SpectralAR: Spectral Autoregressive Visual Generation

Yuanhui Huang Weiliang Chen Wenzhao Zheng✉
Yueqi Duan Jie Zhou Jiwen Lu
Tsinghua University
huangyh22@mails.tsinghua.edu.cn; wenzhao.zheng@outlook.com

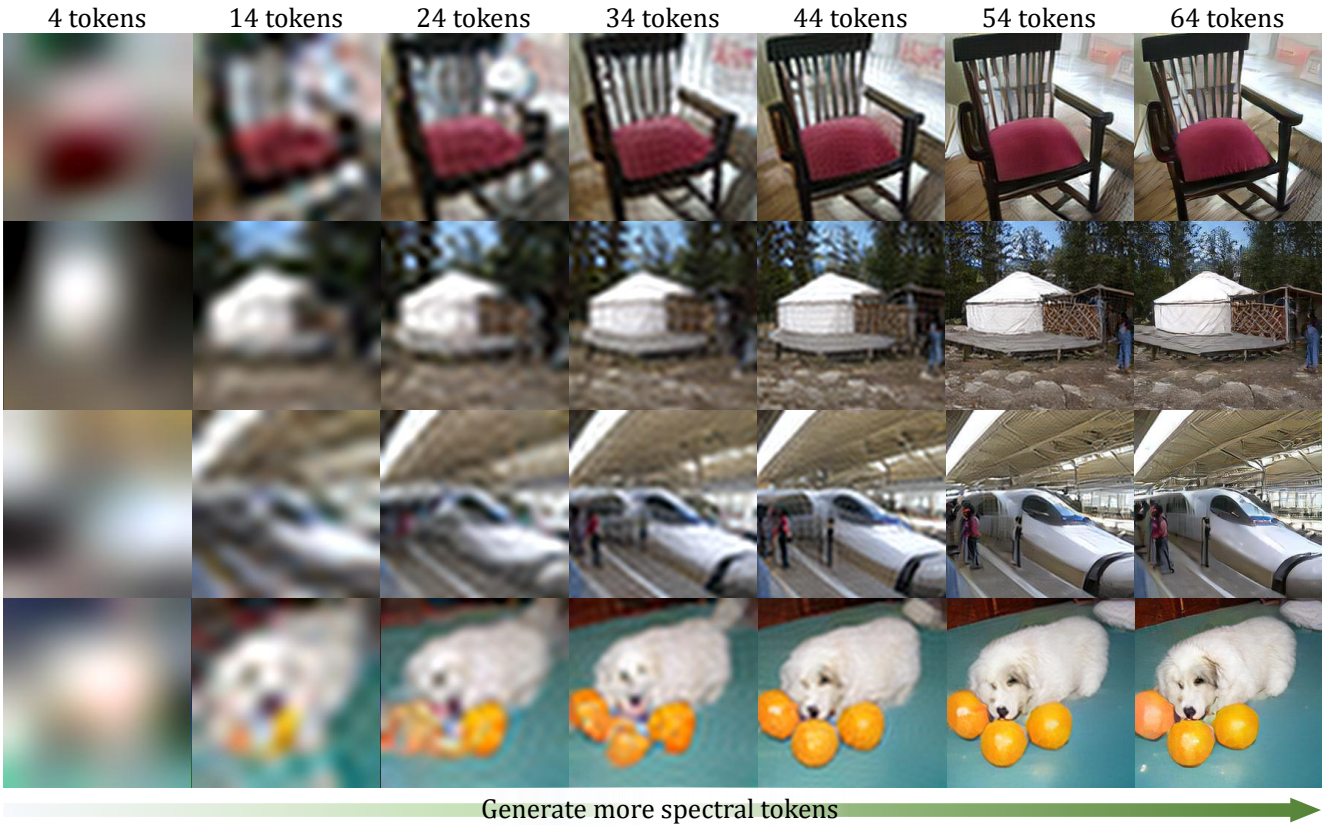

Figure 1. We approach autoregressive visual generation from the spectral perspective and propose **SpectralAR** which converts images into 1D causal sequences with nested spectral tokenization and generates images in a hierarchical coarse-to-fine manner. In the autoregressive process, each generated token improves the quality of the image by introducing new high-frequency components.

Abstract

Autoregressive visual generation has garnered increasing attention due to its scalability and compatibility with other modalities compared with diffusion models. Most existing methods construct visual sequences as spatial patches for autoregressive generation. However, image patches are inherently parallel, contradicting the causal nature of autoregressive modeling. To address this, we propose a Spectral AutoRegressive (SpectralAR) visual generation framework, which realizes causality for visual sequences from the spectral perspective. Specifically, we first transform an image into ordered spectral tokens with Nested Spectral Tokenization, representing lower to higher frequency components. We then perform autoregressive generation in a coarse-to-fine manner with the sequences of spectral tokens. By considering different levels of detail in images, our SpectralAR achieves both sequence causality and token efficiency without bells and whistles. We conduct extensive experiments on ImageNet-1K for image reconstruction and autoregressive generation, and SpectralAR achieves 3.02 gFID with only 64 tokens and 310M parameters. Project page: https://huang-yh.github.io/spectralar/.

✉ Corresponding author.

1. Introduction

Diffusion models [19, 39, 44, 46] have long been the best performing approach to visual generation. Despite their ex-

ceptional generation quality, diffusion models still exhibit deficiencies in multimodal modeling and integration of perception and generation. The advent of autoregressive visual generation methods [10, 29, 36, 42, 48, 50, 65] alleviates these limitations and enables better scalability with the next-token prediction paradigm. It first utilizes a visual tokenizer to convert images into tokens and then generates samples in a sequential manner. This advancement supports a variety of emerging applications, including scalable visual generation [16, 57, 59, 69], mixed-modal foundation models [49, 56, 60], and autoregressive world models [21, 73].

