Autoregressive Image Generation with Frequency Progression

Anonymous authors
Paper under double-blind review

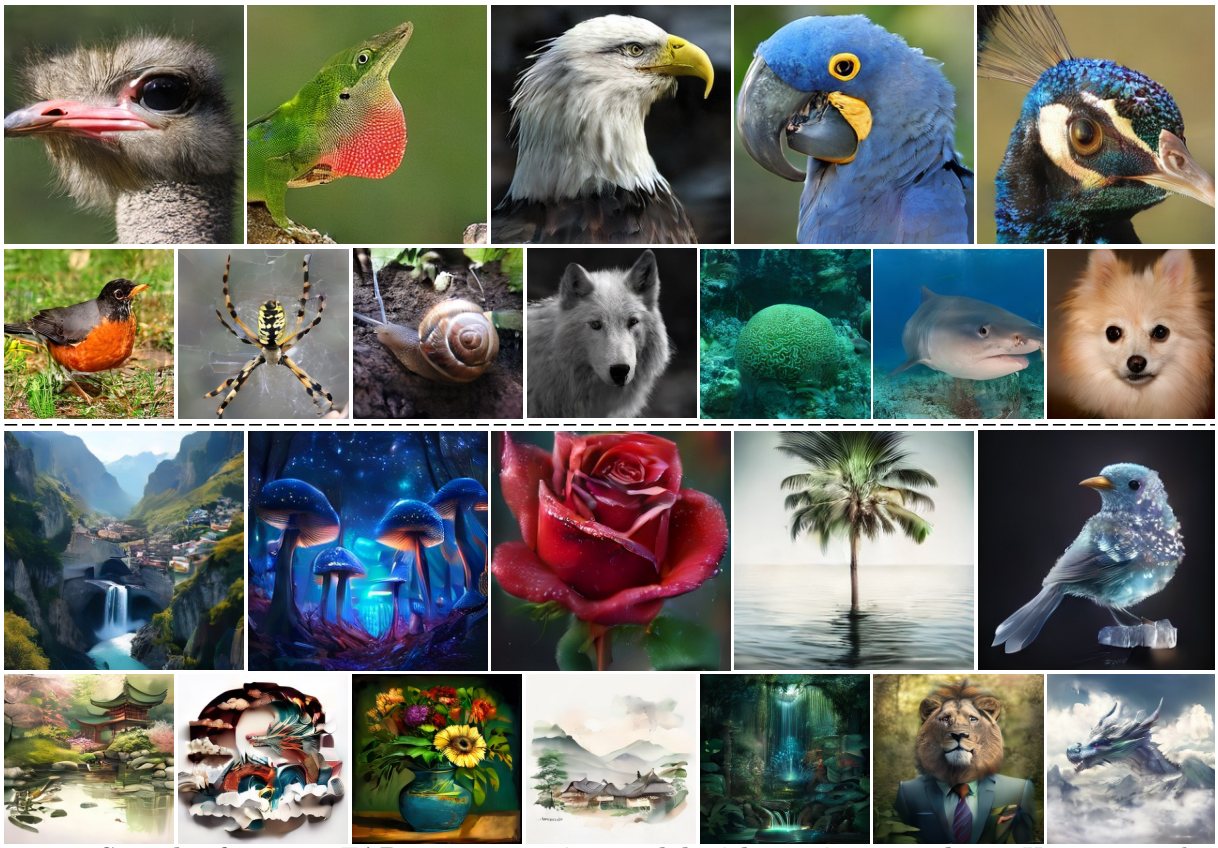

Figure 1: **Samples from our FAR autoregressive model with continuous tokens.** Upper part: class-conditional generation. Lower part: text-to-image generation. All of these samples are generated with only 10 steps.

Abstract

Autoregressive (AR) models for image generation typically adopt a two-stage paradigm of vector quantization and raster-scan "next-token prediction", inspired by its great success in language modeling. However, due to the huge modality gap, image autoregressive models may require a systematic reevaluation from two perspectives: tokenizer format and regression direction. In this paper, we introduce the frequency progressive autoregressive (**FAR**) paradigm and instantiate FAR with the continuous tokenizer. Specifically, we identify spectral dependency as the desirable regression direction for FAR, wherein higher-frequency components build upon the lower one to progressively construct a complete image. This design seamlessly fits the causality requirement for autoregressive models and preserves the unique spatial locality of image data. Besides, we delve into the integration of FAR and the continuous tokenizer, introducing a series of techniques to address optimization challenges and improve the efficiency of training and inference processes. We demonstrate the efficacy

of FAR through comprehensive experiments on the ImageNet dataset and verify its potential on text-to-image generation.

1 Introduction

Building upon autoregressive models, large language models (LLMs) (Devlin, 2018; Raffel et al., 2020; Brown, 2020; OpenAI, 2022) have unified and dominated language tasks with promising intelligence in generality and versatility. Resembling language processing, a typical AR paradigm for image generation involves two stages: **1)** Discretizing image data via vector quantization (VQ) with a finite, discrete vocabulary; and **2)** Flattening the quantized tokens into a 1-D sequence for next-token prediction. Based on this foundational paradigm, recent works Chang et al. (2022); Yu et al. (2023b; 2024b); Tian et al. (2024); Han et al. (2024); Sun et al. (2024); Li et al. (2024) have made inspiring advancements in image generation.

However, due to the huge modality gap between vision and language data, directly inheriting autoregressive generation from language to image is far from optimal. Text and image represent two distinct modalities: **1)** Text is discrete, causal/sequential, and arranged in 1-D; **2)** Image is continuous, non-causal/non-sequential, and arranged in 2-D. These differences introduce two crucial considerations for autoregressive image generation **1)** *Tokenizer format* (concrete vs. continuous); **2)** *Regression direction* (incorporating image-specific causality). Regarding the tokenizer format, continuous tokenizer aligns more naturally with image data and induces less information loss Li et al. (2024); Fan et al. (2024); Rombach et al. (2022). For the regression direction, raster Sun et al. (2024) or random Li et al. (2024) order fails to establish a causal sequence for images and can undermine inherent data priors, such as spatial locality. Consequently, the current AR paradigm for image generation remains sub-optimal and necessitates further investigation.

In this paper, we rethink the autoregressive image generation paradigm from a *spectral dependency* perspective. The rationality lies across three dimensions. **1) For images**, spectral dependency represents the strong correlation between high-frequency image details and low-frequency structures, where higher-frequency components build upon lower-frequency foundations to progressively construct a complete image. **2) For models**, neural networks are verified to first fit the low-frequency information and then the harder high-frequency part Ulyanov et al. (2018). **3) Diffusion model essence.** Diffusion models are proved to be implicit approximate frequency autoregression in nature Dieleman (2024). With these insights, we propose the frequency progressive autoregressive (FAR) paradigm. Specifically, as shown in Figure 3, FAR applies spectral filters to get the corresponding images of different spectral bands. The autoregression is conducted along these spectral-filtered images form lower to higher frequency, thereby inherently satisfying the causality requirements of AR models. For each spectral-filtered image, FAR bidirectionally models the full token sequence, effectively preserving the spatial locality of image data.

To instantiate FAR with continuous tokenizer, we identify the challenges of optimization difficulty and variance when modeling image tokens at different frequency levels, and propose simplifying and re-weighting the diffusion modeling of these levels. Additionally, to enhance training and inference efficiency, we introduce the mask mechanism and frequency-aware diffusion sampling. These proposed techniques, coupled with the intrinsic harmony between spectral dependency and image data, endow FAR with a more compatible autoregressive paradigm for image generation. Comprehensive experiments on the ImageNet dataset demonstrate the efficiency and scalability of FAR, wherein it significantly reduces inference steps to only 10 [1], while maintaining high quality and structural consistency. We also extend FAR to text-to-image generation. FAR achieves promising generation quality and high prompt alignment, utilizing much smaller model size, data scale, training compute, and inference steps, compared to previous text-to-image models.

