Novelty Assessment Report

Paper: Self-Speculative Decoding Accelerates Lossless Inference in Any-Order and Any-Subset Autoregressive Models
PDF URL: https://openreview.net/pdf?id=hZnibTOke7
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

In arbitrary-order language models, it is an open question how to sample tokens in parallel from the correct joint distribution. With discrete diffusion models, the more tokens they generate in parallel, the less their predicted distributions adhere to the originally learned data distribution, as they rely on a conditional independence assumption that only works with infinitesimally small timesteps. We find that a different class of models, any-subset autoregressive models (AS-ARMs), holds the solution. As implied by the name, AS-ARMs can generate tokens in any order, and in parallel. Moreover, AS-ARMs support parallelized joint probability density estimation, allowing them to correct their own parallel-generated token distributions, via our Any-Subset Speculative Decoding (ASSD) algorithm. ASSD provably enables generation of tokens from the correct joint distribution, with the number of neural network calls upper bounded by the number of tokens predicted – notably, previous speculative decoding algorithms lack our efficiency guarantee. We empirically verify that ASSD speeds up language generation, without sacrificing quality. Furthermore, we provide a mathematically justified scheme for training AS-ARMs for generation, and show that AS-ARMs achieve state-of-the-art performance among sub-200M parameter models on infilling benchmark tasks, and nearly match the performance of models 50X larger on code generation. Our theoretical and empirical results indicate that the once-forgotten AS-ARMs are a promising direction of language modeling.

Core Task Landscape

This paper addresses: **Parallel Token Generation from Joint Distributions in Autoregressive Models**

A total of **19 papers** were analyzed and organized into a taxonomy with **13 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Any-Order and Any-Subset Autoregressive Architectures**
- **Diffusion-Based Parallel Generation Methods**
- **Dynamic Multi-Token Prediction Strategies**
- **Autoregressive Models with Parallel Inference Mechanisms**
- **Domain-Specific Autoregressive Applications with Parallel Components**
- **Survey and Methodological Overview Literature**
- **Counterfactual Explanation Methods Using Sampling**

Complete Taxonomy Tree

- Parallel Token Generation from Joint Distributions in Autoregressive Models Survey Taxonomy
- Any-Order and Any-Subset Autoregressive Architectures
  - Any-Subset Autoregressive Models with Speculative Decoding ★ (2 papers)
  - [0] Self-Speculative Decoding Accelerates Lossless Inference in Any-Order and Any-Subset Autoregressive Models (Anon et al., 2026) View paper
  - [3] Reviving Any-Subset Autoregressive Models with Principled Parallel Sampling and Speculative Decoding (Ermon, 2025) View paper
  - Any-Order Generation Without Speculative Correction (1 papers)
  - [1] -GPTs: A New Approach to Autoregressive Models (A Pannatier, 2024) View paper
- Diffusion-Based Parallel Generation Methods
  - Adaptive Parallel Decoding in Diffusion LLMs (1 papers)
  - [2] Accelerating Diffusion LLMs via Adaptive Parallel Decoding (Israel, 2025) View paper
  - Guided Autoregressive Diffusion for Sequential Data (1 papers)
  - [6] Guided autoregressive diffusion models with applications to PDE simulation (Bergamin, 2024) View paper
- Dynamic Multi-Token Prediction Strategies
  - Confidence-Based Dynamic Token Sampling (1 papers)
  - [8] DynaMo: Accelerating language model inference with dynamic multi-token sampling (Tuli, 2024) View paper
  - Pseudo-Autoregressive Span-Based Generation (1 papers)
  - [7] Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis (Yifan Yang, 2025) View paper
- Autoregressive Models with Parallel Inference Mechanisms
  - Langevin Dynamics for Parallel Autoregressive Sampling (1 papers)
  - [17] Parallel and Flexible Sampling from Autoregressive Models via Langevin Dynamics (Jayaram, 2021) View paper
  - Transformer-Based Amortized Joint Inference (1 papers)
  - [14] Efficient autoregressive inference for transformer probabilistic models (Hassan, 2025) View paper

