Novelty Assessment Report

Paper: Locality-aware Parallel Decoding for Efficient Autoregressive Image Generation
PDF URL: https://openreview.net/pdf?id=h06l9w1clt
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

We present Locality-aware Parallel Decoding (LPD) to accelerate autoregressive image generation. Traditional autoregressive image generation relies on next-patch prediction, a memory-bound process that leads to high latency. Existing works have tried to parallelize next-patch prediction by shifting to multi-patch prediction to accelerate the process, but only achieved limited parallelization. To achieve high parallelization while maintaining generation quality, we introduce two key techniques: (1) Flexible Parallelized Autoregressive Modeling, a novel architecture that enables arbitrary generation ordering and degrees of parallelization. It uses learnable position query tokens to guide generation at target positions while ensuring mutual visibility among concurrently generated tokens for consistent parallel decoding. (2) Locality-aware Generation Ordering, a novel schedule that forms groups to minimize intra-group dependencies and maximize contextual support, enhancing generation quality. With these designs, we reduce the generation steps from 256 to 20 (256×256 res.) and 1024 to 48 (512×512 res.) without compromising quality on the ImageNet class-conditional generation, and achieving at least 3.4× lower latency than previous parallelized autoregressive models.

Core Task Landscape

This paper addresses: **Accelerating Autoregressive Image Generation through Parallel Decoding**
A total of **48 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

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

- **Spatial Locality-Based Parallel Decoding**
- **Hierarchical and Multi-Scale Autoregressive Modeling**
- **Block-Based and Semi-Autoregressive Decoding**
- **Random-Order and Flexible-Order Autoregressive Modeling**
- **Speculative and Iterative Parallel Decoding**
- **Masked and Non-Autoregressive Generative Modeling**
- **Retrieval-Augmented and Context-Aware Generation**
- **Variational and Latent Space Autoregressive Models**
- **Bidirectional and Multi-Token Prediction Architectures**
- **Inference Optimization and System-Level Acceleration**
- ... and 2 more categories

Complete Taxonomy Tree

- Accelerating Autoregressive Image Generation through Parallel Decoding Survey Taxonomy
- Spatial Locality-Based Parallel Decoding
  - Column-Wise and Diagonal Parallel Decoding (3 papers)
  - [1] Zipar: Accelerating autoregressive image generation through spatial locality (Yefei He, 2024) View paper
  - [13] ZipAR: Parallel Auto-regressive Image Generation through Spatial Locality (He, 2024) View paper
  - [20] Fast Autoregressive Video Generation with Diagonal Decoding (Ye Yang, 2025) View paper
  - Neighboring and Outpainting-Based Decoding (2 papers)
  - [4] Neighboring autoregressive modeling for efficient visual generation (He, 2025) View paper
  - [28] Learning to Expand Images for Efficient Visual Autoregressive Modeling (Ruiqing Yang, 2025) View paper
  - Flexible Parallelized Autoregressive Modeling ★ (2 papers)
  - [0] Locality-aware Parallel Decoding for Efficient Autoregressive Image Generation (Anon et al., 2026) View paper
  - [2] Parallelized autoregressive visual generation (Yuqing Wang, 2025) View paper
- Hierarchical and Multi-Scale Autoregressive Modeling
  - Coarse-to-Fine Token Prediction (2 papers)
  - [3] Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction (Liu Yiheng, 2025) View paper
  - [8] Improving Autoregressive Image Generation through Coarse-to-Fine Token Prediction (Guo, 2025) View paper
  - Next-Scale and Multi-Scale Prediction (4 papers)
  - [23] MVAR: Visual Autoregressive Modeling with Scale and Spatial Markovian Conditioning (Zhang Jinhua, 2025) View paper
  - [25] Collaborative decoding makes visual auto-regressive modeling efficient (Chen Zigeng, 2025) View paper
  - [26] LSRS: Latent Scale Rejection Sampling for Visual Autoregressive Modeling (Hong-Kai Zheng, 2025) View paper
  - [43] Multi-scale Autoregressive Models are Laplacian, Discrete, and Latent Diffusion Models in Disguise (Steve Hong, 2025) View paper
  - Hierarchical Transformers and Local Parallel Generation (2 papers)
  - [19] Cogview2: Faster and better text-to-image generation via hierarchical transformers (Ding Ming, 2022) View paper

