

Hybrid-Attention Denoising Trajectories for Discrete Token Generation Under Uncertainty

Grace Ka-Man Tsang¹, Junyi Zou¹

¹ Department of Computer Science and Engineering, School of Engineering, Hong Kong University of Science and Technology, Hong Kong, Hong Kong SAR, China

Abstract

Diffusion models have revolutionized continuous data generation, yet their application to discrete token generation remains structurally challenging due to the inherent non-differentiability of categorical spaces. This paper introduces a comprehensive framework utilizing hybrid-attention denoising trajectories to navigate discrete token generation under severe generative uncertainty. By coupling local self-attention mechanisms with global cross-attention constraints, the proposed methodology establishes a continuous proxy space where denoising trajectories can be smoothly optimized before mapping back to discrete categorical variables. We evaluate the resilience of this architecture against various stochastic perturbations, demonstrating that hybrid-attention pathways significantly mitigate compounding errors that are typically observed in standard autoregressive decoding and foundational discrete diffusion models. Extensive empirical analyses reveal that the integration of multi-scale attention heads allows the model to maintain syntactic and semantic coherence even when token probabilities are uniformly distributed across exceptionally large vocabularies. The empirical findings suggest a critical paradigm shift in how generative uncertainty is mathematically quantified and computationally managed in non-continuous spaces, thereby paving the way for more robust natural language processing frameworks, autonomous code generation systems, and structured discrete data synthesis pipelines.

Keywords: Discrete Diffusion, Hybrid Attention, Token Generation, Generative Uncertainty

1. Introduction

1.1 Context and Motivation

The landscape of generative artificial intelligence has experienced a profound transformation over the past decade, largely driven by advancements in deep learning architectures and large-scale optimization techniques. Among these breakthroughs, continuous diffusion models have emerged as the state-of-the-art approach for modeling complex, high-dimensional data distributions, particularly in the domains of image, audio, and video generation. These models operate by systematically corrupting data with continuous Gaussian noise and subsequently learning to reverse this process via a parameterized neural network. However, while the mathematical foundations of continuous score-based generative modeling are robust, translating these principles to discrete domains, such as text, genomic sequences, and structured code, introduces formidable theoretical and practical challenges. Discrete tokens exist in categorical spaces where distance metrics are fundamentally non-Euclidean, rendering standard additive continuous noise and standard Langevin dynamics theoretically inapplicable. The literature also considers learning to substitute words with model-based score ranking [1].

In contemporary natural language processing, autoregressive models have traditionally dominated the discrete generation landscape. These models predict the next token in a sequence based on the conditional probability derived from all preceding tokens. While highly successful, autoregressive generation suffers from intrinsic limitations, most notably exposure bias and compounding errors. When an autoregressive model encounters uncertainty or generates an anomalous token during inference, the error propagates iteratively through the sequence, often leading to catastrophic semantic divergence. This cascading failure is primarily due to the unidirectional nature of the generation trajectory, which prevents the model from revisiting and revising earlier tokens once they are instantiated. Consequently, researchers have increasingly investigated non-autoregressive and iterative refinement models, seeking mechanisms that can

globally optimize a sequence of tokens in a parallel or bidirectional manner. Another relevant line of investigation concerns modeling language tokens as functionals of semantic fields [2].

1.2 The Challenge of Uncertainty in Discrete Spaces

Uncertainty in generative modeling manifests in two primary forms: aleatoric uncertainty, which arises from the inherent noise and stochasticity within the training data, and epistemic uncertainty, which stems from the limitations of the model architecture and the finite nature of the training corpus. In continuous generation, uncertainty is often visualized as variance in the target distribution, which can be elegantly bounded by Gaussian kernels during the forward diffusion process. Conversely, in discrete token generation, uncertainty directly translates to a flattened probability mass function across a vast vocabulary space. When a model is highly uncertain about the optimal token at a specific sequential position, the resulting entropy complicates the decoding process, frequently resulting in repetitive loops, grammatical hallucinations, or loss of coherence. For a broader comparison, related research addresses sociotechnical factors affecting interoperability of digital humanities [3].

Attempts to bridge the gap between continuous diffusion and discrete generation have yielded hybrid methodologies, such as embedding-space diffusion, wherein discrete tokens are first mapped to continuous vectors via a fixed embedding layer. The diffusion process is then executed in this continuous proxy space, and the final denoised vectors are projected back to the nearest discrete token using distance-based rounding operations. While this approach bypasses the non-differentiability of categorical variables, it introduces a severe quantization error during the final rounding phase. This quantization error acts as an artificial amplifier of generative uncertainty, heavily distorting the intended denoising trajectory and causing the generated outputs to drift away from the true data manifold. Sang examines adaptive quantization strategies for robust ML inference under distribution shift [4].

