# Diffusion vs. Autoregressive Models: Core Mechanics in Image Generation

## I. Introduction

Generative models, a cornerstone of modern artificial intelligence, are designed to learn the underlying probability distribution of a dataset, enabling the creation of new data points that resemble the training data. Their significance has grown exponentially, particularly in the domain of image generation, which finds increasing application across diverse sectors such as content creation, scientific visualization, and medical imaging [1]. Within this dynamic landscape, diffusion models and autoregressive models have emerged as two of the most prominent and powerful approaches, each with distinct methodologies for synthesizing novel visual content.

The capabilities of these models have profoundly impacted the field of generative AI, demonstrating remarkable abilities in producing high-quality and diverse images that were once considered the exclusive domain of human creativity [1]. The adoption of diffusion models by major AI research entities underscores their effectiveness and maturity, yet the continued exploration and advancement of autoregressive alternatives indicate an ongoing quest for optimal generative strategies [3]. This report aims to elucidate the fundamental working principles, or core mechanics, of both diffusion and autoregressive models in the context of image generation. By comparing and contrasting their methodologies, strengths, and weaknesses, and by examining recent breakthroughs, this analysis seeks to provide a comprehensive technical understanding of these pivotal generative techniques.

The increasing demand for realistic and creative image synthesis highlights the practical relevance of these models. Their applications extend beyond mere aesthetic generation, offering tools for problem-solving and innovation in various scientific and industrial fields. The fact that leading AI organizations are actively investing in diffusion models suggests a recognition of their current capabilities and future potential. However, the parallel advancements in autoregressive models, as evidenced by ongoing research and the development of novel architectures, indicate that the optimal paradigm for image generation remains an active area of investigation.

## II. Core Mechanics of Diffusion Models

The operational framework of diffusion models centers around a two-stage process: a forward diffusion process that progressively adds noise to an image, and a reverse diffusion process that learns to reconstruct the image from this noise [2].

The forward diffusion process, also known as noising, systematically transforms a clear, original image into a state of seemingly random noise through a series of iterative steps [2]. At each step of this process, a small amount of Gaussian noise is added to the image, gradually degrading its visual coherence [4]. This process can be mathematically formulated, with the amount of noise added at each step *t* being determined by a variance schedule (β<sub>t</sub>) [4]. The equation governing this transformation is often represented as x<sub>t+1</sub> = √(1 - β<sub>t</sub>) x<sub>t</sub> + √β<sub>t</sub> ε, where x<sub>t</sub> is the image at time step *t*, and ε is the Gaussian noise [4]. This iterative addition of noise follows a Markov chain, meaning that the

state of the image at any given step depends only on its state at the previous step [2]. This controlled degradation ultimately leads to an image that is statistically indistinguishable from pure Gaussian noise. This transformation is deterministic in the sense that for a given image and noise schedule, the noising process will always yield the same final noisy image. Furthermore, this process gradually dismantles the inherent structure within the original image's data distribution, ultimately converting it into a well-defined distribution characterized by random noise [2].

The reverse diffusion process is the generative aspect of these models. Starting from an image of random noise, the model learns to invert the forward process, iteratively removing the noise to gradually reveal the underlying image [2]. The core idea is that if the noise added at each step of the forward process is known, the process can be precisely reversed to recover the original image [4]. During training, the model learns to predict the noise that was added at each step of the forward diffusion [4]. In the reverse process, the model takes a noisy image as input and predicts the noise component, which is then subtracted from the image to produce a slightly less noisy version [4]. This denoising step is repeated iteratively, with the model progressively refining the image until a coherent and realistic sample emerges from the initial random noise [4]. The learning in this phase is crucial; the model doesn't simply reverse the noise but rather learns the statistical patterns of the training data distribution. Consequently, when the trained reverse process is initiated with random noise, it generates novel samples that conform to the learned distribution [1].

At the heart of the reverse diffusion process lies a neural network, often employing a U-Net architecture, which is trained to perform the denoising [4]. This network typically consists of an encoder path to downsample the input image to lower resolutions, a bottleneck for processing, and a decoder path to upsample it back to the original size [4]. Residual connections and attention mechanisms are frequently incorporated to enhance the network's ability to capture both fine-grained details and long-range dependencies within the image [4]. The network takes as input a noisy image and the current noise level (represented by the diffusion time step) and outputs a prediction of the noise that should be removed [4]. The training of this denoising network involves minimizing the discrepancy between the predicted clean image and the actual original image that corresponded to the noisy input [5]. The U-Net architecture, with its encoder-decoder structure and attention layers, allows the model to effectively process information at different spatial scales, enabling it to understand and remove noise while preserving or recovering the underlying image structure [4].

