Innovations and Frontiers of Diffusion Models in Natural Language Processing: A Review

Jufang Zhao[1], Zengye Su[1*], Yudan Nie[1]
[1]School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou 511363, China
*Corresponding author: szy@xs.gcc.edu.cn

Abstract

Generative AI (GenAI) has emerged as one of the most transformative forces in artificial intelligence, profoundly impacting content creation, scientific research, and numerous application domains [1, 2]. At its core, these models learn the underlying distributions from existing data to generate novel, high-quality synthetic data [3]. Within this landscape, Foundation Models play a pivotal role. These are typically large-scale models pre-trained on massive datasets, possessing powerful generalization capabilities that serve as a robust baseline for various downstream tasks, thereby significantly reducing the development cost and time for AI applications [1]. Natural Language Processing (NLP), one of the first fields where GenAI achieved major breakthroughs, has largely benefited from the development of the Transformer model [4]. Since its introduction in 2017, the attention-based Transformer architecture has demonstrated outstanding performance on tasks such as machine translation, language understanding, and text generation. This has led to the development of foundation models, particularly Pre-trained Language Models (PLMs), which have greatly enhanced the performance of text generation tasks [5]. However, traditional text generation methods, especially autoregressive (AR) models, suffer from low inference efficiency when processing long texts [5, 6]. This paper provides a comprehensive review of Diffusion Models in NLP, exploring their fundamental principles, applications, and future directions.

Index Terms— Diffusion Models, Natural Language Processing (NLP), Generative AI, Text Generation, Transformer Models, Deep Generative Models, Literature Review.

1 Introduction

In recent years, diffusion models have emerged as a novel class of deep generative models, initially achieving breakthrough success in the image generation domain. They have surpassed previous state-of-the-art (SOTA) models, such as generative adversarial networks (GANs), in their ability to generate high-fidelity and diverse samples [7–10]. The core principle of diffusion models involves a forward process that systematically adds noise to data until it conforms to a simple prior distribution (e.g., random noise), followed by a learned reverse process that gradually denoises the signal to recover a data sample [3, 11, 12].

As research has progressed, the powerful capabilities of diffusion models have been explored for applications in Natural Language Processing, where they have shown immense potential in tasks like text generation [7, 13, 14]. Compared to traditional AR models and other generative frameworks like variational autoencoders (VAEs), GANs, and normalizing flows (NFs), diffusion models exhibit several distinct advantages in NLP [5, 6, 11]. Specifically, diffusion models offer greater flexibility in handling complex conditioning, as they can iteratively refine intermediate outputs based on given inputs to more easily generate high-quality target text [5, 6]. They also demonstrate inherent capabilities for global planning and self-correction, which are crucial for generating coherent and accurate long texts [15]. Furthermore, the training of diffusion models is generally more stable than that of GANs [11]. Although vanilla diffusion models can be slow at sampling, appropriate acceleration methods allow for an effective trade-off between inference time and generation quality [5, 6]. The ascent of diffusion models has even begun to challenge the long-held view that large language models must rely on the autoregressive paradigm, as illustrated in Figure 1, suggesting that the principles of generative modeling may be the true key to language intelligence [16].

Given the rapid development of diffusion models in NLP and their unique advantages, a systematic review of current research progress holds significant academic value and offers practical insights. While some surveys on diffusion models exist, they typically cover foundational principles, algorithmic variants, or applications in specific domains. A comprehensive review focused specifically on the application of diffusion models in NLP—particularly an in-depth analysis of model architectures, training methods, diverse application scenarios, and outstanding challenges—is currently lacking [9, 13].

