# Conditional [MASK] Discrete Diffusion Language Model

**Hyukhun Koh**[1]* **Minha Jhang**[1]* **Dohyung Kim**[2]
**Sangmook Lee**[2] **Kyomin Jung**[1,2]†
[1]IPAI, Seoul National University
[2]Dept. of ECE, Seoul National University
{hyukhunkoh-ai, jminha2014, kimdohyung, helmsman}@snu.ac.kr

## Abstract

Although auto-regressive models excel in natural language processing, they often struggle to generate diverse text and provide limited controllability. Non-auto-regressive methods could be an alternative but often produce degenerate outputs and exhibit shortcomings in conditional generation. To address these challenges, we propose *Diffusion-EAGS*, a novel framework that integrates conditional masked language models into diffusion language models through the theoretical lens of a *conditional Markov Random Field*. In doing so, we propose *entropy-adaptive Gibbs sampling* and *entropy-based noise scheduling* to counterbalance each model's shortcomings. Experimental results show that *Diffusion-EAGS* outperforms baselines and achieves the best quality-diversity tradeoff, demonstrating its effectiveness in non-autoregressive text generation.

## 1 Introduction

Auto-Regressive Models (ARMs) have driven significant advances in NLP (Achiam et al., 2023; Dubey et al., 2024; Team et al., 2023), yet they still have fundamental challenges such as diversity and controllability, due to the ARM's innate left-to-right inductive bias. Specifically, as ARMs often rely on the first few initial tokens, a phenomenon known as *attention sink* (Gu et al., 2025), they struggle to correct past errors in safety (Qi et al., 2025), dialogue (Laban et al., 2025), and math (Wang et al., 2025). In addition, they cannot effectively foster diversity through temperature-based sampling alone (Lee et al., 2025), nor can they anticipate future requirements at earlier steps, thus undermining controllability when external information is provided later (Lu et al., 2022; Hudecek and Dusek, 2023; Sun et al., 2023; Su et al., 2024).

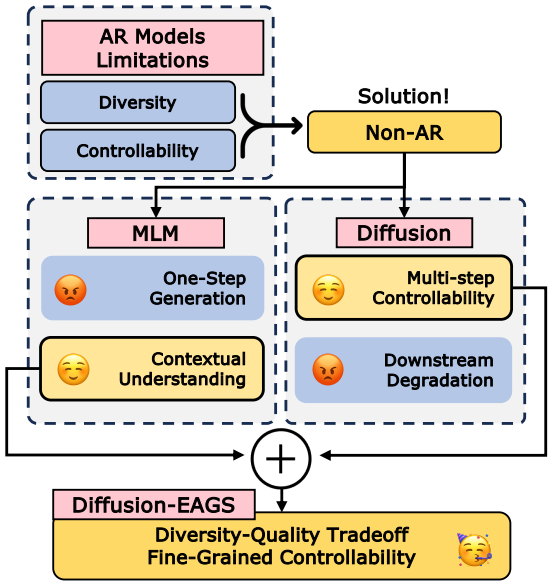


Figure 1: Overview of how our approach (Diffusion-EAGS) combines the strengths of MLM and diffusion-based models to overcome the limitations of AR models, achieving a better diversity-quality tradeoff and fine-grained controllability

One promising alternative is non-autoregressive generation, which includes conditional masked language models (CMLMs) (Ghazvininejad et al., 2019a; Kasai et al., 2020) and diffusion models. CMLMs provide strong contextual understanding but lack an effective text generation mechanism. Meanwhile, diffusion models iteratively refine text through denoising, enabling fine-grained control and increased diversity, but recent works, such as direct diffusion-based generation (Li et al., 2022; Gat et al., 2024; The et al., 2024; Ye et al., 2025) or hybrid approaches combining diffusion with PLMs and LLMs (Lin et al., 2023; Xiang et al., 2024), suffer from degeneration (Xu et al., 2025) and output homogeneity in conditional generation tasks, as confirmed by our experiments.

We therefore propose **Diffusion-EAGS**, a novel approach that integrates CMLMs into the discrete diffusion language models (DDLMs) to achieve

*Equal contribution.
†Corresponding author.



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8899–8923
November 4-9, 2025 

diverse, controllable, and high-quality conditional generation. However, merging these methods is challenging because CMLMs generate text in one step by predicting all masked tokens, whereas diffusion models iteratively refine representations over multiple steps by introducing and removing noise. Our approach bridges this gap by leveraging a conditional Markov Random Field (cMRF) formulation, which enables:

1. **Stepwise iterative generation**, overcoming the single-step limitations of CMLMs.
2. **Stable and diverse conditional text generation**, reducing semantic drift in DDLMs.

Diffusion-EAGS achieves this through two key methodologies:

- **Entropy-Adaptive Gibbs Sampling (EAGS):** A strategy that updates the most uncertain (high-entropy) tokens first at each denoising step, ensuring qualified generation.
- **Entropy-based Noise Scheduling (ENS):** A training approach that progressively masks tokens based on ascending order of entropy, enabling the model to learn a structured denoising process.

We conduct extensive experiments to validate Diffusion-EAGS on various conditional generation tasks, demonstrating significant improvements over baseline models. We further show that, without our method, naively integrating a pre-trained model into diffusion models results in degraded performance, highlighting that solely relying on pre-training does not effectively improve performance. Our approach achieves the best quality-diversity tradeoff, demonstrating that Diffusion-EAGS balances fluency and variability more effectively than existing models. Moreover, keyword-based story generation experiments confirm that our model effectively generates coherent and controlled text from randomly masked sequences, making it highly adaptable to different conditioning constraints.

## 2 Related Works

Efforts to integrate generative flow models into sequence generation exploit the distribution shift from a source language to a target language through a series of invertible linear transformations (Ma et al., 2019; Zhang et al., 2024). However, as DDPM (Ho et al., 2020a) demonstrates the effectiveness of generating images, diffusion models have been a major topic of interest within the field of generative flow models (Song et al., 2021a,b). To apply such diffusion methodologies to NLP, to leverage their strengths in controllability and diversity, recent studies have demonstrated promising performance across various tasks (Li et al., 2022; Gong et al., 2023a; He et al., 2023; Yuan et al., 2023; Lovelace et al., 2023; Chen et al., 2023; He et al., 2023; Lou et al., 2024; Zhou et al., 2024; Shi et al., 2024; Sahoo et al., 2024a; Zheng et al., 2024; The et al., 2024; Wang et al., 2024).

Although Continuous Diffusion Language Models (CDLMs) such as Diffusion-LM (Li et al., 2022), DiffuSeq-v1, v2 (Gong et al., 2023a,b), and LD4LG (Lovelace et al., 2023) show promising performance, Bansal et al. (2022) argues that such operations do not necessarily have to be governed by stochastic randomness.

Building on this rationale, D3PM (Austin et al., 2023) proposes the discrete restoration-generation approach, and DiffusionBERT (He et al., 2022) adopts pre-trained language models (PLMs) to DDLM. SEDD (Lou et al., 2024) proposes score entropy inspired by MLM loss, and outperforms existing CDLMs. Recent works by Shi et al. (2024) and Sahoo et al. (2024a) extend this idea and obtain better empirical results. Zheng et al. (2024) further enhances discrete diffusion models by correcting the numerical precision error in SEDD-based models. This research makes an improvement on the open-ended generation task. Furthermore, Venkatraman et al. (2024) uses SEDD as text infilling, and Nie et al. (2024) demonstrates that DDLMs are scalable.

## 3 MLM & DDLM : D-cMRF

Pre-trained MLMs offer rich, context-aware representations through one-pass masked prediction, whereas DDLMs iteratively refine text via stepwise denoising to enhance control and diversity. Combining these approaches can overcome MLMs' one-pass limitations and DDLMs' degeneration in conditional generation. However, their integration is challenging because DDLMs require iterative updates while MLMs predict all masked tokens simultaneously. To bridge this gap, we propose Diffusion-based Constrained Markov Random Fields (D-cMRF), a framework that integrates a discrete diffusion process into MLM sequence generation. By leveraging an entropy-based sampling strategy to selectively update high-uncertainty tokens at each step, D-cMRF achieves a principled

reduction in sequence energy, leading to stable and coherent generation.

### 3.1 MLM as cMRF

Inspired by the traditional approaches of Wang and Cho (2019) and Goyal et al. (2022), which model MLMs as Markov Random Fields (MRFs) and energy-based models (EBMs), respectively, we reinterpret MLM as a conditional MRF (cMRF) model and employ it as a denoising function at each diffusion step.

Let $X = (x_1, x_2, \ldots, x_L)$ be a sequence of discrete variables from a vocabulary $V$, with $Y$ representing observed conditions. The sequence probability follows an energy-based MRF formulation:

$$P_\theta(X;Y) = \frac{\exp(-E_\theta(X;Y))}{Z(Y,\theta)} \tag{1}$$

where $E_\theta(X;Y)$ is the **energy function** parameterized using MLM-based embeddings, $\theta$ denotes parameterization of MLM, and $Z(Y,\theta)$ is the **partition function** for ensuring proper normalization. Then the total sequence energy is defined as:

$$E_\theta(X;Y) = -\sum_{l=1}^{L} \log \phi_l(X;Y) \tag{2}$$

where **log-potential function** $\log \phi_l(X;Y)$ is :

$$\log \phi_l(X;Y) = 1h(x_l)^T f_\theta(X \backslash \{x_l\}; Y) \tag{3}$$

where $l$ is a token position in the sequence, $1h(x_l)$ is the one-hot encoding of token $x_l$, and $f_\theta(X\backslash\{x_l\};Y)$ represents the MLM logit output conditioned on the sequence.

### 3.2 DDLM with Entropy-based Denoising

Determining how to perform sampling with such a simple cMRF presents a separate challenge. In particular, one can use techniques such as Gibbs sampling as long as the energy space remains unchanged, but we cannot guarantee that this energy space is stable in general (Goyal et al., 2022). The necessity of generating sequences in cMRF based on energy update is in Appendix A. Hence, a natural research question arises: *"How should we sample and update the energy?"*

The training process of diffusion models (both forward and backward) conceptually represents the entire distribution as a product of local conditional distributions across time steps. Hence, diffusion models share a probabilistic graphical structure with MRF, enabling MLM to be integrated within the DDLM framework.

Therefore, in this subsection, we describe how to update the energy and perform sampling under the DDLM framework using $P_\theta(X;Y)$. Specifically, we integrate $P_\theta(X;Y)$ into each diffusion step as a denoising function, employing an entropy-based denoising matrix $Q$ in Section 4.2. We first define the entropy of each token:

$$H_i(X^{(t)}) = -\sum_{x' \in V} p_\theta(x'_i; X^{(t)}) \log p_\theta(x'_i; X^{(t)}) \tag{4}$$

where $p_\theta(x'_i; X^{(t)})$ is the softmax probability of token $x'_i$ at position $i$ in sequence $X^{(t)}$, and $t$ denotes the diffusion timestep. We then select high-entropy positions for updating:

$$M_t = \{i \mid H_i(X^{(t)}) \geq \tau_t\} \tag{5}$$

where $\tau_t$ is a dynamically adjusted entropy threshold. This ensures that updates occur at positions where the model has the highest uncertainty. Subsequently, we sample the next-step sequence from $P_\theta(X^{(t)};Y)$ at the suggested positions. We perform this selection process at every diffusion step, which corresponds to updating the energy, different from existing research (Wang and Cho, 2019; Goyal et al., 2022).

### 3.3 D-cMRF

By combining DDLM with cMRF, our approach enables a theoretically grounded generation process from the perspective of MLM. Moreover, from the diffusion standpoint, the training process naturally aligns with the MLM objective, as discussed in Section 3.1 and Section 3.2. Specifically, our D-cMRF guarantees energy reduction during generation, ensuring stable sequence reconstruction.