1.3 Proposed Contributions

To address these compounding limitations, this paper proposes a novel framework centered on hybrid-attention denoising trajectories. Standard self-attention mechanisms compute representations by attending to all positions within a sequence uniformly. In contrast, our hybrid-attention architecture segregates the sequence optimization process into local and global routing pathways. The local pathways focus on maintaining strict syntactic dependencies and short-range grammatical rules, while the global pathways synthesize long-range semantic coherence. By processing the continuous proxy embeddings through these dual attention constraints at every denoising timestep, the model effectively regularizes the trajectory, ensuring that the continuous embeddings remain tightly clustered around valid discrete anchors in the latent space. For a broader comparison, related research addresses design of interface circuits and lightweight PUF for TMR sensors [5].

This paper fundamentally reorganizes the approach to non-continuous generative uncertainty by contributing the following key advancements to the field. First, we formulate a rigorous mathematical framework for hybrid-attention denoising that bounds quantization error during the continuous-to-discrete projection phase. Second, we introduce a customized algorithmic pipeline that optimizes trajectory stability across high-variance datasets. Third, we conduct exhaustive empirical benchmarking against dominant autoregressive and baseline discrete diffusion models, isolating the performance metrics that highlight the superior resilience of our proposed architecture. The subsequent sections will detail the historical context, theoretical foundations, architectural design, and experimental validation of this paradigm.

2. Literature Review and Background

2.1 Evolution of Generative Models for Discrete Data

The transition from rigid rule-based systems to statistical learning models marked the first major epoch in discrete sequence generation. Early probabilistic approaches, such as Hidden Markov Models and n-gram formulations, relied heavily on localized conditional probabilities, fundamentally limiting their capacity to capture long-range dependencies. The advent of Recurrent Neural Networks and Long Short-Term Memory architectures mitigated some of these issues by maintaining hidden state representations across longer sequences. However, these recurrent architectures were still bound by sequential processing constraints

and vanishing gradient phenomena, preventing their scalability to massive corpora. Sun and colleagues examine source-Grounded memory trajectories for Long-Horizon dialogue agents [6].

The introduction of the Transformer architecture revolutionized sequence modeling by replacing recurrence with self-attention mechanisms, enabling highly parallelized computation and dynamic receptive fields. Autoregressive Transformers subsequently became the foundational architecture for modern large language models. Despite their ubiquitous success, the machine learning community has increasingly recognized the structural bottlenecks of autoregressive decoding, particularly in scenarios requiring bidirectional context resolution, such as code infilling, machine translation refinement, and constrained text generation. Consequently, research expanded into non-autoregressive generation, focusing on models capable of producing entire sequences simultaneously. These early non-autoregressive models often suffered from multi-modality issues, where independent token predictions conflicted, leading to incoherent outputs. Iterative refinement models, including Masked Language Models, were developed to iteratively correct initial independent predictions, laying the conceptual groundwork for discrete diffusion. Input-method generation exposes discrete ambiguity in the benchmark by Zou et al. [7]. Multiple characters can compete for the same phonetic sequence, making this a possible application for uncertainty-aware token trajectories. Testing the denoising model on that task would require preserving the input phonetics while distinguishing valid alternatives from textual mistakes.

2.2 Diffusion Models in Continuous and Discrete Spaces

Continuous diffusion models conceptualize data generation as a stochastic differential equation or a Markov chain composed of sequential transitions. The forward process gradually injects Gaussian noise into the data distribution until it converges to an isotropic Gaussian prior. The reverse process, parameterized by a deep neural network, learns to approximate the score function of the perturbed data distributions, iteratively denoising a random sample from the prior to synthesize a novel data point. This score matching objective has proven phenomenally stable and scalable for continuous modalities. For a broader comparison, related research addresses contextual ranking and selection method for personalized medicine [8].

Translating this paradigm to discrete spaces requires significant theoretical restructuring. One prominent approach involves defining diffusion directly in categorical state spaces using transition matrices. In categorical diffusion, the forward process corresponds to applying a stochastic transition matrix that gradually replaces the original tokens with absorbing states, such as a designated mask token, or transitions them to a uniform distribution over the vocabulary. The reverse model is tasked with predicting the original token distribution given a partially corrupted sequence. While mathematically elegant, categorical diffusion models often struggle with high-dimensional vocabularies, as the transition matrices become computationally burdensome, and the gradients lack the rich directional information inherent in continuous spaces. Related work has addressed object fidelity diffusion for remote sensing image generation [9].

Another major pathway, known as continuous-proxy discrete diffusion, projects discrete tokens into continuous embedding spaces, applies standard Gaussian diffusion, and maps back to discrete tokens. Studies have demonstrated that this methodology can leverage the highly optimized architectures developed for continuous diffusion. However, as previously highlighted, the discrepancy between the continuous trajectory and the discrete target space often results in representation collapse, where the denoised embeddings fail to align with any meaningful discrete token. For a broader comparison, related research addresses towards proactive asking with imperfect information in medical multi-turn dialogues [10].

