Novelty Assessment Report

Paper: Autoregressive Image Generation with Randomized Parallel Decoding
PDF URL: https://openreview.net/pdf?id=rJdGst0W8s
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We introduce ARPG, a novel visual Autoregressive model that enables Randomized Parallel Generation, addressing the inherent limitations of conventional raster-order approaches, which hinder inference efficiency and zero-shot generalization due to their sequential, predefined token generation order. Our key insight is that effective random-order modeling necessitates explicit guidance for determining the position of the next predicted token. To this end, we propose a novel decoupled decoding framework 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 generation, eliminating the need for bidirectional attention. Consequently, ARPG readily generalizes to zero-shot tasks such as image in-painting, out-painting, and resolution expansion. Furthermore, it supports parallel inference by concurrently processing multiple queries using a shared KV cache. On the ImageNet-1K 256 benchmark, our approach attains an FID of 1.83 with only 32 sampling steps, achieving over a 30 times speedup in inference and and a 75 percent reduction in memory consumption compared to representative recent autoregressive models at a similar scale.

Core Task Landscape

This paper addresses: **Autoregressive Image Generation with Randomized Parallel Decoding**

A total of **31 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Parallel Decoding Strategies Based on Spatial Structure**
- **Random-Order and Flexible-Order Autoregressive Modeling**
- **Speculative and Iterative Parallel Decoding**
- **Masked and Bidirectional Generative Transformers**
- **Coarse-to-Fine and Progressive Autoregressive Generation**
- **Alternative Generative Paradigms and Hybrid Approaches**
- **Application-Specific and Cross-Lingual Extensions**

Complete Taxonomy Tree

- Autoregressive Image Generation with Randomized Parallel Decoding Survey Taxonomy
- Parallel Decoding Strategies Based on Spatial Structure
  - Spatial Locality-Based Parallel Decoding (3 papers)
  - [2] Zipar: Accelerating autoregressive image generation through spatial locality (Yefei He, 2024) View paper
  - [7] ZipAR: Parallel Autoregressive Image Generation through Spatial Locality (He, 2024) View paper
  - [29] Locality-aware Parallel Decoding for Efficient Autoregressive Image Generation (Zhang Zhuo-Yang, 2025) View paper
  - Diagonal and Temporal Decoding Paths (1 papers)
  - [6] Fast Autoregressive Video Generation with Diagonal Decoding (Ye Yang, 2025) View paper
  - Block-Based and L-Shape Parallel Decoding (2 papers)
  - [13] Lformer: Text-to-Image Generation with L-shape Block Parallel Decoding (Li, 2023) View paper
  - [19] Blockwise parallel decoding for deep autoregressive models (Stern, 2018) View paper
  - Hierarchical and Multi-Scale Parallel Decoding (2 papers)
  - [8] Parallel multiscale autoregressive density estimation (Reed Scott, 2017) View paper
  - [12] Cogview2: Faster and better text-to-image generation via hierarchical transformers (Ding Ming, 2022) View paper
- Random-Order and Flexible-Order Autoregressive Modeling
  - Random-Order Training with Position Guidance ★ (3 papers)
  - [0] Autoregressive Image Generation with Randomized Parallel Decoding (Anon et al., 2026) View paper
  - [4] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders (Ziqi Pang, 2024) View paper
  - [31] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders Supplementary Material (Pseudo-Code, n.d.) View paper
  - Annealed Random-Order Autoregressive Training (1 papers)
  - [11] Randomized Autoregressive Visual Generation (Yu, 2024) View paper
  - Dependency-Aware Parallel Generation (1 papers)
  - [1] Parallelized autoregressive visual generation (Yuqing Wang, 2025) View paper
- Speculative and Iterative Parallel Decoding
  - Speculative Jacobi Decoding (3 papers)

- [14] Accelerating Auto-regressive Text-to-Image Generation with Training-free Speculative Jacobi Decoding (Yao Teng, 2024) View paper
- [21] Speculative Jacobi-Denoising Decoding for Accelerating Autoregressive Text-to-image Generation (Yao Teng, 2025) View paper
- [24] SJD++: Improved Speculative Jacobi Decoding for Training-free Acceleration of Discrete Auto-regressive Text-to-Image Generation (Yao Teng, 2025) View paper
- Grouped and Coupled Speculative Decoding (2 papers)
- [15] Grouped Speculative Decoding for Autoregressive Image Generation (So, 2025) View paper
- [27] MC-SJD : Maximal Coupling Speculative Jacobi Decoding for Autoregressive Visual Generation Acceleration (So, 2025) View paper
- Superposed Multi-Draft Decoding (1 papers)
- [3] Superposed decoding: Multiple generations from a single autoregressive inference pass (Alan Fan, 2024) View paper
- Jacobi Forcing for Causal Parallel Decoding (1 papers)
- [28] Fast and Accurate Causal Parallel Decoding using Jacobi Forcing (Lanxiang Hu, 2025) View paper

