Novelty Assessment Report

Paper: ToProVAR: Efficient Visual Autoregressive Modeling via Tri-Dimensional Entropy-Aware Semantic Analysis and Sparsity Optimization

PDF URL: https://openreview.net/pdf?id=s1djcQx3Ak

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Visual Autoregressive (VAR) models enhance generation speed but face a critical efficiency bottleneck in later stages. In this paper, we present a novel optimization framework for VAR models that fundamentally differs from prior approaches such as FastVAR and SkipVAR. Instead of relying on heuristic skipping strategies, our method leverages attention entropy to characterize the semantic projections across different dimensions of the model architecture. This enables precise identification of parameter dynamics under varying token granularity levels, semantic scopes, and generation scales. Building on this analysis, we further uncover sparsity patterns along three critical dimensions—token, layer, and scale—and propose a set of fine-grained optimization strategies tailored to these patterns. Extensive evaluation demonstrates that our approach achieves aggressive acceleration of the generation process while significantly preserving semantic fidelity and fine details, outperforming traditional methods in both efficiency and quality. Experiments on Infinity-2B and Infinity-8B models demonstrate that ToProVAR achieves nearly 3.5× average acceleration with minimal quality loss, effectively mitigating the issues found in prior work. Our code will be made publicly available.

Core Task Landscape

This paper addresses: **Accelerating Visual Autoregressive Image Generation**

A total of **50 papers** were analyzed and organized into a taxonomy with **29 categories**.

Taxonomy Overview

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

- **Parallel and Speculative Decoding Methods**
- **Architectural and Modeling Paradigm Innovations**
- **Visual Tokenizer and Representation Optimization**
- **Training and Optimization Strategies**
- **Masked Autoregressive and Bidirectional Models**
- **Controllability and Conditional Generation**
- **Unified and Multi-Modal Autoregressive Models**
- **Video and Temporal Autoregressive Generation**
- **Domain-Specific Applications and Extensions**
- **Theoretical Analysis and Computational Complexity**
- ... and 1 more categories

Complete Taxonomy Tree

- Accelerating Visual Autoregressive Image Generation Survey Taxonomy
- Parallel and Speculative Decoding Methods
  - Spatial Locality-Based Parallel Decoding (3 papers)
  - [2] Parallelized autoregressive visual generation (Yuqing Wang, 2025) View paper
  - [5] Zipar: Accelerating autoregressive image generation through spatial locality (Yefei He, 2024) View paper
  - [22] ZipAR: Parallel Autoregressive Image Generation through Spatial Locality (He, 2024) View paper
  - Speculative Decoding with Draft Models (3 papers)
  - [3] Grouped speculative decoding for autoregressive image generation (So, 2025) View paper
  - [24] Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding (Yao Teng, 2024) View paper
  - [50] Continuous Speculative Decoding for Autoregressive Image Generation (Wang Zili, 2024) View paper
  - Randomized and Flexible Order Generation (2 papers)
  - [16] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders (Ziqi Pang, 2024) View paper
  - [44] Autoregressive image generation with randomized parallel decoding (Li Haopeng, 2025) View paper
- Architectural and Modeling Paradigm Innovations
  - Scale-Wise and Coarse-to-Fine Generation ★ (4 papers)
  - [0] ToProVAR: Efficient Visual Autoregressive Modeling via Tri-Dimensional Entropy-Aware Semantic Analysis and Sparsity Optimization (Anon et al., 2026) View paper
  - [1] Autoregressive model beats diffusion: Llama for scalable image generation (Sun, 2024) View paper
  - [8] STAR: Scale-wise Text-conditioned AutoRegressive image generation (Ma XiaoXiao, 2024) View paper
  - [10] Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction (Liu Yiheng, 2025) View paper
  - Frequency-Domain Autoregressive Modeling (4 papers)
  - [20] Nfig: Autoregressive image generation with next-frequency prediction (Huang Zhihao, 2025) View paper