Overall, our contributions can be summarized as follows:

[1] 10 steps refer to the outer FAR steps, with one forward pass of the transformer per step. Due to the diffusion head in our method, FAR have additional inner diffusion sampling steps for each outer step. Following the practice in MAR, we only report the outer step. Details of inner diffusion steps and the wall-clock time comparison are available in Table 1 and Table 3.

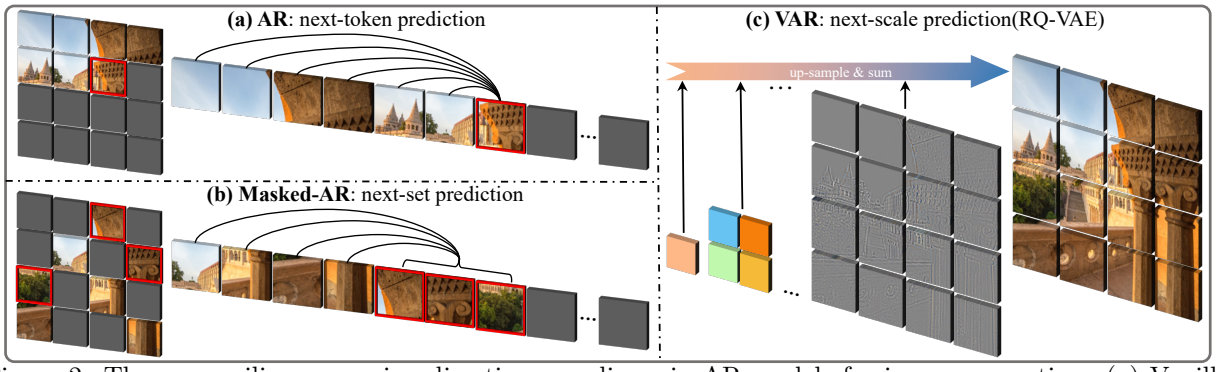

Figure 2: Three prevailing regression direction paradigms in AR models for image generation. (a) Vanilla AR: sequential next-token generation in a raster-scan order; (b) Masked-AR: next-set prediction with random order, generating multiple tokens each step; (c) VAR: combines RQ-VAE and multi-scale, adding all scales to get the final prediction.

- We propose the FAR paradigm, leveraging the spectral dependency of image data. FAR fits the causality requirement of AR models and preserves the spatial locality of image data, while being more sampling efficient.
- We delve into the instantiation of FAR with the continuous tokenizer, introducing a series of techniques to address the optimization challenges and improve the efficiency of both training and inference.
- We demonstrate the effectiveness and scalability of FAR through comprehensive experiments on ImageNet dataset and further extend FAR to text-to-image generation.

2 Related Works

2.1 Tokenizers in Autoregressive Models

Most of existing AR models in the vision domain employ discrete tokens via vector quantization. The pioneering VQVAE Van Den Oord et al. (2017); Razavi et al. (2019) proposes to quantize the latent space with a finite codebook, where each original token is replaced with the nearest discrete token in the codebook. RQ-VAE Lee et al. (2022) proposes residual quantization (RQ). However, due to the significant information loss induced by quantization, the performance upper bound of AR methods may be limited. In contrast, continuous tokenizer Rombach et al. (2022) aligns more naturally with image data and induces less information loss. Some recent works Li et al. (2024); Fan et al. (2024); Deng et al. (2024) also propose to employ a continuous tokenizer for autoregressive generation. In this paper, we also leverage continuous tokenizer and integrate it with our FAR paradigm. Comprehensive analyses of the tokenizer are available in the Appendix.

2.2 Autoregressive Models for Image Generation

Autoregressive model is an important method for image generation, leveraging GPT-style Radford (2018) to predict the next token in a sequence. Raster-scan flattens the 2-D discrete tokens into 1-D sequences in a row-by-row manner. Most of previous image autoregressive models employ this manner, including VQGAN Esser et al. (2021), VQVAE-2 Razavi et al. (2019), Parti Yu et al. (2022), DALL-E Ramesh et al. (2021), LlamaGen Sun et al. (2024), etc.

Besides this classical paradigm, some recent works make inspiring improvements over the raster-scan way. For example, MaskGIT Chang et al. (2022) proposes masked-generation to generate next token set instead of next one token, substantially reducing the inference steps. MAR Li et al. (2024) proposes to combine continuous tokenizers with mask-based generation. Another type of methods adopt residual quantization. For example, RQ-VAE Lee et al. (2022) proposes modeling the residual with vector quantization. VAR Tian et al. (2024) combines RQ-VAE with multi-scale, adding all scales to get the final prediction. This

design enables the scale number to be the inference step. Direct compressing latent into 1-D sequence is also an interesting direction. TiTok Yu et al. (2024b) compresses images into 1-D sequences with a modified autoencoder design.

Some concurrent works Pang et al. (2024b); Yu et al. (2024a); Wang et al. (2024b); Pang et al. (2024a); Ren et al. (2024) propose other interesting AR methods. For example, RAR Yu et al. (2024a) combines randomness and raster-scan and progresses from randomness to raster-scan for sequence generation. FlowAR Ren et al. (2024) combines the multi-scale design and flow model, and proposes multi-scale flow model for image generation. Additionally, some multi-modal large models Zhou et al. (2024); Xie et al. (2024); Wang et al. (2024a) integrate the image generation ability into the AR models for unified understanding and generation.

Different from previous works, we propose FAR for autoregressive image generation with frequency progression, which fits the causality requirement of AR and preserves the unique prior of image data.

3 Preliminaries

3.1 Diffusion Loss for Continuous Tokens

For the tokenizer in autoregressive models, the key is to model the per-token probability distribution, which can be measured by a loss function for training and a token sampler for inference. Following MAR Li et al. (2024), we adopt diffusion models to solve these two bottlenecks for integrating continuous tokenizer into autoregressive models.

Loss function. Given a continuous token z produced by a autoregressive transformer model and its corresponding ground-truth token x, MAR employs diffusion model as loss function, with z being the condition.

$$\mathcal{L}(z, x) = \mathbb{E}_{\varepsilon,t}\left[\|\varepsilon - \varepsilon_\theta\left(x_t \mid t, z\right)\|^2\right]. \tag{1}$$

Here, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $x_t = \alpha_t x_0 + \sigma_t \varepsilon$, with α_t being the noise schedule Ho et al. (2020). The noise estimator ε_θ, parameterized by θ, is a small MLP network.

Token sampler. The sampling procedure totally follows the inference process of diffusion model. Starting from $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the reverse diffusion model iteratively remove the noise and produces $x_0 \sim p(x|z)$, under the condition z.

3.2 Regression Direction

Regression direction plays a crucial role in autoregressive models for image generation. In Figure 2, we illustrate the three prevailing regression direction paradigms for image autoregressive models.

1) Vanilla AR (next-token prediction). The “next-token prediction" approach Esser et al. (2021); Sun et al. (2024) flattens the interdependent 2-D latent tokens via raster-scan. This paradigm, however, violates the causal requirements of AR sequences. For example, the tokens at the front of the next row should depend on the tokens near it, instead of the token at the end of the last row. Another limitation is the inference speed, demanding the token length as step, which is unbearably slow for high-resolution image generation.

2) Masked-AR (next-set prediction). The mask-based generation method Chang et al. (2022); Li et al. (2024) predicts the masked tokens given the unmasked ones. This paradigm enhances vanilla AR by incorporating randomness and predicting multiple tokens at every step. However, similar to the AR approach, Masked-AR violates the unidirectional dependency assumption of autoregressive models and neglects the image prior, limiting its potential.