Image generation using diffusion models is inherently an iterative refinement process [5]. Starting with a completely noisy image, the model progressively applies the learned denoising steps, gradually transforming the random noise into a meaningful visual output [5]. This iterative nature is closely linked to the observed coarse-to-fine generation characteristic of diffusion models [7]. Early denoising steps tend to establish the overall structure and low-frequency components of the image, while subsequent steps add finer details and high-frequency information [7]. While this step-by-step refinement contributes significantly to the high quality of the generated images, it also introduces a key limitation: the large number of iterations required for sampling leads to slower inference speeds compared to some other generative approaches [3].

**III. Core Mechanics of Autoregressive Models**

Autoregressive models, in contrast to the iterative denoising of diffusion models, generate data by predicting one element of a sequence at a time, with each prediction conditioned on all the preceding elements in the sequence [9]. This sequential generation is analogous to how large language models predict the next word in a sentence based on the words that have come before [10]. The fundamental principle underlying autoregressive models is the assumption that a meaningful sequential order exists within the data, and this order can be leveraged to model the conditional dependencies between data points for predictive generation [10].

Early autoregressive models for image generation focused on predicting the intensity of individual pixels in a sequential manner, often following a raster-scan order (top to bottom, left to right) [11]. Models like PixelRNN (Recurrent Neural Network) and PixelCNN (Convolutional Neural Network) were pioneers in this approach [10]. PixelRNN utilized LSTM (Long Short-Term Memory) layers to capture the dependencies between previously generated pixels, while PixelCNN employed masked convolutions to ensure that the prediction of a pixel was only based on the values of pixels that had already been generated [11]. Although these models demonstrated the feasibility of autoregressive image generation, the inherent sequential nature of predicting each pixel individually made them computationally intensive and slow, particularly for generating high-resolution images [11]. The process of generating one pixel at a time implies a direct linear relationship between the number of pixels and the computational cost, rendering it impractical for large images with the computational resources available at the time.

More contemporary autoregressive models have adopted transformer architectures, which have proven highly successful in sequence modeling tasks within natural language processing [15]. To adapt these architectures for image generation, images are first transformed into a sequence of discrete visual tokens using techniques such as VQGAN (Vector Quantized Generative Adversarial Network) [16]. VQGAN employs a vector quantization approach to compress the image into a lower-dimensional discrete representation, which can then be flattened into a one-dimensional sequence of tokens [16]. The transformer model, leveraging its self-attention mechanisms, then learns to predict the next visual token in this sequence, conditioned on the previously generated tokens [13]. This "next-token prediction" paradigm is central to how these modern autoregressive models generate images [13]. The self-attention mechanism within the transformer architecture allows the model to weigh the importance of different parts of the previously generated sequence when predicting the next token, enabling it to capture long-range dependencies across the image [15].

The core mechanic of autoregressive image generation, regardless of the specific architecture, revolves around the prediction of the probability distribution of the subsequent element (be it a pixel or a visual token) given the sequence of elements generated so far [10]. The final image is then constructed by iteratively sampling from these conditional probability distributions [12]. Starting with an initial (often empty or context-providing) sequence, the model predicts the distribution of the next element, samples from this distribution to obtain a concrete value, and appends this value to the sequence. This augmented sequence then serves as the context for predicting the distribution of the subsequent element, and the process continues until the entire image (represented as a sequence of elements) is generated [12]. The quality of the resulting image is heavily dependent on the model's ability to accurately learn and model the complex conditional dependencies that exist between the constituent elements of an image [12].

## IV. Key Differences in Core Mechanics

The fundamental approaches employed by diffusion and autoregressive models for image generation exhibit several key distinctions in their core mechanics. Diffusion models operate through an iterative denoising process, starting from random noise and progressively refining it into a coherent image [7]. This involves a reverse process that learns to undo the gradual addition of noise that occurred in the forward diffusion stage. In contrast, autoregressive models generate images sequentially, predicting one element (pixel or token) at a time, conditioned on the elements that have already been generated [7]. This is akin to building the image step by step, relying on the statistical relationships between consecutive parts. While diffusion models have the potential for parallelizing the denoising steps, allowing for simultaneous processing of different parts of the image, traditional autoregressive generation is inherently sequential, requiring the completion of one step before the next can begin [9].

