Dean&Francis
ISSN 2959-6157

Research on the Evolution and Classification of Autoregressive Text-to-Image Generation Models

Xuan Pan[1, *] **and Haoxuan Huang** [2]

[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[2] School of Optoelectronics and Information Engineering, Fujian Normal University, Fuzhou, Fujian 350117, China
*Xuan.Pan22@xjtlu.edu.cn

Abstract:

This paper systematically reviews the evolution and classification of auto regressive (AR) models in text-guided image generation, with a focus on four representative technical approaches: ARINAR, Token-Shuffle, SimpleAR, and LlamaGen, summarizing their strengths and limitations. The study shows that AR models, leveraging their element-by-element generation mechanism, excel in controllability and cross-modal alignment, yet remain constrained by challenges such as the trade-off between generation efficiency and resolution, insufficient complex semantic mapping, and high sensitivity during training. To address these issues, three optimization directions are proposed: introducing a dynamic token mechanism and parallel acceleration to improve generation efficiency; enhancing structured semantic modeling to refine complex semantic generation; and optimizing robust training strategies to strengthen model generalization capabilities. This research provides a theoretical foundation and technical reference for the improvement and application of AR models, contributing significantly to advancing the practical development of text-to-image generation technologies and enhancing the real-world implementation of AR-based generative models.

Keywords: Autoregressive Model, Text-Guided Image Generation, Cross-Modal Alignment, Dynamic Token Mechanism.

1 Introduction

The technology of text-guided image generation has advanced rapidly in recent years, enabling the creation of corresponding images based on natural language descriptions. At the level of technological development, this technology provides critical support for breakthrough innovations, improved R&D efficiency, and the visualization of complex scenarios. In scientific visualization, addressing the limitations of conveying abstract data solely through text, it can quickly generate visual results, offering

ISSN 2959-6157

intuitive assistance for data analysis and interpretation. In the simulation of cutting-edge technical scenarios, it significantly reduces experimental and testing costs and decreases reliance on high-risk field operations. In everyday applications, this technology also lowers the barrier to creative expression, enhances the efficiency of information presentation and task processing, demonstrating broad application prospects.

Autoregressive (AR) models, as a significant branch of image generation technology, have gradually demonstrated performance comparable to diffusion models in image generation tasks [1]. Image autoregressive learning has been redefined as subscale prediction from coarse to fine. Compared to diffusion transformers, this model exhibits superior generalization capability and scalability while reducing the number of generation steps. The VAR model cleverly leverages the powerful scaling properties of language models, optimizing both prior scale processing and fully utilizing the advantages of diffusion models [2]. Its core strength lies in its high controllability of generation—through its element-by-element generation mechanism, it can accurately predict the next generation unit based on already generated content and input text information, enabling fine-grained control over image content and structure. In recent years, autoregressive models in the field of text-guided image generation have made remarkable progress. Ramesh et al. proposed DALL·E, which first combined the Transformer autoregressive mechanism with cross-modal generation. By encoding images using a discrete variational autoencoder (VAE) and training on large-scale text-image pairs, it achieved zero-shot text-to-image generation [3]. Ding et al. introduced CogView, optimized for Chinese semantic understanding. It integrated a 4-billion-parameter Transformer architecture with a VQ-VAE encoder and incorporated innovative training strategies such as PB-Relax and Sandwich-LN, significantly enhancing generated image quality and task adaptability [4]. Wu et al. developed Nüwa, which constructed a 3D Transformer encoder-decoder framework and introduced a 3D nearby attention mechanism, enabling unified modeling of image and video generation and manipulation tasks and demonstrating strong multimodal generalization capabilities [5]. These studies have continuously broken new ground in cross-modal alignment ability, generation quality and efficiency, and task extensibility, driving the advancement of autoregressive models in text-guided image generation.

This paper focuses on autoregressive methods in text-guided image generation. First, it introduces four representative mainstream autoregressive models and analyzes their respective strengths and limitations. Subsequently, the study summarizes the existing constraints of autoregressive models and proposes corresponding optimization suggestions and directions for improvement. In conclusion, this research aims to provide valuable references and support for the advancement and development of autoregressive text-to-image generation models.