- [30] Locally hierarchical auto-regressive modeling for image generation (T You, 2022) View paper
- Block-Based and Semi-Autoregressive Decoding
  - Next-Block and Next-Set Prediction (2 papers)
  - [5] Next Block Prediction: Video Generation via Semi-Autoregressive Modeling (Ren, 2025) View paper
  - [7] Blockwise parallel decoding for deep autoregressive models (Stern, 2018) View paper
  - L-Shape and Structured Block Decoding (1 papers)
  - [39] Lformer: Text-to-Image Generation with L-shape Block Parallel Decoding (Li, 2023) View paper
- Random-Order and Flexible-Order Autoregressive Modeling (2 papers)
  - [9] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders (Ziqi Pang, 2024) View paper
  - [10] Autoregressive Image Generation with Randomized Parallel Decoding (Li Haopeng, 2025) View paper
- Speculative and Iterative Parallel Decoding
  - Speculative Jacobi Decoding (3 papers)
  - [18] Accelerating Auto-regressive Text-to-Image Generation with Training-free Speculative Jacobi Decoding (Yao Teng, 2024) View paper
  - [32] MC-SJD : Maximal Coupling Speculative Jacobi Decoding for Autoregressive Visual Generation Acceleration (So, 2025) View paper
  - [42] SJD++: Improved Speculative Jacobi Decoding for Training-free Acceleration of Discrete Auto-regressive Text-to-Image Generation (Yao Teng, 2025) View paper
  - Speculative Denoising and Grouped Decoding (2 papers)
  - [34] Speculative Jacobi-Denoising Decoding for Accelerating Autoregressive Text-to-image Generation (Yao Teng, 2025) View paper
  - [36] Grouped Speculative Decoding for Autoregressive Image Generation (So, 2025) View paper
- Masked and Non-Autoregressive Generative Modeling
  - Masked Generative Transformers (4 papers)
  - [6] Muse: Text-To-Image Generation via Masked Generative Transformers (Chang, 2023) View paper
  - [22] Resurrect mask autoregressive modeling for efficient and scalable image generation (Xin Yi, 2025) View paper
  - [38] Improved Masked Image Generation with Knowledge-Augmented Token Representations (Guotao Liang, 2025) View paper
  - [41] Text-Conditioned Sampling Framework for Text-to-Image Generation with Masked Generative Models (JaeWoong Lee, 2023) View paper
  - Masked Autoregressive Models with Caching (1 papers)
  - [37] LazyMAR: Accelerating Masked Autoregressive Models via Feature Caching (Wei Qing-yan, 2025) View paper
  - Discrete Diffusion and Absorbing Processes (2 papers)
  - [14] Muddit: Liberating Generation Beyond Text-to-Image with a Unified Discrete Diffusion Model (Shi Qingyu, 2025) View paper
  - [17] Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes (Sam Bond-Taylor, 2022) View paper
- Retrieval-Augmented and Context-Aware Generation (1 papers)
  - [11] AR-RAG: Autoregressive Retrieval Augmentation for Image Generation (Qi Jingyuan, 2025) View paper
- Variational and Latent Space Autoregressive Models (1 papers)
  - [29] Parallelizing Autoregressive Generation with Variational State Space Models (Lambrechts, 2024) View paper
- Bidirectional and Multi-Token Prediction Architectures (1 papers)
  - [35] BIGFix: Bidirectional Image Generation with Token Fixing (Besnier, 2025) View paper
- Inference Optimization and System-Level Acceleration (3 papers)
  - [16] Flover: A Temporal Fusion Framework for Efficient Autoregressive Model Parallel Inference (Jinghan Yao, 2023) View paper
  - [31] Parallel Vision Token Scheduling for Fast and Accurate Multimodal LMMs Inference (Wengyi Zhan, 2025) View paper
  - [33] Image autoregressive interpolation model using GPU-parallel optimization (Jiaji Wu, 2017) View paper
- Domain-Specific and Cross-Modal Autoregressive Models
  - 3D Shape and Octree-Based Generation (1 papers)
  - [15] Octgpt: Octree-based multiscale autoregressive models for 3d shape generation (Siâ□□Tong Wei, 2025) View paper
  - Video and Temporal Autoregressive Modeling (1 papers)
  - [40] SAMPO: Scale-wise Autoregression with Motion PrOmpt for generative world models (Wang sen, 2025) View paper
  - Multimodal Vision-Language Generation (1 papers)
  - [45] MAGVLT: Masked Generative Vision-and-Language Transformer (Sungwoong Kim, 2023) View paper
- Theoretical Foundations and General Frameworks (8 papers)
  - [12] Beyond tokens: A survey on decoding methods for large language models and large vision-language models (Haoran Wang, 2025) View paper
  - [21] Apar: Llms can do auto-parallel auto-regressive decoding (Zeng, 2024) View paper
  - [24] Parallel multiscale autoregressive density estimation (Reed Scott, 2017) View paper
  - [27] Exploring stochastic autoregressive image modeling for visual representation (Qi Yu, 2023) View paper
  - [44] Anime Generation through Diffusion and Language Models: A Comprehensive Survey of Techniques and Trends (Yujie Wu, 2025) View paper
  - [46] Autoregression with Self-Token Prediction (D Chen, n.d.) View paper
  - [47] Enhancing Image Generation with Diffusion Transformer Architecture (Ruiyang Wu, n.d.) View paper
  - [48] Recursive Autoregressive Depth Estimation with Continuous Token Modeling (J Zhang, n.d.) View paper