Despite their dominance in language modeling [40, 53], the performance of autoregressive models in visual generation [10, 14] is still inferior to that of diffusion models [35, 39] and non-autoregressive models [4, 66]. This distinction can be attributed to the inherent difference between text and image modalities. Text, invented for human communication, is discrete and sequential, while image data is continuous and invariant to translation [28], indicating the intrinsic equality among image pixels. This equality makes it a critical issue for autoregressive visual generation to convert images into one-dimensional sequential tokens [65]. Methods based on spatial scanning [10, 34, 47, 62] attempt to discretize the image locally into tokens according to patches, and then perform autoregressive generation following a certain order based on their locations. However, the resulting spatial sequence violates the equality among image patches, making it suboptimal for causal autoregressive modeling. Another line of work [4, 31, 66] introduces bidirectional interaction in the generator as a workaround. Nonetheless, they still tokenize images spatially, thus assuming a causal order among image patches. In addition, the bidirectional design may deviate from the conventional autoregressive paradigm, complicating their integration into omni-modal frameworks [49, 56]. In contrast to spatial tokenization, VAR [50] explores transforming the image into multiple scales and producing a sequence by concatenating tokens from ordered scales. Although scale-wise autoregressive generation indeed satisfies the equality of image pixels, it suffers from inferior token efficiency and parallel generation of multiple tokens from the same scale.

In this paper, we introduce a spectral autoregressive visual generation framework to achieve causal autoregressive modeling and improve token efficiency, as shown in Figure 2. Frequency is an inherent attribute of all types of signals and has become a significant perspective and methodology that complements the spatial-temporal domain [2]. For visual data, the spectral density often conforms to the power-law distribution [6], with low-frequency components representing the overall structure of an image and high-frequency components focusing on the intricate details. This hierarchical coarse-to-fine nature of the correspondence between spectral and spatial domains indicates a sequential order of images, motivating us to represent images as causal spectral sequences. Specifically, we first transform images into spectral tokens with Nested Spectral Tokenization (NST), which uses varying sequence lengths to represent images across different frequency bands. The causality of the spectral sequences originates from the coarse-to-fine progression characteristic resembling human visual perception and is strengthened with the application of a causal mask. In addition, inspired by the image compression literature [55, 58], we design a non-uniform token-frequency mapping, which allocates more tokens to represent low-frequency components and fewer for high-frequency components. This mapping technique greatly reduces the number of tokens, while maintaining the quality of reconstructed samples. In the autoregressive process, we begin with the token representing the DC component and progressively predict tokens corresponding to higher frequencies conditioned on previous ones. We conduct extensive experiments on the ImageNet-1K dataset for image reconstruction and class-conditional generation. Our SpectralAR demonstrates comparable performance with state-of-the-art methods.

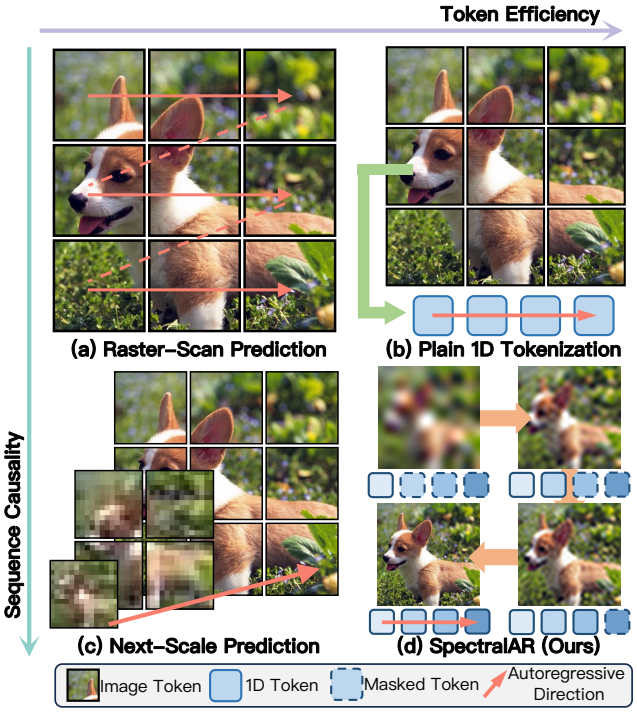

Figure 2. **Comparison between autoregressive visual generation methods.** SpectralAR achieves both token efficiency and sequence causality with nested 1D spectral tokens.

2. Related Work

Autoregressive visual generation. While diffusion models have achieved exceptional performance, they are fundamentally different from the conventional autoregressive framework [49, 56] for cross-modal and cross-task modeling. Therefore, a body of research aims to advance autore-

gressive models for visual generation. Early efforts [5, 51] perform pixel-level generation in the row-major raster-scan order, followed by VQGAN [10] which transfers to the latent feature space of VQVAE [52] for autoregressive modeling. Subsequent work improves based on VQGAN with multiple scales [42], residual quantization [29], ViT architecture [63] or textual conditions [64]. However, raster-scan generation violates the equality between image pixels as discussed in Section 1, contradicting the causality premise of autoregressive modeling. Recently, VAR [50] proposes the scale-wise autoregressive generation that aims to predict the next-scale token map conditioned on the previous ones. Despite achieving causality with the multi-scale design, VAR predicts multiple tokens with bidirectional attention in each step, thus deviating from the standard autoregressive framework. In contrast, our method realizes causal autoregressive generation from the spectral perspective and still conforms to the unidirectional scheme.

Efficient image tokenization. Autoregressive generation requires the conversion of images into token sequences, which is often achieved using autoencoders [18, 54]. Patch-based autoencoders [10, 24, 37, 52, 72] tokenize images spatially, where each token corresponds to a certain patch from the original image. Although this paradigm performs well in image reconstruction and diffusion-based generation [39, 44], it is not suitable for autoregressive modeling due to its spatial design. Also, its token length is proportional to the square of image resolution, which might become the bottleneck in multimodal modeling given limited context length [49]. TiTok [66] proposes a 1D tokenizer which reduces the number of tokens to 32 for 256×256 images. However, TiTok is trained with only an overall reconstruction objective, and thus the precise meaning of these 1D tokens remains unclear. VAR [50] introduces a multi-scale tokenizer for causal autoregressive image generation. Nonetheless, the multi-scale strategy requires an even greater number of tokens compared to patch-based tokenization methods, further diminishing token efficiency. Our method converts images into 1D causal sequences of spectral tokens and enhances token efficiency by leveraging the long-tail distribution of spectral density in image data.