This review aims to fill this gap by providing researchers and practitioners with a clear, comprehensive guide to the innovations and frontiers of diffusion models in the NLP domain. We systematically survey the development history, core model architectures, and training methodologies of diffusion models in NLP. The review focuses on analyzing their application in specific NLP tasks, including text generation (both autoregressive and non-autoregressive), text editing, and cross-modal generation, while also discussing their advantages and

Figure 1: Conceptual Comparison of Text Generation Processes

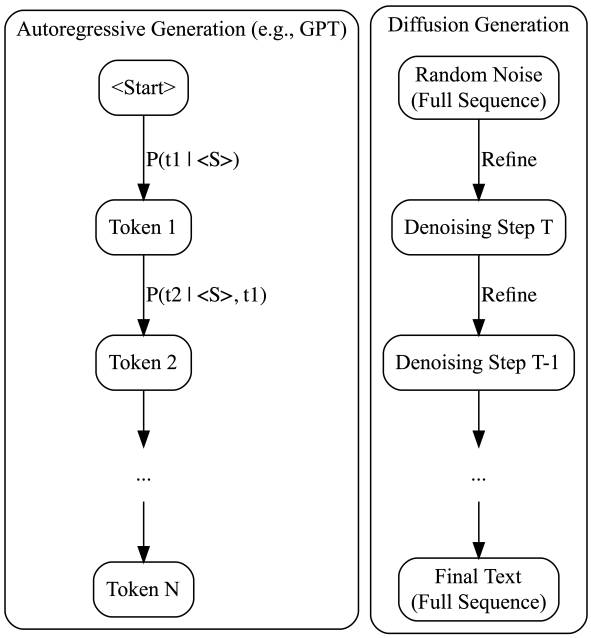

Figure 1: Conceptual comparison of text generation processes. **Left (Autoregressive):** Models like GPT generate text sequentially, predicting one token at a time. **Right (Diffusion):** Models start with random noise and iteratively refine the entire sequence in parallel.

limitations compared to traditional methods. Furthermore, we explore the integration of diffusion models with other prominent NLP models, such as Transformers and PLMs, and discuss current research challenges, including sampling efficiency, the modeling of discrete text data, and controllability. Finally, we provide an outlook on future research directions. By critically analyzing the contributions and shortcomings of existing work, this review seeks to highlight its novelty and value, thereby guiding future research and application of diffusion models in NLP.

The structure of this review is as follows: Section 2 details the foundational theory and principal variants of diffusion models. Section 3 focuses on the application of diffusion models in text generation and other NLP tasks. Section 4 discusses the integration of diffusion models with Transformer-based architectures. Section 5 covers optimization and acceleration techniques. Section 6 analyzes evaluation metrics and performance. Section 7 highlights the challenges and future outlook. Finally, Section 8 concludes the review.

2 Fundamentals of Diffusion Models

Diffusion models, an emerging class of generative models, draw inspiration from non-equilibrium thermodynamics to model complex data distributions by simulating a diffusion process [17]. (In this paper, we adopt the convention that bold lowercase letters, such as $\mathbf{x}$, denote vectors, while bold uppercase letters denote matrices). The core concept involves two processes: a forward process that progressively adds noise to an original data sample until it degenerates into a known prior distribution (typically a standard Gaussian), and a reverse process that learns to invert the forward process, gradually recovering a clean data sample from the noise [5,7,10,11,18,19]. This "corruption-and-reconstruction" paradigm provides a new framework for high-quality data generation [18].

Specifically, the forward diffusion process is modeled as a Markov chain of length T (timesteps) [5, 17]. In this process, the data $\mathbf{x}_0$ gradually evolves into noise $\mathbf{x}_T$. The transition probability $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ at each step, which describes the change from state $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$, is typically defined by the addition of Gaussian noise [11, 19]. By the Markov property, the joint probability of the entire forward process is given by Eq. (1):

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{1}$$

An important property of the forward process is that the noisy data $\mathbf{x}_t$ at any intermediate timestep t can be sampled directly from the original data $\mathbf{x}_0$. For Gaussian diffusion, this has a closed-form solution as shown in Eq. (2):

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\alpha_t = 1 - \beta_t$. The sequence of variances, $\{\beta_t\}_{t=1}^{T}$, is known as the noise schedule and is typically pre-defined to increase with t. As $t \to T$, $\bar{\alpha}_T$ approaches zero, such that $\mathbf{x}_T$ approximates a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ [11].