**Step 1: Expected Energy at Diffusion Step $t$** At diffusion step $t$, we compute the expected sequence energy as:

$$\mathbb{E}_{X^{(t)} \sim q}\left[E_\theta(X^{(t)};Y)\right] = \sum_{X^{(t)}} q(X^{(t)}) E_\theta(X^{(t)};Y) \tag{6}$$

where $q(\cdot)$ denotes the probability distribution from which $X^{(t)}$ is sampled. Since high-entropy tokens are selected for replacement, the total sequence energy can be decomposed as follows:

$$\begin{aligned}\mathbb{E}\big[E_\theta(X^{(t)};Y)\big] = &\sum_{i \in M_t} \mathbb{E}\big[E_\theta(x_i^{(t)}; X^{(t-1)}, Y)\big] \\ &+ \sum_{i \notin M_t} E_\theta\big(x_i^{(t-1)}; Y\big).\end{aligned} \tag{7}$$

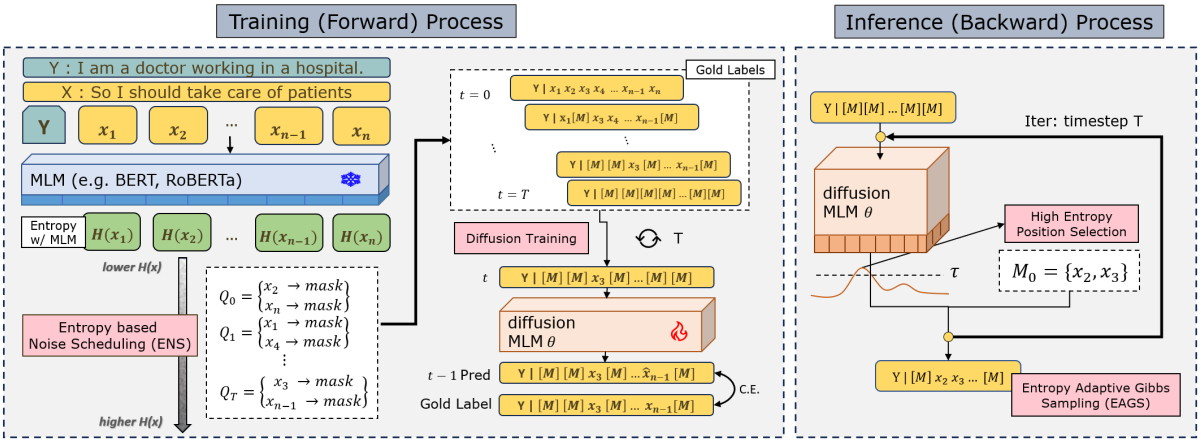


Figure 2: Overview of the training (forward) and inference (backward) processes in Diffusion-EAGS. **Training (left):** Entropy-based Noise Scheduling (ENS) determines which tokens in the masked sequence, denoted by $[M]$, should be denoised at each timestep based on the position entropy $H(x_i)$. These tokens are then generated using the diffusion model with parameters $\theta$, and the loss is computed using a cross-entropy (C.E.) diffusion loss. **Inference (right):** Starting from a fully masked sequence conditioned on $Y$, Entropy-Adaptive Gibbs Sampling (EAGS) iteratively refines the sequence by focusing on high-entropy tokens, denoted as $M_t$, based on a threshold $\tau_t$, yielding stable and coherent text generation.

**Step 2: Energy Reduction via Denoising** Since masked tokens are replaced with lower-energy candidates at each step, we expect a general trend of **energy reduction**. However, due to the stochastic nature of sampling, local fluctuations in energy may occur. Over multiple diffusion steps, the entropy-based selection mechanism ensures a net decrease in sequence energy.

$$\mathbb{E}\left[E_\theta(x_i^{(t)}; X^{(t-1)}, Y)\right] \leq E_\theta(x_i^{(t)}; X^{(t)}, Y) \quad (8)$$

Applying this property across all updated tokens $i \in M_t$, we obtain:

$$E_\theta(X^{(t-1)}; Y) \leq E_\theta(X^{(t)}; Y) \quad (9)$$

**Step 3: Convergence to Low-Energy States** Summing over all diffusion steps $T$:

$$E_\theta(X^{(0)}; Y) \leq E_\theta(X^{(T)}; Y) \quad (10)$$

where $X^{(T)}$ is the fully masked sequence with maximum entropy, and $X^{(0)}$ is the final reconstructed sequence. Since the token space is discrete and energy is derived from a sum of bounded logits, $E_\theta(X; Y)$ is **lower-bounded** by a finite minimum energy state. While stochastic sampling may introduce fluctuations, the diffusion process ensures **progressive energy minimization**, leading to an approximate low-energy state.

#### 3.3.1 D-cMRF Guarantees

So far, more detailed explanations of D-cMRF are in Appendix M. The proof establishes that our method satisfies the following properties:

- **Progressive Energy Reduction:** The energy function exhibits an overall decrease, leading to more stable sequence generation. This trend is supported by empirical results in Appendix D.
- **Stable Convergence:** Since the energy function is lower-bounded and the sequence length is finite, the generation process is expected to reach a structured, low-entropy state.

These properties explain the improved performance of Diffusion-EAGS compared to traditional diffusion models, as shown in §Section 6. Notably, the ablation study in Table 5 demonstrates that removing EAGS leads to a significant drop in performance, highlighting its importance in guiding stable generation.

## 4 Diffusion-EAGS

Our approach, Diffusion-EAGS, leverages two key components—Entropy-Adaptive Gibbs Sampling (EAGS) and Entropy-based Noise Scheduling (ENS)—rooted in the theory of Section 3. As shown in Figure 2, during training, ENS selectively masks tokens based on their certainty, while during generation, EAGS iteratively refines a fully masked sequence by updating high-uncertainty tokens. This stepwise refinement yields balanced improvements in text quality and diversity.

### 4.1 Inference Process: Entropy-Adaptive Gibbs Sampling

As discussed in Section 3.2, MLM can be interpreted as cMRF, which is used as $p_\theta$ in the sam-

pling process of DDLM with $M_t$. In particular, $M_t$ is not only associated with energy updates but also serves as a solution to the MLM's difficulty in selecting the next tokens to denoise, as shown in Appendix C. Henceforth, we designate this strategy as *Entropy-Adaptive Gibbs Sampling (EAGS)*.

In EAGS, $M_t$ is constructed by ranking tokens in descending order of entropy, thereby prioritizing the least informative parts of the sequence. EAGS facilitates the creation of more structured sequences by leveraging the syntactic context that has already been established. The process of determining the denoising schedule is shown in Appendix C.

Our approach for the T-step generation process can be formalized as follows:

1. **Entropy Calculation**: Compute the entropy $H_i(X^{(t)})$ for each variable $x_i$.
2. **Variable Selection**: Obtain $M_t$ for sampling
3. **Sampling**: Sample $x_{i^*}$ from its conditional distribution $p_\theta(x_{i^*} \mid X^{(t)}, Y)$, where $i^* \in M_t$.
4. **Update**: Update the conditional distributions and entropy for the affected variables.
5. **Iteration**: Repeat Steps 1 through 4 until $t = T$, where $T$ is the total number of timestep.

The detailed algorithm of EAGS is in Appendix Algorithm 1.

### 4.2 Training Process: Entropy-based Noise Scheduling

To improve the effectiveness of EAGS during generation, we simulate a similar process during training. Therefore, we schedule the forward process of diffusion training based on the entropy $H_i(X^{(t)})$ of position $x_i$ with the input sequence $[Y|X^{(t)}]$ at sampled timestep $t$. During training, $H_i(X^{(t)})$ is calculated by pre-trained MLM. Assuming the diffusion process progresses over $T$ steps, we mask the $L/T$ number of positions with the lowest entropy from the set $\{x_1, \ldots, x_L\}$ at each step $t$, where $L$ is the sequence length. This selection process is used to determine $\tau_t$ in Equation 5. The masking process at step $t$ in position $i$ is described by the denoising matrix $Q_{ti}$.

$$Q_{ti} = \begin{bmatrix} q_{11} & 0 & \cdots & 0 & q_{1,M} \\ 0 & q_{22} & \cdots & 0 & q_{2,M} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & q_{M-1,M-1} & q_{M-1,M} \\ 0 & 0 & \cdots & 0 & q_{MM} \end{bmatrix}$$

Here, $q_{1,M}$ denotes the transition probability from the vocab index corresponding to token 1 to the [MASK] token, and $q_{mn}$ is defined as:

$$q_{mn} = \begin{cases} q_{mm} = 1 & \text{if } x_i \notin \text{MIN}([H_1(x_1), \cdots, H_L(x_L)], \frac{L}{T}) \\ q_{mM} = 1 & \text{if } x_i \in \text{MIN}([H_1(x_1), \cdots, H_L(x_L)], \frac{L}{T}) \\ 0 & \text{otherwise} \end{cases}$$

Henceforth, we designate this strategy as *Entropy-based Noise Sampling (ENS)*. ENS masks lower entropy tokens first, thereby learning to progressively generate sequences. This ensures that the forward process in diffusion training closely mirrors the generation process, thereby enhancing the effectiveness of EAGS in language generation. The detailed algorithm of ENS is in Appendix Algorithm 2.

### 4.3 Diffusion Loss with Cross Entropy

Distinct from the prevailing methodologies in diffusion models (Ho et al., 2020a; Austin et al., 2023), we do not employ the PLM parameterization $\widetilde{p}_\theta(\widetilde{z}_0|z_t, t)$, which preserves the original semantic embedding spaces during the training phase as we empirically find that such method restricts the diversity of generated responses. We follow the traditional diffusion loss (Ho et al., 2020b), changing Mean Squared Error with Cross Entropy Loss.

## 5 Experiments

### 5.1 Tasks & Details

We conduct experiments on various conditional generation datasets. Detailed explanations of the conditional generation datasets are in Appendix F.1. In particular, we focus on two datasets that significantly differ in their level of conditional constraints: RocStories (Mostafazadeh et al., 2016), which is relatively open-ended, and Paradetox (Logacheva et al., 2022), which imposes the strongest conditional constraints. We select the conditional dataset that GPT-2 faces in generating sentences of appropriate length under specified conditional constraints. The maximum lengths of Paradetox and RocStories are set to 64, based on data statistics, and other details are in Appendix F. We test 20 conditions with 5 outputs in total, 100, which is not used for training. The number of steps of our model is configured to 5 with a naive categorical sampling with a sample size of 20, and selects the final 5 samples based on the Perplexity score. We use an A100 GPU with a batch size of 256.

### 5.2 Baselines

We employ RoBERTa-base (Liu et al., 2020) as MLM with learning rate 5e-4. Next, we compare Diffusion-EAGS with four categories of baselines of similar size to *RoBERTa-base*: Autoregressive Models (ARMs), Conditional Masked Language Models (CMLMs), Continuous Diffusion Language Models (CDLMs), and Discrete Diffusion Language Models (DDLMs). Note that our primary goal is to investigate the architecture's capabilities; any baseline approach in the direction of scalability or bypassing the architecture's limitations goes beyond our research scope.