Spectral visual analysis. Spectral analysis [2] has been a common technique in computer vision, complementing the spatial and temporal domains. Representative applications include image enhancement and denoising [15, 68], texture analysis and feature extraction [41], compression and super-resolution [13, 55], visual generation [61, 67], adversarial attacks and defenses [12, 32, 33]. For autoregressive image generation, CART [43] and SIT [11] propose to transform an image into multiple causal sets of tokens with base-detail decomposition and discrete wavelet transform, respectively. However, these methods still adhere to the multi-scale 2D tokenization paradigm similar to VAR, resulting in suboptimal token efficiency and bidirectional attention to predict multiple tokens per autoregressive step. In contrast, SpectralAR leverages the discrete cosine transform to capture the global information of an image, and compress it into a 1D sequence with high efficiency.

3. Proposed Approach

3.1. Revisiting Images from the Spectral Domain

Discrete cosine transform. Spectral analysis investigates how complex signals can be represented with simpler basis functions [2], producing a spectral density distribution that represents the magnitude of corresponding basic components. This spectral density distribution is an equivalent representation of the original signal and provides a distinct perspective from the spatial domain depending on the properties of the basis functions. We employ the Discrete Cosine Transform (DCT) [1] to convert images into the spectral domain. The DCT result $\mathbf{D}$ shares the same shape with the transformed image $\mathbf{I} \in \mathbb{R}^{H \times W}$ (for simplicity, we omit the channel dimension):

$$\mathbf{D} = \{F(u,v)\}_{u,v=1}^{W,H}, \quad \mathbf{I} = \{f(x,y)\}_{x,y=1}^{W,H}. \tag{1}$$

Each $F(u,v)$ represents the intensity of the corresponding basis function $g_{u,v}(x,y)$ in the image $\mathbf{I}$, which writes:

$$g_{u,v}(x,y) = \frac{2C(u)C(v)}{\sqrt{HW}} cos\frac{(2x+1)u\pi}{2W} cos\frac{(2y+1)v\pi}{2H}, \tag{2}$$

where $C(u) = 1/\sqrt{2}$ if $u = 0$ and $C(u) = 1$ otherwise. This family of basis functions has the following properties: (1) Given u and v, the basis function $g_{u,v}(x,y)$ exhibits a checkerboard-like pattern in the spatial domain, with periods along the x- and y-axis of $2W/u$ and $2H/v$, respectively. This pattern suggests that the basis function characterizes the rate of variation of images in the spatial domain. (2) The $F(u,v)$s also form a 2D matrix together, where the top-left corner represents the low-frequency components (small u, v), while the bottom-right corner corresponds to the high-frequency components (large u, v).

Causality from the spectral domain. Since low- and high-frequency components describe overall structures and intricate details, respectively, we can decompose an image into a sequence of sub-images $\{\mathbf{I}'_i\}_{i=1}^{L}$ with increasing levels of detail by applying inverse-DCT on partially masked spectral density distributions $\{\mathbf{D}'_i\}_{i=1}^{L}$. With more high-frequency components, the sub-images will gradually transition from blurred to sharp, as shown in Figure 2. This hierarchical coarse-to-fine sequence aligns with human visual perception and artistic drawing, thus enhancing causality.

Efficiency from the spectral domain. As the image compression literature [1] pointed out, the DCT result $\mathbf{D}$ of images conforms to the power-law distribution with the absolute values of $F(u,v)$s on the top-left corner substantially

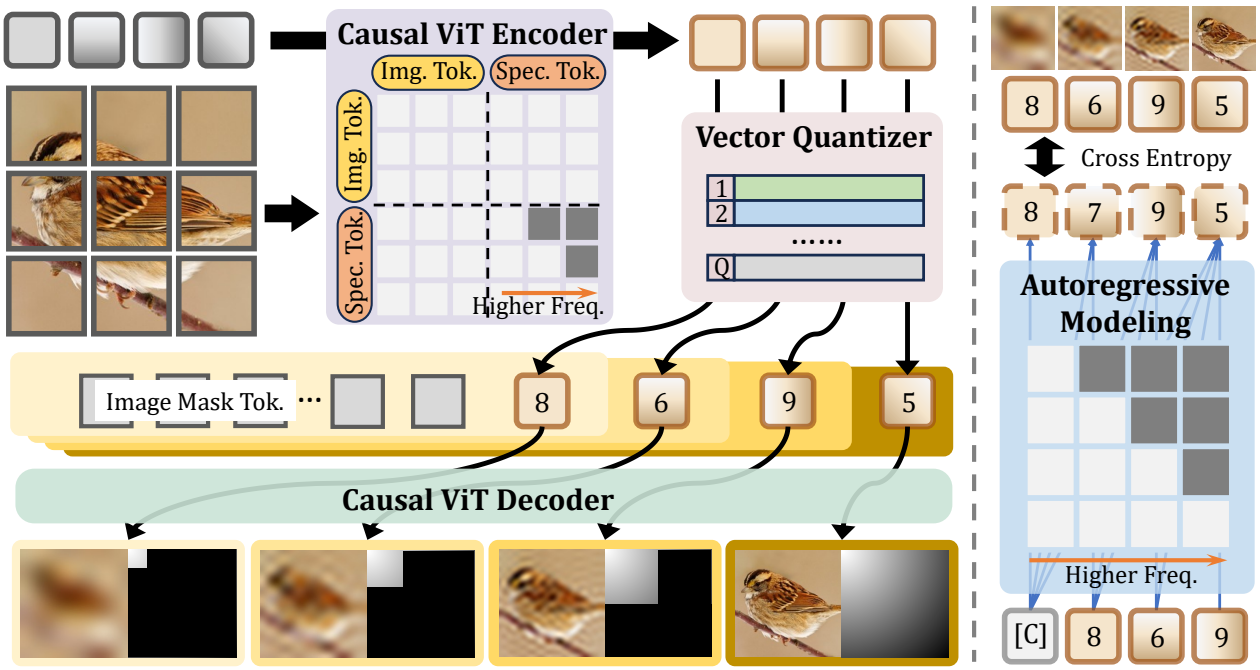

Figure 3. **Overall pipeline of the proposed SpectralAR.** Left: We convert an input image into a 1D causal sequence efficiently with nested spectral tokenization. Each nested sequence is supervised with distinct reconstruction target in a coarse-to-fine manner, which endows each token with an explicit spectral interpretation. We also apply the causal mask to the spectral tokens in the encoder and decoder to enhance the causality. Right: We use the groundtruth sequences from the tokenization process to train an autoregressive generation model.

larger than the bottom-right corner, indicating that most energy of an image is concentrated in the low-frequency components. In addition, human visual perception is less sensitive to high-frequency signals, which have little influence on the visual quality of images. Based on these distinctions, we can encode the high-frequency components of an image with coarser granularity to improve token efficiency, similar to the JPEG algorithm [55] which saves more than 90% storage for images by suppressing high-frequency signals.

