Novelty Assessment Report

Paper: Partition Generative Modeling: Masked Modeling Without Masks
PDF URL: https://openreview.net/pdf?id=vEh1ceS154
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Masked generative models (MGMs) are widely used to capture complex data and enable faster generation than autoregressive models (AR) through parallel decoding. However, MGMs typically operate on fixed-length inputs, which can be inefficient: early in sampling, most tokens are masked and carry little information, leading to wasted computation. In contrast, AR models process only tokens generated previously, making early iterations faster. In this work, we introduce the ``Partition Generative Model'' (PGM), a novel approach that combines the strengths of AR and MGMs. Rather than masking, PGM partitions tokens into two groups and employs sparse attention to block information flow between them. Since there is no information flow between partitions, the model can process the previously-generated tokens only during sampling, while retaining the ability to generate tokens in parallel and in any order. On OpenWebText, PGMs offer at least $5\times$ improvements in sampling latency and throughput, while producing samples with superior generative perplexity, compared to Masked Diffusion Language Models. In the ImageNet dataset, PGMs achieve up to $7\times$ better throughput compared to MaskGIT with only a small change in FID. Finally, we show that PGMs are compatible with distillation methods for MGMs, enabling further inference speedups.

Core Task Landscape

This paper addresses: **Efficient Parallel Token Generation Through Partitioning**

A total of **33 papers** were analyzed and organized into a taxonomy with **32 categories**.

Taxonomy Overview

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

- **Attention Mechanism Optimization**
- **Token-Level Parallelization**
- **Speculative Decoding**
- **Model Distribution Strategies**
- **Hardware Architectures**
- **Diffusion-Based Generation**
- **Domain-Specific Applications**
- **General Parallel Processing Frameworks**

Complete Taxonomy Tree

- Efficient Parallel Token Generation Through Partitioning Survey Taxonomy
- Attention Mechanism Optimization
  - Memory-Efficient Attention (1 papers)
  - [1] Flashattention-2: Faster attention with better parallelism and work partitioning (Dao, 2023) View paper
  - Multi-Dimensional Partitioning for Inference (1 papers)
  - [2] Efficiently scaling transformer inference (Pope, 2023) View paper
  - Sparse Attention Acceleration (2 papers)
  - [12] SOFA: A compute-memory optimized sparsity accelerator via cross-stage coordinated tiling (Huizheng Wang, 2024) View paper
  - [24] Designing Spatial Architectures for Sparse Attention: STAR Accelerator via Cross-Stage Tiling (Huizheng Wang, 2025) View paper
- Token-Level Parallelization
  - Spatial Locality Exploitation (1 papers)
  - [3] ZipAR: Parallel Autoregressive Image Generation through Spatial Locality (He, 2024) View paper
  - Parallel Token Prediction (1 papers)
  - [19] Parallel Token Prediction for Language Models (Felix Draxler, 2025) View paper
  - Partition-Based Generative Modeling ★ (1 papers)
  - [0] Partition Generative Modeling: Masked Modeling Without Masks (Anon et al., 2026) View paper
- Speculative Decoding
  - Layer-Parallel Speculation (1 papers)
  - [8] Easyspec: Layer-parallel speculative decoding for efficient multi-gpu utilization (Gao Ke., 2025) View paper
  - KV-Cache Parallelization (1 papers)
  - [15] Kv-runahead: Scalable causal llm inference by parallel key-value cache generation (Cho, 2024) View paper
  - Dynamic Token Tree Pruning (1 papers)
  - [21] ProPD: Dynamic Token Tree Pruning and Generation for LLM Parallel Decoding (Shuzhang Zhong, 2024) View paper
  - Resource-Aware Speculation (1 papers)