- Domain-Specific Autoregressive Applications with Parallel Components
  - Vision and Multimodal Autoregressive Generation (5 papers)
  - [11] NEP: Autoregressive Image Editing via Next Editing Token Prediction (Wu Huimin, 2025) View paper
  - [13] AR-RAG: Autoregressive Retrieval Augmentation for Image Generation (Qi Jingyuan, 2025) View paper
  - [15] Improving Autoregressive Image Generation through Coarse-to-Fine Token Prediction (Guo, 2025) View paper
  - [18] Recursive Autoregressive Depth Estimation with Continuous Token Modeling (J Zhang, n.d.) View paper
  - [19] SMART-3D: Scaling Masked AutoRegressive Transformer for Efficient 3D Shape Generation (S Mo, n.d.) View paper
  - Motion Prediction and Temporal Sequence Modeling (1 papers)
  - [12] AMP: Autoregressive Motion Prediction Revisited with Next Token Prediction for Autonomous Driving (Jia, 2024) View paper
  - Database Query Estimation and Switching Dynamics (3 papers)
  - [5] Asm: Harmonizing autoregressive model, sampling, and multi-dimensional statistics merging for cardinality estimation (Kyoung-Min Kim, 2024) View paper
  - [9] Bayesian nonparametric inference of switching dynamic linear models (E. Fox, 2011) View paper
  - [16] Discrete Autoregressive Switching Processes in Sparse Graphical Modeling of Multivariate Time Series Data (Hadj-Amar, 2024) View paper
- Survey and Methodological Overview Literature (1 papers)
  - [4] Make every token count: A systematic survey on decoding methods for foundation models (Haoran Wang, 2025) View paper
- Counterfactual Explanation Methods Using Sampling (1 papers)
  - [10] MCCE: Monte Carlo sampling of valid and realistic counterfactual explanations for tabular data (Annabelle Redelmeier, 2021) View paper

Narrative

Core task: parallel token generation from joint distributions in autoregressive models. The field addresses the fundamental tension between the sequential nature of autoregressive generation and the desire for faster, parallel inference. The taxonomy reveals several complementary strategies: Any-Order and Any-Subset Autoregressive Architectures explore flexible factorizations that permit generating tokens in arbitrary orders or subsets, enabling speculative or adaptive decoding schemes. Diffusion-Based Parallel Generation Methods blend diffusion processes with autoregressive structures to sample multiple tokens jointly. Dynamic Multi-Token Prediction Strategies focus on predicting variable numbers of tokens per step, adapting generation granularity on the fly. Autoregressive Models with Parallel Inference Mechanisms encompass techniques like speculative decoding and lookahead methods that maintain the autoregressive framework while parallelizing computation. Domain-Specific Applications demonstrate these ideas in specialized contexts such as audio codecs or retrieval-augmented generation, while Survey and Methodological Overview Literature and Counterfactual Explanation Methods provide broader perspectives and niche applications of sampling-based reasoning.

A particularly active line of work centers on any-subset autoregressive models, which train networks to handle arbitrary token subsets and enable speculative decoding without auxiliary draft models. Self-Speculative Decoding[0] exemplifies this approach by using the model's own any-subset capabilities to propose and verify multiple tokens in parallel, closely aligning with Any-Subset Autoregressive[3], which formalizes the theoretical underpinnings of subset-based factorizations. In contrast, Adaptive Parallel Decoding[2] and DynaMo[8] emphasize dynamic adjustment of the number of tokens predicted per step, trading off between parallelism and accuracy based on model confidence. Meanwhile, diffusion-inspired methods like Guided Autoregressive Diffusion[6] and pseudo-autoregressive approaches such as Pseudo-autoregressive Codec[7] explore hybrid generation paradigms that relax strict left-to-right ordering. Self-Speculative Decoding[0] sits squarely within the any-subset branch, sharing the flexible factorization philosophy of Any-Subset Autoregressive[3] but distinguished by its self-contained speculative mechanism, avoiding the overhead of separate draft models seen in some parallel inference strategies.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Reviving Any-Subset Autoregressive Models with Principled Parallel Sampling and Speculative Decoding

Authors: Ermon, Stefano, Gabe Guo, Stefano Ermon | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

In arbitrary-order language models, it is an open question how to sample tokens in parallel from the correct joint distribution. With discrete diffusion models, the more tokens they generate in parallel, the less their predicted distributions adhere to the originally learned data distribution, as they rely on a conditional independence assumption that only works with infinitesimally small timesteps. We find that a different class of models, any-subset autoregressive models (AS-ARMs), holds the s...