- Masked and Bidirectional Generative Transformers
  - Masked Generative Transformers for Image Synthesis (2 papers)
  - [5] Maskgit: Masked generative image transformer (Huiwen Chang, 2022) View paper
  - [20] Resurrect Mask AutoRegressive Modeling for Efficient and Scalable Image Generation (Xin Yi, 2025) View paper
  - Masked Text-to-Image Generation (1 papers)
  - [10] Muse: Text-To-Image Generation via Masked Generative Transformers (Chang, 2023) View paper
  - Contrastive Attention Guidance for Masked Transformers (1 papers)
  - [22] UNCAGE: Contrastive Attention Guidance for Masked Generative Transformers in Text-to-Image Generation (Kang, 2025) View paper
- Coarse-to-Fine and Progressive Autoregressive Generation (2 papers)
  - [9] DetailFlow: 1D Coarse-to-Fine Autoregressive Image Generation via Next-Detail Prediction (Liu Yiheng, 2025) View paper
  - [25] Learning to Expand Images for Efficient Visual Autoregressive Modeling (Ruiqing Yang, 2025) View paper
- Alternative Generative Paradigms and Hybrid Approaches
  - Discrete Diffusion for Vector-Quantized Tokens (1 papers)
  - [16] Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes (Sam Bond-Taylor, 2021) View paper
  - Variational State Space Models for Parallel Generation (1 papers)
  - [18] Parallelizing Autoregressive Generation with Variational State Space Models (Lambrechts, 2024) View paper
  - Retrieval-Augmented Autoregressive Generation (1 papers)
  - [23] AR-RAG: Autoregressive Retrieval Augmentation for Image Generation (Qi Jingyuan, 2025) View paper
  - Unified Multimodal Consistency Models (1 papers)
  - [26] UniCMs: A Unified Consistency Model For Efficient Multimodal Generation and Understanding (Xu, 2025) View paper
- Application-Specific and Cross-Lingual Extensions
  - Style-Specific and Sketch-to-Image Synthesis (1 papers)
  - [17] Styledrop: Text-to-image synthesis of any style (K Sohn, 2023) View paper
  - Multilingual Text-to-Image Generation (1 papers)
  - [30] Bridging Languages through Images: A Multilingual Text-to-Image Synthesis Approach (Tarun S, 2024) View paper

Narrative

Core task: autoregressive image generation with randomized parallel decoding. The field of autoregressive image generation has evolved beyond strictly sequential token prediction, exploring diverse strategies to accelerate inference while preserving or improving generation quality. The taxonomy reveals several major branches: some methods exploit spatial structure to decode multiple tokens in parallel (e.g., Zipar Spatial Locality[2], Parallelized Autoregressive Visual[1]), while others relax the fixed raster-scan order entirely, training models on random or flexible orderings (e.g., RandAR Random Orders[4], Randomized Autoregressive[11]). A third line of work borrows ideas from speculative execution and iterative refinement (Speculative Jacobi[14], Grouped Speculative[15]), and a fourth branch investigates masked or bidirectional transformers that predict multiple tokens simultaneously (Maskgit[5], Muse[10]). Additional branches address coarse-to-fine hierarchies (Cogview2[12], DetailFlow[9]), hybrid paradigms blending autoregressive and diffusion-like dynamics (Discrete Absorbing Diffusion[16]), and application-specific extensions such as multilingual or cross-domain generation.

Among these directions, random-order and flexible-order modeling has attracted growing interest as a way to enable parallel decoding without committing to a single spatial traversal. Randomized Parallel Decoding[0] sits squarely in this branch, training on randomized orderings augmented with position guidance to allow the model to decode multiple tokens concurrently at inference time. This approach contrasts with purely spatial methods like Zipar Spatial Locality[2], which rely on fixed locality patterns, and with speculative techniques like Superposed Decoding[3], which draft and verify tokens in parallel but do not randomize training order. Compared to RandAR Random Orders[4], which also explores random-order training, Randomized Parallel Decoding[0] emphasizes the integration of position cues to steer parallel generation. The central trade-off across these branches remains balancing inference speed, sample quality, and training complexity, with random-order methods offering a flexible middle ground between fully sequential and fully parallel paradigms.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

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

Authors: Ziqi Pang, Tianyuan ZHANG, Fujun Luan, Yunze Man, Hao Tan, et al. (8 authors total) | **Year/Venue**: 2024 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

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. 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 rand...