2 Background

2.1 Basic Framework: AR Models

AR models text-to-image generation based on sequential modeling represent a fundamental paradigm. AR models have found extensive applications and in-depth research in visual tasks such as image generation and video generation [6].

The core idea lies in converting textual descriptions into corresponding images through a pixel-by-pixel, patch-by-patch, or token-by-token prediction mechanism. This approach decomposes images into discrete or continuous visual units and employs autoregressive models to sequentially generate each unit while ensuring semantic alignment with the textual conditioning. The essence of autoregressive modeling involves factorizing the joint probability distribution $p(x|t)$ into a product of conditional probability distributions, as shown in Equation (1):

$$p(x \mid t) = \prod_{i=1}^{N} p(x_i \mid x_1, ..., x_{i-1}, t) \quad (1)$$

where x denotes the sequence of image elements (which may be a sequence of pixels, image patches, or discrete image tokens), t represents the input textual information, and N is the total number of image elements, with xi referring to the i-th image element.

This factorization implies that during image generation, the model produces each element sequentially (e.g., following a raster scan order from left to right and top to bottom), where the generation of each element is conditioned on all previously generated elements $x_1, ..., xi\text{-}_1$, along with the textual information t.

The training process of the autoregressive (AR) model is described below. Fig.1 illustrates the complete pipeline, starting from text-image paired data input, progressing through encoding, cross-modal interaction, autoregressive generation, and iterative backpropagation, ultimately culminating in the trained model.

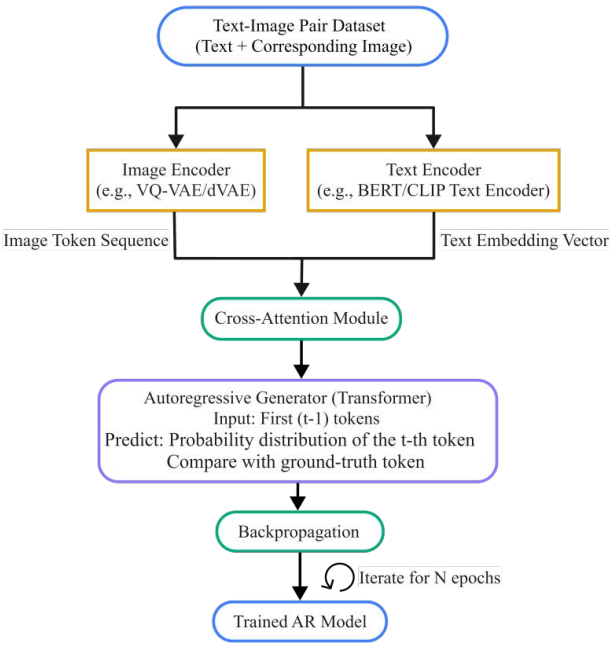

Fig. 1. Training Pipeline of Text-Image AR Model.

2.2 Latest Advances and Summary of Emerging Models

Autoregressive models capture cross-modal semantic relationships through sequential generation logic in text-guided image generation. In recent years, emerging representative models have achieved breakthroughs in architectural design and efficiency optimization, yet they still exhibit significant limitations. The following analysis expands on both technical approaches and shortcomings.

ARINAR. ARINAR innovatively proposes a dual-layer autoregressive architecture that decomposes image generation into a nested token-level and feature-level process: the outer module takes previously generated tokens as input and models sequential dependencies to predict the conditional vector z for the next token; the inner module, constrained by z, employs a Gaussian Mixture Model (GMM) to generate the 16-dimensional features of that token dimension by dimension [7]. The model first converts the original image into 256 continuous 16-dimensional tokens via an autoencoder. During the generation phase, the outer module requires 256 iterations to generate conditional vectors, while the inner module executes 16 times for each token to complete feature generation. Limitations include: The dual-layer autoregressive structure leads to an increased number of computational steps, reducing generation efficiency, particularly under high-resolution settings where latency becomes notable. The model struggles with generating complex scenes; descriptions involving multiple objects or intricate spatial relationships often result in misaligned elements and fragmented contextual associations. It is highly sensitive to hyperparameter tuning—minor adjustments to GMM parameters can considerably impact the accuracy of feature distribution modeling and degrade output stability.