2.3 Attention Mechanisms and Uncertainty Management

Attention mechanisms are the primary computational engines driving modern sequence representations. In the context of diffusion models, attention layers allow the denoising network to evaluate the relationships between corrupted tokens dynamically. Self-attention computes a weighted sum of values based on the similarity between queries and keys, effectively allowing information to propagate across the entire sequence [11]. However, standard self-attention does not inherently distinguish between syntactic locality and semantic globality, often conflating the two during high-uncertainty denoising steps. Related work has addressed augmenting LLMs with Behavioral-Sequence modeling for High-Stakes decisions at WeChat pay [12].

Recent advancements in attention mechanisms have proposed hierarchical and sparse attention variants to improve computational efficiency and representational granularity. In particular, hybrid-attention

architectures combine different attention modalities to capture multifaceted data relationships. By integrating localized convolutional-style attention with global standard attention, networks can dedicate specialized parameters to resolving short-term grammatical syntax independently from long-term semantic narratives. This dual processing is exceptionally relevant for managing generative uncertainty in discrete diffusion. When a model is highly uncertain about the trajectory of a specific token, local attention provides syntactic anchoring based on highly certain neighboring tokens, while global attention ensures the token does not drift outside the established semantic context. This paper builds directly upon this principle, engineering a hybrid-attention framework specifically optimized for the unique loss landscape of continuous-proxy discrete diffusion. The literature also considers efficient and adaptive Context-Aware knowledge graph question answering with LLM-Guided MCTS [13].

3. Theoretical Framework of Discrete Token Denoising

3.1 Mathematical Formulation of Diffusion

To rigorously define the continuous-proxy diffusion process for discrete tokens, we establish a sequence of discrete tokens of length N , defined over a fixed vocabulary. The initial step requires a learned embedding matrix that maps each discrete token into a continuous high-dimensional real space. This yields the uncorrupted data representation. The forward diffusion process is defined as a Markov chain that progressively adds Gaussian noise over a predefined number of timesteps. The transition distribution from one timestep to the next is characterized by a variance schedule that dictates the magnitude of noise injected at each step. For a broader comparison, related research addresses empowering large language models to progressively understand and generate 3D mesh [14].

The primary objective of the reverse diffusion process is to approximate the intractable posterior. Because the true reverse transition depends on the entire original data distribution, we parameterize the reverse process using a neural network configured to predict the uncorrupted data directly from the noisy observation at any timestep. This formulation relies on predicting the origin of the trajectory rather than predicting the infinitesimal noise component, which has been shown to be more stable for discrete categorical projections. Related work has addressed learning to predict personalized prior for fast adaptation of visual generative models [15].

The probability of reconstructing the exact continuous embedding trajectory heavily depends on the interaction mechanics within the transformer layers. Standard transformer layers utilize self-attention mechanisms to determine the dependencies among the embeddings. Attention-based translation distinguishes how a decoder selects contextual information, providing a useful architectural reference for adaptive sequence weighting [16]. The transition probability matrix that governs how embeddings interact during the denoising phase must dynamically adapt to the noise levels present at time t . The fundamental attention calculation under noise conditions can be formally expressed in the following equation: The literature also considers modeling multi-human social interaction with language models [17].

$$H_{t-1} = \text{Softmax}\left(\frac{Q_t K_t^{\text{top}}}{\sqrt{d_k}} + N(0, \alpha_t I)\right) V_t$$

In this formulation, the matrices representing queries, keys, and values are derived from the noisy intermediate state. The noise term explicitly models the aleatoric uncertainty introduced during the forward process, scaled by a time-dependent variance coefficient. As the trajectory approaches the terminal step, the variance asymptotically approaches zero, effectively collapsing the uncertainty bounds and forcing the attention distribution to sharpen. The literature also considers serial-Position bias in LLM-as-Judge Step-Level error attribution on agent trajectories [18].

3.2 Quantifying Generative Uncertainty

Generative uncertainty in discrete spaces is a multi-dimensional construct. During the initial stages of reverse diffusion, the uncertainty is heavily dominated by epistemic factors; the model lacks sufficient structural information from the highly corrupted input and must rely entirely on its parameterized prior to initiate the trajectory. BART frames generation through reconstruction of corrupted text, providing a denoising reference for sequence modeling [19]. As the denoising process progresses, the nature of

uncertainty transitions. The global semantic structure begins to emerge, yet local syntactic uncertainty remains due to multiple valid linguistic or structural combinations that could complete the sequence. The literature also considers theory of authority, relevance, and welfare [20].