Relationship Analysis

Both papers belong to the Random-Order Training with Position Guidance category, using positional information to enable flexible-order autoregressive generation. RandAR inserts position instruction tokens directly into the sequence before each image token to be predicted, enabling random-order training and parallel decoding with KV-Cache. ARPG differs by decoupling positional guidance from

content representation through a two-pass decoder architecture, where position-aware [MASK] tokens serve as queries in cross-attention rather than being interspersed in the sequence, reducing memory overhead and computational cost.

---

2. RandAR: Decoder-only Autoregressive Visual Generation in Random Orders Supplementary Material

Authors: A Pseudo-Code | **URL**: View paper

Abstract

â¦ investigates enabling decoder-only transformers to generate image tokens in random ordersâ¦ bi-directional contexts from images, random-order generation so far achieves comparable â¦

Relationship Analysis

Both papers belong to the Random-Order Training with Position Guidance category, using position information to enable flexible-order autoregressive generation. RandAR intersperses position instruction tokens throughout the sequence alongside content tokens, training with causal attention on this interleaved sequence. The original paper (ARPG) differs by decoupling position and content into separate decoder passes: Pass-1 learns content representations via self-attention, while Pass-2 uses position-aware [MASK] tokens as queries in cross-attention, reducing sequence length and memory overhead compared to RandAR's interleaved approach.

Contributions Analysis

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

Contribution 1: ARPG: Visual autoregressive model with randomized parallel generation

Description: The authors propose ARPG, a visual autoregressive framework that supports fully random-order training and parallel token generation. Unlike conventional raster-order methods, ARPG eliminates sequential constraints and enables efficient inference while maintaining zero-shot generalization capabilities for tasks like inpainting and outpainting.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. -GPTs: A New Approach to Autoregressive Models

URL: View paper

Brief Assessment

The candidate paper context is too fragmentary to assess novelty claims. The provided excerpts mention random order training and parallel evaluation but lack sufficient technical detail about the specific framework, architecture, or methodology to determine whether it refutes ARPG's novelty.

---

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

URL: View paper

Prior Art Analysis

RandAR Random Orders[4] demonstrates that prior work exists for visual autoregressive models with randomized parallel generation and non-raster-order training. Both papers propose decoder-only autoregressive frameworks that enable random-order training and parallel token generation. RandAR Random Orders[4] explicitly states it is 'capable of generating images in arbitrary token orders' and 'removes this inductive bias' of predefined generation order, which directly addresses the same core problem as ARPG. Both methods support parallel decoding with KV-cache and zero-shot generalization tasks like inpainting and outpainting, indicating substantial overlap in the fundamental contribution.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce novel visual autoregressive models that address the same fundamental limitation: the predefined generation order of conventional approaches. RandAR Random Orders[4] explicitly removes the 'predefined generation order' constraint, which is the exact same problem ARPG claims to solve. - **Original**: we introduce arpg, a novel visual autoregressive model that enables randomized parallel generation, addressing the inherent limitations of conventional raster-order approaches, which hinder inference efficiency and zero-shot generalization due to their sequential, predefined token generation order. - **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 explicit positional guidance mechanisms to enable random-order generation. RandAR Random Orders[4] uses 'position instruction token' while ARPG uses 'decoupled decoding framework' with positional queries, but both address the same core technical challenge of providing position information for random-order prediction. - **Original**: our key insight is that effective random-order modeling necessitates explicit guidance for determining the position of the next predicted token. to this end, we propose a novel decoupled decoding framework that decouples positional guidance from content representation, encoding them separately as qu... - **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...

---

3. Diffusion-based Large Language Models Survey

URL: View paper

Brief Assessment

Diffusion LLMs Survey[40] focuses on diffusion-based language models and their arbitrary-order autoregressive properties, not visual autoregressive frameworks for image generation with randomized parallel decoding.

---

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

URL: View paper

Brief Assessment

Any-Subset Autoregressive[41] focuses on language modeling with any-order token generation, while ARPG addresses visual autoregressive modeling with a novel two-pass decoder architecture specifically designed for image generation tasks.