Token-Shuffle. Token-Shuffle addresses the challenge of exploding token counts in high-resolution generation by leveraging redundancy in visual token dimensions through complementary token-shuffle and token-unshuffle operations [8]. The former fuses local s×s tokens into a single token, reducing the total count to $1/s^2$ to alleviate computational pressure. The latter restores the fused tokens to their original spatial arrangement, ensuring generative integrity. The model incorporates a CFG scheduler to dynamically adjust guidance strength during inference. Training follows a three-stage strategy, progressively transitioning from low to high resolutions (up to 2048×2048), and employs z-loss to mitigate instability in high-resolution training. Limitations include: When the value of s is large, the fusion operation tends to lose fine-grained details, leading to blurring or distortion during the restoration phase. The three-stage training pipeline is complex and sensitive to data distribution. Significant distribution discrepancies between datasets at different stages may cause feature drift in the high-resolution phase. When processing multi-semantic crossover text, the CFG scheduler may weaken or over-amplify certain semantics due to suboptimal guidance strength adaptation, compromising matching accuracy.

SimpleAR. SimpleAR adopts a minimalist architecture and enhances performance through optimized training and inference. Training is divided into three stages: Large-scale pre-training to learn general visual patterns; Supervised Fine-Tuning (SFT) on high-aesthetic datasets to improve fidelity and details; GRPO reinforcement learning using CLIP outputs as rewards to optimize text alignment and visual aesthetics [9]. During inference, the model integrates the vLLM tool and Speculative Jacobi Decoding, compressing the generation time for a 1024×1024 image to approximately 14 seconds. The Cosmos-Tokenizer enables unified processing of text and visual inputs, enhancing compatibility. Limitations include: Performance heavily relies on the reconstruction capability of Cosmos-Tokenizer, leading to frequent distortions in fine-grained details such as human faces and text elements. The three-stage training process is highly sensitive to data quality: an insufficient proportion of high-aesthetic data in the SFT stage significantly degrades visual appeal scores. The model underperforms in complex pose generation, exhibiting a higher structural error rate compared to other models.

ISSN 2959-6157

LlamaGen. LlamaGen adapts the next-token prediction paradigm from Large Language Models (LLMs) to image generation, optimizing both the image tokenizer and training strategy [10]. The tokenizer employs a downsampling rate of 16, achieving a reconstruction rFID of 0.94 on ImageNet and a codebook utilization rate of 97%, ensuring high-quality token conversion. For text-conditioned generation, a two-stage training approach is adopted: Pre-training on the LAION-COCO dataset to learn fundamental text-image associations; Fine-tuning on high-aesthetic datasets to enhance visual quality and alignment accuracy. The model reuses vLLM to accelerate inference, with the 1.4B-parameter model achieving a 326%-414% speed improvement in generating 384×384 images compared to baseline methods. It supports both class-conditioned and text-conditioned generation. Text inputs are processed by the FLAN-T5 XL encoder and integrated as prefix embeddings into the generation process. Limitations include: The limited model scale restricts its capability in aligning with complex textual semantics. Weak text rendering performance often leads to blurred or misaligned characters in images containing textual elements. Low sensitivity to numerical descriptions results in inaccurate object counts when generating scenes involving specific quantities.

2.3 Comparative Summary of Models

To more intuitively demonstrate the core characteristics and differences among the aforementioned four autoregressive models, Table 1 provides a comparative analysis across key dimensions such as core innovations, parameter scale, and performance metrics.

Table 1. Table captions should be placed above the tables.