Narrative

Core task: accelerating autoregressive image generation through parallel decoding. The field addresses the inherent sequential bottleneck of autoregressive models by exploring diverse strategies to predict multiple tokens simultaneously. The taxonomy reveals a rich landscape organized around twelve major branches. Spatial Locality-Based Parallel Decoding exploits the natural correlation among neighboring image patches to enable concurrent predictions, as seen in works like Zipar[1] and Parallelized Autoregressive Visual[2]. Hierarchical and Multi-Scale Autoregressive Modeling decomposes generation into coarse-to-fine stages, while Block-Based and Semi-Autoregressive Decoding groups tokens into chunks for batch processing. Random-Order and Flexible-Order approaches relax strict raster-scan dependencies, and Speculative and Iterative Parallel Decoding methods draft multiple candidates in parallel before verification. Masked and Non-Autoregressive branches draw inspiration from diffusion and masked language models, whereas Retrieval-Augmented and Context-Aware Generation incorporates external knowledge. Additional branches cover variational latent models,

bidirectional architectures, system-level optimizations, domain-specific extensions, and theoretical foundations, reflecting the breadth of innovation in this space.

Several active lines of work highlight contrasting trade-offs between generation quality, speed, and architectural complexity. Spatial locality methods such as Neighboring Autoregressive Modeling[4] and Next Block Prediction[5] achieve strong speedups by predicting contiguous regions, yet must carefully balance parallelism with maintaining coherence across boundaries. Speculative techniques and iterative refinement offer flexible acceleration but introduce verification overhead. Locality Parallel Decoding[0] sits within the Spatial Locality-Based branch under Flexible Parallelized Autoregressive Modeling, emphasizing adaptive parallel prediction guided by local dependencies. Compared to neighbors like Parallelized Autoregressive Visual[2], which also leverages spatial structure, Locality Parallel Decoding[0] appears to focus on dynamic locality-aware scheduling rather than fixed block partitioning. This positioning reflects a broader trend toward flexible, content-adaptive parallelization strategies that aim to preserve autoregressive quality while unlocking substantial inference speedups.

Related Works in Same Category

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

1. Parallelized autoregressive visual generation

Authors: Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Autoregressive models have emerged as a powerful approach for visual generation but suffer from slow inference speed due to their sequential token-by-token prediction process. In this paper, we propose a simple yet effective approach for parallelized autoregressive visual generation that improves generation efficiency while preserving the advantages of autoregressive modeling. Our key insight is that parallel generation depends on visual token dependenciesâ□□tokens with weak dependencies can be ...

Relationship Analysis

Both papers belong to the Flexible Parallelized Autoregressive Modeling category, addressing parallel decoding for autoregressive image generation through flexible generation orderings. The original paper (LPD) introduces learnable position query tokens and a locality-aware scheduling algorithm to enable arbitrary generation orders with mutual visibility among concurrent tokens, while the candidate paper (PAR) focuses on a simpler cross-region token grouping strategy that generates initial tokens sequentially then predicts spatially distant tokens in parallel using group-wise bidirectional attention. The key difference is that LPD employs a specialized architecture with position queries and an algorithmic scheduling approach based on proximity metrics, whereas PAR maintains standard decoder-only transformers with a fixed region-based partitioning scheme and learnable transition tokens.

Contributions Analysis

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

Contribution 1: Flexible Parallelized Autoregressive Modeling

Description: The authors introduce a novel architecture that decouples context representation from token generation by using learnable position query tokens. This design enables arbitrary generation order and degrees of parallelization while maintaining mutual visibility among concurrently generated tokens through specialized attention mechanisms, and inherits KV caching to avoid redundant computation.

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. Ar-diffusion: Auto-regressive diffusion model for text generation

URL: View paper

Brief Assessment

Ar Diffusion[57] focuses on text generation using diffusion models with position-dependent timesteps, not on parallelized autoregressive image generation with learnable position query tokens and flexible generation ordering as in the original paper.

---

2. Qwen2.5-Omni Technical Report

URL: View paper

Brief Assessment

Qwen2.5-Omni[56] focuses on multimodal perception and generation with streaming capabilities, not on parallelized autoregressive modeling with flexible generation ordering and learnable position tokens for image generation.

---

3. Larp: Tokenizing videos with a learned autoregressive generative prior

URL: View paper

Brief Assessment

Larp[60] focuses on video tokenization with learned queries for holistic representations, not on parallelized autoregressive modeling with flexible generation ordering. The candidate addresses a different problem domain (video tokenization) rather than parallel decoding architectures for image generation.

---

4. Leapformer: Enabling linear transformers for autoregressive and simultaneous tasks via learned proportions

URL: View paper

Brief Assessment

Leapformer[63] focuses on linearizing transformer attention mechanisms through learned proportions for re-weighting, not on parallelized autoregressive image generation with flexible ordering and learnable position query tokens for concurrent token generation.