The reverse diffusion process aims to learn the inverse path, starting from the noise distribution $p(\mathbf{x}_T) = \mathcal{N}(0, \mathbf{I})$ and progressively denoising it to generate a data sample [11, 15]. This process is also modeled as a Markov chain, where the transition probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is approximated by a parameterized model, typically a deep neural network (the denoising network) [5, 11]. The training objective is to enable the learned reverse process to effectively recover the original data distribution from noise. This is typically achieved by maximizing the variational lower bound (VLB) on the data log-likelihood [16, 20]. In practice, the training objective is often simplified to minimizing the mean squared error (MSE) between the true added noise $\boldsymbol{\epsilon}$ and the noise predicted by the model $\boldsymbol{\epsilon}_\theta$, as shown in Eq. (3):

$$L_{\text{simple}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2\right] \tag{3}$$

This loss function trains the model $\boldsymbol{\epsilon}_\theta$ to predict the noise that was added to the original data $\mathbf{x}_0$ to produce the noisy sample $\mathbf{x}_t$.

Since the proposal of Denoising Diffusion Probabilistic Models (DDPM) [19], diffusion models have garnered widespread attention. DDPM and its variants are among the most widely used diffusion frameworks today [14, 21]. In parallel, Score-Based Generative Models (SGM) and works that

unify both approaches within a Stochastic Differential Equation (SDE) framework have provided alternative mathematical perspectives on the data perturbation and recovery process [7, 20, 22].

2.1 Principles and Variants of Diffusion Models

Diffusion Models model complex probability distributions by simulating a dual-stage process: a Forward Diffusion Process and a Reverse Diffusion Process [3, 7, 12, 23].

In the forward process, a data sample $\mathbf{x}_0$ is progressively perturbed by adding noise over T timesteps, eventually transforming it into pure noise [5, 6, 17, 19]. This is modeled as a Markov chain, where the transition probability is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{4}$$

The reverse diffusion process aims to generate new samples by starting from pure noise and progressively denoising it [5, 17, 19]. This process is also modeled as a Markov chain, implemented via a parameterized reverse transition probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which is also modeled as a Gaussian:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{5}$$

where the mean $\boldsymbol{\mu}_\theta$ and variance $\boldsymbol{\Sigma}_\theta$ are parameterized by a neural network, which typically adopts a U-Net or Transformer architecture [5, 6, 8, 11].

Variants of diffusion models differ in their principles and implementation. Denoising Diffusion Probabilistic Models (DDPM) are the canonical implementation [14, 18, 21]. Score-Based Generative Models (SGM) learn the data distribution's score function [7, 23]. Latent Diffusion Models (LDM) represent a significant advance in computational efficiency by operating in a compressed latent space [10, 18, 21, 24]. Other notable variants include Denoising Diffusion Implicit Models (DDIM), which accelerate sampling by defining a non-Markovian forward process [25].

2.2 Handling of Text Data

A core challenge in applying diffusion models to text generation is bridging the gap between the continuous-space formulation of diffusion and the discrete nature of text data [14]. Researchers have primarily pursued two categories of approaches: discrete text diffusion models and continuous text diffusion models [5, 13, 14].

The fundamental principle of a text diffusion model involves recovering a target text from a noisy input through a progressive denoising process [5, 6]. The reverse process can be generally expressed as:

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{c}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}) \tag{6}$$

where $\mathbf{c}$ represents the input condition [5, 6].

Discrete text diffusion models operate directly at the token level, generalizing the diffusion process to a discrete state space [14]. Continuous text diffusion models, conversely, encode discrete text into a continuous space where diffusion and denoising are performed [14]. Each approach has trade-offs regarding faithfulness to the data versus training stability and semantic richness [13].