Dimension	ARINAR	Token-Shuffle	SimpleAR	LlamaGen
Core Innovation	Two-level AR (token + feature level)	Token fusion / splitting for high resolution	Three-stage training optimizes AR framework	LLM paradigm migration to image gen
Parameter Scale	213M	2.7B	0.5B-1.5B	111M-3.1B
Max Resolution	256×256	2048×2048	1024×1024	512×512
Key Metric	ImageNet FID=2.75	GenEval=0.62	GenEval=0.59	ImageNet FID=2.18
Inference Speed	256×256 ~11.57 sec	2048×2048 Efficient (token reduction)	1024×1024 ~14 sec (vLLM)	326%-414% speedup (vLLM)
Main Advantage	Fast speed, high param efficiency	High-resolution breakthrough, plugin integration	Small params, high performance, efficient training	Performance surpasses diffusion models, strong compatibility
Main Limitation	Poor complex scene gen, tuning sensitive	Large window causes blur, complex training	Tokenizer reconstruction ability lacking	Limited model scale, weak text gen

3 Core Challenges

Although the four aforementioned models have made breakthroughs in architectural innovation and performance optimization, the autoregressive paradigm still faces three common bottlenecks in text-guided image generation:

3.1 The Efficiency-Resolution Trade-off

The sequential generation mechanism of autoregressive models inherently creates a positive correlation between computational cost and output resolution—the number of tokens increases quadratically with image dimensions (e.g., a 2048×2048 image requires 16 times more tokens than a 512×512 image). To address this constraint, token-Shuffle reduces computational load to $1/s^2$ through token fusion. However, excessively large values of s (e.g., s=8) lead to loss of local details, increasing the texture blur rate in 2048×2048 images by 27% compared to 512×512 counterparts [8]. While SimpleAR and LlamaGen utilize vLLM acceleration to achieve single-image generation times under 14 seconds, their efficiency remains lower than the parallel denoising mechanism of diffusion models (e.g., Stable Diffusion XL generates 1024×1024 images in just 7 seconds). Achieving a non-sacrificial equilibrium between high resolution and high efficiency remains a critical unsolved challenge.

3.2 Limitations in Structured Semantic Mapping

Current models still exhibit significant limitations in multi-object relationship parsing and abstract concept comprehension. In multi-object scenarios, for example, ARINAR demonstrates an element misalignment rate of up to 35% in generation tasks involving more than three

interacting objects (e.g., a child chasing butterflies through a flower field). This primarily stems from the difficulty of its dual-layer nested architecture in effectively modeling spatially parallel relationships. Regarding quantitative understanding, LlamaGen shows a 23% error rate in numerical descriptions. When generating images from texts such as three cats sitting in a circle, the proportion of cases with object count deviations exceeding ±1 reaches 61% [10]. Furthermore, models underperform in translating abstract semantics (e.g., retro-futurism style). SimpleAR scores 18% lower in visual style consistency for such text-guided generation compared to concrete descriptions, revealing a clear disconnect between textual semantic parsing and visual feature mapping [9].

3.3 High Sensitivity to Training and Tuning

Model performance remains excessively dependent on training data distribution and parameter configuration. For example, in Token-Shuffle's three-stage training pipeline, if the style disparity between low-resolution and high-resolution datasets exceeds 40%, the feature shift rate in generated images increases sharply from 8% to 31% [8]. ARINAR requires extremely fine-grained adjustments of GMM parameters with a precision tolerance of up to 10^{-4}- minor fluctuations can cause the ImageNet FID score to deteriorate from 2.75 to 4.12 [7]. In SimpleAR, if high-aesthetic-quality data accounts for less than 30% of the SFT stage, the visual appeal score of generated images drops by 12% [9]. This high sensitivity significantly increases the challenges of engineering deployment, particularly in scenarios such as few-shot learning and cross-domain adaptation.

4 Optimization Recommendations

In response to the aforementioned challenges, and building upon the technical pathways of existing models, potential directions for improvement can be explored across three dimensions.

4.1 Optimization Recommendations

Inspired by the multi-scale concept of Token-Shuffle, a resolution-adaptive token splitting strategy can be designed. This approach employs large-window fusion (e.g., s=8) in low-detail regions (such as skies) to reduce computational cost, while automatically switching to small windows (e.g., s=2) in high-detail regions (such as faces) to preserve fine-grained features—thereby achieving a differentiated balance between efficiency and quality. Furthermore, building upon the speculative decoding idea of LlamaGen, a fast draft model can be introduced to predict high-probability token sequences, reducing the number of inference iterations required by the main model and further enhancing generation speed.