---

5. Symbol-rooted cascade propagation in contextual memory routing for large language models

URL: View paper

Brief Assessment

Symbol Rooted Cascade[62] focuses on contextual memory routing mechanisms for semantic thread management in LLMs, not on parallelized autoregressive image generation with learnable position query tokens and flexible generation ordering.

---

6. RandAR: Decoder-only Autoregressive Visual Generation in Random Orders

URL: View paper

Prior Art Analysis

RandAR[9] demonstrates prior work on flexible generation ordering in decoder-only autoregressive models. Both papers address the same core problem: enabling arbitrary generation orders in autoregressive visual generation. RandAR[9] achieves this through position instruction tokens that specify spatial locations before each predicted token, while the original paper uses learnable position query tokens. Both approaches decouple positional information from the generation process to enable non-raster ordering. RandAR[9] was trained on randomly permuted token sequences and supports parallel decoding with KV-cache, demonstrating that flexible ordering with parallelization was already established in prior work.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce decoder-only models with arbitrary generation orders. RandAR[9] explicitly states it removes the predefined generation order constraint, directly challenging the original paper's novelty claim of being the first to enable flexible ordering. - **Original**: we introduce locality-aware parallel decoding(lpd), a framework that consists of a novel flexible parallelized autoregressive modeling architecture and a novel localityaware generation order schedule. we design a new modeling architecture as conventional decoder-only autoregressive models struggle w... - **Candidate**: we introduce randar, a decoder-only visual autoregressive (ar) model capable of generating images in arbitrary token orders. unlike previous decoder-only ar models that rely on a predefined generation order, randar removes this inductive bias, unlocking new capabilities in decoder-only generation.

Evidence 2 - **Rationale**: Both papers use position-based tokens to enable flexible ordering. RandAR[9]'s position instruction tokens serve the same fundamental purpose as the original paper's position query tokens - specifying target positions for generation in arbitrary orders. - **Original**: This is achieved by using learnable position query tokens to guide the model in generating tokens at target positions. Moreover, the generation is parallel-aware, as we leverage specialized attention mechanism to ensure mutual visibility among tokens generated concurrently. notably, our design also ... - **Candidate**: our essential design enables random order by inserting a "position instruction token" before each image token to be predicted, representing the spatial location of the next image token. trained on randomly permuted token sequences - a more challenging task than fixed-order generation, randar achieve...

---

7. STAR: Scale-wise Text-conditioned AutoRegressive image generation

URL: View paper

Brief Assessment

STAR[61] focuses on scale-wise autoregressive generation with multi-scale token representations, not flexible generation ordering with learnable position query tokens for arbitrary parallelization within a single flat token space.

---

8. VideoGPT: Video Generation using VQ-VAE and Transformers

URL: View paper

Brief Assessment

VideoGPT[58] uses standard autoregressive modeling with sequential token generation, not flexible parallelized generation with learnable position query tokens or arbitrary generation ordering.

---

9. Insertion transformer: Flexible sequence generation via insertion operations

URL: View paper

Brief Assessment

Insertion Transformer[59] focuses on insertion-based sequence generation for machine translation with arbitrary orderings, while the original paper addresses image generation with learnable position query tokens and specialized attention mechanisms for parallel decoding with KV caching.

---

10. Autoregressive Image Generation with Randomized Parallel Decoding

URL: View paper

Prior Art Analysis

Randomized Parallel Decoding[10] demonstrates that similar architectural innovations for flexible parallelized autoregressive modeling with learnable position tokens existed prior to the original paper. Both papers decouple context representation from token generation using position-aware query mechanisms and enable arbitrary generation ordering with mutual visibility among concurrently generated tokens. The candidate paper explicitly describes using data-independent [mask] tokens with positional information as target-aware queries that attend to content key-value pairs, achieving the same core functionality of flexible generation order and parallelization while maintaining causal attention patterns and KV caching capabilities.

Evidence

Evidence 1 - **Rationale**: Both papers describe decoupling positional guidance from content representation using separate query tokens to enable arbitrary generation ordering and parallelization. - **Original**: flexible parallelized autoregressive modeling, a novel architecture that enables arbitrary generation ordering and degrees of parallelization. it uses learnable position query tokens to guide generation at target positions while ensuring mutual visibility among concurrently generated tokens for cons... - **Candidate**: we propose a noveldecoupled decodingframework that decouples positional guidance from content representation, encoding them separately as queries and key-value pairs. By directly incorporating this guidance into the causal attention mechanism, our approach enables fully random-order training and gen...