Another significant difference lies in how these models handle data. Diffusion models naturally operate in the continuous space of pixel values, directly manipulating and refining images represented as continuous data [9]. Autoregressive models, particularly the more modern transformer-based ones, often rely on discretizing the image into a sequence of tokens [9]. This discretization step involves converting the continuous pixel data into a finite vocabulary of visual tokens, which the autoregressive model then predicts sequentially. This difference in data handling can have implications for the types of tasks each model is best suited for. For instance, diffusion models have faced challenges when directly generating discrete textual data, as seen in tasks like image captioning, due to the inherent discrepancy between the continuous diffusion process and the discrete nature of text [9].

Finally, the directionality of information flow differs between the two approaches. Traditional autoregressive models, especially those following a raster-scan generation order, exhibit a unidirectional flow of information [9]. The prediction of a given pixel or token is conditioned only on the pixels or tokens that precede it in the defined sequence. This can limit the model's ability to capture bidirectional dependencies present in natural images, where relationships between different parts of the image can be complex and non-sequential. Diffusion models, on the other hand, through their iterative denoising process, can consider the entire image context more holistically at each step [9]. The denoising network has access to all parts of the noisy image and learns to refine it based on the global statistical patterns of the training data, potentially mitigating some of the limitations associated with unidirectional information flow. These fundamental distinctions in their core mechanics ultimately shape the strengths and weaknesses of each model type in terms of image quality, generation speed, and controllability.

## V. Advantages and Disadvantages

Both diffusion and autoregressive models possess unique sets of advantages and disadvantages when applied to the task of image generation.

Diffusion models have demonstrated a remarkable ability to generate images of very high quality, often achieving results that are perceived as photorealistic [1]. Their iterative denoising process allows for a holistic consideration of the image context, enabling them to capture complex details and global coherence [9]. Furthermore, the inherent structure of the denoising process lends itself to potential parallelization, which can be exploited to accelerate the

inference speed [9]. Diffusion models also exhibit flexibility in various image manipulation tasks, such as inpainting (filling in missing regions) and editing, by conditioning the denoising process on specific constraints or modifications [1]. However, a significant drawback of traditional diffusion models is their slow inference speed, primarily due to the large number of iterative denoising steps required to generate a single image [3]. They also face challenges when tasked with directly generating discrete data, such as text, due to the mismatch between their continuous operational space and the discrete nature of language [9]. Additionally, under certain conditions, particularly when using high classifier-free guidance weights to improve image quality, diffusion models can sometimes suffer from a lack of diversity in the generated samples [22].

Autoregressive models, with their roots in sequence modeling, naturally excel at generating sequential data like text, and this paradigm can be adapted to image generation by treating images as sequences of pixels or tokens [10]. They have shown strong performance in tasks that require adherence to specific rules or patterns, such as completing missing parts of an image based on context [23]. The unified tokenized representation employed by many modern autoregressive models simplifies the creation of foundational models that can be applied to various tasks [24]. However, a primary disadvantage of autoregressive models is their inherently sequential generation process, which can lead to slow inference speeds, especially for high-resolution images [20]. They are also susceptible to error propagation, where mistakes made in the early stages of generation can compound and negatively impact the quality of subsequent parts of the image [9]. Traditional autoregressive models that follow a fixed generation order, like raster scan, can be limited by their unidirectional constraints in capturing the complex, often bidirectional, dependencies present in natural images [9]. Moreover, generating high-resolution images pixel by pixel or token by token can be computationally very demanding [17].

The strengths of one model type often counterbalance the weaknesses of the other. For instance, diffusion models are known for their high image quality but suffer from slow speed, while autoregressive models are naturally suited for sequential data but can be prone to error accumulation. This suggests that the choice between the two paradigms often involves a trade-off depending on the specific requirements and priorities of the application.

## VI. Recent Advancements and Hybrid Approaches

The field of generative image modeling is rapidly advancing, with significant innovations occurring in both diffusion and autoregressive approaches. These advancements are often aimed at addressing the inherent limitations of each paradigm and pushing the boundaries of image quality, generation speed, and controllability.