    - [33] NFIG: Multi-Scale Autoregressive Image Generation via Frequency Ordering (Huang Zhihao, 2025) View paper
    - [39] Frequency-Aware Autoregressive Modeling for Efficient High-Resolution Image Synthesis (Chen Zhuo-kun, 2025) View paper
    - [49] SkipVAR: Accelerating Visual Autoregressive Modeling via Adaptive Frequency-Aware Skipping (Li JiaJun, 2025) View paper
    - Patch and Region-Level Prediction (3 papers)
    - [17] ImageFolder: Autoregressive Image Generation with Folded Tokens (Li Xiang, 2024) View paper
    - [34] Next Patch Prediction for Autoregressive Visual Generation (Pang, 2024) View paper
    - [45] Neighboring Autoregressive Modeling for Efficient Visual Generation (He, 2025) View paper
    - Non-Quantized Continuous Token Modeling (4 papers)
    - [12] Autoregressive image generation without vector quantization (Mingyang Deng, 2024) View paper
    - [29] Autoregressive Video Generation without Vector Quantization (Pan Ting, 2024) View paper
    - [35] Fast Autoregressive Models for Continuous Latent Generation (Hang, 2025) View paper
    - [47] E-CAR: Efficient Continuous Autoregressive Image Generation via Multistage Modeling (Yuan, 2024) View paper
    - Alternative Backbone Architectures (1 papers)
    - [26] Scalable Autoregressive Image Generation with Mamba (Li Haopeng, 2024) View paper
- Visual Tokenizer and Representation Optimization
    - Scaling and Architecture of Visual Tokenizers (1 papers)
    - [9] GigaTok: Scaling Visual Tokenizers to 3 Billion Parameters for Autoregressive Image Generation (Xiong Tian-wei, 2025) View paper
    - Residual and Multi-Stage Quantization (1 papers)
    - [28] Autoregressive image generation using residual quantization (Doyup Lee, 2022) View paper
    - Foundation Model-Based Tokenization (1 papers)
    - [30] Vision Foundation Models as Effective Visual Tokenizers for Autoregressive Image Generation (Zheng, 2025) View paper
- Training and Optimization Strategies
    - Supervised Fine-Tuning and Reinforcement Learning (1 papers)
    - [14] Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl (Wang Jun-ke, 2025) View paper
    - Entropy-Guided and Adaptive Sampling (2 papers)
    - [19] Towards better & faster autoregressive image generation: From the perspective of entropy (Ma XiaoXiao, 2025) View paper
    - [31] Improving Chain-of-Thought Efficiency for Autoregressive Image Generation (Gu, 2025) View paper
    - Diffusion Step Annealing and Hybrid Sampling (1 papers)
    - [32] DiSA: Diffusion Step Annealing in Autoregressive Image Generation (Zhao Qinyu, 2025) View paper
- Masked Autoregressive and Bidirectional Models
    - Masked Autoregressive Frameworks (2 papers)
    - [15] Resurrect mask autoregressive modeling for efficient and scalable image generation (Xin Yi, 2025) View paper
    - [37] LazyMAR: Accelerating Masked Autoregressive Models via Feature Caching (Wei Qing-yan, 2025) View paper
    - Hybrid Tokenization for Masked AR (1 papers)
    - [41] DC-AR: Efficient Masked Autoregressive Image Generation with Deep Compression Hybrid Tokenizer (Wu Yecheng, 2025) View paper
- Controllability and Conditional Generation
    - Multi-Conditional and Context-Aware Generation (2 papers)
    - [11] Car: Controllable autoregressive modeling for visual generation (Yao, 2024) View paper
    - [43] Context-Aware Autoregressive Models for Multi-Conditional Image Generation (Chen, 2025) View paper
    - Subject-Driven Fine-Tuning (1 papers)
    - [6] Fine-Tuning Visual Autoregressive Models for Subject-Driven Generation (Hyun, 2025) View paper
- Unified and Multi-Modal Autoregressive Models
    - Unified Spacetime Autoregressive Modeling (1 papers)
    - [13] Infinitystar: Unified spacetime autoregressive modeling for visual generation (Liu Jin-lai, 2025) View paper
    - Multimodal Generative Pretraining (1 papers)
    - [38] Lumina-mGPT: Illuminate Flexible Photorealistic Text-to-Image Generation with Multimodal Generative Pretraining (Liu Dong-yang, 2024) View paper
- Video and Temporal Autoregressive Generation
    - Autoregressive Video Diffusion Models (3 papers)
    - [4] From slow bidirectional to fast autoregressive video diffusion models (Tianwei Yin, 2025) View paper
    - [18] Vid-gpt: Introducing gpt-style autoregressive generation in video diffusion models (Gao, 2024) View paper
    - [21] Art-v: Auto-regressive text-to-video generation with diffusion models (Wen-ming, 2024) View paper
    - Real-Time Interactive Video Generation (1 papers)
    - [42] Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation (Chen Ming, 2025) View paper
    - Multi-View and 3D Novel View Synthesis (1 papers)
    - [40] Causnvs: Autoregressive multi-view diffusion for flexible 3d novel view synthesis (Kong Xin, 2025) View paper
- Domain-Specific Applications and Extensions
    - Image Super-Resolution (1 papers)
    - [36] Visual Autoregressive Modeling for Image Super-Resolution (Qu, 2025) View paper
    - Image-to-Image Translation (1 papers)
    - [25] CycleVAR: Repurposing Autoregressive Model for Unsupervised One-Step Image Translation (Liu Yi, 2025) View paper
    - Image-to-Text Generation (1 papers)
    - [23] LaDiC: Are Diffusion Models Really Inferior to Autoregressive Counterparts for Image-to-Text Generation? (Wang Yu-Chi, 2024) View paper
    - Infinite and Arbitrary-Size Visual Synthesis (2 papers)
    - [46] Yume: An Interactive World Generation Model (Mao, 2025) View paper
    - [48] Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis (Wu, 2022) View paper

