Diffusion Language Models: Foundations and Frontiers of Non-Autoregressive Language Generation

Daisuke Oba
Institute of Science Tokyo, Japan
daisuke.oba@nlp.comp.isct.ac.jp

Danushka Bollegala
University of Liverpool, United Kingdom
danushka@liverpool.ac.uk

Abstract

Diffusion language models (DLMs) have recently emerged as a non-autoregressive approach to language generation, enabling parallel, iterative, and refinement-based decoding beyond the conventional left-to-right paradigm. This tutorial offers the NLP community a technical map of this emerging field: starting from diffusion formulations for text, we clarify their connections to autoregressive language modeling, masked language modeling, and earlier iterative generation methods. We then survey recent progress in large-scale DLMs, efficient inference, training objectives, and post-training, including reinforcement learning for reasoning and alignment. Throughout the tutorial, we treat DLMs not as a solved replacement for autoregressive language models, but as a rapidly developing paradigm whose efficiency claims, controllability, practical assessment, and deployment implications require careful examination. We close by discussing representative open directions, from broader formulations such as flow-based language modeling to reliability, safety, and social bias as key requirements for the social acceptability of DLMs.

1 Motivation and Relevance

Autoregressive language models have become the dominant paradigm for open-ended text generation, driven by Transformer architectures and their successful scaling to modern LLMs (Vaswani et al., 2017; Grattafiori et al., 2024). Yet their left-to-right factorization imposes a fundamental sequential structure: each token is generated after the previous one, even when many parts of an output could in principle be predicted, revised, or refined in parallel. Diffusion language models (DLMs) revisit this limitation in the LLM era by replacing single-pass next-token continuation with iterative denoising and refinement, allowing multiple tokens to be updated within the same generation step. This makes DLMs attractive not only for controllable or infilling-style generation, but also as a possible route toward few-step, parallel, and ultra-fast language generation.

The topic is timely because DLM research has rapidly moved from early text diffusion and discrete generative modeling toward scalable language modeling. Early work introduced continuous diffusion for controllable text generation and discrete diffusion in structured spaces (Li et al., 2022; Austin et al., 2021). More recent work has clarified how diffusion can be applied to discrete text by studying different corruption processes, denoising objectives, and interpretations of masked or discrete diffusion models (Lou et al., 2024; Ou et al., 2024; Shi et al., 2024; Sahoo et al., 2024, 2025). Together, these developments connect DLMs to masked language modeling and iterative refinement while raising broader questions about the right formulation for text generation (Devlin et al., 2019; Ghazvininejad et al., 2019).

Recent scaling results have further raised the stakes. Masked DLMs have been scaled on text, large language diffusion models such as LLaDA and Dream have attracted broad attention, and follow-up work has explored long-context generation, code generation, and hybrid autoregressive–diffusion architectures (Nie et al., 2025a,b; Ye et al., 2025; Liu et al., 2025a; Gong et al., 2025; Arriola et al., 2025; Wang et al., 2025b). In parallel, the literature has rapidly expanded along two practical axes: making DLM generation faster through improved decoding, caching, early stopping, and distillation (Ma et al., 2025; Liu et al., 2025b; Wu et al., 2025; Israel et al., 2025; Wei et al., 2025; Luxembourg et al., 2025; Deschenaux and Gulcehre, 2025; Zheng et al., 2026; Zhang et al., 2026; Oba et al., 2026a), and adapting training or post-training methods to diffusion trajectories (He et al., 2025; Zhao et al., 2025; Tang et al., 2026; Wang et al., 2026; Oba et al., 2026b). This rapid growth makes DLMs an important but difficult area for

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 44–50
November 6-10, 2026 ©2026 Association for Computational Linguistics

NLP researchers to enter.

At the same time, the central promise of DLMs is not yet settled. Practical speedups depend on the number of denoising steps, wall-clock latency, memory use, caching, sampling schedules, and quality degradation under few-step generation. These limitations are already motivating broader formulations of non-autoregressive generation, including flow-based language modeling, which suggests that current DLMs should be understood as part of a still-evolving design space rather than as a completed replacement for autoregressive LMs (Lipman et al., 2023; Lee et al., 2026; Hu et al., 2026; Deschenaux and Gulcehre, 2026). Thus, a tutorial is needed not only to explain how DLMs work, but also to clarify what they currently enable, where they remain weak, and what kinds of future formulations may address their limitations.