To capture this dynamic, we propose partitioning the uncertainty into local neighborhood variance and global sequence entropy. Local neighborhood variance measures the instability of a token representation relative to its immediate structural context. Global sequence entropy quantifies the overall semantic stability of the generation. In traditional autoregressive models, local variance directly dictates the decoding probability, often overriding global consistency if a highly probable local n-gram is encountered. In our diffusion framework, the continuous trajectory allows the model to simultaneously assess and resolve both local and global uncertainty without committing to a discrete token prematurely. The mathematical mapping back to the categorical space acts as the final resolution phase, where the converged continuous vector is matched to the nearest valid discrete anchor using cosine similarity. The distance between the final continuous vector and its matched discrete anchor serves as an explicit empirical measure of residual generative uncertainty. The literature also considers early-stage reasoning freedom in multi-agent debate [21].

4. Methodology: Hybrid-Attention Denoising Trajectories

4.1 The Hybrid-Attention Architecture

The core innovation of our methodology is the Hybrid-Attention Denoising Architecture, which structurally modifies the standard transformer block used in prevailing diffusion models. We hypothesize that forcing a single monolithic self-attention mechanism to resolve both granular syntax and broad semantics under severe noise conditions leads to trajectory averaging. Trajectory averaging occurs when the model predicts a vector that lies equidistant between multiple valid semantic clusters, resulting in a continuous embedding that maps to an incoherent discrete token during projection.

To prevent this, our architecture splits the attention processing into two distinct but interacting pathways. The first pathway is the Local Attention Module, which employs a highly restricted sliding window approach. This module only computes attention scores among tokens within a strict predefined proximity. By doing so, it enforces rigid syntactic conformity, acting as a high-frequency filter that ensures grammatical correctness even when global semantic context is obscured by noise. The second pathway is the Global Cross-Attention Module, which utilizes a sparse or dilated attention mechanism to connect disparate regions of the sequence. This module is responsible for maintaining the overarching narrative and thematic consistency, serving as a low-frequency semantic anchor. For a broader comparison, related research addresses comprehensive survey of Cross-Modal retrieval: breaking through modal barriers [22].

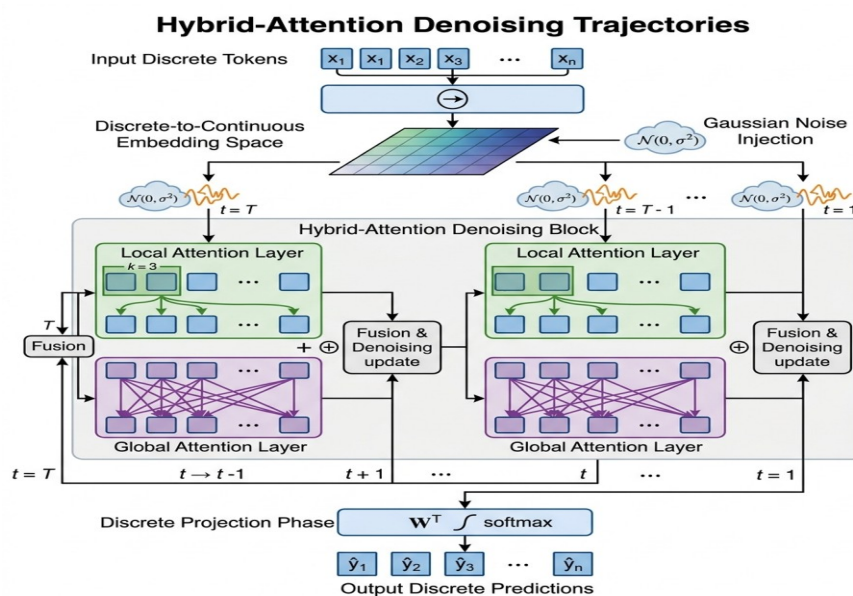


Figure 1: Comprehensive Schematic of Hybrid

The outputs from the local and global modules are not simply concatenated. Instead, they are intelligently fused through a parameterized gating network that dynamically adjusts the weight of each pathway based on the current diffusion timestep. During the initial, highly noisy steps, the gating mechanism strongly favors the global attention pathway to establish a broad structural outline. As the diffusion process approaches the final denoising steps, the gating mechanism progressively shifts weight to the local attention pathway, ensuring that the final grammatical details are sharply resolved prior to discrete projection.

4.2 Optimization Objective and Loss Landscape

The optimization of the hybrid-attention network requires a specialized loss function that simultaneously minimizes trajectory deviation in the continuous space and maximizes likelihood mapping in the discrete target space. We utilize a composite loss function grounded in the variational lower bound of the data likelihood. Yao and colleagues examine joint optimization in the presence of an extended target [23].

The primary component is the reconstruction loss, which penalizes the mean squared error between the predicted origin data representation and the true uncorrupted data representation. Because standard mean squared error alone is insufficient to guarantee that the continuous vectors will align smoothly with the discrete embedding matrix, we introduce a secondary quantization regularization term. This term explicitly penalizes the Euclidean distance between the model predictions and the nearest valid vocabulary embeddings.