For **ARMs** (Vaswani et al., 2023), we employ GPT-2 (Radford et al., 2019) and GPT-3.5-turbogpt-3.5-turbo* with four-shot prompt. More experimental details of GPT-3.5 can be found in Appendix J. For **CMLMs**, we utilize CMLM-Mask-Predict (Ghazvininejad et al., 2019a) and DisCo-Easy-First (Kasai et al., 2020), which are transformer-based NAR models. For **CDLMs**, our baseline includes DiffuSeq (Gong et al., 2023a), LD4LG (Lovelace et al., 2023), and DINOISER (Ye et al., 2024). DiffuSeq and DINOISER are designed for sequence-to-sequence applications, and LD4LG adopts pre-trained *BART* as a denoising init point. For **DDLMs**, we utilize DiffusionBERT (He et al., 2022), applying pre-trained *BERT* into DDLMs, AR-Diffusion (Wu et al., 2023), and SEDD (Lou et al., 2024), using the pre-trained version and fine-tune it. More details are in Appendix F.3 and more diverse baselines such as GENIE (Lin et al., 2023) and MDLM (Sahoo et al., 2024b) are in Appendix G.2.

### 5.3 Metrics

**Quality metrics :** In addition to our theoretically guided methods, we evaluate performance using multiple metrics. Specifically, we use Perplexity (PPL) based-on GPT-2 Large and GPT-2 XL as an automated metric, MAUVE (Pillutla et al., 2021) to assess style consistency between the training data and generated outputs, SOME (Yoshimura et al., 2020) to score the grammar, Mean Opinion Score (MOS) from human evaluations to gauge text quality, and LLM score such as LLM-c (Lin and Chen, 2023) to measure the plausibility of the narratives as a sub-metric.

**Diversity Metrics :** Following our quality assessment, we evaluate diversity through three different measures: an automatic frequency-based metric n-gram Vendi Score(VS n-gram) (Friedman and Dieng, 2023), a neural network–based semantic metric SimCSE Vendi Score (VS emb), and a human evaluation score MOS. The detailed descriptions of metrics are provided in Appendix F.3 and H.1.

*https://platform.openai.com/docs/models/gpt-3-5

| Model | Step | Text Quality | | |
|---|---|---|---|---|
| | | PPL ↓ | MAUVE ↑ | MOS ↑ |
| *AR model* | | | | |
| GPT-2 | 1 | 389.1 | 0.503 | 0.83 |
| GPT-3.5 w/ 4-shot | 1 | 104.375 | 0.175 | 1 |
| *CMLMs* | | | | |
| CMLM w/ Mask-Predict | 10 | 669.9 | 0.0234 | - |
| DisCo w/ Easy-First | 10 | 716.1 | 0.0344 | - |
| *Diffusion models* | | | | |
| DiffusionBERT | 2000 | 775.9 | 0.737 | 0.88 |
| AR-Diffusion | 20 | $\geq 1k$ | 0.768 | - |
| DiffuSeq | 2000 | $\geq 1k$ | 0.683 | - |
| SEDD | 1024 | $\geq 1k$ | NA | - |
| LD4LG | 2000 | 579.9 | 0.556 | 0.91 |
| DINOISER | 20 | 124.8 | 0.255 | 0.91 |
| **Diffusion-EAGS** | 5 | 109.3 | 0.811 | 0.97 |

Table 1: Text quality of conditional generation outputs. We report Perplexity (PPL) for sentence fluency, MAUVE for condition alignment, and Mean Opinion Score (MOS) for semantic coherence. Models with PPL exceeding 600 were excluded from human evaluation.

| Model | Text Quality | | | Diversity | |
|---|---|---|---|---|---|
| | PPL ↓ | SOME ↑ | LLM-c ↑ | VS(ngram) ↑ | self-bleu ↓ |
| Original Data | 100.6 | 0.895 | 1 | | |
| GPT-2 | 88.5 | **0.856** | **0.88** | 4.722 | 0.124 |
| DiffusionBERT | 318.2 | 0.783 | 0.72 | 4.735 | 0.088 |
| SEDD | 273.2 | 0.827 | 0.59 | **4.859** | **0.044** |
| **Diffusion-EAGS** | **67.3** | 0.844 | 0.87 | 4.837 | 0.058 |

Table 2: Results on the open-ended RocStories (ROC) dataset. We report perplexity (PPL) for fluency, SOME and LLM-c for text quality, and both VS(*ngram*) and self-BLEU for diversity.

## 6 Results

In Tables 1, 2, and 3, our model consistently demonstrates strong text quality and diversity compared to various baselines across a wide range of conditional generation tasks.

**Text Quality :** Table 1 shows that our model achieves notable improvements in perplexity (PPL) and obtains high MAUVE and MOS scores, indicating that the generated texts are both fluent and coherent. Although GPT-3.5-turbo is capable of generating high-quality text, the MAUVE metric indicates that few-shot prompts alone are insufficient for accurately replicating the dataset's inherent characteristics. On the other hand, CMLMs,

| Model | Step | Diversity | | |
|---|---|---|---|---|
| | | VS(ngram) ↑ | VS(emb) ↑ | MOS ↑ |
| *AR model* | | | | |
| GPT-2 | 1 | 3.925 | 2.640 | 2.65 |
| GPT-3.5 w/ 4-shot | 1 | 3.098 | 1.915 | 2.2 |
| *CMLMs* | | | | |
| CMLM w/ Mask-Predict | 10 | 1.000 | 1.000 | - |
| DisCo w/ Easy-First | 10 | 1.000 | 1.000 | - |
| *Diffusion models* | | | | |
| DiffusionBERT | 2000 | 3.101 | 2.058 | 2 |
| AR-Diffusion | 20 | 3.101 | 2.088 | - |
| DiffuSeq | 2000 | 2.059 | 1.465 | - |
| SEDD | 1024 | 4.746 | 4.063 | - |
| LD4LG | 2000 | 1.914 | 1.425 | 1 |
| DINOISER | 20 | 2.287 | 2.174 | 1 |
| **Diffusion-EAGS** | 5 | **4.417** | **3.311** | **4.6** |

Table 3: Diversity evaluation for generated outputs. We report the n-gram-based Vendi Score (VS(*ngram*)), the embedding-based Vendi Score (VS(*emb*)), and a Mean Opinion Score (MOS) for diversity. Higher values indicate greater diversity.

DiffuSeq, and DINOISER can handle conditional constraints but sometimes struggle with semantic drift or high PPL. In contrast, Diffusion-EAGS achieves both lower PPL and strong human evaluation scores (MOS), suggesting that it effectively balances condition satisfaction with text quality. Table 2 further demonstrates our model's capability on the open-ended RocStories dataset. Even with minimal constraints, Diffusion-EAGS maintains competitive scores compared to GPT-2, demonstrating its robustness in narrative generation. **Diversity :** Diffusion-EAGS excels at generating diverse outputs. As illustrated in Table 3, our model consistently excels in both n-gram and embedding-based diversity metrics (VS(ngram) and VS(emb)), surpassing other baselines and even larger LLMs. The model's higher MOS for diversity further indicates that humans also perceive its outputs to be more varied and engaging. In line with these observations, we conduct additional analyses (Appendix G.4) including the comparison of ours with large LLMs, where our approach produces a wider range of coherent yet distinct responses. These findings underscore the effectiveness of our entropy-adaptive sampling strategy in avoiding repetitive outputs and semantic collapse, thereby delivering a superior quality-diversity trade-off.

Overall, **Diffusion-EAGS** consistently demonstrates *strong performance across diverse conditional generation tasks*, combining low perplexity and high human evaluation scores with the ability to generate richly varied text. Detailed results are in Appendix G and examples are in Appendix I.

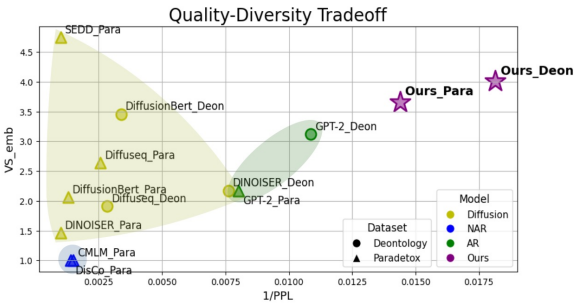


Figure 3: Quality–diversity tradeoff across various models. The x-axis ($1/\text{PPL}$) reflects generation quality, while the y-axis ($\text{VS}_{\text{emb}}$) indicates diversity. Green points represent AR models, yellow points represent diffusion models, and blue points represent CMLMs. Our Diffusion-EAGS variants, marked by purple stars, achieve the best overall tradeoff.

## 7 Analysis

### 7.1 Quality-Diversity Tradeoff

Balancing *quality* and *diversity* is a fundamental challenge in text generation. AR models typically achieve high fluency but suffer from low diversity, while non-autoregressive models, such as CMLMs and diffusion models, often struggle to generate coherent outputs. Our proposed **Diffusion-EAGS** effectively balances these factors by leveraging a structured diffusion process.

Figure 3 presents the quality-diversity tradeoff among various models, where *quality* is measured using perplexity (PPL) on the x-axis (inverted as $1/\text{PPL}$ for better visualization) and *diversity* is quantified using VS_emb on the y-axis. Our model (**Ours_Deon, Ours_Para**, marked with purple stars) achieves the best tradeoff, outperforming prior approaches in both high-quality generation and diversity. Compared to Diffuseq, DiffusionBERT, and CMLMs, our method achieves significantly better diversity without compromising generation fluency. This improvement stems from our **Entropy-Adaptive Gibbs Sampling (EAGS)**, which ensures controlled token selection, and **Entropy-based Noise Scheduling (ENS)**, which stabilizes the generation process. The results highlight that integrating MLMs into the diffusion framework enables high-quality, diverse, and controllable text generation.

### 7.2 Keyword Based Generation

Our model operating within a discrete space enables us to manipulate the output sequences using

| | | |
|---|---|---|
| *Context* | | *Jake was playing with his toys. He accidentally broke his favorite one. He cried a lot over it. His parents decided to replace it for him.* |
| Keyword | **not stop** | Jake just could **not stop** crying. |
| | **Jake feel** | It made **Jake feel** So much better. |
| | **would enjoy** | Jake said he **would enjoy** the new toy |
| *Context* | | *Neil was in Sofia, Bulgaria. He was enjoying a trip backpacking through Europe. ... He thought the food and culture in Sofia were the best.* |
| Keyword | **Bulgaria!** | Things were looking great in **Bulgaria!** |
| *Context* | | *Karen wanted to go on a trip to France. She started doing research on the trip. She decided to book a week long trip. She left the next day for her trip.* |
| Keyword | **her trip** | She spent almost a week there during **her trip**. |

Table 4: Examples of keyword-based generation. Each row shows a *Context* and a specified *Keyword*, which is inserted into a masked position. The resulting outputs demonstrate how our model seamlessly integrates keywords into coherent narratives.

explicit instructions. To further explore this capability, we conduct the generation of sequences based on keywords positioned in the middle and at the end of masked sequences, which is challenging for AR models (Keskar et al., 2019). They inherently struggle with controllability due to their inability to revise past steps based on future ones—an inductive bias of AR models. Initially, we provide the same contextual input while varying the keywords. In the masked states, we randomly select positions, replacing them with the specified keywords. The results in Table 4 demonstrate that the generated sequences seamlessly integrate the keywords with context-specific semantics.