Finally, if DLMs become practical generation systems, their social acceptability must be studied as seriously as their efficiency. Fairness, safety, and reliability have become central concerns for autoregressive LMs, but they remain largely underexplored for diffusion-based generation. This is concerning because DLMs generate through different mechanisms—masking, denoising, confidence estimation, and iterative refinement—which may create different bias propagation and safety failure modes. This tutorial therefore positions social bias, safety, and reliability as key open problems for the NLP community, alongside technical questions about inference, training, post-training, and evaluation (Kaneko et al., 2022; Kaneko and Bollegala, 2022; Oba et al., 2024).

2 Tutorial Overview and Learning Goals

This tutorial focuses on diffusion language models as an emerging approach to non-autoregressive language generation. The core scope is DLMs themselves: their foundations, scalable model families, inference algorithms, training and post-training methods, and unresolved limitations. We discuss related frontiers such as flow-based language modeling, fairness, and safety as open directions motivated by the current limits of DLMs, rather than as separate primary topics (Lipman et al., 2023; Lee et al., 2026; Hu et al., 2026).

The tutorial has four learning goals. *First*, participants will place DLMs in context: how they depart from left-to-right language models, inherit ideas from masked language modeling and iterative refinement, and differ from earlier non-autoregressive decoding methods (Vaswani et al., 2017; Devlin et al., 2019; Ghazvininejad et al., 2019; Gu et al., 2019). *Second*, they will learn the main design choices behind diffusion over text, including absorbing-mask and uniform noising processes, denoising objectives, sampling procedures, and recent interpretations of discrete diffusion for language generation (Austin et al., 2021; Lou et al., 2024; Ou et al., 2024; Shi et al., 2024; Sahoo et al., 2024, 2025). *Third*, they will survey the current landscape: scaled DLMs, hybrid autoregressive–diffusion architectures, long-context and code models, efficient inference through caching, scheduling, and early stopping, and RL-based post-training for reasoning and alignment (Nie et al., 2025b; Ye et al., 2025; Arriola et al., 2025; Wang et al., 2025b; Liu et al., 2025a; Gong et al., 2025; Ma et al., 2025; Wu et al., 2025; Israel et al., 2025; Oba et al., 2026a; Zhao et al., 2025; Tang et al., 2026; Wang et al., 2026; Oba et al., 2026b). *Fourth*, they will learn to evaluate DLM claims critically, including speed–quality trade-offs, controllability, evaluation protocols, reliability, and broader formulations beyond current diffusion models (Li et al., 2022; Schiff et al., 2024; Zheng et al., 2023; Lipman et al., 2023; Lee et al., 2026; Hu et al., 2026). They will also examine social bias and safety as underexplored requirements for practical DLM deployment, building on prior work on bias evaluation and mitigation in language models (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024).

By the end of the tutorial, participants will be able to navigate the DLM literature, assess claims about its advantages and limitations, and identify open problems in diffusion-based text generation.

3 Technical Coverage and Reading List

The citations below indicate the representative reading list for the tutorial. A small subset will be marked as recommended pre-tutorial readings in the final tutorial materials.

Foundations and Formulations. We first situate DLMs within the broader landscape of language generation. We contrast left-to-right autoregressive modeling (Vaswani et al., 2017) with earlier non-autoregressive and iterative refinement methods, such as masked parallel decoding and edit-based generation (Ghazvininejad et al., 2019; Gu et al., 2019). We then introduce diffusion-based generation for text as iterative denoising over discrete

token sequences or latent language representations. This part covers the main design choices behind DLMs: noising processes such as absorbing-mask and uniform corruption, denoising objectives, sampling procedures, and connections to masked language modeling and discrete diffusion theory (Devlin et al., 2019; Ho et al., 2020; Austin et al., 2021; Lou et al., 2024; Ou et al., 2024; Shi et al., 2024; Sahoo et al., 2024, 2025).

Scaled DLMs and Model Families. We next survey how diffusion-based text generation has evolved from early controllable generation and masked diffusion foundations into scalable DLM families (Li et al., 2022; Sahoo et al., 2024; Nie et al., 2025a). We cover large language diffusion models such as LLaDA (Nie et al., 2025b) and Dream (Ye et al., 2025), extensions to long-context (Liu et al., 2025a) and code generation (Gong et al., 2025), and hybrid autoregressive–diffusion approaches such as Block Diffusion (Arriola et al., 2025). The goal is to clarify what capabilities current DLMs demonstrate, where they are competitive with autoregressive baselines, and where important gaps remain.