- Theoretical Analysis and Computational Complexity (1 papers)
  - [27] On computational limits and provably efficient criteria of visual autoregressive models: A fine-grained complexity analysis (Li, 2025) View paper
- Bitwise and High-Resolution Scaling (1 papers)
  - [7] Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis (Jian Han, 2025) View paper

Narrative

Core task: Accelerating visual autoregressive image generation. The field has organized itself around several complementary strategies for making autoregressive image models faster and more practical. At the highest level, one finds branches dedicated to Parallel and Speculative Decoding Methods, which aim to predict multiple tokens simultaneously or verify draft sequences in parallel, and Architectural and Modeling Paradigm Innovations, which rethink the generation order or introduce coarse-to-fine hierarchies. Other major directions include Visual Tokenizer and Representation Optimization, which seeks better discrete or continuous representations to reduce sequence length, and Training and Optimization Strategies, which tune learning procedures for efficiency. Meanwhile, Masked Autoregressive and Bidirectional Models explore relaxing strict left-to-right ordering, and branches like Video and Temporal Autoregressive Generation or Domain-Specific Applications extend these ideas beyond static images. Works such as Grouped Speculative Decoding[3] and Parallelized Autoregressive Visual[2] exemplify efforts to decode in parallel, while STAR[8] and Detailflow[10] illustrate scale-wise generation paradigms.

Within the Architectural and Modeling Paradigm Innovations branch, a particularly active line of work focuses on scale-wise and coarse-to-fine generation, where models first produce low-resolution or abstract structure and then refine details progressively. ToProVAR[0] sits squarely in this cluster, emphasizing a top-down progressive refinement strategy that balances quality and speed. Nearby, STAR[8] adopts a similar multi-scale philosophy but differs in how it schedules token prediction across resolutions, while Detailflow[10] explores flow-based mechanisms for detail injection at finer scales. These coarse-to-fine approaches contrast with fully parallel methods like Grouped Speculative Decoding[3], which sacrifice ordering structure for maximum parallelism, and with tokenizer-centric efforts such as GigaTok[9], which compress sequences so aggressively that even standard autoregressive decoding becomes faster. The central trade-off across these directions is between preserving hierarchical structure for controllability and quality versus maximizing throughput through parallelism or shorter sequences, with ToProVAR[0] occupying a middle ground that leverages progressive generation to achieve both efficiency gains and fine-grained control.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Autoregressive model beats diffusion: Llama for scalable image generation

Authors: Sun, Peize, Jiang Yi, Peize Sun, Chen, et al. (17 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We introduce LlamaGen, a new family of image generation models that apply original ``next-token prediction'' paradigm of large language models to visual generation domain. It is an affirmative answer to whether vanilla autoregressive models, e.g., Llama, without inductive biases on visual signals can achieve state-of-the-art image generation performance if scaling properly. We reexamine design spaces of image tokenizers, scalability properties of image generation models, and their training data ...