3.2. Nested Spectral Tokenization

Overall framework. In contrast to the 2D spatial tokenization that captures the local correlation between image patches, the basis function $g_{u,v}(x, y)$ of DCT encodes a global frequency pattern. Therefore, we convert images into 1D tokens in spectral tokenization to reflect the global nature of the basis functions. We start with the general framework of a 1D image tokenizer [66]. Given an image $\mathbf{I}$, we aim to encode it into N discrete vectors $\mathbf{S}$, and also reconstruct the original image with $\mathbf{S}$. We first patchify the image into $\mathbf{P} \in \mathbb{R}^{hw \times C}$, and concatenate the image features with the initial query vectors $\mathbf{S}_0$ to form $[\mathbf{P}; \mathbf{S}_0] \in \mathbb{R}^{(hw+N) \times C}$, where h, w, C denote the resolution of image features and the channel dimension, respectively. We then employ the vision transformer $\mathcal{E}$ to enable feature extraction and interaction between the image features and 1D query tokens, resulting in the informative 1D representation $\hat{\mathbf{S}}$. In the vector quantizer $\mathcal{Q}$, we match these continuous vectors with the codebook embeddings to derive the discrete representation $\mathbf{S}$, which could serve as the groudtruth for the autoregressive training. At last, we append $\mathbf{S}$ to a set of mask tokens $\mathbf{M} \in \mathbb{R}^{hw \times C}$ and process them with the decoder network $\mathcal{D}$ similar to $\mathcal{E}$, in order to reconstruct the original image. This overall framework could be formulated as:

$$\hat{\mathbf{S}} = \mathcal{E}([\mathbf{P}; \mathbf{S}_0]), \quad \mathbf{S} = \mathcal{Q}(\hat{\mathbf{S}}), \quad \hat{\mathbf{I}} = \mathcal{D}([\mathbf{M}; \mathbf{S}]), \tag{3}$$

where $\hat{\mathbf{I}}$ denotes the reconstructed image. The training objective typically consists of multiple loss functions:

$$\mathcal{L}_{tok} = \|\hat{\mathbf{I}} - \mathbf{I}\|_2^2 + \|\hat{\mathbf{S}} - \mathbf{S}\|_2^2 + \mathcal{L}_P(\hat{\mathbf{I}}, \mathbf{I}) + \mathcal{L}_A(\hat{\mathbf{I}}, \mathbf{I}), \tag{4}$$

where $\mathcal{L}_P$ and $\mathcal{L}_A$ denote the perceptual loss [22, 70] and the adversarial loss [10, 14], respectively.

Nested spectral decoding. Different from the plain 1D tokenization in Figure 2, we aim to represent an image as its spectral decompositions $\{\mathbf{I}'_i\}_{i=1}^{L}$ for sequence causality, which requires establishing a mapping from 1D tokens to these sub-images. One naive way to achieve this would be dividing $\mathbf{S}$ into disjoint subsets and assigning them to model different sub-images independently. Similar to the multi-scale tokenization [50], this strategy would inevitably involve bidirectional interaction and diminish token efficiency because sub-images with finer detail would require increasingly more tokens to represent. In contrast, we propose a nested mapping scheme for efficient tokenization, as shown in the bottom of Figure 3. We first construct a

Algorithm 1: Nested Spectral Tokenization Training

Inputs: raw image $\mathbf{I}$, initial spectral tokens $\mathbf{S}_0 \in \mathbb{R}^{N \times C}$;
Hyperparameters: spectral levels $\{\omega_i | i = 1, ..., N\}$;
$\mathbf{P} = \text{patchify}(\mathbf{I})$, $\hat{\mathbf{S}} = \mathcal{E}([\mathbf{P}; \mathbf{S}_0])$, $\mathbf{S} = \mathcal{Q}(\hat{\mathbf{S}})$;
$idx = \text{random\_choice}(N)$;
$\mathbf{S}' = \mathbf{S}[: idx]$, $\hat{\mathbf{I}}' = \mathcal{D}([\mathbf{M}; \mathbf{S}'])$;
$\mathbf{D} = \text{DCT}(\mathbf{I})$, $\mathbf{D}' = \mathbf{D} \circ \mathbf{1}_{\omega_i}$, $\mathbf{I}' = \text{DCT}^{-1}(\mathbf{D}')$;
$loss = \mathcal{L}_{tok}(\hat{\mathbf{I}}', \mathbf{I}')$;
Return: $loss$ for optimization;

sequence of sub-images with increasing detail by progressively preserving larger regions in the spectral density $\mathbf{D}$:

$$\mathbf{I}'_i = \text{DCT}^{-1}(\mathbf{D}'_i), \quad \mathbf{D}'_i = \mathbf{D} \circ \mathbf{1}_{\omega_i}, \quad \omega_{i-1} < \omega_i, \quad (5)$$

where DCT^{-1}, $\circ$, $\mathbf{1}_{\omega_i}$ denote the inverse DCT operation, element-wise multiplication and a $H \times W$ matrix with the top-left corner of size $\omega_i \times \omega_i$ filled by ones and the remaining parts being zeros. Therefore, the sub-image $\mathbf{I}'_i$ contains all frequency components present in the sub-image $\mathbf{I}'_{i-1}$. Based on this inclusion property, we can reuse the tokens representing the previous sub-image $\mathbf{I}'_{i-1}$ to represent the next sub-image $\mathbf{I}'_i$. To avoid bidirectional attention, we make L equal to N so that each token $\mathbf{s}_i$ in the sequence $\mathbf{S}$ corresponds to a unique sub-image $\mathbf{I}'_i$:

$$\hat{\mathbf{I}}'_i = \mathcal{D}([\mathbf{M}; \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_i]), \quad (6)$$

where $\hat{\mathbf{I}}'_i$ is the reconstruction of sub-image $\mathbf{I}'_i$ given the nested 1D sequence $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_i\}$. Nested spectral decoding compresses an image into a causal 1D sequence where each token $\mathbf{s}_i$ corresponds to a disjoint set of frequencies and achieves token efficiency by reusing previous tokens.