Efficient Inference and Practical Speedups. A central focus of the tutorial is whether the parallelism of DLMs can translate into practical speedups. We organize recent inference work around reducing denoising steps, reusing computation through caching or early stopping, and selecting which tokens to update in parallel (Ma et al., 2025; Liu et al., 2025b; Wu et al., 2025; Israel et al., 2025; Wei et al., 2025; Luxembourg et al., 2025; Huang and Tang, 2025; Li et al., 2025; Deschenaux and Gulcehre, 2025; Zhang et al., 2026; Zheng et al., 2026; Oba et al., 2026a). We will emphasize that DLM efficiency should be assessed by wall-clock latency, memory use, throughput, and quality degradation, not only by the number of sampling steps.

Training and Post-training. We discuss how training and post-training methods must be adapted to the generation dynamics of DLMs. Topics include the training–inference gap and refinement-oriented objectives (He et al., 2025; Oba et al., 2026c), as well as reinforcement-learning methods for reasoning and alignment over diffusion trajectories (Zhao et al., 2025; Tang et al., 2026; Wang et al., 2025a; Zhu et al., 2025; Wang et al., 2026; Zekri and Boulle, 2025; Xie et al., 2025; Oba et al., 2026b). We highlight why DLMs cannot simply inherit post-training recipes from autoregressive LMs: intermediate denoising states, non-left-to-right trajectories, and iterative refinement introduce different credit-assignment and optimization challenges.

Open Problems and Emerging Frontiers. We conclude with open problems that will shape the next stage of DLM research. Technically, current DLMs still face unresolved challenges in reliable few-step generation, speed–quality tradeoffs, length control, and controllability, motivating broader non-autoregressive formulations such as flow-based and flow-map language modeling (Lipman et al., 2023; Stark et al., 2024; Lee et al., 2026; Hu et al., 2026; Potaptchik et al., 2026; Roos et al., 2026). Socially, DLMs remain largely unexplored with respect to safety and fairness, despite the central role of these issues in autoregressive and masked language models (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024).

4 Tutorial Roadmap

Foundations and Landscape.

Why diffusion language models? Sequential bottlenecks in autoregressive decoding, parallel token updates, iterative refinement, and the promise of few-step generation.

Foundations. Diffusion over text: discrete and masked noising processes, denoising objectives, sampling procedures, and links to masked language modeling and iterative generation.

Model families and scaling. Masked diffusion LMs, large-scale DLMs, long-context and code-oriented models, and hybrid autoregressive–diffusion architectures.

Inference and Training.

Efficient inference. Turning DLM parallelism into practical speedups through adaptive decoding, caching, token updates, early stopping, and careful latency–quality measurement.

Training objectives. Denoising and refinement objectives, masking schedules, diffusion trajectories, and training–inference mismatch.

Post-training and alignment. Adapting RL and preference optimization to DLM trajectories, with emphasis on reasoning, intermediate denoising states, and credit assignment beyond left-to-right generation.

Open Frontiers and Responsible Deployment.

Beyond DLMs. Few-step degradation, speed–quality trade-offs, limitations of current diffusion formulations, and emerging flow-based formulations for language modeling.

Social acceptability and responsible deployment. Reliability, safety, social bias, fairness, multilinguality, and responsible deployment beyond autoregressive generation.

5 Diversity Considerations

The tutorial is designed to be accessible to a broad NLP audience, including participants without prior background in diffusion models. The instructor team combines junior and senior perspectives and spans institutions in Japan and the UK, bringing complementary expertise in diffusion language models, efficient generation, NLP, and fairness. In content, we will connect technical questions about DLMs to open problems in social bias, safety, and responsible deployment, rather than treating DLMs only as an efficiency technique. To support broad participation, we will make the tutorial slides and curated reading list publicly available.

6 Ethics Statement

DLMs aim to make text generation faster, more parallel, and potentially cheaper; these benefits may improve accessibility, but they may also lower the cost of misuse, spam, misinformation, and other harmful generation. In addition, the reliability, safety, and fairness properties of DLMs remain underexplored: their sampling dynamics differ from autoregressive generation, so existing findings and mitigation methods for AR LMs may not directly transfer. The tutorial itself does not introduce new datasets, user studies, or deployed systems, but it will cover research directions whose responsible development requires careful evaluation of safety, fairness and multilingual impact (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024; Kaneko et al., 2025, 2026).