    - [20] Collaborative Large Language Model Inference via Resource-Aware Parallel Speculative Decoding (Yang Hyun Jong, 2025) View paper
    - Concurrent Attention for Collaboration (1 papers)
    - [5] Hogwild! inference: Parallel llm generation via concurrent attention (Rodionov, 2025) View paper
- Model Distribution Strategies
    - Edge Device Partitioning (1 papers)
    - [7] Large Language Model partitioning for low-latency inference at the edge (Dimitrios Kafetzis, 2025) View paper
    - Distributed Edge-Cloud Inference (1 papers)
    - [29] Model-Distributed Inference for Large Language Models at the Edge (Davide Macario, 2025) View paper
    - Distributed Retrieval-Augmented Generation (2 papers)
    - [11] Efficient Distributed Retrieval-Augmented Generation for Enhancing Language Model Performance (Liu Shang-yu, 2025) View paper
    - [17] DRAGON: Enhancing On-Device Model Performance with Distributed Retrieval-Augmented Generation (Shangyu Liu, 2025) View paper
- Hardware Architectures
    - Processing-In-Memory Architectures (1 papers)
    - [27] HPIM: Heterogeneous Processing-In-Memory-based Accelerator for Large Language Models Inference (Yang Jianlei, 2025) View paper
    - Near-Memory Processing Architectures (1 papers)
    - [10] Throughput Maximization for Transformer Inference on Processing Near-Memory Architectures (Mengke Ge, 2025) View paper
    - Vision Transformer Accelerators (1 papers)
    - [22] SASDenSebLE: A Compact Vision Transformer Inference Architecture With Saturation-Approximate Softmax Dataflow Enabling Sequence-Parallelism Boosted Layer â¦ (L He, 2025) View paper
- Diffusion-Based Generation
    - Discrete Diffusion for 3D Meshes (1 papers)
    - [6] Topology Sculptor, Shape Refiner: Discrete Diffusion Model for High-Fidelity 3D Meshes Generation (Song Kai-yu, 2025) View paper
    - Next-Frame Diffusion (1 papers)
    - [13] Playing with Transformer at 30+ FPS via Next-Frame Diffusion (He Tianyu, 2025) View paper
    - Discrete Speech Generation (1 papers)
    - [18] DiSTAR: Diffusion over a Scalable Token Autoregressive Representation for Speech Generation (Song, 2025) View paper
    - Multi-Scale Autoregressive Diffusion (1 papers)
    - [25] Multi-scale Autoregressive Models are Laplacian, Discrete, and Latent Diffusion Models in Disguise (Steve Hong, 2025) View paper
- Domain-Specific Applications
    - Video Coding Partition Prediction (1 papers)
    - [4] Efficient Partition Map Prediction via Token Sparsification for Fast VVC Intra Coding (Xin-Min Feng, 2024) View paper
    - Parallel Entropy Decoding (1 papers)
    - [23] A Bin-Based Bitstream Partitioning Approach for Parallel CABAC Decoding in Next Generation Video Coding (Philipp Habermann, 2019) View paper
    - Parallel rANS Decoding (1 papers)
    - [32] Recoil: Parallel rANS Decoding with Decoder-Adaptive Scalability (Fangzheng Lin, 2023) View paper
    - Industrial Quality Detection (1 papers)
    - [14] Multi-layer parallel transformer model for detecting product quality issues and locating anomalies based on multiple timeâseries process data in Industry 4.0 (Jiewu Leng, 2023) View paper
    - Video Temporal Localization (1 papers)
    - [26] Measure Twice, Cut Once: Grasping Video Structures and Event Semantics with LLMs for Video Temporal Localization (Pang, 2025) View paper
    - Blockchain Scalability (1 papers)
    - [9] Scalable Blockchain Architecture: Leveraging Hybrid Shard Generation and Data Partitioning (Praveen M Dhulavvagol, 2023) View paper
- General Parallel Processing Frameworks
    - Multi-Objective Optimization Partitioning (1 papers)
    - [28] Efficient Projection Partitioning for parallel multi-objective integer optimisation (Pettersson, 2022) View paper
    - Complex Event Processing (1 papers)
    - [31] Partition and compose: Parallel complex event processing (Martin Hirzel, 2012) View paper
    - Parallel Expert Systems (1 papers)
    - [33] A parallel expert system tool used in real time planning (M Lignon, 1993) View paper
    - OR-Parallel Token Machines (1 papers)
    - [30] An OR-Parallel Token Machine (Seif Haridi, 1983) View paper
    - Lock-Free Concurrent Data Structures (1 papers)
    - [16] No Cords Attached: Coordination-Free Concurrent Lock-Free Queues (Motiwala, 2025) View paper

Narrative

Core task: efficient parallel token generation through partitioning. The field addresses the challenge of accelerating sequence generation by dividing computational workloads across multiple processing units or temporal stages. The taxonomy reveals several complementary strategies: Attention Mechanism Optimization focuses on reducing the quadratic complexity of self-attention operations (e.g., Flashattention-2[1], Scaling Transformer Inference[2]), while Token-Level Parallelization explores methods that generate multiple tokens simultaneously rather than strictly sequentially. Speculative Decoding introduces draft-and-verify pipelines to overlap computation, and Model Distribution Strategies partition large models across devices or memory hierarchies (e.g., LLM Edge Partitioning[7], Topology Sculptor[6]). Hardware Architectures and Domain-Specific Applications tailor these ideas to specialized processors or tasks such as video coding (VVC Intra Coding[4], CABAC Partitioning[23]), while General Parallel Processing Frameworks provide foundational concurrency primitives (Coordination-Free Queues[16], Parallel Event Processing[31]).