3) VAR (next-scale prediction). VAR Tian et al. (2024) combines RQ-VAE Lee et al. (2022) with multi-scale, aggregating all scales to produce the final prediction. VAR maintains the spatial locality and adheres to the causality requirement. However, its multi-scale discrete residual-quantized tokenizer deviates from the commonly used tokenizers, necessitating specialized training. Our method did not demonstrate such resolution constraints. Compared to VAR, which requires resolution-specialized training for both the tokenizer and

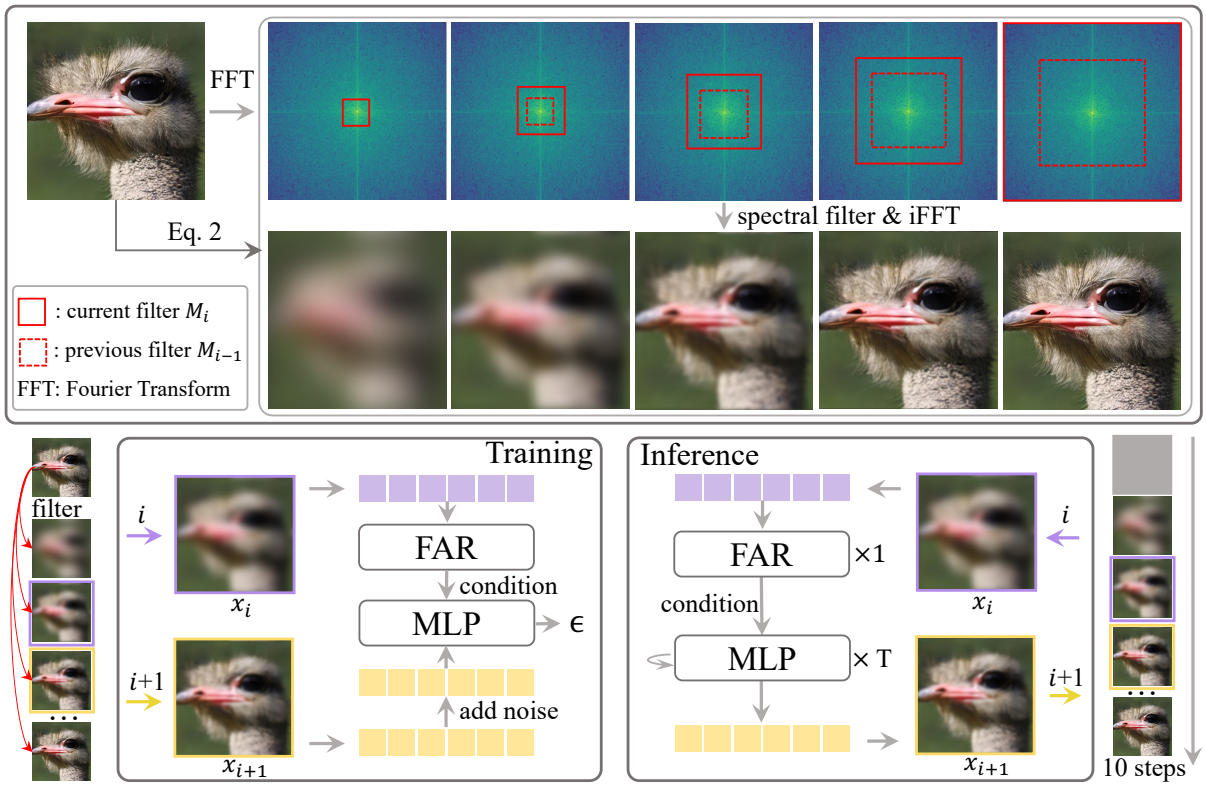

Figure 3: **Next-frequency prediction paradigm for autoregressive image generation via spectral dependency prior**. The upper part presents the spectral dependency and spectral filter process of images. The bottom part illustrates the training and inference stages of FAR.

generation backbone, our method can directly reuse existing VAEs (no constraints on resolution) and the frequency regression design is also flexible to resolutions. More importantly, we reveal that VAR paradigm demonstrates poor compatibility with the continuous tokenizer. Specifically, experiments combining VAR with the continuous tokenizer resulted in poor generation performance. Comprehensive results and analyses are provided in Section 5.

4 Methodology

4.1 Spectral dependency

The pivotal challenge for the regression direction of AR lies in harmonizing the causal sequence requirement with the inherent image prior. In this paper, we identify spectral dependency as a distinctive image prior tailored to this context. *The rationality of such design lies across three dimensions.* **1) For image**, it consist of low-frequency components that capture overall brightness, color, and shapes, alongside the high-frequency part that convey edges, details, and textures Russ (2006); Wornell (1996). The generation of higher-frequency information intrinsically relies on the prior establishment of the lower one. This hierarchical process also mirrors human artistic painting, where an initial sketch outlines the overall structure, followed by the progressive addition of details.

2) For models, neural networks inherently exhibit similar spectral dependencies. DIP Ulyanov et al. (2018) found that neural networks demonstrate high impedance to high-frequency components while allowing low-frequency components to pass with low impedance. This indicates that neural networks prioritize learning low-frequency before progressing to the more complex high-frequency. **3) Diffusion model essence.** Diffusion model is known to model the distribution transition between noise and images via iterative noising and denoising. The recent work Dieleman (2024) further reveals that noising implicitly functions as spectral filter

and diffusion is approximate spectral autoregression. This essence also inherently supports the optimality of our explicit FAR.

4.2 FAR: next-frequency prediction

Leveraging the spectral dependency, we introduce the innovative next-frequency prediction for autoregressive image generation. As shown in Figure 3, for each image x, its intermediate input at frequency level $i \in \{1, 2, .., F\}$ is formed as:

$$x_i = \mathcal{F}^{-1} M_i \mathcal{F} x. \tag{2}$$

Here, F denotes the number of frequency levels. For the latent resolution of 16×16, we can obtain 16 spectral bands (F=16). $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the Fourier transform and the inverse Fourier transform, respectively. M_i denotes the spectral filter within level i, and larger i represents more clear image. x_i is the corresponding spectral-filtered image.

During training, for each frequency level i and corresponding input x_i, FAR employs bidirectional attention and predicts all tokens simultaneously at each step. This effectively models the dependencies between tokens in the 2-D plane, thereby preserving the spatial locality of images. The output of FAR works as the condition of the following MLP network, wherein x_{i+1} functions as target.

During inference, the initial input x_0 to the outer autoregressive loop is blank/zero input. This zero initialization is used to predict the lowerest frequency level x_1, capturing only the most global structure (e.g., dominant color/brightness). FAR conducts autoregressive generation along spectral levels, progressively enhancing image clarity. The total steps for generating an image is reduced to only 10 (The inference step can be flexible as described in the following section). For each step, FAR forwards only once to get the condition of diffusion MLP, which then iterates T times (e.g. 100 in this paper) to get x_{i+1}.

4.3 FAR with Continuous Tokens

In this section, we delve into the combination of FAR and continuous tokens. We first identify and solve two primary challenges: optimization difficulty and variance in modeling token distributions across different frequency levels. Then, we present two techniques to enhance training and inference efficiency. Visualization of the implementation details is available in the Appendix.

Optimization difficulty. The diffusion loss in the continuous tokenizer models the distribution of per token. For FAR, diffusion loss needs to model $p\left(x_{i+1} \mid x_i\right)$ for $i \in [1, F-1]$, encompassing F frequency levels. This multi-level distribution modeling is challenging for the relatively small MLP. To mitigate this, we propose to directly model $p\left(x \mid x_i\right)$, and then filter x to get x_{i+1}. This simplifies the optimization complexity by normalizing the diffusion loss to only model x.