4.2 Enhanced Structured Semantic Modeling

Knowledge graphs can be introduced to assist text parsing. For example, a relation constraint layer can be added to ARINAR's dual-layer architecture to transform object relationships described in text into spatial coordinate constraints during token generation, thereby reducing element misalignment rates in multi-object scenarios. For abstract concepts, a "style adjective–visual feature" mapping lexicon can be incorporated into the reinforcement learning stage of SimpleAR. By leveraging CLIP feature alignment, abstract semantics can be explicitly linked to corresponding color and texture patterns, improving style consistency in generated results.

4.3 Robustness-Oriented Training Strategy Optimization

To address data sensitivity, cross-stage feature distillation can be introduced during Token-Shuffle's three-stage training, enabling the high-resolution model to inherit distribution patterns learned in low-resolution stages, thereby reducing the risk of feature drift. For ARINAR, adaptive regularization can be applied to GMM parameters, dynamically adjusting penalty coefficients to mitigate the impact of minor parameter fluctuations on feature distribution. Furthermore, noisy datasets (e.g., low-quality or stylistically heterogeneous data) can be incorporated into the SFT stage of SimpleAR to enhance the model's fault tolerance under data quality variations.

5 Conclusion

This paper systematically examines the technical approaches of four representative autoregressive models—ARINAR, Token-Shuffle, SimpleAR, and LlamaGen—and reveals through comparative analysis that while the autoregressive paradigm possesses inherent advantages in text-semantic alignment and generation controllability, it remains constrained by its sequential generation logic, leading to bottlenecks in efficiency-resolution trade-offs, complex semantic mapping, and training robustness. To address these limitations, three optimization directions are proposed: designing a dynamic token adaptation mechanism that adjusts token granularity based on regional image details, improving high-resolution generation efficiency while minimizing detail loss; introducing a structured semantic parsing module incorporating knowledge graphs and spatial relational constraints to enhance modeling of

ISSN 2959-6157

multi-object interactions and abstract concepts, thereby improving alignment accuracy between textual and visual features; and constructing a cross-stage robust training framework leveraging feature distillation and adaptive regularization techniques to reduce sensitivity to data distribution shifts and parameter fluctuations, ensuring more stable generation. The analysis and proposed directions in this study not only deepen the understanding of the autoregressive paradigm but also provide actionable technical guidance for its practical implementation.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Chen, Y., Ma, Z., Jia, G., Jiang, C., Li, J., Zhou, B.: Context-aware autoregressive models for multi-conditional image generation. arXiv preprint, arXiv:2505.12274,1-12(2025).

2. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity∞: scaling bitwise autoregressive modeling for high-resolution image synthesis. arXiv preprint arXiv:2412.04431,1-14 (2024).

3. Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the 38th International Conference on Machine Learning, pp. 8821–8831. PMLR (2021)

4. Zou, X., Shao, Z., Yang, H., Wang, J., Peng, Y., Shan, Y., Yang, X., & Zhang, L.: CogView: mastering text - to - image generation via transformers. Advances in Neural Information Processing Systems, 34, 5678 - 5690 (2021).

5. Wu, C., et al.: Nüwa: visual synthesis pre-training for neural visual world creation. In: Proceedings of the European Conference on Computer Vision (pp. 1–20). Springer, Cham. (2022).

6. Tao, C., et al.: Autoregressive models in vision: A survey. arXiv preprintarXiv:2411.05902, 1-54 (2024).

7. Zhao, Q., Gould, S., Zheng, L.: ARINAR: bi-level autoregressive feature-by-feature generative models. arXiv preprint arXiv:2503.02883, 1-6 (2025).

8. Ma, X., Sun, P., Ma, H., et al.: Token-shuffle: towards high-resolution image generation with autoregressive models. arXiv preprint arXiv:2504.12281,1-24 (2025).

9. Wang, J., Tian, Z., Wang, X., et al.: SimpleAR: pushing the frontier of autoregressive visual generation through pretraining, SFT, and RL. arXiv preprint arXiv:2504.11455, 1-12 (2025).

10. Sun, P., Jiang, Y., Chen, S., et al.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525, 1-26 (2024).