One notable advancement in the autoregressive domain is Visual Autoregressive Modeling (VAR), which introduces a "next-scale prediction" or "next-resolution prediction" strategy [16]. This approach deviates from the traditional raster-scan method of predicting the next token in a sequence. Instead, VAR generates multi-scale token maps, starting from a coarse representation and progressively refining it to higher resolutions [16]. This coarse-to-fine generation allows the model to learn visual distributions more quickly and generalize more effectively [16]. VAR has demonstrated significant improvements in both image quality (as measured by FID) and inference speed, in some cases surpassing the performance of diffusion transformer models [16]. Notably, VAR models exhibit power-law scaling laws, similar to those observed in large language models, indicating a predictable relationship between model size

and performance [16]. Furthermore, VAR showcases strong zero-shot generalization capabilities in downstream tasks such as image inpainting, outpainting, and editing, suggesting a robust understanding of visual concepts [16]. This "next-scale prediction" strategy represents a substantial step forward for autoregressive image generation, effectively tackling the limitations of earlier sequential prediction methods.

Another important direction in autoregressive models is the development of techniques for incorporating control signals into the generation process. Frameworks like ControlAR enable users to guide the image generation using various conditioning inputs, such as text descriptions, object masks, and edge maps [24]. ControlAR typically employs a control encoder to transform these spatial or textual inputs into a sequence of control tokens [28]. These control tokens are then used during a conditional decoding phase, where the prediction of the next image token is conditioned not only on the previously generated image tokens but also on the current control token [28]. This approach allows for precise control over the generated image content and style. Surprisingly, ControlAR has also been shown to empower autoregressive models with the ability to generate images at arbitrary resolutions, overcoming a traditional limitation of these models [28]. The development of controllable autoregressive models significantly broadens their applicability by allowing for targeted image synthesis based on user-specified conditions.

In parallel, significant efforts have been made to address the slow sampling speed of diffusion models. Parallel sampling techniques, such as ParaDiGMS and ParaTAA, aim to accelerate the denoising process by performing multiple denoising steps concurrently [8]. These methods often leverage iterative techniques like Picard iterations or fixed-point iteration to guess the full denoising trajectory and iteratively refine it until convergence, allowing for the parallel computation of what were previously sequential steps [8]. These techniques have demonstrated significant speedups in the sampling process without substantial degradation in the quality of the generated images [32]. Furthermore, methods like ParaDiGMS are often compatible with existing fast sequential sampling techniques like DDIM and DPMSolver, allowing for a combination of approaches to achieve even greater efficiency [8]. Overcoming the speed bottleneck is crucial for the wider adoption of diffusion models in real-time applications.

Interestingly, diffusion models are also making inroads into tasks traditionally dominated by autoregressive models. Recent work, such as the development of LaDiC, has shown the potential of diffusion models for image-to-text generation, specifically image captioning [9]. LaDiC employs a novel architecture that includes a split BERT to create a dedicated latent space for captions and integrates a regularization module to manage varying text lengths [9]. The model leverages the holistic context modeling and parallel decoding capabilities inherent to diffusion models for this task [9]. LaDiC has achieved state-of-the-art performance among diffusion-based methods on standard image captioning datasets, demonstrating strong competitiveness with autoregressive models in this domain [9]. This success challenges the notion that diffusion models are inherently less suitable for discrete sequence generation.

Finally, research has also explored the underlying connections between diffusion and autoregressive models. Spectral analysis has revealed that diffusion models can be interpreted as performing approximate autoregression in the frequency domain [7]. This perspective suggests that the coarse-to-fine generation observed in diffusion models is related to the processing of different spatial frequency components of the image [7]. Understanding this relationship can provide deeper insights into the fundamental mechanisms of both approaches and potentially

inspire the development of novel hybrid models or optimization strategies that leverage the strengths of both paradigms.

**VII. Image Quality and Performance Benchmarks**

Evaluating the performance of image generation models requires the use of quantitative metrics that can assess the quality and diversity of the generated images. Common evaluation metrics in this field include the Fréchet Inception Distance (FID) and the Inception Score (IS) [13]. The FID measures the distance between the distributions of generated images and real images from the training dataset, with a lower FID score indicating higher fidelity and better similarity to real images [21]. The IS, on the other hand, assesses the sharpness and clarity of the generated images and their diversity, with a higher IS generally indicating better quality and diversity [21].