Optimization variance. Different frequency levels present varying optimization difficulties. Higher-frequency inputs are easier to predict, resulting in the optimization process being dominated by the more challenging low-frequency levels. To counteract this, we implement a frequency-aware training loss strategy that assigns higher loss weights to higher-frequency levels, ensuring balanced learning across all frequencies. Specifically, the loss weight is implemented in a sine curve, $w_i = 1 + \sin(\frac{\pi}{2} \times \frac{i}{F})$, where w_i is the loss weight of frequency level i.

Training efficiency. During the early steps, FAR primarily needs to learn the low-frequency components, which are information-sparse. Consequently, utilizing all tokens is redundant. To this end, we propose to incorporate the mask mechanism into FAR, leveraging only a subset of tokens. Specifically, we devise a frequency-aware mask strategy that progressively increases the mask ratio for lower-frequency levels. The mask mechanism randomly masks $[r_i, 1]$ input tokens for the input tokens at frequency level i, where r_i linearly transforms from 0.7 to 0. This design effectively reduces the training cost and we find it also contributes to improving generation diversity.

Inference efficiency. Diffusion model is known to first generate blurry structures first and then refines the details at later steps. The frequency progression property of FAR also inspires us to employ fewer diffusion sampling steps for lower frequency levels. Therefore, we devise the frequency-aware diffusion sampling step

strategy that allocates progressively fewer steps to earlier frequency levels. Specifically, we linearly shift the sampling steps for $T = 40$ to $T = 100$, achieving an average sampling step of $T = 70$. This saves 30% inference time of the diffusion model.

5 Experiments

5.1 Setup

Datasets. For class-conditioned generation, we adopt ImageNet Deng et al. (2009) dataset. For text-to-image generation, we employ the JourneyDB Sun et al. (2023) dataset with ∼4.19M image-text pairs and ∼3.57M internal data. By default, all images are processed to 256×256 resolution.

Training setup. We use the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$) Loshchilov (2017) with a weight decay of 0.02. Unless otherwise specified, for class condition, we train for 400 epochs with a batch size of 1024, and at an exponential moving average (EMA) rate of 0.9999. For text condition, we train for 100 epochs with a batch size of 512 and EMA rate 0.99.

Low-pass filters. We explore two frequency filtering types: (a) first down-sample then up-sample in the spatial domain, (b) low-pass filter in the Fourier domain. We find that they yield similar performance, as shown in the Appendix. We empirically adopt type (a) for simplicity.

For a latent representation of resolution $H \times W$ (e.g., 16×16), the maximum number of frequency levels is $F = \min(H, W) = 16$. At frequency level $i \in \{1, 2, \ldots, F\}$, the spectral filter M_i is constructed as follows:

(1) Type (a): The image is down-sampled to resolution $\lceil H \cdot i/F \rceil \times \lceil W \cdot i/F \rceil$, then up-sampled back to $H \times W$. This is equivalent to a low-pass filter that retains frequencies up to band i. Formally:

$$x_i = \text{Upsample}(\text{Downsample}(x, s_i), (H, W)), \quad s_i = \lceil H \cdot i/F \rceil. \tag{3}$$

Note that spatial down-sampling is merely a computationally efficient approximation of a low-pass frequency filter, which is why we validated it against the rigorous Fourier-domain filter (Type b).

(2) Type (b): M_i is a binary square mask in the 2D frequency domain that retains all frequency components within a square of side length i, centered at the DC component:

$$M_i(u, v) = \mathbf{1}\left[\max(|u|, |v|) \leq \frac{i}{2}\right], \tag{4}$$

where (u, v) coordinates are relative to the frequency center (DC component). The filtered image x_i is obtained via: $x_i = \mathcal{F}^{-1}(M_i \odot \mathcal{F}(x))$.

Models. We basically follow MAR Li et al. (2024) to construct our model, containing the AR transformer, diffusion MLP, and tokenizer. Specifically, for model architecture, we follow the two stage encoder-decoder setting of MAR for the backbone transformer design and MLP for the diffusion head. For the tokenizer, we adopt the same KL-VAE as MAR, with a downscaling factor of $f = 16$ and latent channel dimension $C = 16$. MAR pretrain this VAE on ImageNet. We keep it frozen during FAR training. Each 256×256 image is encoded into a $16 \times 16 \times 16$ latent representation, yielding 256 continuous tokens of dimension 16. The AR transformer has three model sizes: *FAR-B* (172M), *FAR-L* (406M), and *FAR-H* (791M). The diffusion MLP is much smaller as shown in Table 2.

For the class label conditioning, the class embedding is repeated 64 times and appended before the token sequence. For text-to-image generation, we employ Qwen2-1.5B Yang et al. (2024) as our text encoder. Specifically, we extract the last hidden layer representations of all text tokens. These are injected into the transformer backbone via cross-attention layers interleaved with the self-attention layers.

We also provide **complete source code** for reproducibility in the revised submission. We plan to release code and pre-trained checkpoints upon acceptance.

Evaluation. We evaluate FAR on ImageNet with four main metrics, Fréchet inception distance (FID), inception score (IS), precision and recall, by generating 50k images. For text-to-image generation, we adopt

Table 1: **Performance comparisons on class-conditional ImageNet 256×256 benchmark**. "↓" or "↑" indicate lower or higher values are better. FAR achieves comparable generation quality in nearly all the evaluated metrics compared to the sota methods, with only 10 inference steps. The only exception of FID is attributed to the slighter lower diversity, which we find to greatly influence the FID metric. The *Avg. Rank* in the last column represents the average ranking on the five indicators (including the additional inference steps) among the AR methods and our FAR-H. MAR (step=10) is evaluated using code and pretrained weights from their official GitHub repository. Note that, due to the diffusion head in both MAR and our method, they have additional inner diffusion sampling steps for each outer step. For MAR, it has 100 diffusion steps per outer step. For FAR, we have an average of 70 diffusion steps per outer step. **Following the setting in MAR, we only report the outer step in this table, and the wall-clock time comparison is available in Table 3**

Tokenizer	Model	FID↓	IS↑	Precision↑	Recall↑	Params	Steps	Avg. Rank
Diffusion Models								
Continuous	ADM Dhariwal & Nichol (2021)	10.94	101.0	0.69	0.63	554M	250	-
	LDM-4 Rombach et al. (2022)	3.60	247.7	0.87	0.48	400M	250	-
	U-ViT-H/2-G Bao et al. (2023)	2.29	263.9	0.82	0.57	501M	50	-
	DiT-XL/2 Peebles & Xie (2023)	2.27	278.2	0.83	0.57	675M	250	-
Autoregressive Models								
Discrete	VQGAN Esser et al. (2021)	15.78	74.3	-	-	1.4B	256	-
	ViT-VQGAN Yu et al. (2021)	4.17	175.1	-	-	1.7B	1024	-
	RQTran. Lee et al. (2022)	7.55	134.0	-	-	3.8B	68	-
	MaskGIT Chang et al. (2022)	6.18	182.1	0.80	0.51	227M	8	8
	LlamaGen Sun et al. (2024)	2.81	311.6	0.84	0.54	3.1B	576	5
	VAR Tian et al. (2024)	3.30	274.4	0.84	0.51	310M	10	2
	PAR Wang et al. (2024b)	2.88	262.5	0.82	0.56	3.1B	51	6
	RandAR Pang et al. (2024b)	2.55	288.8	0.81	0.58	343M	88	1
	Open-MAGVIT2 Luo et al. (2024)	3.08	258.3	0.85	0.51	343M	256	4
Continuous	MAR-H Li et al. (2024)	1.55	303.7	0.81	0.62	943M	256	2
	MAR-H Li et al. (2024)	9.32	207.4	0.71	0.47	943M	10	9
	FAR-B	4.26	248.9	0.79	0.51	208M	10	7
	FAR-L	3.45	282.2	0.80	0.54	427M	10	3
	FAR-H	3.21	300.6	0.81	0.55	812M	10	2

MS-COCO and GenEval Ghosh et al. (2024) dataset. FID is computed over 30K randomly selected image-text pairs from the MS-COCO 2014 training set. The GenEval benchmark measures the alignment with the given prompt.