Within Token-Level Parallelization, a particularly active line of work investigates partition-based generative modeling, where the sequence is divided into chunks that can be processed in parallel or with reduced dependencies. Partition Generative Modeling[0]

exemplifies this approach by structuring generation around explicit partitioning schemes, aiming to balance parallelism with the need to maintain coherent cross-partition dependencies. This contrasts with speculative methods like Easyspec[8] or Collaborative Speculative Decoding[20], which rely on draft models to predict multiple tokens speculatively, and with attention-centric optimizations like Flashattention-2[1] that accelerate existing autoregressive pipelines without altering the generation order. Nearby works such as ZipAR[3] and Parallel Token Prediction[19] also explore token-level parallelism but differ in how they handle inter-token dependencies and verification overhead. The original paper sits squarely in this partition-based cluster, emphasizing structured decomposition over speculative guessing or purely hardware-level acceleration.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

All three subtopics address parallel token generation to accelerate inference beyond sequential autoregressive decoding. The original Partition-Based Generative Modeling uses explicit token partitioning with sparse attention patterns, while Parallel Token Prediction focuses on joint prediction frameworks that model token dependencies, and Spatial Locality Exploitation leverages domain-specific independence in visual data. Each represents a distinct architectural approach to the same core efficiency goal.

Similarities: - All aim to generate multiple tokens simultaneously rather than one-at-a-time - All exclude traditional speculative decoding with separate draft models - All require modeling or exploiting some form of conditional independence structure among tokens

Differences: - Partition-Based uses explicit grouping with sparse attention mechanisms; Parallel Token Prediction uses unified frameworks incorporating sampling; Spatial Locality uses domain-specific visual independence - Partition-Based explicitly excludes masked diffusion approaches; the siblings don't mention diffusion exclusions - Spatial Locality is vision-specific (visual tokens), while Partition-Based and Parallel Token Prediction appear modality-agnostic - Parallel Token Prediction emphasizes joint dependency modeling, while Partition-Based emphasizes sparsity patterns, and Spatial Locality emphasizes locality assumptions

Suggested Search Directions: - Investigate whether partition-based methods could incorporate spatial locality for vision tasks - Explore hybrid approaches combining explicit partitioning with learned dependency structures - Examine whether sampling-integrated frameworks (Parallel Token Prediction) could benefit from partition-based sparse attention

Sibling Subtopics

- **Parallel Token Prediction** (leaves: 1, papers: 1)
- Scope: Universal frameworks jointly predicting multiple dependent tokens by incorporating sampling into the model.
- Exclude: Excludes speculative decoding with separate draft models; see Speculative Decoding.
- **Spatial Locality Exploitation** (leaves: 1, papers: 1)
- Scope: Methods leveraging spatial independence in visual tokens for parallel next-set prediction.
- Exclude: Excludes text-based parallel decoding and speculative methods; see Speculative Decoding and Parallel Token Prediction.

Contributions Analysis

Overall novelty summary. The paper introduces Partition Generative Models (PGM), which partition tokens into groups and use sparse attention to block information flow between partitions, enabling parallel generation without masking. According to the taxonomy, this work resides in the 'Partition-Based Generative Modeling' leaf under 'Token-Level Parallelization'. Notably, this leaf contains only the original paper itself—no sibling papers are listed—suggesting this specific approach to partition-based generation with sparse attention blocking represents a relatively unexplored direction within the broader token-level parallelization landscape.

The taxonomy reveals that PGM sits within a moderately populated parent branch ('Token-Level Parallelization') containing three leaves: Spatial Locality Exploitation (visual tokens), Parallel Token Prediction (joint multi-token prediction), and the original paper's leaf. Neighboring branches include Speculative Decoding (five leaves, draft-verify pipelines) and Attention Mechanism Optimization (three leaves, memory-efficient and sparse attention). The scope notes clarify that PGM differs from masked diffusion methods (excluded from this leaf) and from speculative approaches that use separate draft models. This positioning indicates PGM occupies a niche between pure autoregressive methods and masked generative models, leveraging partitioning rather than speculation or masking.