Relationship Analysis

Both papers belong to the Scale-Wise and Coarse-to-Fine Generation category, focusing on autoregressive visual generation models that progressively generate images from coarse to fine scales. While LlamaGen explores the fundamental design space of autoregressive image generation (tokenizer design, model scaling, and training data quality) to demonstrate that vanilla autoregressive models can achieve state-of-the-art performance, ToProVAR focuses specifically on accelerating the inference of existing VAR models through tri-dimensional entropy-aware optimization across token, layer, and scale dimensions. The key difference is that LlamaGen establishes a foundational framework for scalable autoregressive image generation, whereas ToProVAR addresses the computational efficiency bottleneck in later generation stages through fine-grained sparsity analysis and optimization.

---

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

Authors: Ma XiaoXiao, Zhou Mohan, Xiaoxiao Ma, Liang Tao, Mohan Zhou, et al. (17 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We introduce STAR, a text-to-image model that employs a scale-wise auto-regressive paradigm. Unlike VAR, which is constrained to class-conditioned synthesis for images up to 256$\times$256, STAR enables text-driven image generation up to 1024$\times$1024 through three key designs. First, we introduce a pre-trained text encoder to extract and adopt representations for textual constraints, enhancing details and generalizability. Second, given the inherent structural correlation across different sc...

Relationship Analysis

Both papers belong to the Scale-Wise and Coarse-to-Fine Generation category, employing multi-scale autoregressive approaches for progressive image synthesis. While ToProVAR focuses on accelerating existing VAR models through tri-dimensional entropy-aware optimization (token, layer, and scale pruning), STAR addresses fundamental architectural improvements for text-to-image generation, including normalized 2D RoPE for positional encoding and stable sampling methods to handle inter-token relationships at high resolutions. The key distinction is that ToProVAR optimizes inference efficiency of pre-trained models, whereas STAR proposes novel training and sampling techniques to enable high-resolution text-conditioned generation up to 1024×1024.

---

3. Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction

Authors: Liu Yiheng, Qu, Liao, Yiheng Liu, Zhang Huichao, et al. (31 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

This paper presents DetailFlow, a coarse-to-fine 1D autoregressive (AR) image generation method that models images through a novel next-detail prediction strategy. By learning a resolution-aware token sequence supervised with progressively degraded images, DetailFlow enables the generation process to start from the global structure and incrementally refine details. This coarse-to-fine 1D token sequence aligns well with the autoregressive inference mechanism, providing a more natural and efficien...

Relationship Analysis

Both papers belong to the Scale-Wise and Coarse-to-Fine Generation category, employing progressive generation strategies from coarse to fine scales. While ToProVAR focuses on accelerating existing VAR models through tri-dimensional entropy-aware optimization (token, layer, and scale pruning) to achieve 3.5× speedup, DetailFlow proposes a fundamentally different 1D autoregressive architecture with next-detail prediction and parallel inference mechanisms to reduce token count from 680 to 128 tokens. The key distinction is that ToProVAR optimizes inference of existing multi-scale VAR models through sparsity analysis, whereas DetailFlow redesigns the generation paradigm itself with a compact 1D token sequence and self-correction mechanism.

Contributions Analysis

Overall novelty summary. The paper proposes ToProVAR, an optimization framework for visual autoregressive models that uses attention entropy to identify parameter dynamics across token, layer, and scale dimensions, enabling fine-grained sparsity-based acceleration. It resides in the 'Scale-Wise and Coarse-to-Fine Generation' leaf, which contains four papers including the original work. This leaf sits within the broader 'Architectural and Modeling Paradigm Innovations' branch, indicating a moderately populated research direction focused on progressive refinement strategies. The taxonomy shows this is an active but not overcrowded area, with sibling papers like STAR and Detailflow exploring related multi-scale generation paradigms.

The taxonomy reveals several neighboring research directions that contextualize this work. Adjacent leaves include 'Frequency-Domain Autoregressive Modeling' (four papers decomposing generation by frequency rather than spatial scale) and 'Patch and Region-Level Prediction' (three papers aggregating tokens spatially). The 'Parallel and Speculative Decoding Methods' branch (seven papers across three leaves) represents an alternative acceleration philosophy emphasizing simultaneous token prediction rather than hierarchical refinement. ToProVAR's entropy-driven approach distinguishes it from these neighbors by focusing on dynamic parameter selection within a coarse-to-fine framework, rather than changing generation order or token granularity.