CFG. We also adopt Classifier-Free Guidance (CFG) for evaluation following standard practice. Following MAR, we set CFG = 3 by default in our experiment. For the transformer backbone, we compute both the conditional and unconditional transformer outputs. These two outputs then work as condition for the MLP to get the corresponding diffusion output. Then, we interpolate them as $x = (1 + w) \cdot x_{\text{cond}} - w \cdot x_{\text{uncond}}$. The unconditional output is obtained by replacing the class/text embedding with a zero-initialized learnable embedding (probability 10% during training).

5.2 Class-conditional Image Generation

Main results. In Table 1, we list the comprehensive performance comparison with previous methods. We explore various model sizes and train for 400 epochs. Compared to most of the AR methods, our FAR is more efficient requiring fewer inference steps. Our method is superior to the VQGAN series with much smaller model size and inference steps. For recent works, like VAR and MAR, our method is also comparable in visual quality (indicated by IS and Perception metrics). Note that the lag in the FID metric is attributed to the slightly lower diversity (indicated by the Recall metric), which we find the FID metric is very sensitive to. We argue that AR/mask-based AR fail to exploit the strong correlation between high-frequency image detail and low-frequency structures, leading to random behaviors in the early steps of generation. Figure 1 shows qualitative results. We leave more visual results on ImageNet in the Appendix.

Scaling of the autoregressive transformers and denoising MLP. We investigate the scaling of both the autoregressive transformer and the diffusion loss model in Table 1 and Table 2. The autoregressive transformer takes the main burden of modeling the frequency dependency and mapping, thus also accounting

Table 2: **Scaling of denoising MLP in Diffusion Loss**. The denoising MLP is small and efficient, modeling the per-token distribution. Settings: FAR-L, 400 epochs, ImageNet 256x256.

MLP			*Metrics*	
Depth	Width	#Params	FID↓	IS↑
3	256	2M	3.83	278.2
3	512	6M	3.66	280.0
3	1024	21M	3.45	282.2
3	1536	45M	3.38	284.9

for the majority of the parameters. We find that the size of FAR transformer significantly affects the performance. When scaling up the FAR transformer, the performance is consistently improved.

For the denoising MLP, the requirement to model only the per-token distribution, combined with our distribution modeling simplification strategy, allows a small MLP (e.g., 2M) to achieve competitive performance. As expected, increasing the MLP width helps improve generation quality.

Inference Latency. Apart from the inference step comparison, we present the inference time comparison of the representation methods in Table 3. These values are evaluated with batch size 1 on a single H20 GPU. FAR achieves significantly faster inference speed than MAR and LlamaGen, and comparable speed as VAR.

Table 3: Inference time comparison of representative methods with comparable model parameters.

Methods	MAR-H	LlamaGen-XL	VAR-d24	FAR-H
Params/M	943	775	1000	812
Steps	256	256	10	10
Time/s	70.89	13.12	0.92	1.81

Sampling steps of FAR and diffusion loss. The training of the FAR adopts the maximum of F autoregressive step. For the inference, however, we can flexibly change the autoregression step and adopt fewer steps than F. This flexibility is additionally brought by the diffusion loss distribution modeling simplification strategy. Specifically, given x_i, FAR directly model x. In the next autoregressive step, we can filter x to get a flexible next frequency level, i.e. x_{i+2} instead of x_{i+1}, enabling the dynamic autoregression steps of FAR. Figure 4 depicts the generation performance under different FAR autoregression steps, where a higher step consistently achieves better performance.

Note that we chose $T = 10$ as the default step for three practical reasons: *(1) Controllable and fair comparison with baselines.* An important baseline, VAR, is set to 10 steps. VAR is featured extremely fast inference. While, due to its design, the inference step of VAR is fixed to 10 steps. For the explicit alignment with VAR in steps, we also set 10 inference steps. *(2) Efficiency–quality balance.* The primary significance of FAR is to achieve strong generation quality with significantly fewer AR steps than prior methods (which require 256–576 steps). At 10 steps, FAR already achieves strong FID 3.21 (FAR-H) and high image quality, outperforming or matching most AR methods while being substantially faster. Using more steps would improve FID but at higher inference cost. *(3) Diminishing returns.* While performance improves with more steps, the improvement gradually saturates. As shown in Figure 4, the FID improvement from step 6 to 10 is substantial, but from step 10 to 16, the improvement is modest (0.2 FID).

We also note that the number of inference steps can be flexibly adjusted without retraining. Users can choose their preferred operating point on the quality–cost curve.

The training of the denoising MLP employs a 1000-step noise schedule following DDPM Ho et al. (2020). During inference, MAR verifies that fewer sampling steps ($T = 100$) are sufficient for generation. We further demonstrate that our frequency-aware diffusion sampling achieves comparable results with fewer steps. Specifically, we linearly shift the sampling steps for $T = 40$ to $T = 100$, achieving an average sampling

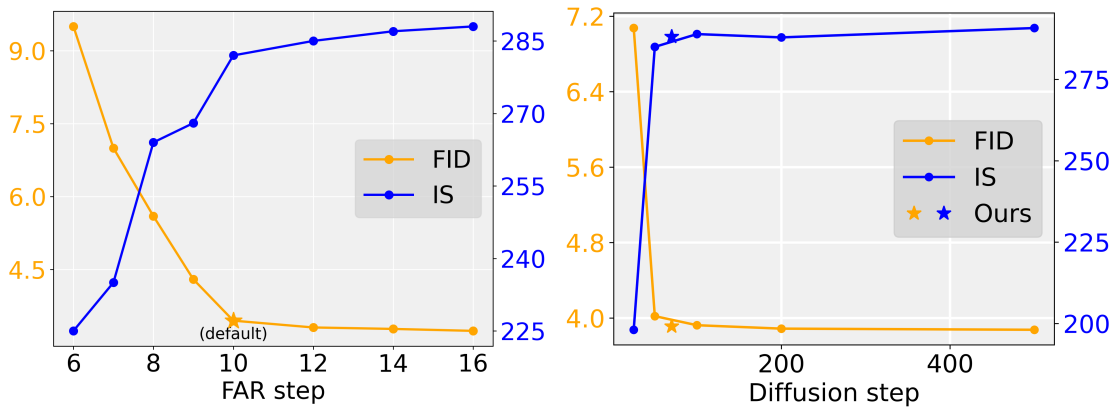

Figure 4: **Sampling steps of FAR and diffusion loss.**

step of $T = 70$. This saves 30% inference time of the diffusion model. Figure 4 shows that our sampling strategy achieves comparable results with fewer steps.

Table 4: **Combining VAR and the continuous tokenizer.**

VAR components		*Metrics*	
RQ	Multi-scale	FID↓	IS↑
✔	✔	75.35	33.2
✗	✔	33.57	96.8

The compatibility of VAR and diffusion loss. As we have noted in Section 3, the VAR paradigm demonstrates poor compatibility with the continuous tokenizer. The reasons are mainly two-fold. 1) The RQ manner is highly accuracy-sensitive for the prediction at every step for the continuous tokenizer. The RQ in VAR up-samples the prediction at each scale to the full latent scale and adds them all to get the final output. This requires highly accurate predictions at each scale. However, the exposure bias problem Bengio et al. (2015); Arora et al. (2022) of AR models induces inevitable error accumulation, deviating from the above requirement. 2) The per-token distribution modeling task in the VAR paradigm is prohibitively challenging for the diffusion loss. The tokens in different scales differ significantly in both the numeric range and receptive field. The numeric range difference is caused by the residual design, and the receptive field gap is induced by the multi-scale design. As shown in Table 4, directly combining VAR (RQ + mulit-scale) with the continuous tokenizer yields poor performance. Besides, we also try to remove the RQ design. The performance is improved due to no error accumulation in the residual paradigm. However, the performance still lags behind sota due to the gap in multi-scale. The visual results are available in the Appendix.