⚠ Similarity Notice

This paper appears to be highly similar to the original paper; both describe Any-Subset Speculative Decoding (ASSD) for AS-ARMs with identical technical contributions including the same theoretical guarantees and empirical results. The titles, abstracts, and core technical content are nearly identical, suggesting this may be a variant or near-duplicate. Please manually verify.

Contributions Analysis

Overall novelty summary. The paper proposes Any-Subset Speculative Decoding (ASSD), an algorithm enabling parallel token generation from correct joint distributions using any-subset autoregressive models (AS-ARMs). It resides in the 'Any-Subset Autoregressive Models with Speculative Decoding' leaf, which contains only two papers including this one. This represents a relatively sparse research direction within the broader taxonomy of parallel generation methods, suggesting the work addresses a focused problem space where few prior solutions exist. The sibling paper in this leaf shares the any-subset architecture philosophy but differs in its speculative correction mechanism.

The taxonomy reveals neighboring approaches in adjacent branches: 'Any-Order Generation Without Speculative Correction' explores flexible factorizations without correction mechanisms, while 'Diffusion-Based Parallel Generation Methods' and 'Dynamic Multi-Token Prediction Strategies' pursue parallel sampling through fundamentally different paradigms—iterative denoising and confidence-based adaptive prediction, respectively. The paper's position bridges architectural flexibility (any-subset capability) with algorithmic guarantees (speculative decoding), distinguishing it from purely architectural contributions in sibling leaves and from diffusion methods that rely on conditional independence assumptions the authors explicitly critique.

Among 30 candidates examined, the ASSD algorithm contribution shows no clear refutation across 10 examined papers, suggesting novelty in the specific speculative decoding formulation with efficiency guarantees. The training scheme contribution similarly lacks refutable prior work among 10 candidates. However, the architectural design criteria contribution encountered one refutable candidate among 10 examined, indicating some overlap with existing AS-ARM architectural principles. The limited search scope (30 total candidates, not hundreds) means these findings reflect top-semantic-match results rather than exhaustive coverage, particularly relevant given the sparse two-paper leaf this work occupies.

Based on the top-30 semantic matches examined, the work appears to introduce novel algorithmic contributions (ASSD with provable efficiency bounds) within an emerging architectural paradigm (AS-ARMs). The single refutable pair for architectural criteria suggests partial overlap with foundational AS-ARM design principles, while the algorithm and training scheme show no clear precedent in the examined literature. The sparse taxonomy leaf and limited sibling papers reinforce that this represents early-stage exploration of a specific solution space.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Any-Subset Speculative Decoding (ASSD) algorithm

Description: The authors introduce ASSD, a novel algorithm that enables parallel token generation from any-subset autoregressive models while maintaining the correct joint distribution. The algorithm is mathematically guaranteed to never increase the number of function evaluations and can handle exponentially more infilling patterns than traditional speculative decoding.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Speculative decoding for multi-sample inference

URL: View paper

Brief Assessment

Multi-Sample Speculative[39] focuses on multi-sample inference scenarios (self-consistency, best-of-n) by exploiting consensus across parallel reasoning paths, whereas ASSD addresses any-subset autoregressive models with arbitrary infilling patterns. The candidate does not challenge ASSD's novelty in enabling parallel token generation from any-subset models with joint distribution guarantees.

---

2. A unified framework for speculative decoding with multiple drafters as a bandit

URL: View paper

Brief Assessment

Multiple Drafters Bandit[43] focuses on selecting among multiple specialized drafters using multi-armed bandit algorithms for speculative decoding, rather than enabling parallel token generation from any-subset autoregressive models with joint distribution guarantees as in ASSD.