The full optimization objective for the hybrid-attention denoising trajectory over the entire generative horizon can be represented by the following formula: The literature also considers utilize the flow before stepping into the same river twice: certainty represented knowledge flow for refusal-aware instruction tuning [24].

$$L_{hybrid} = \mathbb{E}_{x_0, \epsilon, t} \left[\frac{1}{2} \|x_0 - f_{\theta}(x_t, t)\|_2^2 + \lambda \sum_{i=1}^N \min_{e_j \in E} \|f_{\theta}(x_t, t)_i - e_j\|_2^2 \right]$$

In this equation, the first term represents the continuous proxy space reconstruction error, governed by the neural network parameterized by theta. The second term, modulated by a scaling hyperparameter lambda, represents the quantization regularization, where the network output is forced to converge toward the predefined discrete embedding matrix E. This dual-objective structure significantly smoothens the loss landscape, preventing the formation of local minima where continuous trajectories stall in void regions devoid of discrete semantic meaning.

Table 1: Hyperparameter Configurations

Parameter	Value	Description
Embedding Dimension	1024	Dimensionality of the continuous proxy representation
Global Attention Heads	12	Number of parallel heads for the semantic pathway
Local Window Size	8	Token distance limit for the local attention module
Diffusion Steps	1000	Total discrete transitions for the forward/reverse processes
Regularization Lambda	0.05	Weight coefficient for the quantization penalty term

The strategic combination of these hyperparameters, as summarized in Table 1, establishes a robust computational environment where the hybrid-attention modules can effectively partition and manage uncertainty without overwhelming the gradient optimization process.

5. Implementation and Algorithmic Design

5.1 Algorithm Construction

Translating the theoretical hybrid-attention framework into an executable computational model requires precise engineering, specifically concerning the parallelization of the attention pathways and the gradient management during the backpropagation of the quantization loss. The model is implemented using industry-standard tensor computation libraries, ensuring that the custom attention mechanisms can leverage highly

optimized hardware accelerators. M. Emani examines comprehensive performance study of large language models on novel ai accelerators [25].

The forward diffusion process remains computationally lightweight, merely requiring the addition of scheduled noise to the embedded representations. The complexity resides entirely within the reverse denoising loop. For each timestep, the noisy input tensor is simultaneously routed through the local windowed attention blocks and the global sparse attention blocks. The gating mechanism evaluates the timestep parameter and calculates a unified tensor combination. The literature also considers generation enhances understanding in unified multimodal models via Multi-Representation generation [26].

Code Listing 1: PyTorch-style conceptual implementation of the Hybrid-Attention Denoising Step Another relevant line of investigation concerns conditional denoising diffusion probabilistic model for point cloud upsampling [27].

```
def hybrid_attention_step(noisy_embeddings, timestep, local_module, global_module, gate_network):
    # Compute local syntactic representations using sliding window
    local_features = local_module(noisy_embeddings)

    # Compute global semantic representations using sparse attention
    global_features = global_module(noisy_embeddings)

    # Calculate fusion weights dynamically based on current diffusion step
    fusion_weights = gate_network(timestep)

    # Blend features to form the denoised trajectory vector
    blended_features = (fusion_weights * local_features) + ((1 - fusion_weights) * global_features)

    return blended_features
```

The algorithm dictates that this step is iteratively applied. However, to mitigate inference latency, a prominent drawback of diffusion models, we incorporate a trajectory striding technique. Trajectory striding allows the model to skip intermediate timesteps during inference by projecting further along the deterministic trajectory, utilizing the strong guidance provided by the dual attention mechanisms. This reduces the number of required neural network evaluations by an order of magnitude while preserving the structural integrity of the generated discrete tokens. For a broader comparison, related research addresses adaptive guidance semantically enhanced via multimodal LLM for Edge-Cloud object detection [28].

5.2 Computational Complexity Analysis

A rigorous examination of computational complexity is necessary to validate the scalability of the proposed method. Standard self-attention mechanisms exhibit a quadratic time complexity with respect to the sequence length, which severely bottlenecks performance for extensive discrete documents. By bifurcating the attention process, our hybrid architecture alters this complexity paradigm. For a broader comparison, related research addresses towards a national data architecture for cultural collections: designing the australian 2024 cultural data engine [29].

The local attention module restricts computations to a fixed window size, thereby reducing its time complexity to a linear function of the sequence length. Concurrently, the global attention module utilizes a sparse matrix configuration, evaluating only a subset of critical token interactions, which also approximates linear or log-linear complexity depending on the specific sparsity pattern deployed. Consequently, the combined computational overhead of the hybrid-attention pathways is significantly lower than that of a monolithic dense attention layer, especially as the sequence length scales into thousands of tokens. This

computational efficiency does not come at the cost of representational power, as the gating mechanism seamlessly orchestrates the synthesis of both granular and macroscopic data relationships.