Tutor Biographies

Daisuke Oba is a Specially Appointed Assistant Professor at Institute of Science Tokyo, Japan. He brings hands-on experience in large-scale autoregressive LLM development in both academia and industry. He also has recent publications at top-tier ML venues, including ICLR, ICML, and NeurIPS (Oba et al., 2026a; Komiyama et al., 2026; Oi et al., 2026; Miyamoto et al., 2026; Moon et al., 2026; Oba et al., 2026b). Most relevant to this tutorial, he has contributed directly to the emerging DLM literature through work on early stopping for DLM decoding (Oba et al., 2026a), reinforcement learning over diffusion states (Oba et al., 2026b), and refinement objectives for DLMs (Oba et al., 2026c). He has also worked on fairness in LLMs (Oba et al., 2024; Hida et al., 2026), which informs the tutorial's discussion of social bias and responsible deployment. In this tutorial, he will lead the technical coverage of DLM foundations, efficient inference, and training/post-training methods, contrasting DLMs with autoregressive LLM pipelines.
Homepage: `https://daioba.github.io/`

Danushka Bollegala is a Professor of Natural Language Processing at the University of Liverpool, UK, and an Amazon Scholar at Amazon Search. He has published over 220 papers at leading NLP and ML venues including ACL, EMNLP, NAACL, EACL, AACL, NeurIPS, ICLR, AAAI, and IJCAI, and has extensive experience instructing international audiences through invited talks, keynotes, and tutorials at major conferences and summer schools. His research spans large language models, representation learning, and the safety, fairness, and social bias properties of LLMs, including widely-cited work on evaluating and mitigating social biases in masked and large language models (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024), multilingual bias benchmarks, and ethical evaluation of human–LLM interactions (Kaneko et al., 2025, 2026). Most relevant to this tutorial, he has recently co-authored work with Daisuke Oba on diffusion language models, including an ICLR 2026 paper on efficient inference via early stopping for converged tokens in masked diffusion-LM decoding (Oba et al., 2026a), which directly informs the efficient-generation portion of this tutorial. Through his role as an Amazon Scholar at Amazon Search, he also brings an industrial perspective on how non-autoregressive paradigms such as DLMs may shape future large-scale deployment of language technologies, and is well-positioned to discuss practical implications, deployment trade-offs, and responsible-use considerations of diffusion-based generation alongside the technical content of the tutorial.
Homepage: `https://danushka.net/`

References

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Justin Deschenaux and Caglar Gulcehre. 2025. Beyond autoregression: Fast LLMs via self-distillation through time. In *The Thirteenth International Conference on Learning Representations*.

Justin Deschenaux and Caglar Gulcehre. 2026. Language modeling with hyperspherical flows. *arXiv preprint arXiv:2605.11125*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. 2025. Diffucoder: Understanding and improving masked diffusion models for code generation. *arXiv preprint arXiv:2506.20639*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. *Advances in neural information processing systems*, 32.

Haoyu He, Katrin Renz, Yong Cao, and Andreas Geiger. 2025. Mdpo: Overcoming the training-inference divide of masked diffusion language models. *arXiv preprint arXiv:2508.13148*.

Rem Hida, Masahiro Kaneko, Daisuke Oba, Danushka Bollegala, and Naoaki Okazaki. 2026. DyMT-ESB: Dynamic multi-turn evaluation of social bias in user-llm interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2026*. To appear.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. 2026. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*.

Chihan Huang and Hao Tang. 2025. Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation. *arXiv preprint arXiv:2505.14455*.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. 2025. Accelerating diffusion llms via adaptive parallel decoding. *arXiv preprint arXiv:2506.00413*.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask – evaluating social biases in masked language models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962, Vancouver, Canada. AAAI Press.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2025. An ethical dataset from real-world interactions between users and large language models. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI), AI and Social Good Track*, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2026. A multilingual social bias benchmark incorporating thinking processes. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, San Diego, USA. Association for Computational Linguistics. To appear.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Junpei Komiyama, Daisuke Oba, and Masafumi Oyamada. 2026. Best-of-infinity: Asymptotic performance of test-time LLM ensembling. In *The Fourteenth International Conference on Learning Representations*.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. 2026. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. 2025. Beyond fixed: Training-free variable-length denoising for diffusion large language models. *arXiv preprint arXiv:2508.00819*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Xiaoran Liu, Zhigeng Liu, Zengfeng Huang, Qipeng Guo, Ziwei He, and Xipeng Qiu. 2025a. Longllada: Unlocking long context capabilities in diffusion llms. *arXiv preprint arXiv:2506.14429*.

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, and Linfeng Zhang. 2025b. dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*.

Omer Luxembourg, Haim Permuter, and Eliya Nachmani. 2025. Plan for speed–dilated scheduling for masked diffusion language models. *arXiv preprint arXiv:2506.19037*.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781*.