---

3. Fast inference from transformers via speculative decoding

URL: View paper

Brief Assessment

Fast Speculative Decoding[41] focuses on left-to-right autoregressive models (standard transformers), not any-subset autoregressive models. The candidate addresses a fundamentally different problem space with different architectural requirements and guarantees.

---

4. SWIFT: On-the-Fly Self-Speculative Decoding for LLM Inference Acceleration

URL: View paper

Brief Assessment

SWIFT[44] focuses on layer-skipping within a single LLM for left-to-right generation, while ASSD addresses any-subset autoregressive models with arbitrary infilling patterns and joint distribution guarantees through a different mathematical framework.

---

5. DySpec: Faster speculative decoding with dynamic token tree structure

URL: View paper

Brief Assessment

DySpec[37] focuses on dynamic token tree structures for standard left-to-right autoregressive models, while ASSD addresses any-subset autoregressive models with arbitrary infilling patterns. The technical approaches and problem domains are fundamentally different.

---

6. Spectr: Fast speculative decoding via optimal transport

URL: View paper

Brief Assessment

Spectr[35] focuses on optimal transport-based speculative decoding for standard left-to-right autoregressive models, not any-subset autoregressive models that support arbitrary infilling patterns with exponentially more tasks.

---

7. Accelerating Large Language Model Decoding with Speculative Sampling

URL: View paper

Brief Assessment

Speculative Sampling[36] focuses on left-to-right autoregressive models using a draft model to generate continuations, while ASSD addresses any-subset autoregressive models with arbitrary-order token generation and exponentially more infilling patterns (O(2^n) vs O(n)).

---

8. DistillSpec: Improving Speculative Decoding via Knowledge Distillation

URL: View paper

Brief Assessment

DistillSpec[38] focuses on improving speculative decoding through knowledge distillation for left-to-right autoregressive models, not on enabling parallel token generation from any-subset autoregressive models with joint distribution guarantees as ASSD does.

---

9. Fast Large Language Model Collaborative Decoding via Speculation

URL: View paper

Brief Assessment

Collaborative Decoding[42] focuses on accelerating collaborative decoding between multiple models during inference, not on parallel token generation from any-subset autoregressive models with joint distribution guarantees as in the original paper.

---

10. Beyond tokens: A survey on decoding methods for large language models and large vision-language models

URL: View paper

Brief Assessment

Beyond Tokens Survey[40] appears to be a survey paper discussing speculative decoding methods broadly. The provided context fragments mention drafting multiple tokens and contrasting logits but lack sufficient technical detail to demonstrate prior work that would refute ASSD's novelty claims regarding any-subset generation with joint distribution guarantees.

Contribution 2: Mathematically justified training scheme for AS-ARMs

Description: The authors develop a principled training objective based on joint conditional probability maximization with expectations over token orderings and prompt lengths. This training scheme is derived from reversing a discrete-time Markov chain and differs from conditionally independent losses used in prior work.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Rethinking Discrete Tokens: Treating Them as Conditions for Continuous Autoregressive Image Synthesis

URL: View paper

Brief Assessment

Discrete Tokens Conditions[33] focuses on image generation using discrete tokens as conditions for continuous autoregressive synthesis, not on training schemes for any-subset autoregressive models with joint conditional probability maximization over token orderings.

2. PRADA: Probability-Ratio-Based Attribution and Detection of Autoregressive-Generated Images

URL: View paper

Brief Assessment

PRADA[34] focuses on detecting autoregressive-generated images through probability ratio analysis, not on training schemes for autoregressive models with joint conditional probability maximization over token orderings.

3. Autoregressive Conditional Neural Processes

URL: View paper

Brief Assessment

Autoregressive Conditional Processes[28] focuses on autoregressive deployment of Conditional Neural Processes at test time without modifying training, using standard maximum likelihood. The original paper's training scheme involves reversing a discrete-time Markov chain with expectations over token orderings and prompt lengths, which is fundamentally different from CNP training objectives.

4. MotionLM: Multi-Agent Motion Forecasting as Language Modeling

URL: View paper

Brief Assessment

MotionLM[29] focuses on multi-agent motion forecasting using discrete motion tokens with a language modeling objective, not on training schemes for any-subset autoregressive models or joint conditional probability maximization over token orderings.