The following table summarizes the performance of various diffusion and autoregressive models on the ImageNet 256x256 benchmark, using FID and IS scores where available, along with inference times where reported.

| Model Type | Model Name | Dataset | FID | IS | Inference Time | Notes | Snippet(s) |
|---|---|---|---|---|---|---|---|
| Autoregressive | VQGAN-re | ImageNet 256x256 | 5.20 | - | - | Baseline AR model | [21] |
| Autoregressive | VAR-re | ImageNet 256x256 | 1.73 | 350.2 | ~0.3s | VAR model | [16] |
| Diffusion | DiT | ImageNet 256x256 | - | - | Slow | Baseline Diffusion Transformer | [3] |
| Diffusion | Stable Diffusion 3 | - | - | - | - | Outperforms SDXL | [35] |
| Autoregressive | Infinity | - | 0.73 (GenEv | 0.96 (Image | 0.8s (1024x1 | Faster than | [35] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | al) | Reward) | 024) | SD3-Medium, outperforms SDXL | |
| Autoregressive | LlamaGen | - | - | - | - | Competitive with diffusion models | [13] |
| Diffusion | FLUX | - | - | - | - | Strong performance in high-quality generation | [36] |
| Autoregressive | MAR (with Diff Loss) | ImageNet 256x256 | <2.0 | - | <0.3s | Achieves strong FID with fast speed | [19] |
| Autoregressive | VAR | ImageNet 256x256 | 1.80 | 356.4 | 20x faster than baseline AR | Surpasses DiT in multiple dimensions | [17] |
| Diffusion (LaDiC) | LaDiC | MS COCO | 38.2 (BLEU@4) | 126.2 (CIDEr) | - | State-of-the-art for diffusion in image-to-text | [9] |

These benchmark results indicate that recent advancements, particularly in autoregressive modeling with the introduction of VAR, have significantly improved the performance of these models, achieving FID scores comparable to or even better than those of diffusion models on

standard datasets like ImageNet [16]. For instance, VAR has demonstrated a substantial improvement over traditional autoregressive models and exhibits competitiveness with top diffusion models in terms of both image quality and inference speed [16]. The choice between diffusion and autoregressive models may ultimately depend on the specific application and the desired balance between image quality, generation speed, and controllability. Some benchmarks, like T2I-CompBench, are specifically used to evaluate the compositional generation capabilities of text-to-image models, providing insights into how well these models handle prompts with novel or complex compositions [36].

## VIII. Speed and Efficiency Considerations

A crucial aspect in the practical application of image generation models is their speed and efficiency, particularly during inference. Generally, traditional diffusion models have been noted for their slower inference speeds compared to autoregressive models due to the iterative nature of the denoising process [3]. However, recent advancements in both paradigms are addressing these limitations.

The development of models like VAR has led to significant improvements in the inference speed of autoregressive models [16]. For example, VAR has been reported to achieve inference speeds up to 20 times faster than baseline autoregressive models while also improving image quality [16]. Similarly, Infinity, another recent autoregressive model, has demonstrated impressive speed, generating high-resolution images in under a second and outperforming some diffusion models in speed [35]. On the other hand, the speed limitations of diffusion models are being actively tackled through the development of parallel sampling techniques like ParaDiGMS and ParaTAA, which allow for the simultaneous computation of multiple denoising steps, leading to substantial reductions in sampling time [8]. These techniques can achieve speedups of 2 to 4 times across various diffusion models without significant loss in image quality [32].

Computational costs, both for training and inference, are also important considerations. While a detailed comparison of the computational resources required by different models is beyond the scope of this report, it is worth noting that the data efficiency of models like VAR, which can achieve strong performance with fewer training epochs, contributes to overall efficiency [16]. Optimization techniques, such as distillation for diffusion models, where a smaller, faster model is trained to mimic the output of a larger, more computationally intensive model, and the "next-scale prediction" strategy in VAR, which reduces the sequential dependencies, are key strategies for improving the efficiency of both diffusion and autoregressive approaches [3]. The ongoing research in this area underscores the importance of speed and efficiency for the practical deployment of these powerful generative models.