Non-uniform token-frequency mapping. To further enhance token efficiency, we introduce the non-uniform token-frequency mapping technique. Since high-frequency components have low magnitude and minimal impact on the visual quality of images, we can encode them with coarser granularity compared to the low-frequency counterparts. We achieve this by reducing the interval between ω_{i-1} and ω_i when i is small, and increase it otherwise:

$$\omega_i - \omega_{i-1} \leq \omega_{i+1} - \omega_i, \quad (7)$$

which demonstrates the general case. This non-uniform mapping allocates later tokens to broader frequency ranges, enabling precise modeling of crucial low-frequency components while efficiently representing high-frequency details.

Spectral causal mask. Although the sequence $\mathbf{S}$ supervised with (5)(6) already exhibits a certain degree of causality, the encoding and decoding processes, i.e. $\mathcal{E}$ and $\mathcal{D}$, are still bidirectional, which can lead to information leakage from high-frequency to low-frequency components. Therefore, we propose applying causal masks to the spectral tokens $\mathbf{S}$ in both the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$, as shown in Figure 3. This spectral causal mask restricts each token $\mathbf{s}$ to only attend to tokens that represent lower frequencies, enhancing causality from the architectural perspective. We outline the training procedure in Algorithm 1.

Table 1. **Correlation between tokens of different autoregressive paradigms.** We use linear correlation as a proxy metric for the causality of sequences. The spectral sequence demonstrates better causality compared with other methods.

Correlation Type	Raster-scan	Scale-wise	Spectral
$R^2_{avg}(\mathbf{t}_2; \mathbf{t}_1)$	0.471	0.889	**0.916**
$R^2_{avg}(\mathbf{t}_3; \mathbf{t}_1, \mathbf{t}_2)$	0.366	0.953	**0.977**
$R^2_{avg}(\mathbf{t}_4; \mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3)$	0.525	0.943	**0.994**
Average	0.454	0.928	**0.962**

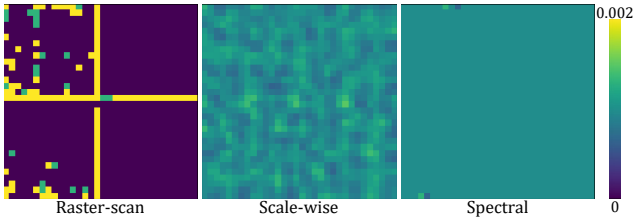

Figure 4. **Frequency of each element having the highest correlation with another one.** It represents the source of reference information for predicting the next token. The raster-scan method exhibits excessive dependency on boundaries of image patches.

3.3. Spectral Autoregressive Generation

Autoregressive modeling has gained prominence in computer vision for its scalability, generalization, and effectiveness across multimodal tasks [25, 49, 56]. While conventional autoregressive generation follows a spatial raster-scan order [10, 51], we propose a hierarchical coarse-to-fine approach in the spectral domain to enhance causality, as shown in the right side of Figure 3. We start with the general framework of autoregressive modeling:

$$\mathbf{p}_{i+1} = \mathcal{M}(\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_i), \quad (8)$$

where $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^N$ is a sequence of quantized tokens, and $\mathbf{p}_{i+1}$, $\mathcal{M}$ denote the probability logits for the (i+1)th token and the autoregressive model, respectively. The autoregressive process (8) assumes that the generation of token $\mathbf{t}_{i+1}$ depends solely on the previous ones. The spatial autoregressive paradigm violates this premise because of the equality of image pixels (as discussed in Section 1). In contrast, we take the spectral tokens $\mathbf{S}$ from the nested spectral tokenziation as the autoregressive targets $\mathbf{T}$. Since the spectral tokens are trained in a nested manner to reconstruct sub-images of increasing levels of detail as in (5)(6), each spectral token $\mathbf{s}_i$ is expected to enhance the quality of the sub-image $\mathbf{I}'_{i-1}$ represented by previous tokens from the spectral domain. This progressive refinement process aligns with human visual perception and artistic painting, both of which start with the overall structure and gradually focus on

Table 2. **Comparison between generative models on class-conditional ImageNet 256×256 benchmark.** "↓" or "↑" indicate lower or higher values are better. "#Token": the number of tokens used in transformer architectures. "#Step": the number of model runs needed to generate an image. We compute our wall-clock inference time and scale the Time for other methods accordingly. †: trained on larger datasets including OpenImages [27]. ‡: implemented with the official tokenizer weight [66] and the scripts from VAR.