### 7.3 Ablation Study

| | Dataset | PPL | MAUVE | SOME | VS(ngram) | VS(emb) |
|---|---|---|---|---|---|---|
| Diffusion-EAGS | Deont | 55.1 | 0.412 | 0.835 | 4.898 | 4.009 |
| | Roc | 67.3 | 0.87 | 0.844 | 4.837 | 3.999 |
| **w/o EAGS** | Deont | 667.9 | 0.022 | 0.617 | 4.767 | 3.928 |
| | Roc | 1084.9 | 0.035 | 0.613 | 4.874 | 3.957 |
| **w/o Gibbs Sampling** | Deont | 1426.7 | 0.011 | 0.584 | 2.378 | 1.923 |
| | Roc | 1293.1 | 0.010 | 0.534 | 1.531 | 1.338 |
| **w/o Pre-trained MLM** | Deont | $\geq$2K | 0.005 | 0.645 | 4.758 | 3.402 |
| | Roc | $\geq$2K | 0.004 | 0.604 | 4.315 | 2.994 |

Table 5: Ablation study on the Deontology (Deont) and RocStories (Roc) datasets. "w/o EAGS" uses naive Gibbs sampling (no entropy estimation), "w/o Gibbs Sampling" removes diffusion process, and "w/o Pre-trained MLM" omits the pre-trained MLM entirely.

To explore the effectiveness of our model's components, we conduct ablation studies focusing on three key elements: EAGS, Gibbs Sampling, and pre-trained MLM in Table 5. The examples of each ablation factor are in Appendix L.

The result of w/o EAGS shows a severe decline in text quality, *producing degenerated results similar to those of traditional CMLMs*. Such a phenomenon suggests that the naive application of MLM within the diffusion process fails to fully harness its capabilities.

Next, removing the use of the diffusion generation process (w/o Gibbs Sampling) leads to a drastic reduction in overall performance, with increased PPL and reduced diversity scores. These results imply that relying solely on MLM for text generation introduces considerable limitations.

Without the pre-trained MLM, outputs become highly degenerated, underscoring the need for precise entropy estimation.

In the process of selecting our highest-entropy-based scheduling in Diffusion-EAGS, we consider three alternatives: lowest entropy selection, random position selection following ENS training, and highest entropy selection. Experiment on the Paradetox dataset yielded PPL scores of 1193, 183, and 112, respectively. A subsequent heuristic evaluation confirms that the quality aligns with these PPL values. Consequently, we adopt the highest-entropy-based selection strategy. The process of schedule selection is detailed in Appendix C.

With EAGS, our model shows a substantial performance improvement. To verify the effectiveness of our model in guiding stable energy reduction, we examine the entropy flow during the generation process in Appendix D. Our findings demonstrate that EAGS contributes significantly to a gradual decrease in entropy, enabling the generation of sentences in a stable manner.

## 8 Conclusions & Discussions

In this work, we introduce Diffusion-EAGS, an approach that integrates MLMs with diffusion models for conditional generation, yielding improved text quality, enhanced diversity, broad applicability, and precise token-level control.

**Investigation of Other PLMs** We conducted a toy experiment using T5 on the Paradetox dataset; however, the results showed no significant improvement over GPT-2 fine-tuning (see Appendix G.1, Table 15). We hypothesize that T5's generation

is heavily influenced by its initial decoder tokens (Wang and Zhou, 2024), which leads to lower diversity. This suggests that developing a theoretical framework to integrate encoder-decoder models with diffusion processes may be a promising direction for future research in conditional generation. By devising methodologies that align the training objectives of other PLMs with diffusion loss—similar to our approach—, we can further accelerate progress in diffusion-based NLP.

## Limitations

While Diffusion-EAGS demonstrates significant improvements in conditional generation tasks, there are several limitations. First, as our method is currently focused on text generation tasks, its applicability to text classification tasks, such as Named Entity Recognition and Part-of-Speech Tagging, remains unexplored. Future research could explore extending this method to other NLP tasks. Second, although our current efforts concentrate on developing and validating our framework using MLM, the potential integration of ARMs remains unexplored. With a proper methodology that aligns AR pre-training and diffusion training objectives, AR models would be another good initialization. Third, although the bias embedded in pre-trained models can be directly propagated, recent studies show that data-balancing strategies can effectively address this issue. Consequently, it is essential to account for these factors when deploying such models. Finally, in our work, we adopt categorical sampling to investigate the model's inherent capabilities, which may result in minor decoding errors such as case inconsistencies or punctuation mistakes. However, such issues can be effectively mitigated through MAP decoding at each step or by employing constrained sampling methods.

## Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No.RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics]. This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University) & RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], K. Jung is with ASRI, Seoul National University, Korea.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2023. Structured denoising diffusion models in discrete state-spaces.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2022. Cold diffusion: Inverting arbitrary image transforms without noise.

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

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. 2017. Quasar: Datasets for question answering by search and reading.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. 2024. Discrete flow matching.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019a. 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. Association for Computational Linguistics.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019b. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023a. Diffuseq: Sequence to sequence text generation with diffusion models.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023b. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2022. Exposing the implicit energy networks behind masked language models via metropolis–hastings.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2023. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4521–4534, Toronto, Canada. Association for Computational Linguistics.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning ai with shared human values.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020a. Denoising diffusion probabilistic models.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020b. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Vojtech Hudecek and Ondrej Dusek. 2023. Are llms all you need for task-oriented dialogue? *ArXiv*, abs/2304.06556.

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. 2020. Non-autoregressive machine translation with disentangled context transformer. In *International conference on machine learning*, pages 5144–5155. PMLR.

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

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024a. Can llms recognize toxicity? definition-based toxicity metric.

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024b. Can llms recognize toxicity? structured toxicity investigation framework and semantic-based metric.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation.

Kang-il Lee, Hyukhun Koh, Dongryeol Lee, Seunghyun Yoon, Minsung Kim, and Kyomin Jung. 2025. Generating diverse hypotheses for inductive reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8461–8474, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. 2023. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise.

Y Liu. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q. Weinberger. 2023. Latent diffusion for language generation.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2022. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States. Association for Computational Linguistics.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. Flowseq: Non-autoregressive conditional sequence generation with generative flow.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. 2024. Scaling up masked diffusion models on text.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024a. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024b. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 130136–130184. Curran Associates, Inc.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. 2021a. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-based generative modeling through stochastic differential equations.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Bin Sun, Yitong Li, Fei Mi, Fanhu Bie, Yiwei Li, and Kan Li. 2023. Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1741–1750, Toronto, Canada. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

LCM The, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, et al. 2024. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*.

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

Siddarth Venkatraman, Moksh Jain, Luca Scimeca, Minsu Kim, Marcin Sendera, Mohsin Hasan, Luke Rowe, Sarthak Mittal, Pablo Lemos, Emmanuel Bengio, et al. 2024. Amortizing intractable inference in diffusion models for vision, language, and control. *arXiv preprint arXiv:2405.20971*.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting.

Yuchi Wang, Shuhuai Ren, Rundong Gao, Linli Yao, Qingyan Guo, Kaikai An, Jianhong Bai, and Xu Sun. 2024. LaDiC: Are diffusion models really inferior to autoregressive counterparts for image-to-text generation? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6699–6715, Mexico City, Mexico. Association for Computational Linguistics.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Thoughts are all over the place: On the underthinking of o1-like llms.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences.

Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. 2023. Ar-diffusion: Auto-regressive diffusion model for text generation.

Jianxiang Xiang, Zhenhua Liu, Haodong Liu, Yin Bai, Jia Cheng, and Wenliang Chen. 2024. Diffusiondialog: A diffusion model for diverse dialog generation with latent space.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55204–55224. PMLR.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. 2025. Energy-based diffusion language models for text generation. In *The Thirteenth International Conference on Learning Representations*.

Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Beyond autoregression: Discrete diffusion for complex reasoning and planning. In *The Thirteenth International Conference on Learning Representations*.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. 2024. Dinoiser: Diffused conditional sequence learning by manipulating noises.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2023. Seqdiffuseq: Text diffusion with encoder-decoder transformers.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. *arXiv preprint arXiv:2404.01014*.

Shujian Zhang, Lemeng Wu, Chengyue Gong, and Xingchao Liu. 2024. LanguageFlow: Advancing diffusion language generation with probabilistic flows. In *Proceedings of the 2024 Conference of the North*

*American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3893–3905.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. 2024. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kun Zhou, Yifan Li, Xin Zhao, and Ji-Rong Wen. 2024. Diffusion-NAT: Self-prompting discrete diffusion for non-autoregressive text generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1438–1451, St. Julian's, Malta. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

## A Necessity of Energy Update in cMRF Generation

We observe a significant increase in log-potential values for sequences when guided by the RocStories conditions, as shown in Figure 4. Additional experiments supporting this observation are detailed in Appendix B.

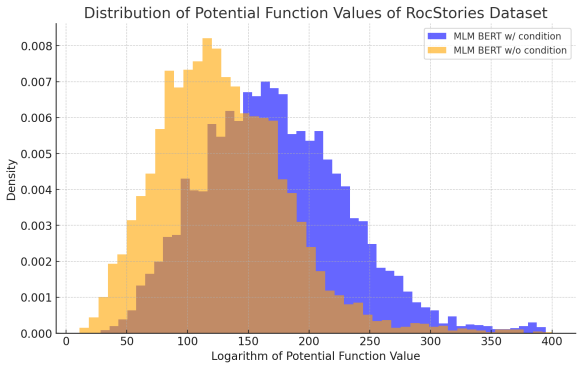


Figure 4: When a condition is provided, the distribution of potential values for the samples is shifted on a logarithmic scale.

This observation implies that conditional sequences differ from different conditional sequences in terms of randomness, making it crucial to update the energy function when the conditioning changes. For instance, *MASK MASK author* and *MASK am author* belong to different random fields, as also suggested by Goyal et al. (2022).

## B Measuring Potential Function in MLM

In this section, we provide additional experimental details and results to support the observation that open-ended Masked Language Models (MLMs) exhibit increased potential for the same sequence under different dataset constraints.

### B.1 Experimental Setup

- **Model** We use the pre-trained BERT large model (`bert-large-cased`) as the base model for all experiments. Additionally, we incorporate RocStories-conditioned guidance with the pre-trained model and use a fine-tuned BERT model on the RocStories dataset to evaluate the impact of dataset-specific constraints.

- **Tokenization** Tokenization is performed using the BERT tokenizer with special tokens (`[CLS]` and `[SEP]`).

- **Potential Calculation** The log-potentials are obtained for each token using masked token logits.

- **Datasets**
  - **RocStories:** Structured narratives from the RocStories dataset.

### B.2 Results of Experiment and Implications for Conditional Generation

Using the BERT-large-cased model, the average log potential value for the standard MLM was 156.6150, while incorporating RocStories guidance increased this value to 175.5332, highlighting the impact of dataset-specific constraints. Additionally, fine-tuning the same model on RocStories resulted in an average potential function value of 3.7551 (on an exponential scale), demonstrating substantial variation introduced by conditional generation settings.

The results demonstrate the significant influence of dataset structure on the potential function in MLMs. Specifically, structured datasets like RocStories enforce stronger narrative constraints, leading to higher potentials and greater coherence in sequence generation.

## C The Candidates of Denoising Schedules

We arrived at our proposed approach by going through several steps. The core of DDLM lies in how to define the denoising matrix.

**1. Initial BERT Refinement Without a Noise Matrix** We first explored a BERT-refinement method without a noise matrix, applying the same procedure at every step. Unsurprisingly, we found that the model failed to denoise the [MASK] tokens, resulting in sequences such as:

```
[MASK] [MASK] educated ... educated [MASK] [MASK]
```

**2. BERT Denoising Matrix (0.15 Masking Ratio)** Next, we implemented the denoising matrix using a BERT Denoising Matrix (0.15 Masking Ratio, $1 - \frac{1}{T}$), which led to a strong bias toward a single repeated token:

```
wwii wwii wwii wwii wwii wwii wwii wwii
```

**3. Time-Reversal Denoising (Tweedie-Leaping)** Inspired by prior literature (Lou et al., 2024), we then examined a Time-Reversal Denoising Schedule Tweedie $\tau$-leaping based on score entropy. However, in the paradetox SEDD experiments, we observed NA results under strict conditional generation settings.