## IX. Conclusion

In summary, diffusion models and autoregressive models represent two distinct yet powerful paradigms for image generation, each with its own core mechanics. Diffusion models excel through an iterative process of denoising, allowing for holistic context consideration and the generation of high-quality images. Their main drawback has been the slow inference speed due to the sequential nature of the denoising steps. Autoregressive models, on the other hand, generate images sequentially, predicting one element at a time, a method that aligns well with sequence data and tasks requiring rule-consistent generation. However, they have traditionally

suffered from slow speed, error propagation, and limitations in capturing bidirectional dependencies.

Recent advancements in both fields are rapidly addressing these limitations. The emergence of Visual Autoregressive Modeling (VAR) with its "next-scale prediction" strategy has significantly improved the speed and quality of autoregressive image generation, in some cases surpassing diffusion models in performance metrics. Techniques like ControlAR are also enhancing the controllability of autoregressive models, making them more versatile for various applications. Simultaneously, parallel sampling methods for diffusion models, such as ParaDiGMS and ParaTAA, are effectively tackling the issue of slow inference, making diffusion models more viable for real-time applications. Furthermore, the successful application of diffusion models to tasks like image captioning demonstrates their growing versatility beyond continuous data generation. The spectral analysis revealing a connection between these two approaches suggests a deeper relationship that could inspire future innovations.

The choice between diffusion and autoregressive models for image generation is increasingly nuanced and depends on the specific requirements of the task. Factors such as the desired image quality, the importance of generation speed, and the need for controllability will likely dictate which paradigm is most suitable. The rapid evolution of both technologies suggests a future where these lines may become even more blurred, potentially leading to hybrid models that leverage the strengths of both approaches to achieve even more remarkable results in generative AI.

## Works cited


1. How A.I. Creates Art - A Gentle Introduction to Diffusion Models | Weaviate, accessed March 15, 2025, https://weaviate.io/blog/how-ai-creates-art
2. Diffusion Models for Image Generation – A Comprehensive Guide - LearnOpenCV, accessed March 15, 2025, https://learnopencv.com/image-generation-using-diffusion-models/
3. Diffusion versus Auto-regressive models for image generation. Which is better? [D] [R], accessed March 15, 2025, https://www.reddit.com/r/MachineLearning/comments/1c53pc5/diffusion_versus_autoregressive_models_for_image/
4. Generate Images Using Diffusion - MathWorks, accessed March 15, 2025, https://www.mathworks.com/help/deeplearning/ug/generate-images-using-diffusion.html
5. Diffusion Models for Image Generation: Reversing the Degradation - DataForest, accessed March 15, 2025, https://dataforest.ai/blog/diffusion-models-for-image-generation-reversing-the-degradation
6. Understanding Image Generation with Diffusion | by Deven Joshi - Medium, accessed March 15, 2025, https://medium.com/@dev.n/understanding-image-generation-with-diffusion-78eea7e7d6f8
7. Diffusion is spectral autoregression - Sander Dieleman, accessed March 15, 2025, https://sander.ai/2024/09/02/spectral-autoregression.html
8. proceedings.neurips.cc, accessed March 15, 2025, https://proceedings.neurips.cc/paper_files/paper/2023/file/0d1986a61e30e5fa408c81216a616e20-Paper-Conference.pdf
9. LaDiC: Are Diffusion Models Really Inferior to Autoregressive Counterparts for Image-to-Text Generation? - ACL Anthology, accessed March 15, 2025,