---

5. Locality-aware Parallel Decoding for Efficient Autoregressive Image Generation

URL: View paper

Brief Assessment

Locality-aware Parallel[29] focuses on locality-aware generation ordering with learnable position query tokens for parallel decoding, while ARPG uses a decoupled two-pass decoder architecture with [mask] tokens for random-order generation. The architectural approaches and core mechanisms differ fundamentally.

Contribution 2: Decoupled decoding framework with positional guidance

Description: The authors introduce a two-pass decoder architecture that separates content representation learning from position-guided prediction. The first pass uses causal self-attention to build content representations as key-value pairs, while the second pass uses position-aware mask tokens as queries that attend to these representations via causal cross-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. Learning from Next-Frame Prediction: Autoregressive Video Modeling Encodes Effective Representations

URL: View paper

Prior Art Analysis

Next-Frame Prediction[38] demonstrates prior work on decoupling semantic representation from target decoding in autoregressive visual models. The candidate paper explicitly describes a 'context-isolated autoregressive predictor to decouple semantic representation from target decoding,' which directly addresses the same architectural principle as the original paper's two-pass decoder that separates content representation learning from position-guided prediction. Both approaches fundamentally separate the representation learning phase from the prediction phase, though applied in different domains (video vs. image generation).

Evidence

Evidence 1 - **Rationale**: Both papers describe decoupling semantic/content representation from the decoding/prediction process. The original paper uses a two-pass architecture where content representations are built separately from position-guided prediction, while the candidate explicitly states it decouples 'semantic representation from target decoding' using a 'context-isolated autoregressive predictor,' demonstrating the same fundamental architectural principle existed in prior work. - **Original**: our approach decouples the prediction process into two distinct passes: (1) the content refinement pass utilizes causal self-attention over a random-order token sequence to construct label-leakage-free content representations as key-value pairs, without directly predicting tokens. (2) in the positio... - **Candidate**: next-vid introduces a context-isolated autoregressive predictor to decouple semantic representation from target decoding, and a conditioned flow-matching decoder to enhance generation quality and diversity.

Evidence 2 - **Rationale**: While the original paper claims insight into the need for explicit positional guidance as a novel finding, the candidate paper demonstrates that autoregressive visual pretraining methods with decoupled architectures already existed, addressing similar challenges of semantic localization in autoregressive models, suggesting the architectural principle was not entirely novel. - **Original**: based on the characteristics of the methods discussed, we find: insight 1: breaking the order-specific constraints of ar model requires explicit positional guidance. this insight stems from the fundamental difference in how models determine the prediction target. - **Candidate**: the few existing autoregressive visual pretraining methods suffer from issues such as inaccurate semantic localization and poor generation quality, leading to poor semantics. in this work, we propose next-vid, a novel autoregressive visual generative pretraining framework that utilizes masked next-f...

---

2. Design of a Modified Transformer Architecture Based on Relative Position Coding

URL: View paper

Brief Assessment

Relative Position Coding[32] focuses on modifying the self-attention mechanism within a standard transformer architecture by replacing absolute position encoding with relative position encoding. It does not propose a two-pass decoder architecture that separates content representation learning from position-guided prediction as described in the original paper.

---

3. Architectural entanglement via sequential convergence anchors: A novel framework for latent synchronization in large language models

URL: View paper

Brief Assessment

Architectural Entanglement[33] focuses on latent synchronization mechanisms in LLMs rather than visual autoregressive generation. The minimal context provided does not demonstrate prior work on decoupled two-pass decoder architectures for image generation.

---

4. SSD: Spatial-Semantic Head Decoupling for Efficient Autoregressive Image Generation

URL: View paper

Brief Assessment

Spatial-Semantic Head[37] focuses on KV cache compression for efficient inference by identifying spatial locality and semantic sink patterns in attention heads, not on decoupling positional guidance from content representation in the decoder architecture.

---

5. Disentangled sequential autoencoder

URL: View paper

Brief Assessment

Disentangled Sequential[34] focuses on separating time-invariant content (e.g., object identity) from time-varying dynamics in sequential data using a VAE framework with separate latent variables f and z_t. The original paper's two-pass decoder architecture with causal self-attention followed by position-guided cross-attention for autoregressive image generation represents a fundamentally different approach and application domain.