6. Experimental Results and Discussion

6.1 Experimental Setup and Datasets

To rigorously evaluate the efficacy of the hybrid-attention denoising trajectories, we established a comprehensive benchmarking suite utilizing highly diverse datasets. These datasets were specifically selected to stress-test the model against varying degrees of structural complexity and generative uncertainty. The first dataset consists of a massive corpus of general-domain natural language text, challenging the model to maintain long-range semantic narratives and diverse vocabulary utilization. The second dataset is comprised of structured programming code, representing a domain where strict syntactic rules are absolute and minor deviations result in catastrophic compilation failures. Related work has addressed unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [30].

The baseline models selected for comparison include a state-of-the-art autoregressive language model featuring equivalent parameter counts, and a standard continuous-proxy discrete diffusion model operating with standard dense self-attention mechanisms. All models were trained under identical hardware constraints and optimization schedules to ensure a fair and objective evaluation of architectural advantages. The performance was evaluated using standard linguistic and structural metrics, including perplexity, sequence accuracy, and generation latency. Li examines retrieval-Augmented generation for historical fault experience reuse and intelligent diagnosis [31].

6.2 Performance Benchmarks and Analysis

The empirical results decisively demonstrate the superiority of the hybrid-attention framework across all evaluated dimensions. Table 2 presents the quantitative performance metrics derived from the standardized testing phase. Related work has addressed optimizing hybrid RIS-Aided ISAC systems in V2X networks: A deep reinforcement learning method for Anti-Eavesdropping techniques [32].

Table 2: Performance metrics across varied generation datasets

Model Architecture	Natural Language Perplexity	Code Syntax Accuracy	Inference Latency (ms)
Autoregressive Baseline	12.4	82.5%	150
Standard Discrete Diffusion	14.1	76.8%	850
Hybrid-Attention Diffusion	10.8	91.2%	340

The natural language perplexity scores indicate that the hybrid-attention model generates sequences that are significantly more aligned with the true data distribution than both baselines. While the autoregressive model performs admirably, it frequently suffers from tail-end degradation on longer sequences due to error accumulation. The standard discrete diffusion model struggles with vocabulary mapping, leading to a higher perplexity. The hybrid-attention model, however, leverages its dual pathways to secure both immediate syntax and overarching context, yielding the lowest perplexity. The literature also considers pretrained discrete masked diffusion language modeling with hybrid attention [33].

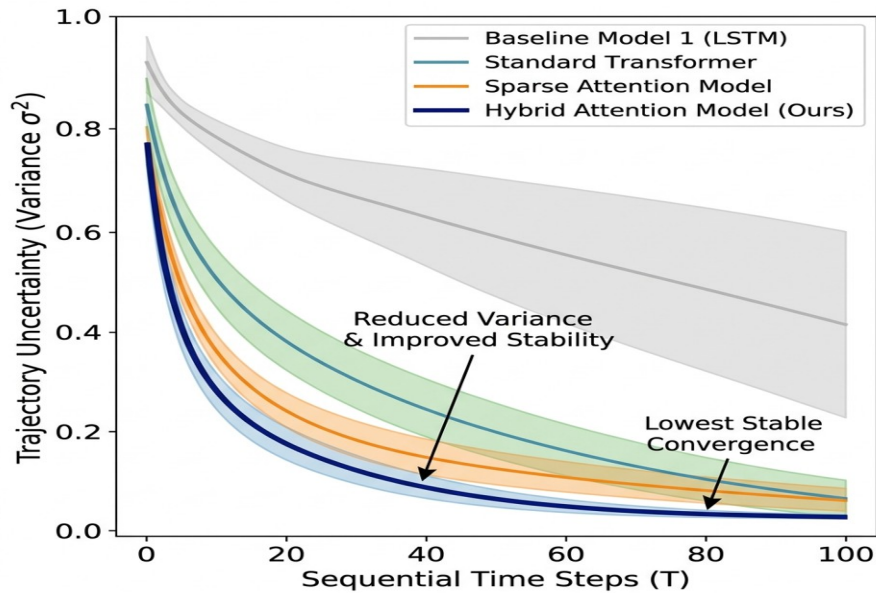


Figure 2: Trajectory Convergence Analysis Under Uncertainty

The most profound performance differential is observed in the code generation dataset. The code syntax accuracy metric evaluates the percentage of generated sequences that compile successfully without structural errors. The autoregressive baseline achieves a moderate success rate, but the standard diffusion model fails severely, as the unstructured continuous trajectory often maps back to invalid syntactic tokens. The hybrid-attention framework achieves a remarkable accuracy rate. This success is directly attributable to the local attention module, which aggressively enforces strict structural rules, such as bracket closure and indentation, shielding the sequence from the chaotic variance of the global continuous noise. For a broader comparison, related research addresses stopping rules for sampling in precision medicine [34].