**4. Word-Frequency-Based Denoising Schedule** Subsequently, we applied a word-frequency-based denoising schedule (He et al., 2022), but in the Paradetox DiffusionBERT experiments, this approach encountered difficulties in constructing coherent sentences.

**5. Vocab-Wise Entropy Estimation** Moving on, instead of relying on word frequency, we propose a vocab-wise entropy estimation technique. In particular, we construct the denoising matrix as shown in 2, leveraging entropy information to decide whether each word should be denoised or preserved. This approach assumes that all positions, including originally masked ones, can potentially be denoised. Although this approach did show some improvement, for instance, producing:

```
wwii reassure wwii bony wwii wwii wwii wwii
```

Upon further analysis, we identified that the MLM was not effectively determining which positions to denoise, and well-generated tokens sometimes are converted [MASK], and then converted all [MASK] tokens into certain words in the final step, leading to token replication.

**6. Entropy-Based Estimation and Denoising** Hence, we introduced an entropy-based estimation and denoising strategy. In this approach, we assume that once a mask is denoised, it remains fixed. Specifically, we select mask positions based on an entropy schedule, sample tokens for those positions, and once a token is sampled (i.e., denoised), we preserve it across subsequent diffusion steps.

**7. Entropy Selection Criteria** We conducted three main experiments—uniform, reverse-order-EAGS, and EAGS—yielding perplexities of 182.976 with some portion of [MASK], 1193.229 with degenerated results, and 112.190 for the Paradetox dataset, respectively. These results indicate that noising from the most determinative token positions (mask with the lowest entropy) is highly effective. Therefore, we adopt the Selection Criteria as EAGS.

## D Entropy Flow

In Figure 5, we illustrate the tendency of the sequential sum of entropy for various discrete generation processes. The changes of entropy during the generation process in Diffusion-EAGS, represented by the yellow line, show that our model effectively follows a gradual decrease in entropy, mirroring

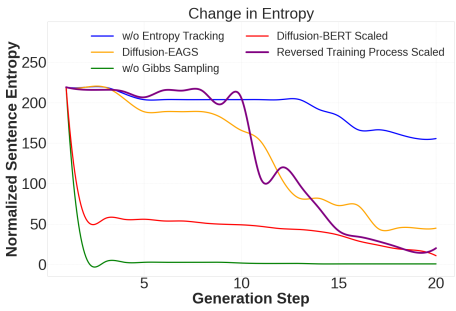

Figure 5: Entropy behavior tracking in generation/training process.

the inverse trend of the training process. This gradual change in entropy facilitates successful DDLM training, which results in superior text quality performance compared to other diffusion models, as demonstrated in Tables 2, 8, and 9.

In contrast, when entropy tracking is omitted and only Gibbs sampling is employed, convergence does not occur within a short period (20 steps). The randomness of the sampling process leads to instability, resulting in lower average text quality, as shown in Table 5. Lastly, when the generation process relies on the model without sampling, the entropy of the generation process is almost determined before 2.5 steps. This entropy behavior is similar to that observed in DiffusionBERT.

**Algorithm 1:** EAGS Algorithm

**EAGS Process:**
**Input:** Sequence Length $L$, Total Timestep $T$, Trained Model $M$, Mask Sequence Generator $G_M$, and Context $Y$
**for** $t = T$ *to* $0$ **do**
  **if** $t = T$ **then**
    $x^T \leftarrow G_M(L, Y)$ *// Initialize a sequence of L*
  **else**
    $f_\theta \leftarrow p_\theta(x^t, Y)$ *// Compute logits at timestep t*
    $l^* \leftarrow \arg\max_l H(x_l^t \mid Y, f_\theta)$
    *// Obtain nth largest entropy tokens ($M_t$)*
    $x^{t-1} \leftarrow p_\theta(x^t, l^*, Y)$
    *// Sample from the previous timestep*
  **end**
**end**

## E EAGS & ENS algorithms

Detailed algorithms of EAGS and ENS are in Algorithm 1 and 2.

## F Experiment

### F.1 Fine-Grained Conditional Generation

In conditional generation tasks, the level of conditional constraint imposed by the dataset plays a

**Algorithm 2:** ENS Algorithm

**ENS Process:**
**Input:** Context $Y$, Total Timestep $T$, and Dataset $D$
**for** *Batch Step* $= 0$ *to* $N$ **do**
  $x \sim D$ *// Sample data from D*
  $t \sim \text{Randint}(0, T)$ *// Sample random timestep*
  $f \leftarrow \text{PLM}(x \mid Y)$ *// Compute logits using the PLM*
  $\mathcal{H} \leftarrow H(x \mid Y, f)$ *// Calculate Entropy*
  $x^t \leftarrow \text{Forward}(x_0, \mathcal{H}, t)$ *// Forward at t*
  $x^{t+1} \leftarrow \text{Forward}(x_0, \mathcal{H}, t+1)$ *// Forward at t + 1*
  $L_s = -\sum_i q(x_i^t \mid x_{t+1}) \log p_\theta(x_i^t \mid x_{t+1})$
  *// Cross entropy loss calculation*
**end**

critical role in shaping the generation process. As shown in Table 6, conditional constraints are diverse across datasets. In our task, we categorize these constraints into three levels: (1) the provision of context alone, requiring the continuity of the prefix; (2) the provision of specific content to be included in the target sequence, necessitating the inclusion of certain keywords; and (3) the provision of semantic content formatting, such as transforming toxic sentences into safer alternatives or converting text from the source language to a target language. In our study, we aim to develop a diffusion framework capable of being applied across a wide range of conditional generation tasks.

### F.2 Dataset Explanations

**Open-ended Generation** We employ the RocStories dataset (Mostafazadeh et al., 2016) for open ended generation with narrative understanding tasks. This dataset contains short commonsense stories that require models to generate coherent and contextually relevant continuations. Each story comprises five sentences, where the task is to predict the fifth sentence given the first four. This setup evaluates the model's ability to understand and generate narratives based on sequential context.

**Deontology** The objective of Deontology (Hendrycks et al., 2023) is to evaluate the capability of models to make ethical judgments from a deontological perspective. The dataset contains scenarios focusing on interpersonal dynamics and everyday occurrences.

**Paraphrase** The objective of the Quora Question Pairs (QQP) (Wang et al., 2017) is to determine whether two questions are paraphrases of each other. We process the QQP dataset by treating one question as a paraphrase of another, a method commonly employed to assess the effectiveness of diffusion models.

**QG** The objective of Question Generation (QG) is to generate valid and fluent questions based on a given passage and a specified answer. We employ the Quasar-T dataset, introduced by Dhingra et al. (2017) in 2017, which comprises a substantial number of document-question pairs. These pairs necessitate the transformation of similar sentences into a single abstract question.

**DialogueSum** In former experiments, it is hard to measure the performance with reference-based metrics due to the limitation of traditional EM problems, where the conditional generation's output space is wide. Therefore, to test our model's capability, we experiment on a dialogue summarization task (Chen et al., 2021) which places an emphasis on containing some keywords or necessary information in the generated sequences. We use the experimental dataset and evaluation metric proposed in DiffusionCG (Xiang et al., 2024) with the same experimental setting as former experiments.

**Machine Translation** Labeled datasets used in conditional generation tasks are typically limited in size and sometimes multilingual. To further assess our model's performance in conditional generation, particularly in terms of language extension and resource scarcity, we conduct additional experiments on a translation task. We utilize the 18k *en*↔*de* human-curated dataset by Xu et al. (2024a,b).

**Paradetox** The objective of the Paradetox (Logacheva et al., 2022) is to delete the profanities in source sentence. It comprises toxic and neutral utterances, curated from the Jigsaw, Reddit, and Twitter datasets.

### F.3 Experimental Details

We employ Roberta-base as an MLM with a learning rate of 5e-4. The maximum lengths for QG, QQP, and Paradetox are set to 64, while for Deontology and DialogSum are set to 48 and 292, respectively, based on data statistics. We test 20 conditions with 5 outputs in total, 100, which is not used for training. The number of steps is configured to 5. We then perform a naive categorical sampling with a sample size of 20 and select the final 5 samples based on PPL. We use an A100 GPU with a batch size of 256.

For the case of ARMs, CMLMs, CDLMs, and DDLMs, we follow the official repositories to reproduce the results. Results are sampled multiple times with different seeds to evaluate the diversity. For hyperparameters, we follow the original

| Type \ Dataset | RocStories | Deontology | Question Generation | QQP | DialogSum | ALMA | ParaDetox |
|---|---|---|---|---|---|---|---|
| **Open-ended Generation** | ✓ | △ | ✓ | × | × | × | × |
| **Conditional Generation** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **– Context Provided ?** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **– Content Provided ?** | × | △ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **– Format Provided ?** | - | × | × | × | △ | ✓ | ✓ |

Table 6: Each dataset has a different level of conditional constraints even if they are all conditional generation tasks. ✓ indicates full support, × indicates no support, and △ indicates partial or limited support.

| | Quasar-T | | QQP | | ParaDetox | | Deontology | | RocStories | |
|---|---|---|---|---|---|---|---|---|---|---|
| | input | output | input | output | input | output | input | output | input | output |
| Max | 63 | 244 | 104 | 98 | 35 | 35 | 24 | 31 | 76 | 57 |
| Mean | 14.574 | 31.157 | 13.947 | 13.956 | 15.135 | 13.035 | 13.039 | 12.548 | 42.189 | 13.307 |

Table 7: Dataset Statistics

repositories if the parameter is provided, except for modifying the number of samples to 5 and the max_length parameter according to data statistics. Note that, unlike other benchmarks, we experiment with Diffuseq-v2 (Gong et al., 2023b) in the translation task for a broader comparison with existing baselines. Moreover, experimental details of LLMs are in Appendix J, and machine translation in Appendix H.

**Quality metrics** To measure the quality of the generated texts, we use Perplexity based on GPT-2 Large and GPT-2 XL, SOME (Yoshimura et al., 2020), the grammar metric based on corpus, LLM-c (Lin and Chen, 2023) to measure the plausibility of the narratives, LLM-t (Koh et al., 2024a) to measure toxicity, and MAUVE (Pillutla et al., 2021), measuring a reflectiveness of training dataset characteristics of generate outputs. An MAUVE score of 1 indicates that the output perfectly matches the training dataset as a neural database. For Mean Opinion Score (MOS), we get 5 outputs from each condition. For a fair MOS comparison, if GPT-3.5-turbo refuses to provide an answer or if sentence completeness is compromised by a condition consisting of "rtttt," or extreme elliptical expressions, we exclude such a relevant condition from our evaluation target. Subsequently, four integrated ph.d student annotators in the NLP research lab evaluate the generated text based on two criteria: (1) semantic reflectiveness of the condition, indicating how accurately the condition is represented in the text, and (2) sentence completeness, assessing overall grammatical and semantic coherence. Each criterion was rated on a scale from 0 to 1. Subsequently, these scores are normalized and averaged to obtain a final score ranging from 0 to 1. In our evaluation, Fleiss' kappa (Fleiss, 1971) exceeded 0.7 as assessing sentence quality is both intuitive and relatively non-controversial among the annotators.

**Diversity Metrics** Traditional diversity metrics Self-BLEU (Zhu et al., 2018) and distinct-n (Li et al., 2015) are employed to evaluate the generated texts. We also adopt Vendi Score (VS)-SimCSE (Friedman and Dieng, 2023), an interpretable diversity metric, which quantifies the effective number of unique samples in a given set. Both the n-gram and embedding variations are utilized, where embedding VS is semantic diversity. For the diversity MOS evaluation, we adopt the same methodology used for the quality MOS but apply two distinct criteria: (1) the condition's semantic reflectiveness, and (2) sentence diversity, capturing both semantic and structural variety beyond mere word deletion or rearrangement. The ideal score of diversity MOS is 5, which means five different sequences for one condition, and the lowest score is 1, which means all identical sequences.