Sora Miyamoto, Daisuke Oba, and Naoaki Okazaki. 2026. Aligning tree-search policies with fixed token budgets in test-time scaling of LLMs. In *Forty-third International Conference on Machine Learning*.

Sangwhan Moon, Daisuke Oba, Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. 2026. Beyond perplexity: UTF-8 validity in byte-aware language models. In *Forty-third International Conference on Machine Learning*.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. 2025a. Scaling up masked diffusion models on text. In *The Thirteenth International Conference on Learning Representations*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025b. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Daisuke Oba, Danushka Bollegala, Masahiro Kaneko, and Naoaki Okazaki. 2026a. Stopping computation for converged tokens in masked diffusion-LM decoding. In *The Fourteenth International Conference on Learning Representations*.

Daisuke Oba, Hiroki Furuta, and Naoaki Okazaki. 2026b. Diffusion-state policy optimization for masked diffusion language models. In *Advances in Neural Information Processing Systems*. To appear.

Daisuke Oba, Hiroki Furuta, and Naoaki Okazaki. 2026c. Drifting objectives for refining discrete diffusion language models. *arXiv preprint arXiv:2605.19470*.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1722–1742, St. Julian's, Malta. Association for Computational Linguistics.

Masanari Oi, Koki Maeda, Ryuto Koike, Daisuke Oba, Nakamasa Inoue, and Naoaki Okazaki. 2026. From correspondence to actions: Human-like multi-image spatial reasoning in multi-modal large language models. In *Forty-third International Conference on Machine Learning*.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2024. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*.

Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S Albergo. 2026. Discrete flow maps. *arXiv preprint arXiv:2604.09784*.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, Ismail Ilkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. 2026. Categorical flow maps. *arXiv preprint arXiv:2602.12233*.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T Chiu, and Volodymyr Kuleshov. 2025. The diffusion duality. In *Forty-second International Conference on Machine Learning*.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. 2024. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. 2024. Dirichlet flow matching with applications to dna sequence design. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46495–46513.

Xiaohang Tang, Rares Dolga, Sangwoong Yoon, and Ilija Bogunovic. 2026. wd1: Weighted policy optimization for reasoning in diffusion language models. In *The Fourteenth International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chenyu Wang, Paria Rashidinejad, DiJia Su, Song Jiang, Sid Wang, Siyan Zhao, Cai Zhou, Shannon Zejiang Shen, Feiyu Chen, Tommi Jaakkola, Yuandong Tian, and Bo Liu. 2026. SPG: Sandwiched policy gradient for masked diffusion language models. In *The Fourteenth International Conference on Learning Representations*.

Guanghan Wang, Gilad Turok, Yair Schiff, Marianne Arriola, and Volodymyr Kuleshov. 2025a. d2: Improved techniques for training reasoning diffusion language models. *arXiv preprint arXiv:2509.21474*.

Xu Wang, Chenkai Xu, Yijie Jin, Jiachun Jin, Hao Zhang, and Zhijie Deng. 2025b. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing. *arXiv preprint arXiv:2508.09192*.

Qingyan Wei, Yaojie Zhang, Zhiyuan Liu, Dongrui Liu, and Linfeng Zhang. 2025. Accelerating diffusion large language models with slowfast: The three golden principles. *arXiv preprint arXiv:2506.10848*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

Shaoan Xie, Lingjing Kong, Xiangchen Song, Xinshuai Dong, Guangyi Chen, Eric P Xing, and Kun Zhang. 2025. Step-aware policy optimization for reasoning in diffusion large language models. *arXiv preprint arXiv:2510.01544*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Oussama Zekri and Nicolas Boulle. 2025. Fine-tuning discrete diffusion models with policy gradient methods. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tunyu Zhang, Xinxi Zhang, Ligong Han, Haizhou Shi, Xiaoxiao He, Zhuowei Li, Hao Wang, Kai Xu, Akash Srivastava, Vladimir Pavlovic, and 1 others. 2026. T3d: Few-step diffusion language models via trajectory self-distillation with direct discriminative optimization. *arXiv preprint arXiv:2602.12262*.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. 2025. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *arXiv preprint arXiv:2504.12216*.

Haoyang Zheng, Xinyang Liu, Xiangrui Kong, Nan Jiang, Zheyuan Hu, Weijian Luo, Wei Deng, and Guang Lin. 2026. Ultra-fast language generation via discrete diffusion divergence instruct. In *The Fourteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *Preprint*, arXiv:2505.19223.