Among seventeen candidates examined across three contributions, no refutable prior work was identified. The core PGM contribution examined eight candidates with zero refutations, while the distillation compatibility contribution examined nine candidates, also with zero refutations. The encoder-decoder architecture contribution examined no candidates. Given the limited search scope (seventeen papers from semantic search and citation expansion, not exhaustive), these statistics suggest that within the examined literature, no directly overlapping prior work was found. However, the absence of refutations does not confirm absolute novelty—only that the top-K semantic matches did not reveal clear precedents for partition-based generation with sparse attention blocking.

Based on the limited literature search, PGM appears to introduce a distinctive approach within token-level parallelization, occupying an otherwise unpopulated taxonomy leaf. The lack of sibling papers and zero refutations among seventeen candidates examined suggest the specific combination of partitioning and sparse attention for parallel generation may be novel relative to the surveyed literature. However, the analysis covers only top-K semantic matches and does not exhaustively survey all related work in masked generative models, diffusion methods, or parallel decoding strategies.

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

Contribution 1: Partition Generative Model (PGM)

Description: The authors propose a new generative modeling approach that partitions tokens into two groups instead of masking them. This design allows the model to process only unmasked tokens during sampling while retaining parallel generation capabilities, combining advantages of autoregressive and masked generative models.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. MARch\'e: Fast Masked Autoregressive Image Generation with Cache-Aware Attention

URL: View paper

Brief Assessment

MARche[37] focuses on accelerating masked autoregressive (MAR) models through cache-aware attention and selective KV refresh for image generation, not on partitioning tokens into groups as an alternative to masking. The candidate addresses computational efficiency in existing MAR models rather than proposing a new generative modeling paradigm that combines AR and MGM strengths through token partitioning.

---

2. Autoregression with Self-Token Prediction

URL: View paper

Brief Assessment

Self-Token Prediction[40] focuses on continuous modalities (audio, image, video) and addresses identity collapse in regression tasks, while PGM targets discrete token generation (text) and eliminates mask tokens through partitioning. The technical approaches and problem domains are fundamentally different.

---

3. MARchÃ©: Fast Masked Autoregressive Image Generation with Cache-Aware Attention

URL: View paper

Brief Assessment

MARche[38] focuses on accelerating masked autoregressive (MAR) image generation through cache-aware attention and selective KV refresh, not on partitioning tokens as an alternative to masking. The candidate addresses computational efficiency in existing MAR models rather than proposing a new generative modeling paradigm that eliminates masks.

---

4. SMART-3D: Scaling Masked AutoRegressive Transformer for Efficient 3D Shape Generation

URL: View paper

Brief Assessment

SMART-3D[41] focuses on 3D shape generation using masked autoregressive transformers with progressive masked decoding for point clouds. The original paper's PGM addresses general sequence modeling (text/images) through token partitioning with group-wise attention to eliminate mask tokens entirely. These are distinct application domains and architectural approaches.

---

5. Customize your visual autoregressive recipe with set autoregressive modeling

URL: View paper

Brief Assessment

Visual Autoregressive Recipe[34] focuses on visual autoregressive modeling with set-based token generation for images, while PGM addresses masked generative modeling for language and images by partitioning tokens into groups without using mask tokens. The technical approaches differ fundamentally in their masking strategies and application domains.

---

6. Recursive Autoregressive Depth Estimation with Continuous Token Modeling

URL: View paper

Brief Assessment

Recursive Autoregressive Depth[39] focuses on monocular depth estimation using a fractal visual autoregressive framework with diffusion modeling for continuous depth values. It does not address the general masked generative modeling paradigm or token partitioning strategies that PGM proposes for parallel generation across modalities.

---

7. General point model pretraining with autoencoding and autoregressive

URL: View paper

Brief Assessment

Point Model Pretraining[36] focuses on point cloud representation learning by combining autoencoding and autoregressive tasks for 3D point cloud data, not on general token-based generative modeling for text/images. The partitioning approach in GPM is domain-specific to point cloud patches, fundamentally different from PGM's token partitioning strategy for parallel generation.

---

8. Unified Video Generation via Next-Set Prediction in Continuous Domain

URL: View paper

Brief Assessment

Next-Set Prediction[35] focuses on video generation using spatial progressive partitioning and temporal next-frame partitioning in continuous domain with VAE tokenization, while PGM addresses text generation using token partitioning with group-wise attention to avoid processing masked tokens during sampling.