---

6. Generative temporal models with spatial memory for partially observed environments

URL: View paper

Brief Assessment

Spatial Memory[36] uses a two-pass architecture for spatial memory in RL environments (content representation via self-attention, then spatial queries via cross-attention), while the original paper focuses on autoregressive image generation with randomized parallel decoding. The architectural similarities are superficial; the applications and objectives differ fundamentally.

---

7. IAR2: Improving Autoregressive Visual Generation with Semantic-Detail Associated Token Prediction

URL: View paper

Brief Assessment

IAR2[35] focuses on a dual codebook structure for semantic-detail decomposition in visual generation, not on decoupling positional guidance from content representation through a two-pass decoder architecture.

Contribution 3: Parallel inference with shared KV cache

Description: The framework enables efficient parallel decoding by allowing multiple position-aware queries to simultaneously attend to a shared key-value cache. This design achieves significant speedups (30× over raster-order models, 3× over recent parallel AR models) while reducing memory consumption by 75% compared to similar-scale methods.

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. Seesaw: High-throughput llm inference via model re-sharding

URL: View paper

Brief Assessment

Seesaw[42] focuses on LLM inference optimization through model re-sharding techniques, not on parallel decoding for autoregressive image generation with position-aware queries and shared KV cache as described in the original paper.

---

2. Compression Barriers for Autoregressive Transformers

URL: View paper

Brief Assessment

Compression Barriers[44] focuses on theoretical memory lower bounds for KV cache storage in autoregressive transformers, proving that sublinear space is impossible without structural assumptions. It does not propose or implement a parallel decoding framework with shared KV cache for speedup.

---

3. Memory-Efficient Visual Autoregressive Modeling with Scale-Aware KV Cache Compression

URL: View paper

Brief Assessment

Scale-Aware KV[43] focuses on KV cache compression for VAR's multi-scale architecture, not on enabling parallel decoding through shared cache mechanisms. The candidate addresses memory optimization in existing VAR models rather than proposing parallel inference frameworks.

---

4. MiniCache: KV Cache Compression in Depth Dimension for Large Language Models

URL: View paper

Brief Assessment

MiniCache[48] focuses on compressing KV cache across layers to reduce memory footprint in LLMs, not on enabling parallel decoding with multiple position-aware queries attending to shared cache as in the original paper's image generation framework.

---

5. Autoregressive Image Generation Needs Only a Few Lines of Cached Tokens

URL: View paper

Brief Assessment

Cached Tokens[46] focuses on progressive KV cache compression for memory reduction in AR image generation, not on enabling parallel decoding through shared cache. The original paper's contribution is about architectural design for simultaneous multi-query processing, while this candidate addresses cache management optimization.

---

6. Longlive: Real-time interactive long video generation

URL: View paper

Brief Assessment

Longlive[49] focuses on frame-level autoregressive video generation with KV caching for temporal sequences, not parallel decoding of multiple position-aware queries attending to shared KV cache for image generation as in the original paper.

---

7. Why Are Positional Encodings Nonessential for Deep Autoregressive Transformers? A Petroglyph Revisited

URL: View paper

Brief Assessment

Positional Encodings Nonessential[50] focuses on the theoretical property that multi-layer autoregressive transformers can process sequences without explicit positional encodings, not on parallel inference architectures or shared KV cache mechanisms for speedup.

---

8. Scope: Optimizing key-value cache compression in long-context generation

URL: View paper

Brief Assessment

Scope[51] focuses on KV cache compression for long-context generation by separating prefill and decoding phases, not on parallel decoding with shared KV cache for speedup as in the original paper's contribution.

---

9. Apar: Llms can do auto-parallel auto-regressive decoding

URL: View paper

Brief Assessment

Apar[47] focuses on parallel auto-regressive text generation in LLMs through hierarchical planning, while the original paper addresses visual autoregressive image generation with position-aware queries and cross-attention mechanisms. The technical approaches and application domains are fundamentally different.

---

10. GEAR: An Efficient KV Cache Compression Recipe for Near-Lossless Generative Inference of LLM

URL: View paper

Brief Assessment

GEAR[45] focuses on KV cache compression techniques for memory efficiency in LLM inference, not on parallel decoding architectures or position-aware query mechanisms for image generation.