The experiment in the second row of Table 4 also serves as the *resolution-progressive coarse-to-fine baseline*. Specifically, it replace the residual quantization (RQ) design of VAR with spatial down-sample and adopt continuous tokens. This surpasses the direct combination of VAR and continuous tokens (the first row of Table 4), but still performs worser than our method. The reason is that the per-token distribution modeling task in the VAR paradigm is prohibitively challenging for the diffusion loss.

Note that FlowAR Ren et al. (2024) combines VAR with continuous tokenizer, via directly modeling the distribution of the whole image with multi-scale diffusion model (flow model), instead of the token-wise distribution. It is thus more like a multi-scale diffusion model than a AR model.

More ablations. We also conduct extensive ablations to verify the effectiveness of our method, including: **S1**) **DMS**: Diffusion loss Distribution Modeling Simplification strategy. **S2**) **Mask** mechanism. The mask mechanism improves the training efficiency by about 50%. **S3**) **FTL**: Frequency-aware Training Loss strategy. As shown in Table 8, with technique S1, FAR can already generate high-quality images (indicated by IS and Perception). The low FID is attributed to its lower diversity (indicated by Recall). This frequency

Table 5: **Ablations on the effectiveness of the proposed techniques**. Specific meanings of the abbreviations are in the ablation part. Settings: FAR-L, MLP size 21M, 400 epochs.

Ablations			*Metrics*			
DMS	Mask	FTL	FID↓	IS↑	Pre↑	Rec↑
✔	✗	✗	13.47	281.4	0.89	0.11
✔	✗	✔	12.83	283.6	0.89	0.13
✔	✔	✗	4.11	288.9	0.79	0.51
✔	✔	✔	4.05	290.2	0.80	0.52

progression itself is the foundation. Other components, like FTL and Mask, are supplementary techniques applied on top of the framework to better align FAR and continuous tokens. Adding Mask dramatically improves Recall ($0.11 \rightarrow 0.51$), which in turn lowers FID significantly ($13.47 \rightarrow 4.11$). This is because FID is very sensitive to diversity. The concrete explanations are available in the following subsection. FTL provides consistent improvement. Row 2 (DMS + FTL, no Mask) shows FTL improves FID from 13.47 to 12.83. Row 3 vs Row 4 shows a similar incremental gain ($4.11 \rightarrow 4.05$). We depict visual comparisons of this ablation in the Appendix.

Different CFG values. Following MAR, we set CFG = 3 by default in our experiment. We additionally list the results under different CFG values in Table 6. The optimal FID is achieved at $w = 3.0$, which we adopt as the default.

Table 6: Results under different CFG values. The optimal FID is achieved at $w = 3.0$, which we adopt as the default.

CFG	FID↓	IS↑	Pre↑	Rec↑
2.0	5.13	284.1	0.76	0.46
3.0	4.05	290.2	0.80	0.52
4.0	4.22	297.5	0.82	0.48

Visual comparison with MAR and VAR. As shown in Figure 5, the mask mechanism in MAR induces poor architecture under small inference steps. Adopting larger step elevates the generation quality but sacrifices the inference efficiency of this paradigm. The discrete tokenizer in VAR may also limit the performance limit and has difficulty in generating images with complex composition. In contrast, due to the intrinsic harmony with image data, FAR can generate high-quality images with consistent structures and fine details with only 10 steps.

Cause and solution analyses of diversity. The diversity of FAR is highly relevant to the result of the first step generation, as later steps progressively add finer details. For the first step, the randomness mainly stems from the noise in diffusion process. However, the condition of the diffusion MLP is the transformer output, which is identical for the same class label. This highly restricts the diversity brought by the random noise given the strong and identical condition information. Note that, this diversity issue only exists for the corner class-condition evaluaiton task, where we need to generate multiple (50) samples for the same class label. Consequently, text-to-image is free from this issue.

The mechanism by which masking improves diversity can be understood as follows: for the result of each inference step, mask mechanism randomly mask the result and keep certain tokens according to a cosine mask schedule. The remaining token number is very small in early stage, and the specific tokens are different and diverse across different generation due to random selection. This effectively amplifies the diversity. Instead, without masking, the model sees all tokens (entire image) and the generation falls into refining the result of the very similar first generation result.

Other techniques contribute limitedly to the diversity, e.g., the CFG strategy. We list the results under different CFG values in Table 6, where different CFG values differ slightly.

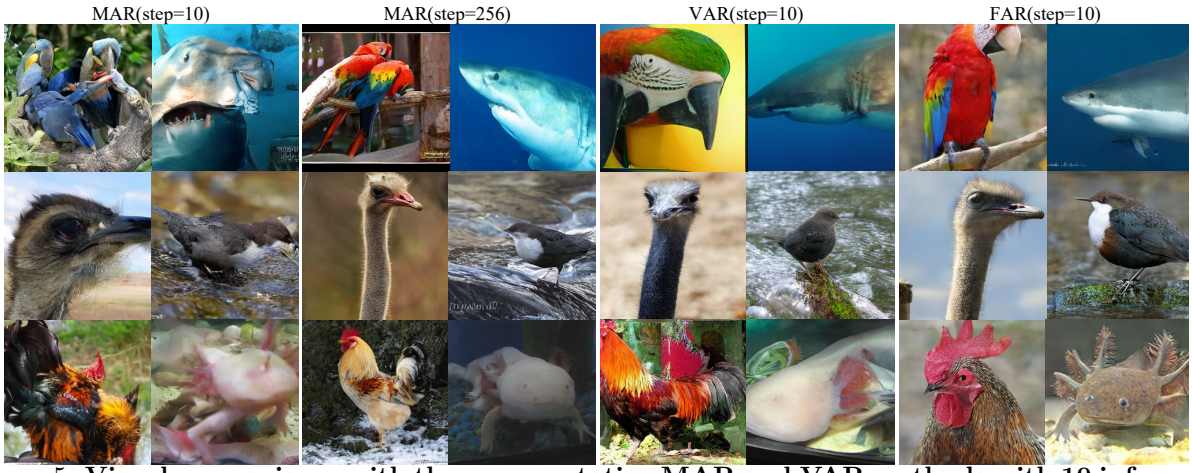

Figure 5: **Visual comparisons with the representative MAR and VAR methods with 10 inference steps**. Thanks to the intrinsic harmony with image data, our FAR can generate high-quality images with consistent structures and fine details with only 10 steps.

Table 7: **Performance comparisons on text-to-image task.**. Metrics include MS-COCO zero-shot FID-30K and GenEval benchmark. Please note that FAR employs much smaller model size, training data, GPU costs, and inference steps. We do not intend to demonstrate that FAR achieves cutting-edge performance, but rather to verify its potential in achieving high efficiency and promising results.