Evidence 2 - **Rationale**: Both papers describe using position-aware query tokens (learnable position query tokens vs. [mask] tokens with positional information) to decouple context representation from token generation. - **Original**: Our core idea is to decouple the context representation and token generation by leveraging separate tokens. we illustrate this in figure 3 (b). in this formulation, previously generated tokens are encoded to provide context and the generation is driven by learnable position query tokens correspondin... - **Candidate**: in theposition-guided prediction pass, data-independent[mask]tokens endowed with positional information corresponding to a right-shift of the input, act as target-aware queries. These queries use causal cross-attention to predict their respective target tokens based on the content key-value pairs pr...

Evidence 3 - **Rationale**: Both papers describe mechanisms ensuring mutual visibility among concurrently generated tokens to enable consistent parallel decoding. - **Original**: query attention ensures mutual visibility among the position query tokens within the same step, and prevents any subsequent tokens from attending to the query tokens. - **Candidate**: this design allows the model to be trained within a fully causal paradigm while enabling generalization to block-wise parallel decoding with flexible token orders, as multiple queries can be processed independently in a single step.

Evidence 4 - **Rationale**: Both papers explicitly describe inheriting KV caching mechanisms to avoid redundant computation during parallel generation. - **Original**: notably, our design also inherits the kv caching mechanism, avoiding redundant computation. - **Candidate**: at inference time, we first compute the kv cache from the known tokens using the self-attention in pass-1. next, we select multiple target-aware queries. these queries can simultaneously attend to the kv cache via cross-attention in pass-2, thereby enabling multi-token prediction within a single inf...

Contribution 2: Locality-aware Generation Ordering

Description: The authors propose a generation order schedule guided by two principles: selecting target positions spatially close to existing context for strong conditioning, and ensuring concurrently generated tokens are spatially distant to reduce mutual dependency. This schedule leverages spatial locality patterns observed in autoregressive image generation attention.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Next patch prediction for autoregressive visual generation

URL: View paper

Brief Assessment

Next Patch Prediction[51] focuses on grouping image tokens into patches to reduce training costs through a multi-scale coarse-to-fine strategy, not on generation ordering schedules using spatial locality patterns. The candidate addresses training efficiency via patch aggregation, while the original contribution concerns scheduling concurrent token generation based on spatial proximity principles.

---

2. Mimt: Masked image modeling transformer for video compression

URL: View paper

Brief Assessment

Mimt[52] focuses on video compression using masked image modeling for entropy coding with iterative decoding based on token confidence/entropy. The original paper addresses autoregressive image generation with spatial locality principles for parallel decoding groups, which is a fundamentally different application domain and technical approach.

---

3. Toward Improving the Generation Quality of Autoregressive Slot VAEs

URL: View paper

Brief Assessment

Autoregressive Slot VAEs[50] focuses on learning object generation orders for compositional scene generation in slot-based VAEs, not on spatial locality patterns for general autoregressive image generation. The technical domains and objectives differ fundamentally.

---

4. RichControl: Structure- and Appearance-Rich Training-Free Spatial Control for Text-to-Image Generation

URL: View paper

Brief Assessment

RichControl[55] focuses on spatial control in text-to-image diffusion models through feature injection schedules, not autoregressive image generation ordering. The paper addresses sampling schedules for condition features in diffusion models, which is fundamentally different from the autoregressive patch generation ordering proposed in the original paper.

---

5. Autoregressive Image Generation with Linear Complexity: A Spatial-Aware Decay Perspective

URL: View paper

Brief Assessment

Spatial Aware Decay[54] focuses on linear attention mechanisms with spatial-aware decay for computational efficiency in autoregressive image generation, not on generation ordering schedules that leverage spatial locality patterns. The candidate addresses attention computation complexity rather than the scheduling of token generation order.

---

6. HMAR: Efficient Hierarchical Masked Auto-Regressive Image Generation

URL: View paper

Brief Assessment

HMAR[53] focuses on hierarchical masked autoregressive modeling with next-scale prediction conditioned on immediate predecessors, not on spatial locality-based generation ordering schedules for flat token representations as in the original paper.

---

7. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis

URL: View paper

Brief Assessment

Nuwa Infinity[49] focuses on patch-level autoregressive generation for infinite visual synthesis using a 'render-and-optimize' strategy with nearby context pools. While it discusses ordering patches and spatial relationships, it does not propose a generation ordering schedule based on spatial locality principles (maximizing proximity to existing context while minimizing proximity among concurrent tokens) as the original paper does.