Type	Model	rFID↓	gFID↓	IS↑	Pre↑	Rec↑	#Token	#Para	#Step	Time
GAN	BigGAN [3]	75.24	6.95	224.5	**0.89**	0.38		112M	1	–
GAN	GigaGAN [23]	–	3.45	225.5	0.84	0.61	N.A.	569M	1	–
GAN	StyleGan-XL [45]	7.06	2.30	265.1	0.78	0.53		166M	1	0.4
Diff.	ADM [8]	125.78	10.94	101.0	0.69	**0.63**	N.A.	554M	250	235
Diff.	CDM [20]	–	4.88	158.7	–	–	N.A.	–	8100	–
Diff.	LDM-4-G [44]	0.27†	3.60	247.7	–	–	N.A.	400M	250	–
Diff.	DiT-L/2 [39]	0.62†	5.02	167.2	0.75	0.57	256	458M	250	43
Diff.	DiT-XL/2 [39]	0.62†	2.27	278.2	0.83	0.57	256	675M	250	63
Diff.	L-DiT-3B [71]	–	2.10	304.4	0.82	0.60	256	3.0B	250	>63
Diff.	L-DiT-7B [71]	–	2.28	316.2	0.83	0.58	256	7.0B	250	>63
Mask.	MaskGIT [4]	2.28	6.18	182.1	0.80	0.51	256	227M	8	0.7
Mask.	RCG (cond.) [30]	–	3.49	215.5	–	–	–	502M	20	2.7
Mask.	TiTok-B64 [66]	1.70	2.48	214.7	–	–	64	177M	8	0.4
2D Scan	VQVAE-2 [42]	–	31.11	∼45	0.36	0.57	N.A.	13.5B	5120	–
2D Scan	VQGAN [10]	7.94	15.78	74.3	–	–	N.A.	1.4B	256	34
2D Scan	ViTVQ [63]	1.28	4.17	175.1	–	–	1024	1.7B	1024	>34
2D Scan	RQTran. [29]	3.20	7.55	134.0	–	–	64, 4	3.8B	68	29
VAR	VAR-$d16$ [50]	0.90†	3.30	274.4	0.84	0.51	680	310M	10	0.6
VAR	VAR-$d20$ [50]		2.57	302.6	0.83	0.56		600M	10	0.7
VAR	VAR-$d24$ [50]		2.09	312.9	0.82	0.59		1.0B	10	0.8
VAR	VAR-$d30$ [50]		**1.92**	**323.1**	0.82	0.59		2.0B	10	1.4
1D AR	TiTok-B64-$d16$‡ [66]	1.70	6.30	190.1	0.85	0.47	64	310M	64	1
1D AR	**SpectralAR**-$d16$	4.03	3.02	282.2	0.81	0.55	64	310M	64	1
1D AR	**SpectralAR**-$d20$	4.03	2.49	305.4	-	-	64	600M	64	1.2
1D AR	**SpectralAR**-$d24$	4.03	2.13	307.7	-	-	64	1.0B	64	1.4
1D AR	**SpectralAR**-$d16$-p4	4.03	3.13	276.1	-	-	16×4	310M	16	0.4

details. This similarity qualitatively validates the rationale for performing causal autoregressive generation in the spectral domain. We further provide some quantitative analysis through a toy experiment in Section 4.2.

Potential applications. The frequencies represented by the token $\mathbf{s}_i$ become higher as i increases, while its influence on the image quality diminishes accordingly (check Section 3.1 for details). Therefore, we can control the visual quality of sampled images by adjusting the length of generated sequences, similar to the image compression algorithms [55]. In addition, we can achieve super-resolution by dividing images into disjoint parts smaller than $H \times W$, and conducting individual spectral autoregressive generation on each part. We provide further results in Section 4.4.

4. Experiments

4.1. Dataset and Implementation Details

We train and evaluate our SpectralAR on the ImageNet-1K [7] benchmark, which contains 1,281,167 and 50,000 images for training and validation, respectively. We train the tokenizer and generator on the training split. We evaluate the reconstruction performance on the validation set with reconstruction Fréchet inception distance [17] (rFID), and the generation results with generation FID (gFID) using pre-computed statistics and scripts from ADM [8].

For tokenizer training, we follow the exactly same settings of TiTok [66] for a fair comparison. We also employ the two-stage training strategy with proxy codes [4, 66]. We use the ViT-B [9] as the encoder and decoder, and set the number of spectral tokens as $N = 64$ in our main experiments and the sequence ωs for 256×256 images as:

$$\omega_i = \begin{cases} i, & if \ \ i \in (0, 32], \\ 2i - 32, & if \ \ i \in (32, 48], \\ 12i - 512, & if \ \ i \in (48, 64]. \end{cases} \tag{9}$$

For generator training, we adopt the same architecture and training recipe as VAR [50], which leverages a GPT-2-like transformer architecture [40] for autoregressive modeling.