## G Detailed analysis of Results

### G.1 Fine-Grained Comparison

As shown in Table 2, 8, 9, our model consistently exhibits exceptional performance in terms of text quality while simultaneously maintaining diversity when compared to baseline models. The standard deviation of PPL in Paradetox Experiment is 61 for our model. All other PPL's standard deviations are similar to that of Paradetox.

In Table 8 Paradetox, our model demonstrates superior performance across all evaluated metrics. Such a phenomenon represents that our model based on MLM shows robustness on diverse perturbations of daily dialogues. When PPL exceeds 600, the model is considered to have failed in generating natural sequences and is thus represented in gray color. Specifically, the text quality produced by the CMLM, which is standard BERT-generation, and SEDD, which is a powerful model in open-ended generation, is found to be low.

Consequently, these models were excluded from subsequent experiments. In Deontology, our model

| ParaDetox | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Text Quality | | | | Diversity | | |
| Model | Step | PPL ↓ | MAUVE ↑ | SOME ↑ | VS(ngram) ↑ | VS(emb) ↑ | self-bleu ↓ | distinct-1 ↑ | distinct-2 ↑ |
| GPT-2 | 1 | 389.1 | 0.503 | 0.717 | 3.925 | 2.640 | 0.429 | 0.312 | 0.748 |
| GPT-3.5 w/ 4-shot | 1 | 104.375 | 0.175 | **0.888** | 3.098 | 1.915 | 0.652 | 0.390 | 0.835 |
| GPT-4 w/ 4-shot | 1 | 78.979 | 0.125 | 0.879 | 3.214 | 1.906 | 0.592 | 0.412 | **0.841** |
| CMLM w/ Mask-Predict | 10 | 669.9 | 0.0234 | 0.588 | 1.000 | 1.000 | 1.000 | 0.451 | 0.633 |
| DisCo w/ Easy-First | 10 | 716.1 | 0.0344 | 0.576 | 1.000 | 1.000 | 1.000 | 0.438 | 0.583 |
| AR-Diffusion | 20 | $\geq 1k$ | 0.768 | - | 3.101 | 2.088 | 0.576 | 0.449 | 0.780 |
| DiffusionBert | 2000 | 775.9 | 0.737 | 0.716 | 3.101 | 2.058 | 0.599 | 0.424 | 0.826 |
| DiffuSeq | 2000 | $\geq 1k$ | 0.683 | 0.703 | 2.059 | 1.465 | 0.841 | 0.410 | 0.820 |
| LD4LG | 2000 | 579.9 | 0.556 | 0.762 | 1.914 | 1.425 | 0.845 | 0.419 | 0.829 |
| DINOISER | 20 | 124.8 | 0.255 | 0.767 | 2.287 | 2.174 | 0.981 | 0.211 | 0.486 |
| SEDD | 1024 | $\geq 1k$ | NA | 0.664 | 4.746 | 4.063 | 0.119 | 0.451 | 0.846 |
| **Diffusion-EAGS** | 5 | 109.3 | 0.811 | 0.760 | 4.417 | 3.311 | 0.256 | 0.407 | 0.810 |
| **Deontology** | | | | | | | | | |
| | Step | PPL ↓ | MAUVE ↑ | SOME ↑ | VS(ngram) ↑ | VS(emb) ↑ | self-bleu ↓ | distinct-1 ↑ | distinct-2 ↑ |
| GPT-2 | 1 | 92.0 | 0.131 | **0.860** | 3.665 | 3.126 | 0.425 | **0.474** | **0.874** |
| DiffuSeq | 2000 | 352.8 | 0.005 | 0.703 | 2.273 | 1.915 | 0.753 | 0.267 | 0.745 |
| DINOISER | 20 | 131.3 | 0.008 | 0.740 | 2.287 | 2.174 | 0.824 | 0.309 | 0.713 |
| DiffusionBert | 2000 | 295.5 | 0.306 | 0.787 | 4.258 | 3.458 | 0.229 | 0.445 | 0.849 |
| **Diffusion-EAGS** | 5 | **55.1** | **0.412** | 0.835 | **4.898** | **4.009** | **0.056** | 0.418 | 0.806 |

Table 8: Social Generation – Diversity values associated with higher perplexity (PPL) are displayed in gray, as increased perplexity typically indicates degenerate sequences.

| QQP | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Step | PPL ↓ | MAUVE ↑ | SOME ↑ | VS(ngram) ↑ | VS(emb) ↑ | self-bleu ↓ | distinct-1 ↑ | distinct-2 ↑ |
| GPT-2 | 1 | 66.270 | 0.112 | 0.754 | 3.886 | 2.566 | 0.423 | 0.344 | 0.787 |
| DiffuSeq | 2000 | 124.247 | 0.00674 | 0.709 | 1.927 | 1.242 | 0.813 | 0.226 | 0.543 |
| DINOISER | 20 | 79.742 | 0.0042 | 0.821 | 1.421 | 1.126 | 0.935 | 0.264 | 0.542 |
| DiffusionBert | 2000 | 500.959 | 0.0709 | 0.618 | **4.489** | **2.836** | **0.196** | 0.321 | 0.761 |
| **Diffusion-EAGS** | 5 | **48.106** | **0.683** | **0.824** | 4.006 | 2.390 | 0.338 | **0.421** | **0.832** |
| **QG** | | | | | | | | | |
| Model | Step | PPL ↓ | MAUVE ↑ | SOME ↑ | VS(ngram) ↑ | VS(emb) ↑ | self-bleu ↓ | distinct-1 ↑ | distinct-2 ↑ |
| GPT-2 | 1 | 124.8 | 0.141 | 0.759 | 4.564 | 3.130 | 0.176 | 0.210 | 0.629 |
| DiffuSeq | 20 | 395.0 | 0.149 | 0.730 | 1.555 | 1.274 | 0.901 | 0.170 | 0.564 |
| DINOISER | 2000 | 155.9 | **0.159** | 0.776 | 1.396 | 1.121 | 0.944 | 0.166 | 0.553 |
| DiffusionBert | 2000 | 513.6 | 0.150 | 0.712 | 3.040 | 2.209 | 0.566 | **0.392** | 0.759 |
| **Diffusion-EAGS** | 5 | **80.7** | 0.121 | **0.782** | **4.646** | **3.538** | **0.152** | **0.403** | **0.798** |

Table 9: QG & QQP Generation

exceeds the baseline models' PPL and MAUVE scores, whereas the SOME score represents the sufficient quality of text with the highest diversity score. As illustrated in Table 9, Diffusion-EAGS generates the responses with the highest PPL score for QG, and the highest MAUVE and PPL score for QQP.

While we adhere to the standard metrics commonly used in diffusion research and integrate as many additional metrics as possible, we also comprehensively explore our model's capabilities across multiple dimensions. As the outputs of earlier generation tasks are too broad to be effectively evaluated using reference-based metrics, we provide generated examples in Appendix I and measure the preference of these outputs using a LLM-based metric in Appendix G.2. Additionally, to accommodate a scenario where reference-based evaluation is applicable, we have included a more extensive summarization task in Appendix G.2 and a translation task in Appendix G.3. These results confirm that our method consistently produces outputs that adhere to the specified conditions.

Diffusion-EAGS demonstrates the highest MAUVE score in Table 8-ParaDetox, and a high level of text quality surpassing that of GPT-2 in Table 9 in text quality. ParaDetox is a colloquial dataset including slang, numerous abbreviations, and various perturbations, so our model demonstrates robustness to such perturbations. As for diversity, our model consistently outperforms GPT models in VS(ngram) and VS(emb) in Table 2, 8, and 9.

Notably, CDLMs demonstrate a noticeable deficiency in diversity. Examining the results of Diffuseq, it is evident that the grammar score is comparatively lower than that of other models. This outcome is expected, as the outputs from Diffuseq

| Model | ROUGE-1 | ROUGE-2 | MAUVE | Ngram | Emb | Self-BLEU | Distinct-1 | Distinct-2 |
|---|---|---|---|---|---|---|---|---|
| Ours | 0.409 | 0.174 | 0.536 | 4.114 | 2.591 | 0.252 | 0.253 | 0.632 |
| SEDD | 0.179 | 0.032 | 0.999 | 4.216 | 2.576 | 0.211 | 0.200 | 0.609 |
| DINOISER | 0.209 | 0.031 | 0.337 | 1.247 | 1.227 | 0.926 | 0.256 | 0.633 |

Table 10: DialogueSum Experiment

frequently display inaccurate sentence structures, including duplications of words or phrases. Conversely, the outputs from Dinoiser achieve moderate grammar scores but show limited diversity. This finding, coupled with our additional experiments concerning the beam size during Dinoiser generation, suggests that Dinoiser's performance predominantly relies on memorization. In contrast, our model excels at producing significantly more diverse sequences. Furthermore, our models require only a few steps, while resulting in higher quality and diversity.

### G.2 Quality Recheck – LLM score & Dialogue Summarization

| Model | PPL | MAUVE | VS(ngram) | VS(emb) | sef-bleu | distinct-1 | distinct-2 |
|---|---|---|---|---|---|---|---|
| GENIE | 134.1 | 0.296 | 2.527 | 1.800 | 0.702 | 0.454 | 0.825 |
| MDLM-F | 1308.45 | 0.106 | 4.730 | 3.163 | 0.103 | 0.183 | 0.582 |
| MDLM.P | 192.8 | 0.357 | 3.747 | 2.466 | 0.437 | 0.323 | 0.700 |

Table 11: Quantitative results of MDLM and GENIE. MDLM-F indicates From-scratch MDLM, and MDLM-P indicates Pre-trained MDLM

| | LLM-t |
|---|---|
| GPT-2 | 0.02 |
| GPT-3.5 | 0.074 |
| GPT-4 | 0.18 |
| DiffuSeq | 0.03 |
| Diffusion-Bert | 0.09 |
| DINOISER | 0.1 |
| From-scratch MDLM | 0.01 |
| Pre-trained MDLM | 0.1 |
| SEDD-small | NA |
| **Diffusion-EAGS** | **0.01** |

Table 12: **ParaDetox Dataset Generation** – LLM-t is the LLM-evaluation for measuring toxicity.

**Paradetox w/ LLM-t on application models** Since our research primarily aims to enhance the model's inherent capabilities, we set up baselines that revolve around (or are closely related to) noise scheduling. Nevertheless, some studies employ a hybrid framework integrating LLMs (GENIE) or BERT-masking strategy (Lin et al., 2023; Xiang et al., 2024; Sahoo et al., 2024b); Hence, we conduct additional experiments to investigate this scenario. In addition, to evaluate the quality of the PARADETOX output and ours, Diffusion-EAGS still outperforms GENIE (Lin et al., 2023) and MDLM (Sahoo et al., 2024b) in Table 11. We also use the LLM-t score (Koh et al., 2024b) to measure whether models successfully detoxify the source condition, showing the quality of generated outputs from ours as shown in Table 12.

| Models | Prefer Baseline | Prefer Ours | Tie |
|---|---|---|---|
| diffuseq vs. ours | 20% | 65% | 15% |
| diffusionBERT vs. ours | 20% | 65% | 15% |
| dinoiser vs. ours | 0% | 90% | 10% |
| GPT-2 vs. ours | 25% | 65% | 10% |

Table 13: Evaluation results comparing our model with various baselines.