Contribution 3: Locality-aware Parallel Decoding Framework

Description: The authors present a complete framework combining flexible parallelized autoregressive modeling with locality-aware generation ordering to significantly reduce generation steps (from 256 to 20 for 256×256 resolution and 1024 to 48 for 512×512 resolution) while maintaining generation quality and achieving at least 3.4× lower latency than previous parallelized autoregressive models.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Parallel multiscale autoregressive density estimation

URL: View paper

Prior Art Analysis

Parallel Multiscale Autoregressive[24] demonstrates that parallel decoding frameworks for autoregressive image generation existed prior to the original paper's work. The candidate paper presents a multiscale approach that reduces generation steps from O(N) to O(log N) by modeling pixel groups as conditionally independent, achieving orders of magnitude speedup while maintaining quality. This directly challenges the novelty claim of the original paper's locality-aware parallel decoding framework, as both papers address the same fundamental problem of accelerating autoregressive generation through parallelization and both achieve significant step reduction (the candidate reduces from sequential pixel-by-pixel to O(log N), while the original reduces from 256 to 20 steps). The candidate's approach of forming pixel groups to minimize dependencies and maximize contextual support parallels the original's locality-aware generation ordering principle.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same core problem: sequential autoregressive generation is slow and memory-bound. The candidate paper already proposed parallelization through conditional independence modeling, which is the fundamental approach

also used in the original paper. - **Original**: we present locality-aware parallel decoding(lpd) to accelerate autoregressive image generation. traditional autoregressive image generation relies on nextpatch prediction, a memory-bound process that leads to high latency. existing works have tried to parallelize next-patch prediction by shifting to... - **Candidate**: pixelcnn achieves state-of-the-art results in density estimation for natural images. although training is fast, inference is costly, requiring one network evaluation per pixel; o(n) for n pixels. this can be sped up by caching activations, but still involves generating each pixel sequentially. in th...

Evidence 2 - **Rationale**: The candidate paper already achieved orders of magnitude speedup and enabled 512×512 generation through parallel decoding, demonstrating that the acceleration benefits claimed by the original paper were previously achieved. - **Original**: with these designs, we reduce the generation steps from 256 to 20 (256 x256 res.) and 1024 to 48 (512 x512 res.) without compromising quality on the imagenet class-conditional generation, and achieving at least 3.4x lower latency than previous parallelized autoregressive models. - **Candidate**: our new pixelcnn model achieves competitive density estimation and orders of magnitude speedup -o(log n) sampling instead ofo(n) - enabling the practical generation of 512→512 images.

Evidence 3 - **Rationale**: Both papers recognize the same locality principle: nearby pixels are highly dependent and should not be generated independently. The candidate paper already identified this insight and designed their grouping strategy around it. - **Original**: we observe strong spatial locality in image generation attention where tokens predominantly attend to nearby regions as shown in figure 2. this indicates a high dependency among nearby tokens, meaning that spatially closer tokens provide stronger conditioning. recent works (wang et al., 2024b; besni... - **Candidate**: ideally we would generate multiple pixels in parallel, which could greatly accelerate sampling. in the autoregressive framework this only works if the pixels are modeled as independent. thus we need a way to judiciously break weak dependencies among pixels; for example immediately neighboring pixels...

Evidence 4 - **Rationale**: The candidate paper's approach of forming pixel groups to break weak dependencies while maintaining strong contextual support is conceptually equivalent to the original paper's locality-aware generation ordering. - **Original**: with these insights, we introduce a locality-aware generation order schedule that selects parallel decoding groups to maximize contextual support while minimizing intra-group dependencies, enabling higher degrees of parallelization. - **Candidate**: multiscale image generation provides one such way to break weak dependencies. in particular, we can model certain groups of pixels as conditionally independent given a lower resolution image and various types of context information, such as preceding frames in a video.

---

2. Macro-from-micro planning for high-quality and parallelized autoregressive long video generation

URL: View paper

Brief Assessment

Macro from Micro[64] focuses on long video generation through hierarchical planning (micro and macro planning) for temporal coherence across video segments, not on accelerating autoregressive image generation through parallel decoding with locality-aware ordering as in the original paper.

---

3. Grouped Speculative Decoding for Autoregressive Image Generation

URL: View paper

Brief Assessment

Grouped Speculative Decoding[36] focuses on speculative decoding techniques for accelerating autoregressive image generation through token clustering and acceptance rate optimization, not on locality-aware generation ordering or flexible parallelized autoregressive modeling architectures.

---

4. Maskgit: Masked generative image transformer

URL: View paper

Brief Assessment

Maskgit[65] focuses on masked generative modeling with bidirectional transformers for parallel token prediction, not on locality-aware generation ordering or flexible parallelized autoregressive modeling as proposed in the original paper.

---

5. Speculative Jacobi-Denoising Decoding for Accelerating Autoregressive Text-to-image Generation

URL: View paper

Brief Assessment

Speculative Jacobi Denoising[34] focuses on accelerating autoregressive text-to-image generation through denoising-based parallel decoding in the embedding space, not on locality-aware generation ordering for image patch sequences.

---

6. Collaborative decoding makes visual auto-regressive modeling efficient

URL: View paper

Brief Assessment

Collaborative Decoding[25] focuses on accelerating VAR's next-scale prediction through model collaboration (large drafter + small refiner), not on parallelizing next-patch prediction with locality-aware ordering as in the original paper.