5. Joint Document-Level Event Extraction via Token-Token Bidirectional Event Completed Graph

URL: View paper

Brief Assessment

Token-Token Event Graph[32] focuses on document-level event extraction using graph-based methods for entity-event relationships, not on training autoregressive models with joint conditional probability maximization over token orderings.

6. Xlnet: Generalized autoregressive pretraining for language understanding

URL: View paper

Brief Assessment

XLNet[26] focuses on permutation language modeling for bidirectional pretraining in NLP, not on training schemes for any-subset autoregressive models with joint conditional probability maximization over token orderings and prompt lengths as described in the original paper.

7. Reviving Any-Subset Autoregressive Models with Principled Parallel Sampling and Speculative Decoding

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

8. Randomized Autoregressive Visual Generation

URL: View paper

Brief Assessment

Randomized Visual Generation[27] focuses on visual generation with randomized autoregressive modeling using annealing schedules for factorization orders, while the original paper develops a training scheme for any-subset autoregressive models based on reversing discrete-time Markov chains with joint conditional probability maximization. The technical approaches and mathematical foundations differ substantially.

9. Token-Shuffle: Towards High-Resolution Image Generation with Autoregressive Models

URL: View paper

Brief Assessment

Token-Shuffle[30] focuses on reducing image token redundancy in autoregressive image generation through spatial token merging, not on training objectives for any-order autoregressive models with joint conditional probability maximization over token orderings.

10. SutraNets: Sub-series Autoregressive Networks for Long-Sequence, Probabilistic Forecasting

URL: View paper

Brief Assessment

SutraNets[31] focuses on autoregressive forecasting for time series using sub-series decomposition, not on training schemes for any-order/any-subset autoregressive models with joint conditional probability maximization over token orderings.

Contribution 3: Architectural design criteria for AS-ARMs supporting parallel sampling and density estimation

Description: The authors establish architectural requirements for AS-ARMs that enable both parallel token generation through arbitrary positional queries and single-step joint density estimation via causal-like attention masking. These design principles allow AS-ARMs to serve as both draft and oracle models simultaneously.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation

URL: View paper

Brief Assessment

GraphAF[20] focuses on molecular graph generation using flow-based autoregressive models for chemistry applications, not on general architectural design principles for any-subset autoregressive models that enable parallel token generation and joint density estimation in language modeling.

---

2. Set block decoding is a language model inference accelerator

URL: View paper

Brief Assessment

Set Block Decoding[24] focuses on combining next token prediction with masked token prediction in a single transformer architecture for inference acceleration, not on establishing architectural requirements for any-subset autoregressive models that enable both parallel generation and single-step joint density estimation.

---

3. Accelerating Diffusion LLMs via Adaptive Parallel Decoding

URL: View paper

Brief Assessment

Adaptive Parallel Decoding[2] focuses on diffusion language models (DLLMs) with adaptive parallel token generation using multiplicative mixtures, not on any-subset autoregressive model (AS-ARM) architectural design criteria for attention masking and positional queries.

---

4. Efficient autoregressive inference for transformer probabilistic models

URL: View paper

Brief Assessment

Efficient Autoregressive Inference[14] focuses on transformer-based probabilistic models for meta-learning tasks with set-based conditioning, not on any-subset autoregressive models (AS-ARMs) for language generation with arbitrary-order token prediction.

---

5. -GPTs: A New Approach to Autoregressive Models

URL: View paper

Prior Art Analysis

GPTs Autoregressive[1] demonstrates that prior work exists on architectural designs enabling both parallel token generation and single-step density estimation in any-order autoregressive models. The candidate paper presents σ-GPT, which uses double positional encodings to enable arbitrary-order generation with parallel sampling capabilities and conditional density estimation in a single forward pass. This directly addresses the same architectural requirements claimed as novel by the original paper: supporting parallel token generation through arbitrary positional queries and enabling joint density estimation via causal-like attention masking.