Appendix: Text Similarity Detection

Textual similarity detection checked 23 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. RandAR: Decoder-only Autoregressive Visual Generation in Random Orders

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **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] Autoregressive Image Generation with Randomized Parallel Decoding View paper
- [1] Parallelized autoregressive visual generation View paper
- [2] Zipar: Accelerating autoregressive image generation through spatial locality View paper
- [3] Superposed decoding: Multiple generations from a single autoregressive inference pass View paper
- [4] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders View paper
- [5] Maskgit: Masked generative image transformer View paper
- [6] Fast Autoregressive Video Generation with Diagonal Decoding View paper
- [7] ZipAR: Parallel Autoregressive Image Generation through Spatial Locality View paper
- [8] Parallel multiscale autoregressive density estimation View paper
- [9] DetailFlow: 1D Coarse-to-Fine Autoregressive Image Generation via Next-Detail Prediction View paper
- [10] Muse: Text-To-Image Generation via Masked Generative Transformers View paper
- [11] Randomized Autoregressive Visual Generation View paper
- [12] Cogview2: Faster and better text-to-image generation via hierarchical transformers View paper
- [13] Lformer: Text-to-Image Generation with L-shape Block Parallel Decoding View paper
- [14] Accelerating Auto-regressive Text-to-Image Generation with Training-free Speculative Jacobi Decoding View paper
- [15] Grouped Speculative Decoding for Autoregressive Image Generation View paper
- [16] Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes View paper
- [17] Styledrop: Text-to-image synthesis of any style View paper
- [18] Parallelizing Autoregressive Generation with Variational State Space Models View paper
- [19] Blockwise parallel decoding for deep autoregressive models View paper
- [20] Resurrect Mask AutoRegressive Modeling for Efficient and Scalable Image Generation View paper
- [21] Speculative Jacobi-Denoising Decoding for Accelerating Autoregressive Text-to-image Generation View paper
- [22] UNCAGE: Contrastive Attention Guidance for Masked Generative Transformers in Text-to-Image Generation View paper
- [23] AR-RAG: Autoregressive Retrieval Augmentation for Image Generation View paper
- [24] SJD++: Improved Speculative Jacobi Decoding for Training-free Acceleration of Discrete Auto-regressive Text-to-Image Generation View paper
- [25] Learning to Expand Images for Efficient Visual Autoregressive Modeling View paper
- [26] UniCMs: A Unified Consistency Model For Efficient Multimodal Generation and Understanding View paper
- [27] MC-SJD : Maximal Coupling Speculative Jacobi Decoding for Autoregressive Visual Generation Acceleration View paper
- [28] Fast and Accurate Causal Parallel Decoding using Jacobi Forcing View paper
- [29] Locality-aware Parallel Decoding for Efficient Autoregressive Image Generation View paper
- [30] Bridging Languages through Images: A Multilingual Text-to-Image Synthesis Approach View paper
- [31] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders Supplementary Material View paper
- [32] Design of a Modified Transformer Architecture Based on Relative Position Coding View paper
- [33] Architectural entanglement via sequential convergence anchors: A novel framework for latent synchronization in large language models View paper
- [34] Disentangled sequential autoencoder View paper
- [35] IAR2: Improving Autoregressive Visual Generation with Semantic-Detail Associated Token Prediction View paper
- [36] Generative temporal models with spatial memory for partially observed environments View paper
- [37] SSD: Spatial-Semantic Head Decoupling for Efficient Autoregressive Image Generation View paper
- [38] Learning from Next-Frame Prediction: Autoregressive Video Modeling Encodes Effective Representations View paper
- [39] -GPTs: A New Approach to Autoregressive Models View paper
- [40] Diffusion-based Large Language Models Survey View paper
- [41] Reviving Any-Subset Autoregressive Models with Principled Parallel Sampling and Speculative Decoding View paper
- [42] Seesaw: High-throughput llm inference via model re-sharding View paper
- [43] Memory-Efficient Visual Autoregressive Modeling with Scale-Aware KV Cache Compression View paper
- [44] Compression Barriers for Autoregressive Transformers View paper
- [45] GEAR: An Efficient KV Cache Compression Recipe for Near-Lossless Generative Inference of LLM View paper
- [46] Autoregressive Image Generation Needs Only a Few Lines of Cached Tokens View paper
- [47] Apar: Llms can do auto-parallel auto-regressive decoding View paper
- [48] MiniCache: KV Cache Compression in Depth Dimension for Large Language Models View paper
- [49] Longlive: Real-time interactive long video generation View paper
- [50] Why Are Positional Encodings Nonessential for Deep Autoregressive Transformers? A Petroglyph Revisited View paper
- [51] Scope: Optimizing key-value cache compression in long-context generation View paper