Contribution 2: Encoder-decoder architecture with group-wise attention

Description: The authors design a specialized transformer architecture featuring an encoder with partition-wise self-attention, a novel GroupSwap layer, and a decoder with cross-attention but no self-attention. This architecture ensures predictions at position i never depend on the token at position i, enabling efficient processing without masked tokens.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Compatibility with distillation methods for MGMs

Description: The authors demonstrate that PGMs can be combined with existing distillation algorithms designed for masked generative models (specifically Self-Distillation Through Time), preserving performance on downstream tasks while achieving additional speedups in inference.

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. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning

URL: View paper

Brief Assessment

Efficientvlm[49] focuses on distilling vision-language models (VLMs) using knowledge distillation techniques, not masked generative models (MGMs). The candidate paper does not address masked generative modeling, discrete diffusion, or the specific distillation method (Self-Distillation Through Time) mentioned in the original contribution.

---

2. Distilling temporal knowledge with masked feature reconstruction for 3D object detection

URL: View paper

Brief Assessment

Temporal Knowledge Distillation[47] focuses on distilling temporal knowledge for 3D object detection in autonomous driving, not on masked generative models or their inference speedups. The paper addresses camera-based BEV detection with temporal feature reconstruction, which is a different domain and methodology from PGMs' compatibility with Self-Distillation Through Time for masked generative models.

---

3. Integrating masked generative distillation and network compression to identify the severity of wheat fusarium head blight

URL: View paper

Brief Assessment

Wheat Fusarium Detection[45] applies masked generative distillation to image classification for agricultural disease detection, not to masked generative models for sequence generation or inference speedup as in the original paper.

---

4. Masked autoencoders enable efficient knowledge distillers

URL: View paper

Brief Assessment

Masked Autoencoder Distillers[46] focuses on distilling knowledge from pre-trained masked autoencoders (MAE) for image classification, not on distillation methods for masked generative models (MGMs) used in text/image generation. The original paper discusses combining PGMs with Self-Distillation Through Time for MGMs, which is a different domain and application.

---

5. LKD-YOLOv8: A Lightweight Knowledge Distillation-Based Method for Infrared Object Detection

URL: View paper

Brief Assessment

LKD-YOLOv8[48] applies masked generative distillation (MGD) to infrared object detection using YOLO architecture, not to masked generative models for text/image generation. The candidate focuses on knowledge distillation for object detection networks, while the original contribution concerns distillation methods specifically designed for masked generative models like MDLM that enable inference speedups through parallel token generation.

---

6. Di o: Distilling masked diffusion models into one-step generator

URL: View paper

Brief Assessment

Dio[43] focuses on distilling masked diffusion models into one-step generators for image generation tasks, while the original paper demonstrates compatibility with Self-Distillation Through Time (SDTT) for language modeling. These are different application domains with different distillation objectives.

---

7. Distillspec: Improving speculative decoding via knowledge distillation

URL: View paper

Brief Assessment

Distillspec[44] focuses on knowledge distillation for speculative decoding in autoregressive language models, not on distillation methods for masked generative models. The candidate addresses a different model architecture and inference paradigm.

---

8. V2X-MGHD: A Collaborative Perception Network for Multiview LiDAR Sensors via Masked Generative Heterogeneous Distillation

URL: View paper

Brief Assessment

V2X-MGHD[51] applies distillation in a collaborative perception context for LiDAR sensors, not for accelerating masked generative model inference. The candidate focuses on feature recovery in vehicle networks rather than sampling speedups for generative models.