---

7. Parallelized autoregressive visual generation

URL: View paper

Prior Art Analysis

Parallelized Autoregressive Visual[2] demonstrates that parallel decoding frameworks for autoregressive image generation existed prior to the original paper's submission. Both papers address the same core problem: reducing generation steps in autoregressive visual generation through parallel token prediction while maintaining quality. The candidate paper presents a complete framework (PAR) that achieves 3.6-9.5× speedup by generating multiple tokens in parallel, reducing steps from 576 to 147 (4× parallelization) or 51 (16× parallelization) on ImageNet 256×256. This directly refutes the novelty claim of being the first to present a 'complete framework combining flexible parallelized autoregressive modeling with locality-aware generation ordering' that reduces generation steps significantly.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same fundamental problem: sequential token-by-token prediction in autoregressive visual generation leads to slow inference. Both propose parallelized approaches to address this limitation. - **Original**: we present locality-aware parallel decoding(lpd) to accelerate autoregressive image generation. traditional autoregressive image generation relies on nextpatch prediction, a memory-bound process that leads to high latency. existing works have tried to parallelize next-patch prediction by shifting to... - **Candidate**: autoregressive models have emerged as a powerful approach for visual generation but suffer from slow inference speed due to their sequential token-by-token prediction process. in this paper, we propose a simple yet effective approach for parallelized autoregressive visual generation that improves ge...

Evidence 2 - **Rationale**: Both papers report similar speedup achievements (3.4-3.6× with comparable quality) through parallel generation frameworks, demonstrating that such performance improvements were already achieved by the candidate paper. - **Original**: with these designs, we reduce the generation steps from 256 to 20 (256 x256 res.) and 1024 to 48 (512 x512 res.) without compromising

quality on the imagenet class-conditional generation, and achieving at least 3.4x lower latency than previous parallelized autoregressive models. - **Candidate**: experiments on imagenet and ucf-101 demonstrate that our method achieves a 3.6x speedup with comparable quality and up to 9.5 x speedup with minimal quality degradation across both image and video generation tasks.

Evidence 3 - **Rationale**: Both papers present complete frameworks that combine parallel generation architectures with strategies for organizing token generation order based on dependencies, refuting the claim of being first to present such a combined framework. - **Original**: we introduce locality-aware parallel decoding(lpd), a framework that consists of a novel flexible parallelized autoregressive modeling architecture and a novel localityaware generation order schedule. - **Candidate**: based on these insights, we propose a simple yet effective approach for parallel generation in autoregressive visual models. our key idea is to identify and group weakly dependent visual tokens for simultaneous prediction while maintaining sequential generation for strongly dependent ones.

Evidence 4 - **Rationale**: Both papers identify and leverage the same locality principle: nearby tokens have strong dependencies while distant tokens have weak dependencies, which guides their parallel generation strategies. - **Original**: we observe strong spatial locality in image generation attention where tokens predominantly attend to nearby regions as shown in figure 2. this indicates a high dependency among nearby tokens, meaning that spatially closer tokens provide stronger conditioning. - **Candidate**: for visual data, such dependencies are naturally correlated with spatial distances-while locally adjacent tokens exhibit strong dependencies, spatially distant tokens often have weak correlations. this motivates us to reconsider how to organize tokens for generation: by identifying spatially distant...

---

8. AR-RAG: Autoregressive Retrieval Augmentation for Image Generation

URL: View paper

Brief Assessment

AR-RAG[11] focuses on retrieval-augmented generation at the patch level during autoregressive image generation, not on parallel decoding frameworks or generation ordering strategies to reduce sequential steps.

---

9. SCALAR: Scale-wise Controllable Visual Autoregressive Learning

URL: View paper

Brief Assessment

SCALAR[66] focuses on controllable generation in visual autoregressive models with scale-wise conditional decoding, not on parallel decoding frameworks for accelerating generation steps or locality-aware ordering mechanisms.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 1 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. Maskgit: Masked generative image transformer