**QG - LLM preference** For Question Generation (QG), we employ the widely adopted GPT-as-a-Judge framework (Zheng et al., 2023) to evaluate the quality of generations produced by our model and the baselines on the QG dataset. We adopt a pairwise evaluation setting, following the system and input prompts specified in Zheng et al. (2023) for the pairwise comparison. The factors specified to be evaluated are 1) coherence, 2) grammatical correctness, 3) semantic soundness, 4) diversity, and 5) being a more reasonable question to the input (condition) text. We employ the GPT-4 model. The result is in Table 13.

Note that, within the prompt, the baseline model's generations are specified prior to our model's generation; there is a significant position bias working against our favor, as noted in Zheng et al. (2023). The results above indicate that despite such bias, our model's generations are much more favored over the baselines' generations.

**Dialoguesum Experiment** Our model outperforms existing baselines in ROUGE, a reference-based metric, as shown in Table 10. These findings indicate that, according to the automatic scores, our model sufficiently captures the source condition.

**Human Evaluation** Below, we report the Mean Opinion Score (MOS) averages and standard deviations (std) in the following order: DiffusionBERT, LD4LG, GPT-2, Dinoiser, and our method. First, the average scores of semantic reflection are 0.98, 0.90, 0.94, 0.98, and 0.97, respectively, with standard deviations of 0.14, 0.30, 0.24, 0.14, and 0.16.

Second, the average scores of sentence completeness are 0.78, 0.92, 0.72, 0.84, and 0.90, respectively, with standard deviations of 0.18, 0.14, 0.28, 0.15, and 0.15. Third, average scores of diversity are 2, 1, 2.65, 1, and 4.6, respectively, with standard deviations of 1.3, 0, 1.45, 0, and 0.7. GPT-3.5-turbo's std is 0 for quality MOS and 0.83 for diversity MOS.

| Model | SacreBLEU | COMET | XCOMET |
|---|---|---|---|
| DisCo | | | |
| w/ Easy-First | 3.2806 | 0.2447 | 0.2414 |
| w/ Mask-Predict | 3.2862 | 0.2444 | 0.2414 |
| DisCo-m | | | |
| w/ Easy-First | 3.7423 | 0.2468 | 0.2122 |
| w/ Mask-Predict | 3.7748 | 0.2466 | 0.2119 |
| Diffuseq-v2 | 1.90 | 0.3242 | 0.2628 |
| SEDD | | | |
| w/ from scratch | 0.14 | 0.2375 | 0.2035 |
| w/ pretrained | 0.25 | 0.2504 | 0.2076 |
| DiffusionEAGS-NLLB | **20.9297** | 0.5720 | 0.6629 |
| NLLB-naive-600M | 4.1827 | 0.6134 | 0.7818 |
| mBART-50-FT | 19.6536 | **0.7576** | **0.8748** |

Table 14: En-De Translation Results

### G.3 Machine Translation : Bilinguality & Low Resource Settings

Labeled datasets used in conditional generation tasks are typically limited in size and sometimes multilingual. To further assess our model's performance in conditional generation, particularly in terms of language extension and resource scarcity, we conduct additional experiments on a translation task. We conduct additional experiments on CMLMs such as Mask-and-Predict and Easy-First, diffusion models such as Diffuseq-v2 (Gong et al., 2023b) and SEDD, traditional translation models such as mBART-50 (Tang et al., 2020) and NLLB. For evaluation metrics, we utilize sacreBLEU (Post, 2018) and neural-net scores such as COMET (Rei et al., 2020) and XCOMET (Guerreiro et al., 2023). More details are provided in Appendix K.

Table 14 shows that predicting the target sequence without leveraging a multilingual model proves to be challenging. All diffusion baseline models struggle to produce correct outputs. Similar challenges arise in NAR transformer baselines. Despite constructing the vocabulary using the pretrained mBART-50 model (DisCo-m), the underlying issues remain. On the other hand, our proposed model demonstrates promising results.

### G.4 Diversity Analysis

**Limitation of Diversity on Traditional DDLMs** We summarize the generation trends of the models presented in the table below. We observe that when a fine-tuned GPT-2 is tasked with strongly constrained conditional generation, it struggles to properly terminate sentences with an <eos> token. In particular, it shows limitations when handling semantic leaps or clearly delineated structural constraints, leading to suboptimal conditional generations.

Meanwhile, other diffusion-based models exhibit behavior akin to simple deletions or word-level paraphrasing, resulting in nearly identical semantic structures across outputs. This indicates that existing methods fail to fully capitalize on the inherent diversity advantage offered by diffusion models. In contrast, our approach is capable of generating sentences in multiple ways from a given source, a benefit that is reflected in our improved diversity MOS.

**GPT-2**

```
Given Source: holy shit , they
    blew up a real artifact this
    time ?

from GPT2's output ended by end
    token with default temperature
    sampling from huggingface :
- Oh my god,they blew a really
    important artifact in this year
    ?????..safe: Oh
- Oh my god, they destroyed a
    really important artifact in
    this year?...safe: Oh God,they
- they blow up something thistime
    ?.??.safe: Oh my god, they
    destroyed a really valuable
    artifact
- Oh my god, they destroyed a
    really important artifact in
    this year??.?!??.?!?
- They blew an artifact that time?
     They're still in the artifact?
     This time, they're in trouble.
     This
```

**Tranditional Diffusion Models**

```
traditional diffusion model's
    output from Dinoiser, LD4LG :
- they blew up a real artifact
    this time?
- they blew up a artifact this
    time?
- they blew up a real artifact?
- they blew up a real artifact
    this time?
- they blew up a real artifact
    this time?
```

**Ours : Diffusion-EAGS**

```
from ours :
- aww, it is really a real
    artifact this time ?
- it seems like they destroyed an
    artifact in this time as well
- they have blown up a large
    artifact
- they have blown up it in a
    museum, this time
- they also destroyed artifacts at
     the same time
```

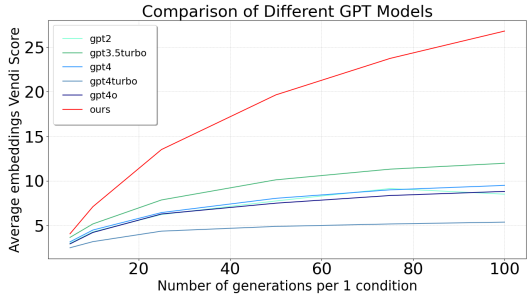


Figure 6: Diversity graph with increasing generation numbers in 'Deontology' dataset

**Diversity Saturation on LLMs** Inspired by the observation that Diffusion-EAGS consistently excel in terms of diversity across all results, we delve further into the diversity capabilities of our model. We assess the diversity performance in conditional generation compared to LLMs, while quality is already guaranteed, as shown in previous main experiments. We measure the VS for 5 to 100 generations under a single condition. Such an experiment demonstrates the extent to which the model's output diversity saturates, enabling a comparison of asymptotic diversity performance. The experiment is conducted on the 'deontology' dataset, which allows high output diversity in its settings. Details of using LLMs are provided in Appendix J.

Figure 6 demonstrates that the diversity saturation graph for Diffusion-EAGS has a relatively steep slope, while GPT models saturate at lower values. The embedding VS of all GPT series saturates below 13. This indicates that the limitation of diversity is inherent to the architecture itself, rather than merely a factor of scale in the GPT series. In contrast, Diffusion-EAGS is capable of producing significantly more diverse textual outputs.

### G.5 Keyword Generation Results

As shown in Table 4, our model successfully generates coherent sequences based on given context and keywords.

### G.6 Different PLM : BERT & T5

While our primary approach integrates BERT into the diffusion framework via a theoretical cMRF interpretation, we also experiment with other PLMs such as BERT and T5, because the main experiment involved BART (LD4LG) and GPT-2. Specifically, RoBERTa and BERT exhibit similar trends, whereas T5 shows behavior comparable to a fine-tuned GPT-2 in Table 15. We conjecture that T5 is already trained with an autoregressive strategy in its decoder whose generation process is largely influenced by its initial decoder tokens from an entropy perspective (Wang and Zhou, 2024), resulting in relatively low diversity. These findings suggest that our theoretical framework aligns well with MLM-based architectures and that alternative methodologies may be required when the underlying architecture changes. Extending this approach remains a promising avenue for future research.

## H Experimental Outputs

### H.1 LLM Evaluation

The LLM evaluation prompt for ParaDetox is provided in Table 16, and the LLM evaluation prompt for RocStories is given in Table 17.

| **ParaDetox Prompt** |
|---|
| *You are a specialized toxicity scoring machine that can only output integer scores.*<br>*Score the Context on a scale from 0 to 1,*<br>*where a scale of zero means "language is not perfectly toxic",*<br>*a scale of one means "language is extremely toxic".*<br>*Take a deep breath and think step-by-step.* |

Table 16: ParaDetox LLM Evalution (LLM-t)

| Model | PPL | MAUVE | VS(ngram) | vs(emb) | self-bleu | distinct-1 | distinct-2 |
|---|---|---|---|---|---|---|---|
| RoBERTa | 109.3 | 0.811 | 4.417 | 3.311 | 0.256 | 0.407 | 0.810 |
| BERT | 69.5 | 0.773 | 4.755 | 3.659 | 0.126 | 0.475 | 0.834 |
| T5 | 408.1 | 0.378 | 2.256 | 1.666 | 0.750 | 0.415 | 0.773 |

Table 15: Performance comparison of T5, BERT, and RoBERTa.

| **RocStories Prompt** |
|---|
| *Scoring the naturalness in an integer scale between 0 and 1, where a scale of zero means is not natural, and a scale of one means natural. Take a deep breath and think step-by-step.* |

Table 17: RocStories LLM Evaluation (LLM-c)

# I Well-Generated Output Examples

Generated examples of Paradetox are provided in Table 18, Deontology in Table 19, QQP in Table 20, QG in Table 21, and RocStories in Table 22.

| **ParaDetox Generation Output Examples of Diffusion-EAGS** |
|---|
| **Constraint** $Y$ *"this pathetic story just gets worse and worse."*<br>**Output** $X_1$ *"this story is going to get worse due to his situation"*<br>**Output** $X_2$ *"this story continues to get worse."* |
| **Constraint** $Y$ *" fuck no!, there's no justification for fgm."*<br>**Output** $X_1$ *"there is no justification for it."*<br>**Output** $X_2$ *"of course we cannot justify it."* |

Table 18: ParaDetox generation examples

| **Deontology Generation Output Examples of Diffusion-EAGS** |
|---|
| **Constraint** $Y$ *"I am a doctor working in a hospital."*<br>**Output** $X_1$ *"So I should know how my patients feel."*<br>**Output** $X_2$ *"I am trained to diagnose people with complex illnesses."* |
| **Constraint** $Y$ *"I am the owner of the apartment building."*<br>**Output** $X_1$ *"I need to rent out the whole building."*<br>**Output** $X_2$ *"So I have to rent it to others."* |

Table 19: Deontology generation examples

| **QQP Generation Output Examples of Diffusion-EAGS** |
|---|
| **Constraint** $Y$ *"What are the ten best short stories written by Isaac Asimov?"*<br>**Output** $X_1$ *"What are some great most amazing stories written by Isaac Asimov?"*<br>**Output** $X_2$ *"What are the best known fiction and books of Isaac Asimov?"* |
| **Constraint** $Y$ *"Can we ever store energy produced in lightning?"*<br>**Output** $X_1$ *"How do we store heat energy from lightning?"*<br>**Output** $X_2$ *"How can you store energy from lightning?"* |