Among sixteen candidates examined across three contributions, none were identified as clearly refuting the proposed methods. The tri-dimensional attention entropy framework examined six candidates with zero refutations, while the fine-grained sparsity optimization strategies examined ten candidates, also with zero refutations. The Flash Attention Entropy optimization had no candidates examined. This limited search scope—sixteen papers from semantic search and citation expansion—suggests the specific combination of entropy-guided analysis and tri-dimensional sparsity patterns may be relatively unexplored in the examined literature. However, the modest search scale means potentially relevant prior work in attention analysis or dynamic pruning may exist beyond these candidates.

Based on the available signals, the work appears to occupy a distinct position within the coarse-to-fine generation paradigm by introducing entropy-based parameter dynamics analysis. The taxonomy structure indicates this is a moderately active research area with clear boundaries from parallel decoding and tokenizer-focused approaches. The absence of refuting candidates among sixteen examined papers suggests novelty in the specific technical approach, though the limited search scope prevents definitive conclusions about the broader landscape of attention-based optimization methods or dynamic sparsity techniques in autoregressive models.

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

Contribution 1: Tri-dimensional attention entropy framework for VAR optimization

Description: The authors propose a novel framework that uses attention entropy to analyze Visual Autoregressive models across three dimensions (token, layer, and scale) rather than relying on heuristic methods. This enables precise identification of parameter dynamics under varying token granularity, semantic scopes, and generation scales.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Fast-ARDiff: An Entropy-informed Acceleration Framework for Continuous Space Autoregressive Generation

URL: View paper

Brief Assessment

Fast-ARDiff[55] focuses on entropy-informed speculative decoding for AR-diffusion hybrid models in continuous space, not on tri-dimensional (token, layer, scale) attention entropy analysis for Visual Autoregressive model optimization.

---

2. A neural autoregressive approach to attention-based recognition

URL: View paper

Brief Assessment

Neural Autoregressive Recognition[56] addresses attention-based visual recognition through sequential fixations in images, not visual autoregressive (VAR) model optimization. The candidate focuses on learning fixation policies for recognition tasks, while the original paper analyzes attention entropy across token, layer, and scale dimensions specifically for VAR generation efficiency.

---

3. Reinforcement Learning for Solving Colored Traveling Salesman Problems: An Entropy-Insensitive Attention Approach

URL: View paper

Brief Assessment

Traveling Salesman Entropy[52] applies attention mechanisms to combinatorial optimization (colored TSP) using RL, not visual autoregressive model optimization across architectural dimensions (token/layer/scale).

---

4. DREAM: Drafting with Refined Target Features and Entropy-Adaptive Cross-Attention Fusion for Multimodal Speculative Decoding

URL: View paper

Brief Assessment

DREAM[51] focuses on speculative decoding for vision-language models using attention entropy for feature selection, not on visual autoregressive model optimization across architectural dimensions (token, layer, scale).

---

5. Group Critical-token Policy Optimization for Autoregressive Image Generation

URL: View paper

Brief Assessment

Group Critical-Token[53] focuses on identifying critical tokens for RL-based policy optimization in AR visual generation, while the original paper analyzes attention entropy across token, layer, and scale dimensions for computational efficiency optimization in VAR models. These are fundamentally different optimization objectives and methodological approaches.

---

6. DPAR: Dynamic Patchification for Efficient Autoregressive Visual Generation

URL: View paper

Brief Assessment

DPAR[54] uses entropy from a lightweight autoregressive model for token merging in image generation, not for analyzing VAR models across token, layer, and scale dimensions as in the original paper's tri-dimensional framework.

Contribution 2: Fine-grained sparsity optimization strategies across three dimensions

Description: The authors identify sparsity patterns in token, layer, and scale dimensions and develop corresponding optimization strategies: token-level pruning of non-essential semantics, layer-level compression distinguishing global from detail representation, and scale-level depth adjustment tailored to object fineness.

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. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference

URL: View paper

Brief Assessment

SparseVLM[59] focuses on vision-language models with text-guided visual token pruning, not visual autoregressive models with tri-dimensional (token-layer-scale) entropy-based optimization for image generation.

---

2. Saliency-driven dynamic token pruning for large language models

URL: View paper

Brief Assessment

Saliency Token Pruning[58] focuses on token-level pruning for LLMs using saliency scores, not on multi-dimensional sparsity patterns across token, layer, and scale dimensions for visual autoregressive models. The candidate addresses a different model architecture (LLMs vs. VAR) and does not analyze layer-level or scale-level semantic patterns.