Tokenizer	Model	MS-COCO FID-30K↓	GenEval							Params	Training Data	A100 Days	Infer Steps
			Sing-O.	Two-O.	Count.	Color	Pos.	Color-A.	Overall				
Diffusion Models													
Continuous	LDM Rombach et al. (2022)	12.64	0.92	0.29	0.23	0.70	0.02	0.05	0.37	1.4B	-	6250	250
	DALL-E 2 Ramesh et al. (2022)	10.39	0.94	0.66	0.49	0.77	0.10	0.19	0.52	4.2B	650M	-	250
	SD3 Esser et al. (2024)	-	0.98	0.84	0.66	0.74	0.40	0.43	0.68	8B	-	-	50
Autoregressive Models													
Discrete	DALL-E Ramesh et al. (2021)	27.50	-	-	-	-	-	-	-	12B	250M	-	256
	CogView2 Ding et al. (2022)	17.50	-	-	-	-	-	-	-	6B	35M	-	-
	Muse Chang et al. (2023)	7.88	-	-	-	-	-	-	-	3B	460M	2688	-
	Parti Yu et al. (2022)	7.23	-	-	-	-	-	-	-	20B	4.8B	-	256
	LlamaGen Sun et al. (2024)	-	0.71	0.34	0.21	0.58	0.07	0.04	0.32	775M	60M	-	256
Continuous	**FAR**	13.91	0.85	0.29	0.31	0.59	0.06	0.09	0.37	564M	7.8M	24	10

Exposure bias in autoregressive rollout. The exposure bias in autoregressive rollout is well-founded and deserves careful analysis. In this part, we clarify several aspects of our design that mitigate this issue.

First, a key architectural feature of FAR is the distribution modeling simplification (DMS) strategy described in Section 4.3. Rather than having the diffusion head model $p(x_{i+1}|x_i)$ directly, FAR models $p(x|x_i)$—the distribution of the full (unfiltered) latent given the current frequency state. The next frequency state x_{i+1} is then obtained by applying the spectral filter to the sampled x. This design serves as a natural self-correction mechanism: even if x_i at inference deviates from the training distribution, the model always targets the same ground-truth distribution $p(x|x_i)$, and filtering the output re-projects it onto the valid frequency subspace. This significantly reduces the compounding of errors across steps compared to directly chaining $p(x_{i+1}|x_i)$.

Second, we note that FAR operates with only 10 outer autoregressive steps—far fewer than the 256–576 steps used in raster-scan AR models (e.g., LlamaGen). With significantly fewer rollout steps, the error accumulation of our method is correspondingly narrower.

5.3 Text-to-Image Generation

Main results. In Table 7, we depict the performance comparison on text-to-image generation task. Previous methods in this task usually employ substantially large model parameters, web-scale datasets, and unbearable computation costs. FAR can beat the classical DALL-E, CogView2 and LlamaGen, and achieve comparable

performance to the recent sotas, with significantly smaller training and inference costs. The total training cost is 24 days with single A100. The concurrent works Fan et al. (2024); Deng et al. (2024) also verify the effectiveness of continuous tokens on text-to-image generation, but with substantially more training resources. In Figure 1, we show the visual results on text-to-image generation. FAR can generate high-quality images with coherent structures and complex composition in 10 steps. Besides, FAR demonstrates high prompt following capacity, even given a complex prompt.

We emphasize that the T2I experiment is intended as a *proof of potential* rather than a SOTA claim. FAR uses 564M parameters and 24 A100-days—orders of magnitude, significantly less than systems like Parti (20B params) or SD3 (8B params).

We also hold that the T2I field needs a fair and controllable experimental setting, normalizing the compute resources/data scale/model parameters. However, T2I field is now more performance and engineering-prioritized, engaging substantial resources as shown in Table 7, including web-scale and polished internal data; massive model size; industry-level GPUs. Therefore, we conduct controllable and fair comparisons in the class-condition image generation. For the T2I field, we try to verify the potential of our method with existing but limited resources.

Among existing autoregressive models, LlameGen employs least training resources. Our method outperforms LlamaGen (FAR 0.37 GenEval vs LlamaGen 0.32 GenEval) with smaller model size (FAR 564M vsLlamaGen 775M), training data (FAR 7.8M vs Llamagen 60M), and inference steps (FAR 10 steps vs LlamaGen 256 steps). Note that even with the lowerest training resources among previous methods, LlamaGen still adopts a mixture of public and internal data, reflecting the general interal data trend in this field. Given the same mixture data set, lower training resources and better performance over LlamaGen, we argue that this can well support the *proof of potential* claim.

5.4 Broader Impact Discussion on Misuse Risk

The field of generative modeling has made remarkable progress, enabling the synthesis of images with unprecedented quality and fidelity. We acknowledge that such advances, including our work on accelerating high-quality image generation, carry potential risks of misuse such as the creation of deceptive content or unauthorized impersonation. It is important to note that these risks are an inherent, community-wide challenge in generative modeling, rather than a vulnerability unique to our specific methodology. Nevertheless, we take these ethical considerations seriously. Our approach is fully compatible with standard community mitigations, including: (1) careful dataset curation to minimize harmful biases and inappropriate content; (2) watermarking and provenance tracking techniques to support authenticity verification of generated images; and (3) responsible deployment guidelines and access controls to promote ethical use within downstream applications. We leave these directions as part of our future work.

6 Conclusion

In this paper, we propose the frequency autoregressive generation paradigm and instantiate FAR with the continuous tokenizer. Specifically, we identify spectral dependency as the desirable regression direction for FAR. Besides, we delve into the integration of FAR and the continuous tokenizer. We demonstrate the efficacy and scalability of FAR through comprehensive experiments on class-conditional generation and further verify its potential on text-to-image generation.

References

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. *arXiv preprint arXiv:2204.01171*, 2022.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Liang Chen, Sinan Tan, Zefan Cai, Weichu Xie, Haozhe Zhao, Yichi Zhang, Junyang Lin, Jinze Bai, Tianyu Liu, and Baobao Chang. A spark of vision-language intelligence: 2-dimensional autoregressive transformer for efficient finegrained image generation. *arXiv preprint arXiv:2410.01912*, 2024.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Sander Dieleman. Diffusion is spectral autoregression, 2024. URL `https://sander.ai/2024/09/02/spectral-autoregression.html`.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

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

OpenAI. Chatgpt, 2022. `https://openai.com/blog/chatgpt`.

Yatian Pang, Peng Jin, Shuo Yang, Bin Lin, Bin Zhu, Zhenyu Tang, Liuhan Chen, Francis EH Tay, Ser-Nam Lim, Harry Yang, et al. Next patch prediction for autoregressive visual generation. *arXiv preprint arXiv:2412.15321*, 2024a.

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*, 2024b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Alec Radford. Improving language understanding by generative pre-training. 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Flowar: Scale-wise autoregressive image generation meets flow matching. *arXiv preprint arXiv:2412.15205*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

John C Russ. *The image processing handbook*. CRC press, 2006.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2023.

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.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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*, 2024a.

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation. *arXiv preprint arXiv:2412.15119*, 2024b.

Gregory Wornell. *Signal processing with fractals: a wavelet-based approach*. 1996.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

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.

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.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023a.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023b.

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

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *arXiv preprint arXiv:2406.07550*, 2024b.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A Appendix

This supplementary document is organized as follows:

Section B shows the comprehensive analyses on tokenizer from the perspective of compression.

Section C shows the details and visualization of the training and inference processes.

Section D demonstrates the poor compatibility of VAR and continuous tokenizer.

Section E shows more visual results of our method on text-to-image generation.

Section F depicts the visual results of ablations on the mask mechanism, which elevates generation diversity.

Section G presents the visual results as model size and inference step scaling up.

Section H shows the visual results at intermediate steps.

Section I shows the results of different low-pass filters.

Section J visualize some generation failure cases.

Section K shows the prompts of the text-to-image generaton for Figure 1 in main manuscript.

B Comprehensive Analyses on Tokenizer

Tokenizer: discrete or continuous. Data compression and reconstruction are vital for image generation, determining the performance upper bound of generation. Given the discrete nature and the mature categorical cross-entropy loss of languages, a commonly adopted strategy for visual autoregressive models is to discretize the data with VQ. However, compared to the discrete human-created language, natural image space is continuous and infinite. Quantization, specifically in VQVAE, inevitably introduces significant information loss.