https://aclanthology.org/2024.naacl-long.373.pdf
10. What are Autoregressive Models? - AR Models Explained - AWS, accessed March 15, 2025, https://aws.amazon.com/what-is/autoregressive-models/
11. Generating High-Resolution Images Using Deep Autoregressive Models - Medium, accessed March 15, 2025, https://medium.com/towards-data-science/generating-high-resolution-images-using-autoregressive-models-3683f9af0db4
12. A Survey of Autoregressive Models for Image and Video Generation - Saqib Azim, accessed March 15, 2025, https://saqib1707.github.io/assets/pubs/autoregressive_generation_survey.pdf
13. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation - Reddit, accessed March 15, 2025, https://www.reddit.com/r/StableDiffusion/comments/1dd93ry/autoregressive_model_beats_diffusion_llama_for/
14. The Diffusion Revolution: How Parallel Processing Is Rewriting the Rules of AI Language Models | by Cogni Down Under | Mar, 2025 | Medium, accessed March 15, 2025, https://medium.com/@cognidownunder/the-diffusion-revolution-how-parallel-processing-is-rewriting-the-rules-of-ai-language-models-d6410f4bb938
15. Scalable Pre-Training of Large Autoregressive Image Models - viso.ai, accessed March 15, 2025, https://viso.ai/deep-learning/autoregressive-image-models/
16. NeurIPS Poster Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction, accessed March 15, 2025, https://neurips.cc/virtual/2024/poster/94115
17. Paper Review: Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction | by Andrew Lukyanenko, accessed March 15, 2025, https://artgor.medium.com/paper-review-visual-autoregressive-modeling-scalable-image-generation-via-next-scale-prediction-059c759139aa
18. viso.ai, accessed March 15, 2025, https://viso.ai/deep-learning/autoregressive-image-models/#:~:text=Autoregressive%20Image%20Modeling%20(AIM)%20uses,given%20the%20previous%20data%20point.
19. Autoregressive Image Generation without Vector Quantization - arXiv, accessed March 15, 2025, https://arxiv.org/html/2406.11838v1
20. [2404.10763] LaDiC: Are Diffusion Models Really Inferior to Autoregressive Counterparts for Image-to-Text Generation? - arXiv, accessed March 15, 2025, https://arxiv.org/abs/2404.10763
21. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction, accessed March 15, 2025, https://openreview.net/forum?id=gojL67CfS8
22. Kaleido Diffusion: Improving Conditional Diffusion Models with Autoregressive Latent Modeling - Apple Machine Learning Research, accessed March 15, 2025, https://machinelearning.apple.com/research/kaleido-diffusion
23. Diverse capability and scaling of diffusion and auto-regressive models when learning abstract rules - arXiv, accessed March 15, 2025, https://arxiv.org/html/2411.07873v1
24. ControlAR: Controllable Image Generation with Autoregressive Models | AI Research Paper Details - AIModels.fyi, accessed March 15, 2025, https://www.aimodels.fyi/papers/arxiv/controlar-controllable-image-generation-autoregressive-models
25. Distilled Decoding 1: One-step Sampling of Image Auto-regressive Models with Flow Matching | OpenReview, accessed March 15, 2025, https://openreview.net/forum?id=zKlFXV87Pp
26. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction - ChatPaper, accessed March 15, 2025, https://chatpaper.com/chatpaper/paper/79609

27. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction, accessed March 15, 2025, https://arxiv.org/html/2404.02905v1
28. ControlAR: Controllable Image Generation with Autoregressive Models - arXiv, accessed March 15, 2025, https://arxiv.org/html/2410.02705v1
29. ControlAR: Controllable Image Generation with Autoregressive Models - OpenReview, accessed March 15, 2025, https://openreview.net/forum?id=BWuBDdXVnH
30. [ICLR 2025] ControlAR: Controllable Image Generation with Autoregressive Models - GitHub, accessed March 15, 2025, https://github.com/hustvl/ControlAR
31. ControlAR: Controllable Image Generation with Autoregressive Models - Paper Details, accessed March 15, 2025, https://www.chatpaper.ai/dashboard/paper/50314bb0-a5f2-4fa2-917c-e8e84b046a95
32. NeurIPS Poster Parallel Sampling of Diffusion Models, accessed March 15, 2025, https://neurips.cc/virtual/2023/poster/71125
33. Parallel Sampling of Diffusion Models - OpenReview, accessed March 15, 2025, https://openreview.net/forum?id=bpzwUfX1UP&referrer=%5Bthe%20profile%20of%20Andy%20Shih%5D(%2Fprofile%3Fid%3D~Andy_Shih1)
34. Accelerating Parallel Sampling of Diffusion Models, accessed March 15, 2025, https://proceedings.mlr.press/v235/tang24f.html
35. [R]Infinity ∞ : Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis : r/MachineLearning - Reddit, accessed March 15, 2025, https://www.reddit.com/r/MachineLearning/comments/1htehx6/rinfinity_scaling_bitwise_autoregressive_modeling/
36. [2410.22775] Diffusion Beats Autoregressive: An Evaluation of Compositional Generation in Text-to-Image Models - arXiv, accessed March 15, 2025, https://arxiv.org/abs/2410.22775
37. NeurIPS Poster Autoregressive Image Generation without Vector Quantization, accessed March 15, 2025, https://neurips.cc/virtual/2024/poster/94905