Generation. *arXiv preprint arXiv:2009.06367*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models.

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

Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A. Smith, and James Henderson. 2020. Plug and play autoencoders for conditional text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6076–6092, Online. Association for Computational Linguistics.

Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. EdiT5: Semi-autoregressive text editing with t5 warm-start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2126–2138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick. 2021. Style pooling: Automatic text style obfuscation for improved classification fairness. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2009–2022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.

Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. 2019. Global autoregressive models for data-efficient sequence learning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 900–909, Hong

Kong, China. Association for Computational Linguistics.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551.

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

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6833–6844.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, Dan Klein, and UC Berkeley. Detoxifying language models risks marginalizing minority voices.

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

Kevin Yang and Dan Klein. 2021. 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. Association for Computational Linguistics.

Michihiro Yasunaga and Percy Liang. 2021. Break-it-fix-it: Unsupervised learning for program repair. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11941–11952. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.