Detected in: 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] Locality-aware Parallel Decoding for Efficient Autoregressive Image Generation View paper
- [1] Zipar: Accelerating autoregressive image generation through spatial locality View paper
- [2] Parallelized autoregressive visual generation View paper
- [3] Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction View paper
- [4] Neighboring autoregressive modeling for efficient visual generation View paper
- [5] Next Block Prediction: Video Generation via Semi-Autoregressive Modeling View paper
- [6] Muse: Text-To-Image Generation via Masked Generative Transformers View paper
- [7] Blockwise parallel decoding for deep autoregressive models View paper
- [8] Improving Autoregressive Image Generation through Coarse-to-Fine Token Prediction View paper
- [9] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders View paper
- [10] Autoregressive Image Generation with Randomized Parallel Decoding View paper
- [11] AR-RAG: Autoregressive Retrieval Augmentation for Image Generation View paper
- [12] Beyond tokens: A survey on decoding methods for large language models and large vision-language models View paper
- [13] ZipAR: Parallel Auto-regressive Image Generation through Spatial Locality View paper
- [14] Muddit: Liberating Generation Beyond Text-to-Image with a Unified Discrete Diffusion Model View paper
- [15] Octgpt: Octree-based multiscale autoregressive models for 3d shape generation View paper
- [16] Flover: A Temporal Fusion Framework for Efficient Autoregressive Model Parallel Inference View paper
- [17] Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes View paper
- [18] Accelerating Auto-regressive Text-to-Image Generation with Training-free Speculative Jacobi Decoding View paper
- [19] Cogview2: Faster and better text-to-image generation via hierarchical transformers View paper
- [20] Fast Autoregressive Video Generation with Diagonal Decoding View paper
- [21] Apar: Llms can do auto-parallel auto-regressive decoding View paper
- [22] Resurrect mask autoregressive modeling for efficient and scalable image generation View paper
- [23] MVAR: Visual Autoregressive Modeling with Scale and Spatial Markovian Conditioning View paper
- [24] Parallel multiscale autoregressive density estimation View paper
- [25] Collaborative decoding makes visual auto-regressive modeling efficient View paper
- [26] LSRS: Latent Scale Rejection Sampling for Visual Autoregressive Modeling View paper
- [27] Exploring stochastic autoregressive image modeling for visual representation View paper
- [28] Learning to Expand Images for Efficient Visual Autoregressive Modeling View paper
- [29] Parallelizing Autoregressive Generation with Variational State Space Models View paper
- [30] Locally hierarchical auto-regressive modeling for image generation View paper
- [31] Parallel Vision Token Scheduling for Fast and Accurate Multimodal LMMs Inference View paper
- [32] MC-SJD : Maximal Coupling Speculative Jacobi Decoding for Autoregressive Visual Generation Acceleration View paper

- [33] Image autoregressive interpolation model using GPU-parallel optimization View paper
- [34] Speculative Jacobi-Denoising Decoding for Accelerating Autoregressive Text-to-image Generation View paper
- [35] BIGFix: Bidirectional Image Generation with Token Fixing View paper
- [36] Grouped Speculative Decoding for Autoregressive Image Generation View paper
- [37] LazyMAR: Accelerating Masked Autoregressive Models via Feature Caching View paper
- [38] Improved Masked Image Generation with Knowledge-Augmented Token Representations View paper
- [39] Lformer: Text-to-Image Generation with L-shape Block Parallel Decoding View paper
- [40] SAMPO: Scale-wise Autoregression with Motion PrOmpt for generative world models View paper
- [41] Text-Conditioned Sampling Framework for Text-to-Image Generation with Masked Generative Models View paper
- [42] SJD++: Improved Speculative Jacobi Decoding for Training-free Acceleration of Discrete Auto-regressive Text-to-Image Generation View paper
- [43] Multi-scale Autoregressive Models are Laplacian, Discrete, and Latent Diffusion Models in Disguise View paper
- [44] Anime Generation through Diffusion and Language Models: A Comprehensive Survey of Techniques and Trends View paper
- [45] MAGVLT: Masked Generative Vision-and-Language Transformer View paper
- [46] Autoregression with Self-Token Prediction View paper
- [47] Enhancing Image Generation with Diffusion Transformer Architecture View paper
- [48] Recursive Autoregressive Depth Estimation with Continuous Token Modeling View paper
- [49] Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis View paper
- [50] Toward Improving the Generation Quality of Autoregressive Slot VAEs View paper
- [51] Next patch prediction for autoregressive visual generation View paper
- [52] Mimt: Masked image modeling transformer for video compression View paper
- [53] HMAR: Efficient Hierarchical Masked Auto-Regressive Image Generation View paper
- [54] Autoregressive Image Generation with Linear Complexity: A Spatial-Aware Decay Perspective View paper
- [55] RichControl: Structure- and Appearance-Rich Training-Free Spatial Control for Text-to-Image Generation View paper
- [56] Qwen2.5-Omni Technical Report View paper
- [57] Ar-diffusion: Auto-regressive diffusion model for text generation View paper
- [58] VideoGPT: Video Generation using VQ-VAE and Transformers View paper
- [59] Insertion transformer: Flexible sequence generation via insertion operations View paper
- [60] Larp: Tokenizing videos with a learned autoregressive generative prior View paper
- [61] STAR: Scale-wise Text-conditioned AutoRegressive image generation View paper
- [62] Symbol-rooted cascade propagation in contextual memory routing for large language models View paper
- [63] Leapformer: Enabling linear transformers for autoregressive and simultaneous tasks via learned proportions View paper
- [64] Macro-from-micro planning for high-quality and parallelized autoregressive long video generation View paper
- [65] Maskgit: Masked generative image transformer View paper
- [66] SCALAR: Scale-wise Controllable Visual Autoregressive Learning View paper