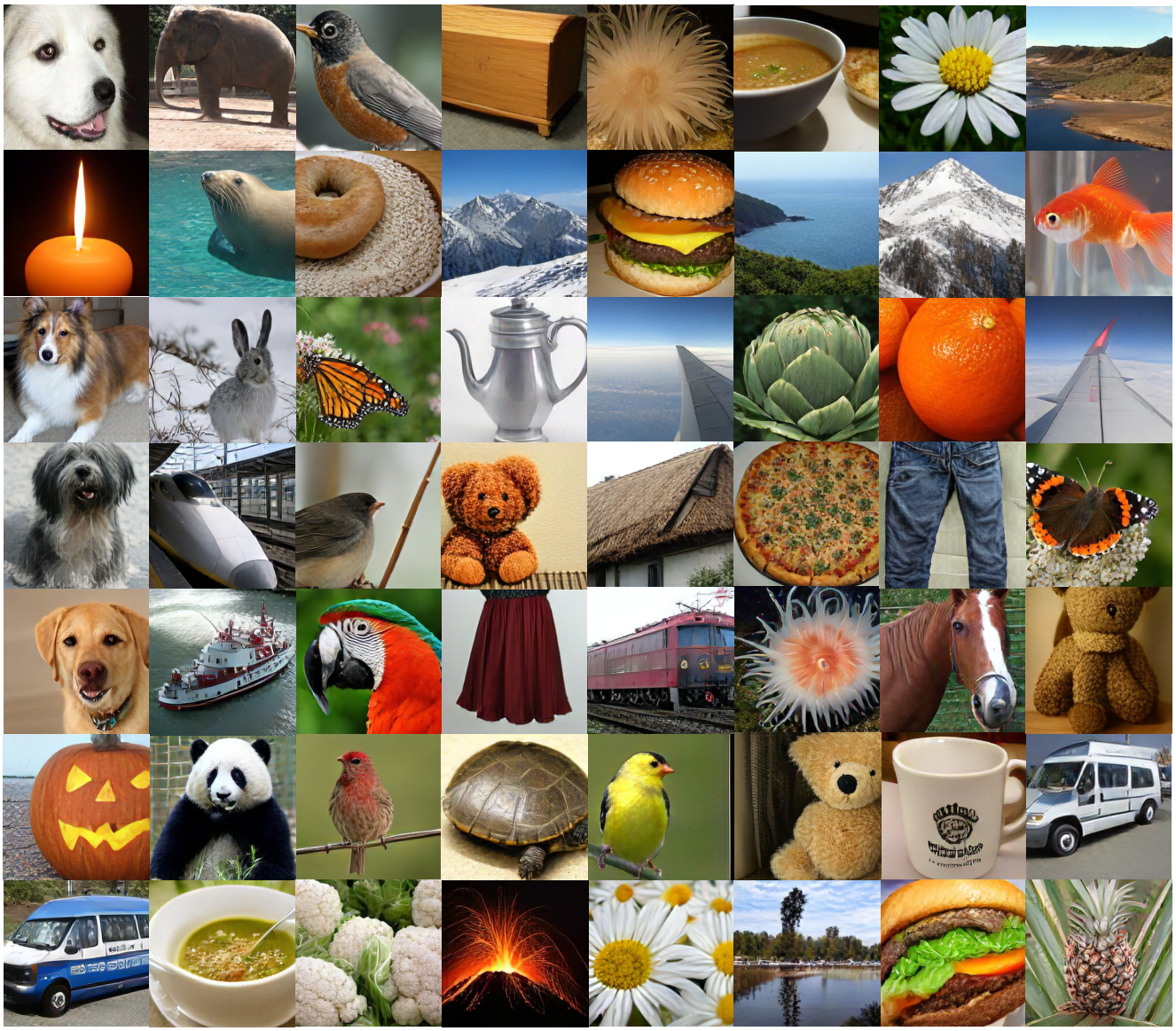

Figure 5. **Visualizations of generated 256×256 samples on ImageNet-1K.** These samples cover a wide range of categories and styles, demonstrating the ability of SpectralAR to generate both diverse and high-quality images.

4.2. Proof of Concept

We conduct a proof-of-concept experiment to validate the causality of images in the spectral domain. We use the validation set of the CIFAR-100 dataset [26] as a proxy for the distribution of natural images, which contains 10,000 32×32 images from 100 semantic classes. We aim to measure the linear correlation (R^2) [38] between one token $\mathbf{t}_i$ and its previous ones $\{\mathbf{t}_j\}_{j=1}^{i-1}$ as a metric for the causality of the sequence $\mathbf{T}$. We consider three typical autoregressive paradigms, including raster-scan [10], scale-wise [50] and spectral methods which construct sequence $\mathbf{T}$ in different ways. Specifically, we divide each image spatially into four patches of size 16×16 and rearrange them in raster-scan order for the first variant. For the scale-wise method, we downsample each image into sizes of 8×8, 16×16, 24×24 and 32×32, respectively. We set the frequency thresholds to 8, 16, 24, 32 and generate the sub-images following (5) for the spectral sequence. Since each token $\mathbf{t}$ is a high dimensional tensor, we calculate R^2 between every element of $\mathbf{t}_i$ and every element of $\{\mathbf{t}_j\}_{j=1}^{i-1}$. For each element of $\mathbf{t}_i$, we then find the element in $\{\mathbf{t}_j\}_{j=1}^{i-1}$ with the highest correlation, and take an average as the final measurement:

$$R^2_{avg} = \underset{m}{\text{AVG}}(\max_{n} R^2_{m,n}), \tag{10}$$

where m, n denote element number of $\mathbf{t}_i$ and $\{\mathbf{t}_j\}_{j=1}^{i-1}$, respectively, and $R^2_{m,n}$ is the correlation between the mth and nth elements. We report the quantitative results in Table 1, where the scale-wise and spectral sequences demonstrate

Table 3. **Applications of SepctralAR.** Super-reso., Trunc. represent super-resolution and truncated, respectively.

App. Type	Model	FID↓	IS↑
Super-reso.	Upsample	3.09	**286.6**
	SpectralAR-Stride	**2.93**	276.4
	SpectralAR-Patch	14.76	170.0
Trunc.	SpectralAR-Trunc.5	3.34	271.5
	SpectralAR-Trunc.10	6.65	211.4
	SpectralAR-Trunc.15	27.69	91.69
	SpectralAR-Trunc.0	**3.02**	**282.2**

considerably higher correlation than the raster-scan counterpart. This could be attributed to the coarse-to-fine nature of the former two paradigms, while the raster scan method lacks adequate reference information to predict the next image patch, as shown in Figure 4.

4.3. Main Results

We report the performance of SpectralAR on the class-conditional ImageNet-1K [7] 256×256 generation benchmark in Table 2. We also implement an autoregressive version of TiTok [66] by using the official tokenizer weight and scripts from VAR [50] for a fair comparison. The reconstruction performance of SpectralAR (4.03 rFID) is inferior compared with TiTok because SpectralAR requires to reconstruct the sub-images corresponding to different frequencies with different lengths of tokens, which is much more difficult compared with the overall reconstruction target in TiTok. Despite the lower reconstruction score, SpectralAR-$d16$ still outperforms TiTok-B64-$d16$ and VAR-$d16$ in autoregressive generation with a clear margin due to better sequence causality which eases autoregressive learning. Furthermore, SpectralAR also achieves better or comparable performance against VAR under the $d20$ and $d24$ settings, respectively. In addition, SpectralAR uses only 64 tokens in both reconstruction and generation, demonstrating superior token efficiency compared to VAR [50] and 2D scan-based methods. We also visualize the generated samples in Figure 5, which shows the diversity and quality of the generation process of SpectralAR. In addition, Figure 1 highlights the hierarchical coarse-to-fine refinement of the images, while SpectralAR generates more tokens in an autoregressive way.