2.3 Key Designs in the Diffusion Process

The performance of text diffusion models is critically influenced by four key design components: the denoising network, the noise schedule, the objective function, and the conditioning strategy [5, 6].

Denoising Network: The denoising network is the core of the reverse process. For sequential data like text, the Transformer architecture is widely adopted, as it allows the model to capture complex, long-range dependencies between tokens during the iterative refinement process [24].

Noise Schedule: The noise schedule defines the magnitude of noise added at each forward diffusion step. A well-designed schedule (e.g., linear or cosine) is crucial for generation quality, as it controls how quickly the original data signal is corrupted [19, 21].

Training Objective: The training objective is to learn the reverse denoising process, typically by minimizing the MSE between the predicted noise and the actual added noise (Eq. (3)). This simplified objective has been shown to be effective and stable for training high-quality generative models.

Conditioning Strategies: Conditioning strategies incorporate external information to guide generation. A powerful and common method is classifier-free guidance, which trains a single model to handle both conditional and unconditional generation, enabling strong, steerable synthesis at inference time [**?**, 5, 11].

3 Applications of Diffusion Models in NLP

Diffusion Models have achieved remarkable success in domains like image synthesis and are now showing significant potential to advance NLP tasks [26]. This section reviews their applications across NLP, detailing implementation methods, performance, and prospects.

3.1 Text Generation

Text generation aims to produce high-quality, coherent, and meaningful text. While traditional methods have been dominated by autoregressive models, diffusion models have recently emerged as a powerful alternative [12, 13]. Diffusion models offer unique advantages, including a natural fit for non-autoregressive (NAR) generation, better controllability, and flexible speed–quality trade-offs.

A variety of diffusion model variants have been developed for text generation. DIFFUSEQ and DIFFUSUM applied conditional diffusion to sequence-to-sequence tasks [13]. DIFFORMER, a Transformer-based model, showed strong perfor-

mance in machine translation and summarization [13]. The masked diffusion language model framework achieved state-of-the-art perplexity scores, demonstrating powerful modeling capabilities [27].

A key evaluation is comparison with autoregressive models like the GPT series. While GPT excels at fluent text generation [1], diffusion models offer complementary strengths. LLaDA, the first 8B-parameter diffusion-based large language model, demonstrates competitive performance with strong LLMs like LLaMA-7B in in-context learning [12, 16]. Notably, LLaDA mitigates the "reversal curse" seen in some AR models. However, limitations remain: some models are restricted to fixed-length text [13], and scaling diffusion models incurs significant computational cost [16].

3.2 Text Editing and Manipulation

Diffusion models are expanding into text editing and manipulation. Unlike autoregressive models, diffusion models treat editing as iterative denoising. For example, DiffusER conceptualizes edit operations as a noising process reversed by a denoising model [13]. The SUNDAE model handles arbitrary infilling within a template, providing a flexible framework for text repair [13]. This non-AR nature is advantageous for edits requiring global context. While direct quantitative comparisons with baseline editors are limited, the ability of diffusion models to perform text-guided image editing is well established, showcasing nuanced, instruction-based manipulation [25, 28, 29].

3.3 Text Representation Learning

In complex cross-modal tasks (e.g., text-to-image generation), text representations must capture fine-grained details. For instance, the Swinv2-Imagen model enhances text understanding by extracting entity and relationship embeddings from scene graphs [8]. However, the specific advantages of using diffusion models for representation learning per se remain unclear from current literature. Future research is needed to clarify the potential of diffusion models in this domain.

3.4 Machine Translation

Diffusion models have been applied to machine translation with promising results [13]. Several variants demonstrate strong translation capabilities. For example, the diffusion-based LLaDA model can effectively translate between Chinese, English, and German [16]. Other models like CDCD and SUNDAE also report high performance [13]. While these studies indicate excellent performance, they provide few details on diffusion's advantages for very long or complex sentences. Moreover, the potential of diffusion models for low-resource languages is intriguing but not yet detailed in available sources [30].