Evidence

Evidence 1 - **Rationale**: Both papers describe architectures that enable parallel token generation and density estimation. The candidate demonstrates this capability was already achieved in σ-GPT. - **Original**: we seek any-subset autoregressive models (as-arms) that can support ourany-subset speculative decoding (assd)algorithm (which is fully described in section 5). to recap, the criteria are: (1) generates arbitrarily-ordered tokens in parallel; (2) evaluates joint density with only one forward pass. - **Candidate**: our method allows making conditional density estimation of the rest of the sequence. it is capable of making predictions all over the task space conditioned on any known subpart of the task. this can be done by prompting the model with the known part of the sequence and then decoding, in parallel an...

Evidence 2 - **Rationale**: Both papers describe mechanisms for arbitrary positional queries. The candidate's double positional encoding enables the same parallel sampling capability through arbitrary position specification. - **Original**: parallel sampling via arbitrary positional queries:with one function evaluation, we should be able to simultaneously predict in parallel (conditionally independent) distributions for all the masked tokens xσ(≥m), conditioned on the prompt xσ(<m). this allows the network to act as a quick "draft" mod... - **Candidate**: to be able to model sequences in any order, each token needs to have information about its position and the one of the next token in the shuffled sequence. specifically, when handling a sequence of tokens alongside a given permutation σ, every token contains three distinct pieces of information: its...

Evidence 3 - **Rationale**: Both papers emphasize the capability for joint density estimation. The candidate explicitly demonstrates this was achievable in σ-GPT, which predates the original paper's claims. - **Original**: density estimation via causal-like attention masking:another crucial ingredient of speculative decoding is density estimation - this allows the "oracle" model to correct the mistakes of the draft model. as such, discrete diffusion models trained with an elbo (lou et al., 2023; sahoo et al., 2024; de... - **Candidate**: our method allows making conditional density estimation of the rest of the sequence. it is capable of making predictions all over the task space conditioned on any known subpart of the task. this can be done by prompting the model with the known part of the sequence and then decoding, in parallel an...

Evidence 4 - **Rationale**: Both papers describe architectural requirements for single-pass density estimation. The candidate's σ-GPT architecture achieves this through causal masking with shuffled sequences, demonstrating prior implementation of these design criteria. - **Original**: care must be taken, however, to pick an architecture that evaluates the joint density of a sequence in one function evaluation,i.e., o(s) time. some architectures (shih et al., 2022; hoogeboom et al., 2021) take o(s∗n) steps, as only logits of masked tokens are predicted at each function evaluation ... - **Candidate**: we propose a novel approach for training autoregressive models, which involves doing next-token prediction on a shuffled input sequence. we presentσ-gpt, where σ denotes the permutation used to shuffle the sequence, and by gpt we mean any causal transformer encoder (or causal transformer decoder wit...

---

6. Reviving Any-Subset Autoregressive Models with Principled Parallel Sampling and Speculative Decoding

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

7. Clarinet: Parallel wave generation in end-to-end text-to-speech

URL: View paper

Brief Assessment

Clarinet[25] focuses on parallel waveform generation for speech synthesis using Gaussian inverse autoregressive flows, not on general autoregressive model architectures supporting arbitrary-order token generation and joint density estimation as in the original paper.

---

8. Parallel sampling via counting

URL: View paper

Brief Assessment

Parallel Sampling Counting[21] addresses parallel sampling via counting oracles for arbitrary distributions on product spaces, not architectural design criteria for autoregressive models. The candidate focuses on algorithmic complexity of sampling through counting queries rather than model architecture requirements.

---

9. Inference Acceleration of Autoregressive Normalizing Flows by Selective Jacobi Decoding

URL: View paper

Brief Assessment

Selective Jacobi Decoding[22] focuses on accelerating autoregressive normalizing flows through iterative optimization methods, not on establishing architectural design criteria for any-subset autoregressive models that enable parallel token generation and density estimation.