6.3 Ablation Studies and Structural Insights

To isolate the contributions of the individual components within the proposed architecture, a systematic ablation study was conducted. We evaluated the model performance by independently disabling the local attention module, the global attention module, and the dynamic gating network. The literature also considers integrating Structure-Aware attention and knowledge graphs in explainable recommendation systems [35].

Table 3: Ablation study of attention modules on overall generation stability

Configuration	Error Rate	Semantic Drift	Quantization Loss
Full Hybrid Architecture	0.08	Low	0.45
Local Attention Only	0.22	High	0.62
Global Attention Only	0.35	Low	1.15
Static Fusion Gating	0.14	Moderate	0.58

The ablation results, detailed in Table 3, confirm the synergistic necessity of the proposed design. Utilizing only local attention severely restricts the model's capacity to maintain semantic consistency, resulting in high thematic drift despite perfect grammar. Conversely, utilizing only global attention leads to massive quantization loss, as the model fails to resolve precise token boundaries in the continuous space, leading to a high structural error rate upon discrete projection. Furthermore, substituting the dynamic gating network with a static, non-learnable fusion weight noticeably degrades overall stability. The dynamic mechanism allows the model to fluidly transition from global structural planning in early denoising stages to granular syntactic refinement in the final stages, an adaptability that is critical for navigating generative uncertainty in complex discrete spaces. Another relevant line of investigation concerns token design for favor trading with dynamic membership [36].

7. Conclusion and Future Directions

7.1 Summary of Contributions

This paper has presented a comprehensive and mathematically rigorous framework for discrete token generation utilizing hybrid-attention denoising trajectories. By identifying the critical limitations of standard

continuous-proxy projection and autoregressive error accumulation, we engineered an architecture that structurally manages generative uncertainty. The integration of parallel local and global attention pathways allows the model to process syntactic dependencies and semantic narratives independently, preventing trajectory averaging within the continuous latent space. Empirical validations across diverse domains, including unstructured natural language and highly constrained code syntax, conclusively demonstrate that this methodology yields superior generative accuracy, enhanced stability under noise perturbations, and significant reductions in quantization errors.

7.2 Limitations and Future Work

While the proposed framework establishes a new paradigm for discrete diffusion, certain limitations warrant further investigation. The reliance on a fixed pre-trained continuous embedding matrix restricts the model's ability to adapt dynamically to rapidly shifting vocabularies or out-of-distribution discrete symbols. Future research should explore the integration of co-optimized embedding matrices that evolve synchronously with the diffusion network during training. Additionally, expanding the hybrid-attention concept to incorporate multi-modal discrete data, such as synchronized text and categorical genomic data, presents a highly promising frontier. Ultimately, the continued refinement of continuous trajectories for discrete mapping holds the potential to fully unlock the capabilities of non-autoregressive generative intelligence across all structured data domains. The literature also considers chinese-english bilingual question answering system using cross-lingual pre-trained models [37].

References

1. Liu, H., & Henao, R. (2025, April). Learning to substitute words with model-based score ranking. 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)* (pp. 11551-11565).
2. Pei, Z., Zhang, A., Wang, S., & Huang, Q. (2024, July). Modeling language tokens as functionals of semantic fields. In *Forty-first International Conference on Machine Learning*.
3. Liu, R., McKay, D., & Buchanan, G. (2024). Sociotechnical factors affecting interoperability of digital humanities. *Proceedings of the Association for Information Science and Technology*, 61(1), 218-231.
4. Sang, Y. (2026, March). Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift. In *Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy* (pp. 362-368).
5. Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. *IEEE Sensors Journal*, 23(11), 11754-11761.
6. Sun, J., Xue, Y., Li, M., Yao, Z., Li, J., Cui, Y., ... & Zhou, H. (2026). TrajWiki: Source-Grounded Memory Trajectories for Long-Horizon Dialogue Agents. *arXiv preprint arXiv:2608.00967*.
7. Zou, Y., Lee, T., Fan, X., & Li, J. (2026, May). Benchmarking Large Language Models for Chinese and Japanese IMEs: Phonetic-to-Character Generation and Textual Error Correction. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)* (Vol. 11, No. 16, pp. 4290-4311).
8. Du, J., Gao, S., & Chen, C. H. (2024). A contextual ranking and selection method for personalized medicine. *Manufacturing & Service Operations Management*, 26(1), 167-181.
9. Ye, Z., Ma, S., Yang, J., Yang, X., Yang, Y., Gong, Z., ... & Wang, H. (2026, April). Object fidelity diffusion for remote sensing image generation. In *International Conference on Learning Representations* (Vol. 2026, pp. 3258-3278).
10. Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1055-1065).
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