3.5 Dialogue Generation

In dialogue generation, diffusion models offer notable advantages, particularly in maintaining context and generating diverse responses [13, 16]. They effectively integrate multi-turn dialogue history; for example, LLaDA accurately captures extended conversation context [16]. The Latent Diffusion Energy-Based Model (LDEBM) is one approach that addresses issues like mode collapse by combining diffusion with an energy-based model [13]. Additionally, integrating external knowledge can further improve dialogue relevance [12].

3.6 Complex Reasoning Tasks

Diffusion models are being applied to complex reasoning in NLP. A notable development is the Diffusion of Thought (DoT) method [15], which introduces a chain-of-thought style reasoning within the diffusion framework. DoT performs reasoning by refining a sequence of latent "thought" variables in parallel over multiple steps. This allows multi-step reasoning to diffuse in parallel, offering a novel approach to tasks requiring several logical steps. DoT has been applied successfully to tasks needing sophisticated math and logic reasoning, demonstrating a powerful and novel reasoning mechanism [15].

3.7 Other Applications and Cross-Modal Fusion

Diffusion models are being extended to drive innovation in NLP and cross-modal tasks. In historical language studies, Oracle Bone Script Decipher (OBSD) uses diffusion to interpret ancient scripts [31]. In computer vision, OVDiff uses a diffusion model for open-vocabulary semantic segmentation without task-specific training [32]. Multimodal Diffusion Models, often within Multimodal LLMs, aim to process and fuse different modalities [9, 33, 34]. A common architecture uses a Transformer to create shared embeddings, which then condition a diffusion model [10, 24]. The Transfusion model enables seamless integration of discrete and continuous modalities within a single model [28].

4 Integration with Transformer Models

The Transformer architecture, with its powerful self-attention mechanism, is dominant in NLP [4, 18]. Given Transformers' prowess in sequence modeling, integrating them with diffusion models promises enhanced performance on complex generative tasks [14].

Diffusion Transformers (DiTs) exemplify this integration, replacing the typical U-Net backbone in vision diffusion models with a Transformer [24, 25]. In the multimodal domain, combining Transformers and diffusion is especially powerful. Latent Diffusion Models also often use Transformers to encode conditioning information (like text) into latent embeddings fed into the U-Net via cross-attention [10].

Table 1: Overview of Representative Diffusion Models in NLP and Related Fields.

Model (Year)	Primary Task / Domain	Key Architectural Feature	Reported Advantage / Contribution
Diffusion-LM (2022)	Unconditional Text Generation	Transformer in continuous embedding space	First to show diffusion LMs can achieve strong perplexity.
Masked Diffusion LM (2024)	Language Modeling	Transformer on discrete tokens with masking	Achieved state-of-the-art perplexity among diffusion models.
Differmer (2023)	Machine Translation, Summarization	Transformer-based denoising backbone	Competitive BLEU/ROUGE scores vs. strong Transformer baselines.
LLaDA (8B) (2024)	Instruction Following, Dialogue	Large-scale (8B param) diffusion LM	On-par with LLaMA-3 8B on dialogue benchmarks; mitigates repetition.
DoT (2024)	Mathematical/Logical Reasoning	Diffusion with parallel "thought" vectors	Outperforms AR chain-of-thought on some reasoning benchmarks.
Stable Diffusion (LDM) (2022)	Text-to-Image Generation	Diffusion in a compressed latent space (U-Net)	High efficiency and quality, enabling widespread use.