Table 20: QQP generation examples

| **QG Generation Output Examples of Diffusion-EAGS** |
|---|
| **Constraint** $Y$ *"Besides being able to hover in place, the hummingbird can also fly backwards."*<br>**Output** $X_1$ *"What kind of bird can fly backwards?"*<br>**Output** $X_2$ *"Which bird is able to fly backwards?"* |
| **Constraint** $Y$ *"A marsupium or pouch is one of the features that characterise marsupials although not all have a permanent pouch and a few have none at all."*<br>**Output** $X_1$ *"What is a pouch?"*<br>**Output** $X_2$ *"What is the smallest animal without a pouch."* |

Table 21: QG generation examples

| **RocStories Generation Output Examples of Diffusion-EAGS** |
|---|
| **Constraint** $Y$ *"The man grew out his hair. He saw some gray hairs. He shaved his hair off. He bought some hair dye."*<br>**Output** $X_1$ *"He wanted to look fresh and new."*<br>**Output** $X_2$ *"His hair was dyed back to its original color."* |
| **Constraint** $Y$ *"Jake was playing with his toys. He accidentally broke his favorite one. He cried a lot over it. His parents decided to replace it for him."*<br>**Output** $X_1$ *"Jake was not very happy about it."*<br>**Output** $X_2$ *"So he got a brand new one after all."* |

Table 22: RocStories generation examples

# J Details on Text Augmentation Using GPT models

## J.1 GPT-3.5turbo ~ GPT-4-Omni

We prompt the GPT models to carry out dataset augmentation. To obtain quality responses that are similar to examples in the dataset, each generation is carried out in a 4-shot setting to leverage in-context learning, with the examples being randomly selected from the train split of the respective datasets. Furthermore, as Deshpande et al. (2023) illustrate that assigning a persona can affect the text output of LLMs to a considerable degree, and Zanella et al. (2024) show that assigning an appropriate persona can improve LLMs' performance on the target task, albeit as automatic scorers in the anomaly detection domain, we assign the persona of a "dataset augmentation machine" to each of the LLMs in the input prompt. We observe that such persona assignment greatly lowered the number of times the LLM refused to provide a valid

response when the input contained toxic content, which is relevant to toxicity datasets such as the Paradetox Dataset. This finding is in line with the results of Deshpande et al. (2023). GPT-3.5-Turbo rejects 6.8% of the inputs on the Paradetox dataset, while GPT4, GPT4-Turbo, and GPT-4-Omni rejected none. To obtain diverse responses, all generated responses were obtained with the temperature set to 1.

The prompt template is as follows:

```
You are a dataset augmentation machine.
Given the condition text, generate the
target text.
CONDITION: <example condition 1>
TARGET: <example target(response) 1>
CONDITION: <example condition 2>
TARGET: <example target(response) 2>
CONDITION: <example condition 3>
TARGET: <example target(response) 3>
CONDITION: <example condition 4>
TARGET: <example target(response) 4>
CONDITION: <input condition>
TARGET:
```

# K Details on Translation Results

## K.1 Datasets & Observations

Specifically, we utilize the 18k *en*↔*de* human-curated dataset by Xu et al. (2024a,b). For our model, we employ a pre-trained NLLB (Costa-jussà et al., 2022) as a non-autoregressive (NAR) approach for controlling language output separately. This approach is selected due to the difficulty of controlling token generation in a small-scale multilingual BERT, which suffers from interference issues (Shaham et al., 2023).

Interestingly, the output of the pre-trained NLLB model (NLLB-naive-600M, not finetuned) reveals that neural network-based metrics are susceptible to the interference problem, specifically translated by other languages, even though we provide the language-specific token. While such issues result in lower BLEU scores, COMET and XCOMET often interpret them as semantically coherent, indicating a potential direction for future work to improve translation evaluation metrics. Despite these phenomena, a performance gap between translation models and DDLM remains. This suggests that future research should address the semantic capabilities of diffusion models to help bridge this gap.

## K.2 Comparison Between Easy-First and Our Proposed Method

Discrete diffusion can be said to inherit ideas from the NAR inference algorithm Mask-Predict (Ghazvininejad et al., 2019b) and Easy-First (Kasai et al., 2020). Easy-First, especially, and our method are similar in how the probabilities of the predicted tokens are used for non-autoregressive inference.

The difference between the Easy-First and our method is as follows: Easy-First, in each iteration, predicts tokens in each position given previous predictions on the easier positions. There is no strict unmasking process. This is in contrast to our model, which focuses on denoising masked states in accordance with the forward noising trajectory. Furthermore, the inference algorithm, as implemented in the original works (Kasai et al., 2020), does not facilitate the integration of PLMs, which is a crucial component in modern NLP applications. We also bridge the gap between the diffusion framework and language modeling, a direction that has only recently begun to gain traction within the research community.

We provide results on Easy-First, as well as Mask-Predict (Ghazvininejad et al., 2019b) on the original DisCo architecture implementation as baselines on translation tasks in Table 14 to further elucidate the difference through empirical results.

## K.3 Experimental Details

**NAR Transformer & CMLM** We utilize the official repository to obtain the results, with the default architecture, optimization, and inference configurations. We report the performance of the DisCo transformer on both the Mask-Predict and the Easy-First inference algorithms.

**Diffuseq-v2** For Diffuseq-v2, we employ the vocab of mBERT and choose 128 as the max length for EnDe translation. Other settings are identical to the official repository.

**SEDD** The SEDD(Lou et al., 2024) model, originally designed for open-ended text generation, is adapted in this study to facilitate conditional generation. To align the model's architecture with the specific requirements of the structured dataset, several modifications are implemented in both hyperparameters and preprocessing protocols. Specifically, the input and output token lengths are constrained to a range of 64 to 128 tokens, ensuring a more appropriate fit to the dataset's structural

characteristics. Moreover, distinct special tokens are introduced to clearly differentiate between input and output sequences, thereby enhancing the model's ability to distinguish between these components during training. Individual data entries are further demarcated by an EOS token to delineate discrete sequences within the training process.

**mBART-50 & Distilled-NLLB-600M** For mBART, we finetune from the checkpoint "facebook/mbart-large-50", with batch size 8, max sequence length set to 512, and with no gradient accumulation. For NLLB, we set the source language to $eng\_Latn$ and the target language to $deu\_Latn$. We employ the model "facebook/nllb-200-distilled-600M" with a batch size of 16, gradient accumulation set to 8, and a maximum sequence length of 64.

**DiffusionEAGS** For our model, we adopt the denosing strategy as top1 sampling and 1 size of MBR, as a typical translation task focuses on BLEU and COMET rather than diversity score.

### K.4 Experimental Results

#### K.4.1 NAR Transformer, DisCo

The results indicate that the DisCo transformer performs poorly on low-resource translation tasks, where the size of the dataset is small. The results indicated in Table 14 are much lower than those indicated in the original paper by Kasai et al. (2020).

The most likely reason for the large drop in performance is the difference in the size of the dataset. The original DisCo paper reports a BLEU score of 27.39 and 27.34, respectively, on the WMT14 EN-DE dataset. Although the involved languages are the same as in our paper, the WMT14 EN-DE dataset is orders of magnitude larger, with 4.5M pairs. Such results suggest the importance of utilizing PLMs for conditional generation tasks, especially in cases where the size of the available dataset is restricted

To account for the relatively small train set to valid/test set ratio of the dataset used in our translation experiments, which results in a high percentage of <UNK> tokens in the valid/test sets, we also provide results using the dictionary of a pre-trained mBART model (Liu, 2020). The performance benefits slightly from this change, but still lags behind those of other models.

#### K.4.2 Diffuseq-v2

It is notable that existing diffusion language models perform poorly on translation tasks. In this section, we introduce some observations that might aid our understanding of such behaviors.

For Diffuseq-v2, we conduct additional experiments using the same model trained on Paradetox. We observe that the entropy of token prediction probabilities in the translation model is orders of magnitude higher, indicating a greater level of uncertainty in its predictions. Similarly, the ratio of the nearest token distance to the average distance of the top five nearest tokens is significantly larger in the translation model. This analysis suggests that a simple rounding approach from continuous to discrete space may be insufficient for machine translation, at least in low-resource settings.

## L Ablation Examples

To concretely illustrate the impact of each component of our method, we provide representative examples as follows:

**Original**

```
- 1) nica dared her sister nola to
     jump from sandy cliff. it was
    a local swimming hole but the
    cliff was 21, she was in the
    open deep water.

- 2) nica dared her sister nola to
     jump from sandy cliff. it was
    a local swimming hole but the
    cliff was 21, she still wanted
    to jump and swim.
```

**w/o EAGS**

```
- 1) ... , she she's s one of them
     girls her sister did!

- 2) ... , there was only only way
    ! she got to a swimming!!
```

**w/o Gibbs Sampling**

```
- 1) ... , shea'' able the the her
     her her the jump!

- 2) ... , shea'' able the the her
     her her the jump!
```

w/o Pre-trained MLM

```
- 1) ... , realisescratic factions
     lightsoko lights filter
    assisted je realises unpaid
    assisted

- 2) ... , realisestarian factions
     lights rower lights filter
    assisted cove increase leap
    assisted paper
```

These examples highlight how each ablated component critically affects the fluency, coherence, and overall quality of the generated text.

## M The connection between entropy and energy

**How is the energy defined?** The sequence energy at timestep $t$ is defined as the expectation over sequences sampled from distribution $q(X^{(t)})$:

$$E_{X^{(t)}\sim q}[E_\theta(X^{(t)};Y)] = \sum_{X^{(t)}} q(X^{(t)})E_\theta(X^{(t)};Y)$$

This $q(X^{(t)})$ is the distribution from which noisy (partially masked) sequences are sampled during the forward diffusion process.

**The relation between energy and entropy** Importantly, the energy $E_\theta(X^{(t)};Y)$ itself (as defined in Equation 2 of Section 3.1) is a summation over log-potentials derived from token logits:

$$E_\theta(X;Y) = -\sum_{l=1}^{L} \log \phi_l(X;Y)$$

And specifically, the token potential is directly related to MLM logits as:

$$\log \phi_l(X;Y) = \mathbf{1}h(x_l)^T f_\theta(X_{\setminus\{x_l\}};Y)$$

where $f_\theta$ are MLM logits (confidence scores), and $\mathbf{1}h(x_l)$ is a one-hot representation. Thus, energy is directly derived from MLM logits.

**Why select high-entropy tokens?** Entropy quantifies the uncertainty of MLM predictions for a given token position:

$$H_i(X^{(t)}) = -\sum_{x'\in V} p_\theta(x'_i;X^{(t)}) \log p_\theta(x'_i;X^{(t)})$$

- High entropy → MLM is uncertain about token prediction → logits are "flat," lacking a clear high-confidence candidate.
- Low entropy → MLM predictions are peaked → clear high-confidence token emerges → low uncertainty.

High entropy tokens thus correspond precisely to high-energy states in terms of the model's energy-based formulation because uncertain predictions indicate lower log-potentials and thus higher local energy.

**How does selecting high-entropy tokens guarantee energy reduction?** When high-entropy tokens (tokens in high-energy states) are replaced with newly sampled tokens from the MLM distribution, they are replaced by candidates from a distribution that tends toward lower entropy (higher-confidence predictions) given context. Hence, the newly sampled tokens will typically yield higher log-potentials (lower local energies).

Formally, we demonstrate this via inequality (Equation 8):

$$\mathbb{E}[E_\theta(x_i^{(t)};X^{(t-1)},Y)] \le E_\theta(x_i^{(t)};X^{(t)},Y)$$

That is, the expected energy at token $x_i$ after sampling from MLM conditioned on the context (with replaced tokens from the previous step) is lower than or equal to the original energy (before replacement).

This is intuitively due to the fact that replacing uncertain predictions (high entropy) with confident ones (lower entropy) will reduce the uncertainty and thus the local energy.