---

3. Hash layers for large sparse models

URL: View paper

Brief Assessment

Hash Layers[61] focuses on sparse routing in transformer feedforward layers using hash functions to assign tokens to expert modules, without analyzing sparsity patterns across token, layer, and scale dimensions for visual autoregressive models as in the original paper.

---

4. The sparse frontier: Sparse attention trade-offs in transformer llms

URL: View paper

Brief Assessment

Sparse Frontier[66] focuses on sparse attention mechanisms in transformer LLMs for long-context processing, examining sparsity trade-offs across different attention strategies. The original paper addresses sparsity in visual autoregressive models across token, layer, and scale dimensions with entropy-based semantic analysis, which is a fundamentally different domain and approach.

---

5. Scaling and evaluating sparse autoencoders

URL: View paper

Brief Assessment

Sparse Autoencoders[57] focuses on sparse bottleneck layers for feature extraction in language models, not on token/layer/scale sparsity patterns in visual autoregressive generation. The technical domains and optimization targets are fundamentally different.

---

6. Spatten: Efficient sparse attention architecture with cascade token and head pruning

URL: View paper

Brief Assessment

Spatten[62] focuses on token and head pruning for attention mechanisms in NLP with hardware acceleration, not on visual autoregressive models across token/layer/scale dimensions for image generation.

---

7. Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed

URL: View paper

Brief Assessment

Efficient LoFTR[65] focuses on efficiency optimization for image matching transformers through aggregated attention and correlation layers, not on tri-dimensional (token, layer, scale) sparsity patterns for visual autoregressive generation models.

---

8. Dynamicvit: Efficient vision transformers with dynamic token sparsification

URL: View paper

Brief Assessment

DynamicViT[64] focuses on token-level pruning in vision transformers for image classification, using attention-based importance scoring. The original paper addresses visual autoregressive (VAR) models with entropy-based analysis across token, layer, and scale dimensions for generation tasks—a fundamentally different architecture and application domain.

---

9. Scaling sparse fine-tuning to large language models

URL: View paper

Brief Assessment

Sparse Fine-Tuning[60] focuses on parameter-efficient fine-tuning of LLMs through sparse parameter updates (maintaining parameter indices and deltas), not on sparsity patterns across token, layer, and scale dimensions for visual autoregressive model optimization during inference.

---

10. Base layers: Simplifying training of large, sparse models

URL: View paper

Brief Assessment

Base Layers[63] focuses on token-to-expert assignment in sparse mixture-of-experts models for language modeling, not on sparsity patterns across token/layer/scale dimensions in visual autoregressive generation. The technical approaches and application domains are fundamentally different.

Contribution 3: Flash Attention Entropy computational optimization

Description: The authors develop an efficient computational mechanism called Flash Attention Entropy that extends FlashAttention to compute attention entropy online without materializing the full attention matrix, ensuring both effectiveness and practicality of the framework.

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.