Beyond direct use, researchers are improving the Transformer architecture for diffusion. The DiffTransformer proposes a "differential attention" to reduce attention noise [4]. Its differential attention operator is calculated as:

$$\begin{aligned}\text{DiffAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \Bigg(&\text{softmax}\frac{\mathbf{Q}_1\mathbf{K}_1^T}{\sqrt{d_k}} \\ &- \lambda \cdot \text{softmax}\frac{\mathbf{Q}_2\mathbf{K}_2^T}{\sqrt{d_k}}\Bigg)\mathbf{V}\end{aligned} \tag{7}$$

where query, key, and value projections are split into two groups, and λ is a learnable scalar.

Will Transformers replace diffusion models? Current consensus is they are complementary rather than replacements [24]. Many SOTA models (LDM, DiT) combine both effectively. However, some work explores pure-Transformer generation, such as Google's Muse, which operates on discrete tokens and achieves SOTA text-to-image results efficiently without continuous diffusion [35].

5 Optimization and Acceleration

Despite their outstanding generative performance, diffusion models face challenges in computational cost and slow sampling [7, 22, 26, 36].

5.1 Sampling Acceleration

Slow sampling is a major bottleneck. Key acceleration strategies include:

- **Discretization Optimization:** Improving numerical solvers for the diffusion SDE/ODE [7, 15].
- **Non-Markovian Sampling:** Relaxing the Markov assumption to allow larger reverse steps. Denoising Diffusion Implicit Models (DDIM) [25] can cut steps from 1000 to as few as 50.
- **Distillation:** Progressive distillation trains a student model to perform two denoising steps of a teacher in one step, recursively halving inference time [7].

Additionally, efficiency improves by performing diffusion in a compressed latent space (as in LDM) [10, 21].

5.2 Maximum Likelihood Estimation Enhancement

Improving log-likelihood is crucial [7]. Methods include:

- **Noise Schedule Optimization:** Nonlinear schedules like cosine can improve performance [19, 21].
- **Objective Design:** Tailoring the loss to the task can yield better results [5].
- **Learnable Reverse Variance:** Learning the reverse process variance can improve likelihoods [7].

5.3 Model Architecture and Inference Acceleration

Refining the network architecture is another key lever [21, 24]. For inference, model compression is widely used [7, 18]:

- **Knowledge Distillation:** A large teacher model trains a smaller student model to speed up inference [7, 25].
- **Pruning and Quantization:** Techniques like QLoRA combine 4-bit quantization with low-rank adaptation for efficient fine-tuning and inference [18].

6 Evaluation Metrics and Performance Analysis

Evaluating diffusion models requires diverse metrics to assess quality, fidelity, diversity, and efficiency. For cross-modal generation (e.g., text-to-image), standard metrics are Fréchet Inception Distance (FID) and CLIP score [8, 24, 28, 35, 37]. For text generation, traditional metrics include BLEU, ROUGE, and perplexity [37]. However, these often miss nuanced qualities like global coherence or logical consistency, indicating a need for more comprehensive evaluation protocols for diffusion-generated text [12]. For tasks requiring precise correctness (like reasoning), accuracy is key [15]. Several factors influence performance including model architecture, data scale/quality, and training strategy. Despite SOTA results, diffusion models have known challenges. They can be sensitive to input noise and remain computationally intensive [33]. More critically, current automatic metrics for text often fail to capture high-level attributes of generated text [12].

7 Challenges and Future Outlook

Despite their potential, diffusion models in NLP face several challenges [6]:

- **Computational Cost and Sampling Speed:** High cost and slow generation remain prominent issues [10, 12, 20, 22, 37].
- **Discrete Data Modeling:** A fundamental mismatch exists between discrete text and continuous diffusion formulations [5, 6].
- **Interpretability and Controllability:** Diffusion processes are less interpretable, and fine-grained control remains challenging [3, 20].
- **Data, Safety, and Bias:** Diffusion models can learn societal biases or produce harmful content. Developing methods for content moderation and "detoxification" is crucial for responsible AI deployment [1, 3, 5, 6].
- **Multilingual and Low-Resource Scenarios:** Extending diffusion models to multilingual or low-resource settings is largely unexplored and will require innovative strategies.