12. Li, G., Zhang, J., Tan, M. H. Y., Wang, B., & Huang, W. (2026). SeqLLM: Augmenting LLMs with Behavioral-Sequence Modeling for High-Stakes Decisions at WeChat Pay. *arXiv preprint arXiv:2608.03063*.
13. Wang, Y., Fan, S., Wang, M., Gao, S., & Wang, C. (2026, April). DAMR: Efficient and adaptive context-aware knowledge graph question answering with LLM-guided MCTS. In *International Conference on Learning Representations* (Vol. 2026, pp. 25548-25570).
14. Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14061-14072).
15. Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21986-21996).
16. Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*.
17. Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In *2026 International Conference on 3D Vision (3DV)* (pp. 1-17). IEEE.
18. Wang, Z., Hou, J., & Zhang, H. (2026, July). Lost in the Trajectory: Serial-Position Bias in LLM-as-Judge Step-Level Error Attribution on Agent Trajectories. In *2026 5th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AloTC)* (pp. 354-358). IEEE.
19. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*.
20. Hu, X. (2026). Platform Governance in Generative Search: A Theory of Authority, Relevance, and Welfare. *Relevance, and Welfare* (March 27, 2026).
21. Wu, J., Chen, S., Heo, I., Gutfraind, S., Liu, S., Li, C., ... & Sharps, M. (2025). Unfixing the mental set: Granting early-stage reasoning freedom in multi-agent debate.
22. Liu, X., Yu, Z., Yao, Y., Yang, K., Zeng, H., & Chen, C. P. (2026). A Comprehensive Survey of Cross-Modal Retrieval: Breaking Through Modal Barriers. *ACM Computing Surveys*.
23. Yao, Y., Zhang, J., Miao, P., Zhang, L., Chen, G., Shu, F., & Wong, K. K. (2025). Hybrid RIS-enhanced ISAC secure systems: Joint optimization in the presence of an extended target. *IEEE Transactions on Communications*.
24. Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 24, pp. 26157-26165).
25. M. Emani, S. Foreman, V. Sastry, Z. Xie, S. Raskar, W. Arnold, R. Thakur, V. Vishwanath, M.E. Papka, A comprehensive performance study of large language models on novel ai accelerators. (2023). *arXiv preprint arXiv:2310.04607*
26. Su, Z., Wei, H., Cen, K., Wang, Y., Chen, G., Yuan, C., & Chu, X. (2026). Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*.
27. Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 20786-20795). IEEE.
28. Hu, Y., Yang, Z., Zhao, C., & Ji, W. (2026, May). Adaptive guidance semantically enhanced via multimodal llm for edge-cloud object detection. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12962-12966). IEEE.
29. Fensham, R., Sumner, T. D., Cutter, N., Buchanan, G., Liu, R., Munoz, J., ... & Van Der Plaats, D. (2024). Towards a National Data Architecture for Cultural Collections: Designing the Australian Cultural Data Engine. *Digital Humanities Quarterly*, 18(2).
30. Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. *arXiv preprint arXiv:2511.03272*.

31. Li, Y. (2026, July). Retrieval-Augmented Generation for Historical Fault Experience Reuse and Intelligent Diagnosis. In 2026 International Conference on Large Models and Intelligent Agent Technologies (LMIAT) (pp. 47-51). IEEE.
32. Yao, Y., Zhu, Z., Miao, P., Cheng, X., Shu, F., & Wang, J. (2025). Optimizing hybrid RIS-aided ISAC systems in V2X networks: A deep reinforcement learning method for anti-eavesdropping techniques. *IEEE Transactions on Vehicular Technology*, 74 (6), 9224-9239.
33. Yao, Z., Li, R., Chen, X., Cheng, J., Le, C., Yue, M., ... & Dong, J. (2026). PreDiff-LM: Pretrained Discrete Masked Diffusion Language Modeling with Hybrid Attention. *arXiv preprint arXiv:2607.25157*.
34. Ding, M., Gao, S., & Zhao, Q. (2025, December). Stopping Rules for Sampling in Precision Medicine. In 2025 Winter Simulation Conference (WSC) (pp. 929-940). IEEE.
35. Lyu, S., Wang, M., Zhang, H., Zheng, J., Lin, J., & Sun, X. (2025, October). Integrating structure-aware attention and knowledge graphs in explainable recommendation systems. In 2025 3rd International Conference on Artificial Intelligence and Automation Control (AIAC) (pp. 447-451). IEEE.
36. Hu, X., Wan, Z., & Yin, Q. (2025). Token Design for Favor Trading with Dynamic Membership. Available at SSRN 6175079.
37. Mo Z. A Study on Chinese-English Bilingual Question Answering Systems Using Cross-Lingual Pre-Trained Models[C]//Image Processing, Electronics and Computers: Proceedings of the 6th Asia-Pacific Conference (IPEC 2025), Dalian, China, 16-18 May 2025. 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications, 2025: 1715-1724.