---

9. On distillation of guided diffusion models

URL: View paper

Brief Assessment

Guided Diffusion Distillation[42] focuses on distilling classifier-free guided diffusion models for image generation, not masked generative models for discrete token sequences. The technical approaches and application domains are fundamentally different.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Partition Generative Modeling: Masked Modeling Without Masks View paper
- [1] Flashattention-2: Faster attention with better parallelism and work partitioning View paper
- [2] Efficiently scaling transformer inference View paper
- [3] ZipAR: Parallel Autoregressive Image Generation through Spatial Locality View paper
- [4] Efficient Partition Map Prediction via Token Sparsification for Fast VVC Intra Coding View paper
- [5] Hogwild! inference: Parallel llm generation via concurrent attention View paper
- [6] Topology Sculptor, Shape Refiner: Discrete Diffusion Model for High-Fidelity 3D Meshes Generation View paper
- [7] Large Language Model partitioning for low-latency inference at the edge View paper
- [8] Easyspec: Layer-parallel speculative decoding for efficient multi-gpu utilization View paper
- [9] Scalable Blockchain Architecture: Leveraging Hybrid Shard Generation and Data Partitioning View paper
- [10] Throughput Maximization for Transformer Inference on Processing Near-Memory Architectures View paper
- [11] Efficient Distributed Retrieval-Augmented Generation for Enhancing Language Model Performance View paper
- [12] SOFA: A compute-memory optimized sparsity accelerator via cross-stage coordinated tiling View paper
- [13] Playing with Transformer at 30+ FPS via Next-Frame Diffusion View paper
- [14] Multi-layer parallel transformer model for detecting product quality issues and locating anomalies based on multiple timeâseries process data in Industry 4.0 View paper
- [15] Kv-runahead: Scalable causal llm inference by parallel key-value cache generation View paper
- [16] No Cords Attached: Coordination-Free Concurrent Lock-Free Queues View paper
- [17] DRAGON: Enhancing On-Device Model Performance with Distributed Retrieval-Augmented Generation View paper
- [18] DiSTAR: Diffusion over a Scalable Token Autoregressive Representation for Speech Generation View paper
- [19] Parallel Token Prediction for Language Models View paper
- [20] Collaborative Large Language Model Inference via Resource-Aware Parallel Speculative Decoding View paper
- [21] ProPD: Dynamic Token Tree Pruning and Generation for LLM Parallel Decoding View paper

- [22] SASDenSebLE: A Compact Vision Transformer Inference Architecture With Saturation-Approximate Softmax Dataflow Enabling Sequence-Parallelism Boosted Layer â¦ View paper
- [23] A Bin-Based Bitstream Partitioning Approach for Parallel CABAC Decoding in Next Generation Video Coding View paper
- [24] Designing Spatial Architectures for Sparse Attention: STAR Accelerator via Cross-Stage Tiling View paper
- [25] Multi-scale Autoregressive Models are Laplacian, Discrete, and Latent Diffusion Models in Disguise View paper
- [26] Measure Twice, Cut Once: Grasping Video Structures and Event Semantics with LLMs for Video Temporal Localization View paper
- [27] HPIM: Heterogeneous Processing-In-Memory-based Accelerator for Large Language Models Inference View paper
- [28] Efficient Projection Partitioning for parallel multi-objective integer optimisation View paper
- [29] Model-Distributed Inference for Large Language Models at the Edge View paper
- [30] An OR-Parallel Token Machine View paper
- [31] Partition and compose: Parallel complex event processing View paper
- [32] Recoil: Parallel rANS Decoding with Decoder-Adaptive Scalability View paper
- [33] A parallel expert system tool used in real time planning View paper
- [34] Customize your visual autoregressive recipe with set autoregressive modeling View paper
- [35] Unified Video Generation via Next-Set Prediction in Continuous Domain View paper
- [36] General point model pretraining with autoencoding and autoregressive View paper
- [37] MARch\'e: Fast Masked Autoregressive Image Generation with Cache-Aware Attention View paper
- [38] MARchÃ©: Fast Masked Autoregressive Image Generation with Cache-Aware Attention View paper
- [39] Recursive Autoregressive Depth Estimation with Continuous Token Modeling View paper
- [40] Autoregression with Self-Token Prediction View paper
- [41] SMART-3D: Scaling Masked AutoRegressive Transformer for Efficient 3D Shape Generation View paper
- [42] On distillation of guided diffusion models View paper
- [43] Di o: Distilling masked diffusion models into one-step generator View paper
- [44] Distillspec: Improving speculative decoding via knowledge distillation View paper
- [45] Integrating masked generative distillation and network compression to identify the severity of wheat fusarium head blight View paper
- [46] Masked autoencoders enable efficient knowledge distillers View paper
- [47] Distilling temporal knowledge with masked feature reconstruction for 3D object detection View paper
- [48] LKD-YOLOv8: A Lightweight Knowledge Distillation-Based Method for Infrared Object Detection View paper
- [49] Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning View paper
- [50] Distribution Matching Distillation Meets Reinforcement Learning View paper
- [51] V2X-MGHD: A Collaborative Perception Network for Multiview LiDAR Sensors via Masked Generative Heterogeneous Distillation View paper