Future Outlook: The future of diffusion models in NLP is promising, with key directions including:

1. **More Powerful and Efficient Models:** Continue scaling model size and exploring novel architectures, while developing training and sampling methods that improve efficiency [8, 11, 25].
2. **Broader NLP Applications:** Apply diffusion models to a wider range of tasks, including analytical tasks, knowledge graph construction, and structured prediction [5, 7, 22].
3. **Synergy with Other Technologies:** Deeper integrate diffusion with PLMs, combine with Transformers and knowledge graphs, and develop unified multimodal models [5, 6, 28].
4. **Advancing Language Representation:** Move toward a continuous language space for representing text, eliminating discrete tokenization limits [12].
5. **Improved Evaluation and Responsible AI:** Create more holistic evaluation benchmarks and focus on reliability, controllability, and bias mitigation to ensure safe deployment [9, 34].

The rise of foundation models and generative AI will continue to shape diffusion models' trajectory in NLP [1, 2].

8 Conclusion

This review has provided a comprehensive overview of diffusion models in NLP. As powerful generative tools [3], diffusion models have shown a remarkable ability to generate high-quality data [23]. Their innovative role and immense potential in NLP are increasingly evident [14].

Current research demonstrates significant advantages on core NLP tasks like text generation and editing [14]. Compared to AR models, diffusion models excel in parallel generation and fine-grained controllability. Advanced applications such as Diffusion-of-Thought show potential to surpass the AR paradigm for complex reasoning tasks [15]. Large-scale models like LLaDA are now competitive with traditional LLMs, while offering unique benefits like bidirectional generation [16].

However, challenges remain, including high computational costs, difficulty modeling discrete text, and interpretability and safety issues [3]. Effectively modeling diffusion for text, optimizing sampling efficiency, and better leveraging PLMs are key open questions.

Looking ahead, the potential of diffusion models in NLP is vast. Central focus will be on more efficient training and inference [37]. Architectural innovation—particularly integrating Transformers [4]—will be critical. The synergy between these technologies will likely spur novel applications in low-resource languages, advanced text analysis, and multimodal fusion [11]. In summary, diffusion models are bringing new vitality to NLP. While challenges persist, ongoing research is poised to overcome these obstacles, fully unleashing their

power to build more intelligent and creative language technologies. Ultimately, by bridging the gap between parallel and sequential processing, diffusion models are not just a new tool for NLP but a step towards more flexible and human-like language intelligence.

Funding

This work is funded by the Guangdong Provincial Sci-Tech Innovation Strategy Fund [pdjh2024a467].

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, and others. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] I. Toumi, I. Rjiba, S. Ben Abdallah, and A. M. Alimi. Generative ai: systematic review of advancements. *Multimedia Tools and Applications*, pages 1–39, 2024.

[3] Chenshuang Zhang, Chaoning Zhang, Meng Zhang, and Hedvig Kjellstrom. Text-to-image diffusion models in generative ai: A survey. *IEEE Transactions on Multimedia*, 2023.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008, 2017.

[5] Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10:e1905, 2024.

[6] Yifan Li, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Diffusion models for non-autoregressive text generation: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6298–6306. ijcai.org, 2023.

[7] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[8] Jialu Sui, Jianzong Wang, Shijing Si, Zhangcheng Huang, and Jing Xiao. Swinv2-imagen: text-to-image generation via hierarchical models. *Neural Computing and Applications*, 36(13):11119–11129, 2024.

[9] Jian-Wei Zhang, Han-Jia Chen, Jian-She Tan, Run-Ze He, Kun Zhang, Tao Qin, and Tie-Yan Liu. A survey on controllable diffusion models. *Journal of Computer Science and Technology*, 39(4):923–955, 2024.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE, 2022.