4.4. Applications

In this section, we provide a quantitative analysis for the potential applications of SpectralAR. For super-resolution, we conduct our experiments based on the 256×256 images generated by SpectralAR-$d16$ in Table 2, and directly upsample them to 512×512 as the baseline. We construct 4 sub-images using strided and patch-based methods for SpectralAR-Stride and -Patch, respectively. We then upsample the 4 sub-images to 256×256 and use SpectralAR to refine them in the spectral domain, and finally reassemble them to generate the final result. According to Table 3, SpectralAR can indeed serve as a spectrum completer for the super-resolution task. In addition, we also experiment with truncated autoregressive generation, where we discard the last few tokens. For example, we discard the last 5 tokens in SpectralAR-Trunc.5 in Table 3. The generation performance worsens slowly when the number of truncated tokens is fewer than 10, and therefore it is possible to further improve token efficiency through truncation according to the requirement for generation quality.

Table 4. **Ablation on design choices.** Spectral Supervision means using sub-image supervision across diffirent frequency bands. Causal Mask and non-uniform mapping refers to the spectral causal mask and the frequency-token mapping, respectively.

Spectral Supervision	Causal Mask	Non-uniform Mapping	FID↓	IS↑
×	×	×	6.30	190.1
✓	×	×	5.64	255.1
✓	✓	×	3.49	222.6
✓	✓	✓	**3.02**	**282.2**

4.5. Ablation Study

We conduct ablation study to validate the effectiveness of our design choices in Table 4. The first row corresponds to the baseline TiTok implementation for autoregressive generation. Note that the spectral supervision alone could improve FID compared with the TiTok counterpart. This is because the spectral design enhances sequence causality compared with the overall reconstruction target which ignores the correlation between tokens and thus complicates the autoregressive modeling process. The spectral causal mask further enhances performance by improving causality in the encoding and decoding architecture. And the non-uniform token-frequency mapping technique guides the model to focus more on the crucial low-frequency components while representing the high-frequency components efficiently, thus further improving performance.

5. Conclusion

In this paper, we have proposed the spectral autoregressive visual generation method for both causal and efficient autoregressive modeling of image data. Specifically, we first convert images into 1D sequences with nested spectral tokenization. In addition, we have adopted causal masks for spectral tokens in the encoder and decoder to further enhance causality from the architectural perspective. We have also designed a non-uniform token-frequency mapping with emphasis on the low-frequency components in order to improve token efficiency. On the ImageNet-1K generation benchmark, our SpectralAR achieves superior performance compared with other autoregressive generation methods.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62125603, Grant 62321005, Grant 62336004, and Grant 62441616, and in part by the Beijing Natural Science Foundation under Grant No. L247009.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93, 2006. 3

[2] Jean Baptiste Joseph Baron Fourier et al. *The analytical theory of heat*. Courier Corporation, 2003. 2, 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 6

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, pages 11315–11325, 2022. 2, 6

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703. PMLR, 2020. 3

[6] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6, 8

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NIPS*, 34:8780–8794, 2021. 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2, 3, 4, 5, 6, 7

[11] Carlos Esteves, Mohammed Suhail, and Ameesh Makadia. Spectral image tokenizer. *arXiv preprint arXiv:2412.09607*, 2024. 3

[12] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, pages 3247–3258. PMLR, 2020. 3

[13] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCVW*, pages 3599–3608. IEEE, 2019. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 27, 2014. 2, 4

[15] Hayit Greenspan, Charles H Anderson, and Sofia Akber. Image enhancement by nonlinear extrapolation in frequency space. *TIP*, 9(6):1035–1048, 2000. 3

[16] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30, 2017. 6

[18] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020. 1

[20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23(47):1–33, 2022. 6

[21] Yuanhui Huang, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Di Zhang, Jie Zhou, and Jiwen Lu. Owl-1: Omni world model for consistent long video generation. *arXiv preprint arXiv:2412.09600*, 2024. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 4

[23] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 6

[24] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 3

[25] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 5

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[27] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. 6

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[29] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, pages 11523–11532, 2022. 2, 3, 6

[30] Tianhong Li, Dina Katabi, and Kaiming He. Self-conditioned image generation via generating representations. *CoRR*, 2023. 6

[31] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NIPS*, 37:56424–56445, 2025. 2

[32] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *ECCV*, pages 549–566. Springer, 2022. 3

[33] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *CVPR*, pages 15315–15324, 2022. 3

[34] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024. 2

[35] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, pages 23–40. Springer, 2024. 2

[36] Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Huaian Chen, and Yi Jin. Star: Scale-wise text-to-image generation via auto-regressive representations. *arXiv preprint arXiv:2406.10797*, 2024. 2

[37] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 3

[38] Daniel J Ozer. Correlation and the coefficient of determination. *Psychological bulletin*, 97(2):307, 1985. 7

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 1, 2, 3, 6

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2, 6

[41] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021. 3

[42] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NIPS*, 32, 2019. 2, 3, 6

[43] Siddharth Roheda. Cart: Compositional auto-regressive transformer for image generation. *arXiv preprint arXiv:2411.10180*, 2024. 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3, 6

[45] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 6

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[47] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2

[48] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2

[49] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 3, 5

[50] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NIPS*, 37:84839–84865, 2025. 2, 3, 4, 6, 7, 8

[51] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *NIPS*, 29, 2016. 3, 5

[52] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NIPS*, 30, 2017. 3

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017. 2

[54] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008. 3

[55] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991. 2, 3, 4, 6

[56] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2, 5

[57] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024. 2

[58] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *TCSVT*, 13(7):560–576, 2003. 2

[59] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 2

[60] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2

[61] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. In *ECCV*, pages 1–17. Springer, 2022. 3

[62] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 2

[63] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3, 6

[64] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 3

[65] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024. 2

[66] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *NIPS*, 37: 128940–128966, 2025. 2, 3, 4, 6, 8

[67] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *ICCV*, pages 14114–14123, 2021. 3

[68] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. Cid: Combined image denoising in spatial and frequency domains using web images. In *CVPR*, pages 2933–2940, 2014. 3

[69] Qian Zhang, Xiangzi Dai, Ninghua Yang, Xiang An, Ziyong Feng, and Xingyu Ren. Var-clip: Text-to-image generator with visual auto-regressive modeling. *arXiv preprint arXiv:2408.01181*, 2024. 2

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 4

[71] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient finetuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 6

[72] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *NIPS*, 35:23412–23425, 2022. 3

[73] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, pages 55–72. Springer, 2024. 2