---

10. Chunked autoregressive gan for conditional waveform synthesis

URL: View paper

Brief Assessment

Chunked Autoregressive GAN[23] focuses on audio waveform synthesis using autoregressive chunking for GANs, not on general architectural requirements for any-subset autoregressive models that enable both parallel token generation and single-step density estimation.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Reviving Any-Subset Autoregressive Models with Principled Parallel Sampling and Speculative Decoding

Detected in: Core Task (sibling), Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Self-Speculative Decoding Accelerates Lossless Inference in Any-Order and Any-Subset Autoregressive Models View paper
- [1] -GPTs: A New Approach to Autoregressive Models View paper
- [2] Accelerating Diffusion LLMs via Adaptive Parallel Decoding View paper
- [3] Reviving Any-Subset Autoregressive Models with Principled Parallel Sampling and Speculative Decoding View paper
- [4] Make every token count: A systematic survey on decoding methods for foundation models View paper
- [5] Asm: Harmonizing autoregressive model, sampling, and multi-dimensional statistics merging for cardinality estimation View paper
- [6] Guided autoregressive diffusion models with applications to PDE simulation View paper
- [7] Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis View paper
- [8] DynaMo: Accelerating language model inference with dynamic multi-token sampling View paper
- [9] Bayesian nonparametric inference of switching dynamic linear models View paper
- [10] MCCE: Monte Carlo sampling of valid and realistic counterfactual explanations for tabular data View paper
- [11] NEP: Autoregressive Image Editing via Next Editing Token Prediction View paper
- [12] AMP: Autoregressive Motion Prediction Revisited with Next Token Prediction for Autonomous Driving View paper
- [13] AR-RAG: Autoregressive Retrieval Augmentation for Image Generation View paper
- [14] Efficient autoregressive inference for transformer probabilistic models View paper
- [15] Improving Autoregressive Image Generation through Coarse-to-Fine Token Prediction View paper
- [16] Discrete Autoregressive Switching Processes in Sparse Graphical Modeling of Multivariate Time Series Data View paper
- [17] Parallel and Flexible Sampling from Autoregressive Models via Langevin Dynamics View paper
- [18] Recursive Autoregressive Depth Estimation with Continuous Token Modeling View paper
- [19] SMART-3D: Scaling Masked AutoRegressive Transformer for Efficient 3D Shape Generation View paper
- [20] GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation View paper
- [21] Parallel sampling via counting View paper
- [22] Inference Acceleration of Autoregressive Normalizing Flows by Selective Jacobi Decoding View paper
- [23] Chunked autoregressive gan for conditional waveform synthesis View paper
- [24] Set block decoding is a language model inference accelerator View paper
- [25] Clarinet: Parallel wave generation in end-to-end text-to-speech View paper
- [26] Xlnet: Generalized autoregressive pretraining for language understanding View paper
- [27] Randomized Autoregressive Visual Generation View paper
- [28] Autoregressive Conditional Neural Processes View paper
- [29] MotionLM: Multi-Agent Motion Forecasting as Language Modeling View paper
- [30] Token-Shuffle: Towards High-Resolution Image Generation with Autoregressive Models View paper
- [31] SutraNets: Sub-series Autoregressive Networks for Long-Sequence, Probabilistic Forecasting View paper
- [32] Joint Document-Level Event Extraction via Token-Token Bidirectional Event Completed Graph View paper
- [33] Rethinking Discrete Tokens: Treating Them as Conditions for Continuous Autoregressive Image Synthesis View paper
- [34] PRADA: Probability-Ratio-Based Attribution and Detection of Autoregressive-Generated Images View paper
- [35] Spectr: Fast speculative decoding via optimal transport View paper
- [36] Accelerating Large Language Model Decoding with Speculative Sampling View paper
- [37] DySpec: Faster speculative decoding with dynamic token tree structure View paper
- [38] DistillSpec: Improving Speculative Decoding via Knowledge Distillation View paper

- [39] Speculative decoding for multi-sample inference View paper
- [40] Beyond tokens: A survey on decoding methods for large language models and large vision-language models View paper
- [41] Fast inference from transformers via speculative decoding View paper
- [42] Fast Large Language Model Collaborative Decoding via Speculation View paper
- [43] A unified framework for speculative decoding with multiple drafters as a bandit View paper
- [44] SWIFT: On-the-Fly Self-Speculative Decoding for LLM Inference Acceleration View paper