[11] Zhaoyu Chen, Yuerong Chen, Zijie Yue, Yihang Luo, Shanshan Li, Pedram Ghamisi, and Beichen Zhang. Diffusion models in remote sensing image processing: A review and outlook. *arXiv preprint arXiv:2404.08926*, 2024.

[12] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 3036–3053, 2022.

[13] Hao Zou, Zae Myung Kim, and Dongyeop Kang. A survey of diffusion models in natural language processing. *arXiv preprint arXiv:2305.14671*, 2023.

[14] Xiao Han, Sheng He, ZipeGASUS Li, and Stan Z. Li. A survey on diffusion models in nlp. *arXiv preprint arXiv:2305.14387*, 2023.

[15] Boming Pang, Chen Meng, Qing Han, and Kun He. Diffusion of thought: A new potential for llms. *arXiv preprint arXiv:2402.07754*, 2024.

[16] Hong-Yi Lin, Zhaowei Zhang, Chenghao Chi, Lichao Yu, Yunchao Wang, Haotian Liu, Chunyuan Wu, Peng Li, and Lijuan Wang. LLaDA: A Foundation Model for Bidirectional Language Generation. *arXiv preprint arXiv:2406.18349*, 2024.

[17] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015.

[18] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative models: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11674–11693, 2023.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 6840–6851, 2020.

[20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[21] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of*

the 38th International Conference on Machine Learning (ICML), pages 8162–8171, 2021.

[22] Bowen Jing, Morteza Eslami, Ezra Miller, Peter J.M. Claes, Tom Sercu, Alexander M. Rush, and Frederick P. Roth. Diffusion models are a new generation of generative ai for biology. *Nature Computational Science*, 3(11):923–933, 2023.

[23] Han Cao, Cheng Tan, Zhangyang Gao, Yitan Li, Siyuan Liu, Pin-Yu Xie, Li Zhang, Jian-Li Li, and Jian Gao. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205. IEEE, 2023.

[25] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image edit instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402. IEEE, 2023.

[26] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[27] Sheng He, Jiacheng Chen, Kang Zhou, and Zhe Zhao. Masked diffusion language models are latent variable models. *arXiv preprint arXiv:2406.11181*, 2024.

[28] Zhaowei Cai, Riza Velioglu, Yuxuan Fang, Bora Alten, Avinash Ravichandran, Anurag Arnab, Chen Sun, Ziad Al-Halah, and Stefano Soatto. Transfusion: A Unified Generative Model for Text, Image, and Multi-Modal Tasks. *arXiv preprint arXiv:2406.13110*, 2024.

[29] Michal Avrahami, Dani Lischinski, and Ohad Fried. Blended-latent-diffusion. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11. ACM, 2023.

[30] Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57, 2014.

[31] Zihan Li, Yixuan Gou, Xiang Fan, Zhiqiang Zuo, Peng Li, Ming-Hsuan Yang, Zhenglin Ye, and Cheng-Chuan Ping. Lost in translation: Reimagining the classic cipher and lost language problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1020–1034, 2024.

[32] Zeren Xu, Zhiyong Zhang, Jin Tang, and Chang Xu. Open-vocabulary semantic segmentation with diffusion models. *arXiv preprint arXiv:2307.03131*, 2023.

[33] Jiawei Zhang, Jialing Jia, Yuan Zhang, Lin Luo, Wei Wang, and En Zhang. VL-3DDet: A new paradigm for vision-language models in 3D object detection. *arXiv preprint arXiv:2405.18738*, 2024.

[34] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Zhan. A survey of multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[35] Huiwen Chang, Han Zhang, Jarred Barber, A. J. Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Gaddy, and others. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, Cham, 2015.

[37] Calvin Gao, Jiaming Li, Jun Zhu, and Maciej Bogun. T-MARS: A-Posteriori-Controllable Text-to-Image Generation. *arXiv preprint arXiv:2403.01824*, 2024.