Appendix: Text Similarity Detection

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

References

- [0] ToProVAR: Efficient Visual Autoregressive Modeling via Tri-Dimensional Entropy-Aware Semantic Analysis and Sparsity Optimization View paper
- [1] Autoregressive model beats diffusion: Llama for scalable image generation View paper
- [2] Parallelized autoregressive visual generation View paper
- [3] Grouped speculative decoding for autoregressive image generation View paper
- [4] From slow bidirectional to fast autoregressive video diffusion models View paper
- [5] Zipar: Accelerating autoregressive image generation through spatial locality View paper
- [6] Fine-Tuning Visual Autoregressive Models for Subject-Driven Generation View paper
- [7] Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis View paper
- [8] STAR: Scale-wise Text-conditioned AutoRegressive image generation View paper
- [9] GigaTok: Scaling Visual Tokenizers to 3 Billion Parameters for Autoregressive Image Generation View paper
- [10] Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction View paper
- [11] Car: Controllable autoregressive modeling for visual generation View paper
- [12] Autoregressive image generation without vector quantization View paper
- [13] Infinitystar: Unified spacetime autoregressive modeling for visual generation View paper
- [14] Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl View paper
- [15] Resurrect mask autoregressive modeling for efficient and scalable image generation View paper
- [16] RandAR: Decoder-only Autoregressive Visual Generation in Random Orders View paper
- [17] ImageFolder: Autoregressive Image Generation with Folded Tokens View paper
- [18] Vid-gpt: Introducing gpt-style autoregressive generation in video diffusion models View paper
- [19] Towards better & faster autoregressive image generation: From the perspective of entropy View paper
- [20] Nfig: Autoregressive image generation with next-frequency prediction View paper
- [21] Art-v: Auto-regressive text-to-video generation with diffusion models View paper
- [22] ZipAR: Parallel Autoregressive Image Generation through Spatial Locality View paper
- [23] LaDiC: Are Diffusion Models Really Inferior to Autoregressive Counterparts for Image-to-Text Generation? View paper
- [24] Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding View paper
- [25] CycleVAR: Repurposing Autoregressive Model for Unsupervised One-Step Image Translation View paper
- [26] Scalable Autoregressive Image Generation with Mamba View paper
- [27] On computational limits and provably efficient criteria of visual autoregressive models: A fine-grained complexity analysis View paper
- [28] Autoregressive image generation using residual quantization View paper
- [29] Autoregressive Video Generation without Vector Quantization View paper
- [30] Vision Foundation Models as Effective Visual Tokenizers for Autoregressive Image Generation View paper
- [31] Improving Chain-of-Thought Efficiency for Autoregressive Image Generation View paper
- [32] DiSA: Diffusion Step Annealing in Autoregressive Image Generation View paper
- [33] NFIG: Multi-Scale Autoregressive Image Generation via Frequency Ordering View paper
- [34] Next Patch Prediction for Autoregressive Visual Generation View paper
- [35] Fast Autoregressive Models for Continuous Latent Generation View paper
- [36] Visual Autoregressive Modeling for Image Super-Resolution View paper
- [37] LazyMAR: Accelerating Masked Autoregressive Models via Feature Caching View paper
- [38] Lumina-mGPT: Illuminate Flexible Photorealistic Text-to-Image Generation with Multimodal Generative Pretraining View paper
- [39] Frequency-Aware Autoregressive Modeling for Efficient High-Resolution Image Synthesis View paper
- [40] Causnvs: Autoregressive multi-view diffusion for flexible 3d novel view synthesis View paper
- [41] DC-AR: Efficient Masked Autoregressive Image Generation with Deep Compression Hybrid Tokenizer View paper
- [42] Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation View paper
- [43] Context-Aware Autoregressive Models for Multi-Conditional Image Generation View paper
- [44] Autoregressive image generation with randomized parallel decoding View paper
- [45] Neighboring Autoregressive Modeling for Efficient Visual Generation View paper
- [46] Yume: An Interactive World Generation Model View paper
- [47] E-CAR: Efficient Continuous Autoregressive Image Generation via Multistage Modeling View paper
- [48] Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis View paper
- [49] SkipVAR: Accelerating Visual Autoregressive Modeling via Adaptive Frequency-Aware Skipping View paper
- [50] Continuous Speculative Decoding for Autoregressive Image Generation View paper
- [51] DREAM: Drafting with Refined Target Features and Entropy-Adaptive Cross-Attention Fusion for Multimodal Speculative Decoding View paper
- [52] Reinforcement Learning for Solving Colored Traveling Salesman Problems: An Entropy-Insensitive Attention Approach View paper
- [53] Group Critical-token Policy Optimization for Autoregressive Image Generation View paper
- [54] DPAR: Dynamic Patchification for Efficient Autoregressive Visual Generation View paper
- [55] Fast-ARDiff: An Entropy-informed Acceleration Framework for Continuous Space Autoregressive Generation View paper
- [56] A neural autoregressive approach to attention-based recognition View paper
- [57] Scaling and evaluating sparse autoencoders View paper
- [58] Saliency-driven dynamic token pruning for large language models View paper
- [59] SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference View paper
- [60] Scaling sparse fine-tuning to large language models View paper
- [61] Hash layers for large sparse models View paper
- [62] Spatten: Efficient sparse attention architecture with cascade token and head pruning View paper
- [63] Base layers: Simplifying training of large, sparse models View paper
- [64] Dynamicvit: Efficient vision transformers with dynamic token sparsification View paper

- [65] Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed View paper
- [66] The sparse frontier: Sparse attention trade-offs in transformer llms View paper