For the tokenizer, the VQ operation induces significant information loss, making compression stage the bottleneck for better generation. Further, the autoregressive paradigm, i.e., "predicting next tokens based on previous ones", is independent of whether the values are discrete or continuous. The only difficulty that restricts the adoption of continuous-valued tokenizer is the lack of proper loss function to model the per-token probability distribution, which is easily done with cross-entropy for discrete-valued tokenizer. To this end, following the pioneering MAR Li et al. (2024), we adopt the diffusion model as loss function. Specifically, the autoregressive method predicts a vector for each token, which then serves as conditioning for the denoising network.

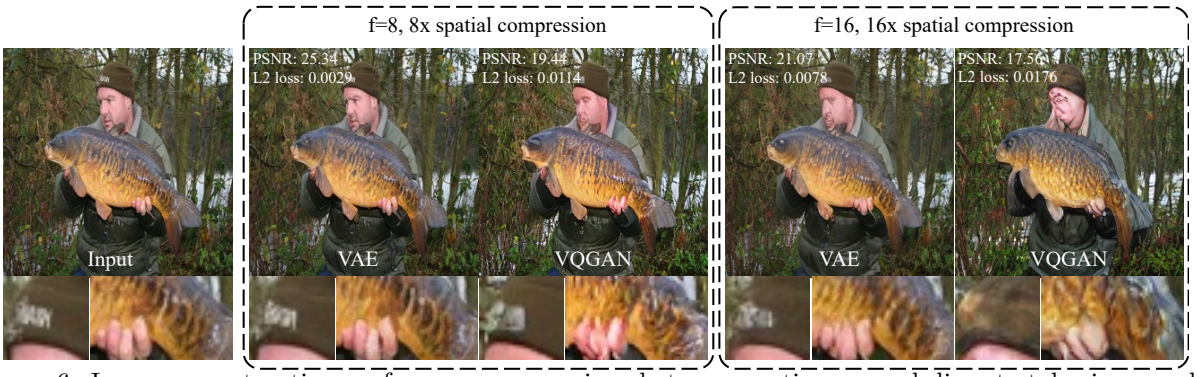

Figure 6: Image reconstruction performance comparison between continuous and discrete tokenizers under different spatial compression ratios (f=8 and f=16). Constrained by their finite vocabulary codebooks, discrete tokenizers suffer from significant information loss, struggling to faithfully reconstruct images with intricate, high-frequency details such as human faces. Note that the reconstruction of continuous tokenizer at f=16 is still better than the discrete one at f=8, which is also consistent with the rate distortion theory.

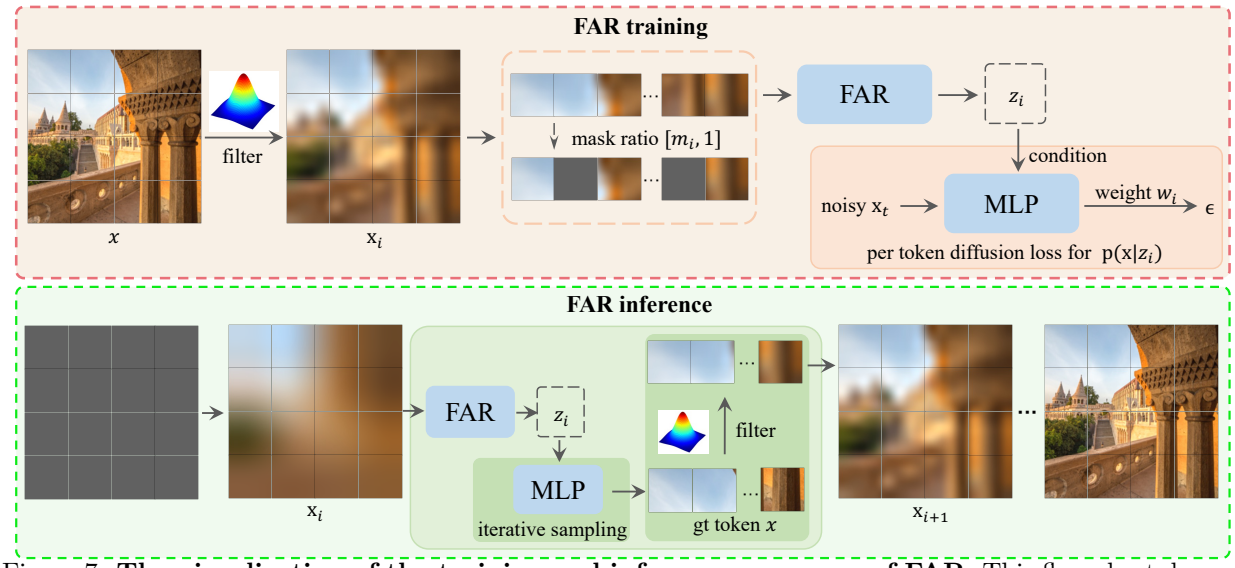

Figure 7: **The visualization of the training and inference processes of FAR.** This flow chart demonstrates the details of FAR and its integration with continuous tokens.

Tokenizer: compression perspective. For images, latent space is crucial for generation for the purpose of reducing computational burden. Thus besides the data format (discrete or continuous), we further analyses the tokenizers, VQGAN or VAE, as compression models from two key aspects: 1) Theoretical compression performance; and 2) Reconstruction visual results.

For measuring the theoretical compression performance, we adopt the information compression ratio (ICR) Chen et al. (2024). For discrete tokenizer, we take VQVAE as example, with downscaling factor f, codebook size N, input image's size of $H \times W$. We assume that the code follows a uniform distribution, so each code has $\log N$ bits information. For continuous tokenizer, we take VAE as example, with downscaling factor f, channel number C. Assume that the latent representation is fp32 tensor precision. The ICR of these two tokenizers are then as follows.

$$\mathrm{ICR}(N, f) = \frac{(H/f) \times (W/f) \times \log N}{H \times W \times 3 \times \log 256} = \frac{\log N}{24 f^2}. \tag{5}$$

$$\mathrm{ICR}(C, f) = \frac{(H/f) \times (W/f) \times C \times 32}{H \times W \times 3 \times \log 256} = \frac{32C}{24 f^2}. \tag{6}$$

Taking compression ration $f = 16$ for example, $\mathrm{ICR}(N, f) = 0.23\%$ for discrete tokenizer with codebook size $N = 16384$ and $\mathrm{ICR}(C, f) = 8.33\%$ for continuous tokenizer with channel number $C = 16$. Further, to achieve same ICR under same f, we need to exponentially enlarge the codebook size N from 16384 to $2^{512}(1.34 \times 10^{154})$. Given that discrete tokenizers are inherently difficult to train Yu et al. (2023a); Mentzer et al. (2023), it is thus prohibitively hard to train codebook at this scale.

In Figure 6, we visualize the reconstructed images comparison between discrete and continuous tokenizers. Compared to continuous tokenizer, the discrete one has difficulty in both detail fidelity and semantic consistency. For instance, the character detail on the hat is poorly reconstructed by discrete tokenizer. For semantic, the face and fingers of the man as well as the fish scales fail to remain semantically identical by discrete tokenizer.

Based on the above analyses, the discrete tokenizer for images suffers from substantially more information loss than the continuous one, indicating that quantization, the shortcut stemming from mimicking languages autoregressive generation, may be a inferior solution for image data. Thus, apart from the commonly adopted VQ paradigm for image autoregressive models, it is quite necessary and promising to employ continuous tokenizer.

Figure 8: **The visual results of combining VAR and continuous tokenizer.** These images correspond to the first 32 class labels. Consistent with our analyses, VAR paradigm demonstrates poor compatibility with the continuous tokenizer. The generation suffers from poor architectures and severe artifacts. Removing the residual quantization helps reducing the artifact, but still suffers from poor image quality.

C Details and visualization of the training and inference processes

In Figure 7, we depict the flow chart of the training and inference processes of FAR, demonstrating the implementation details. *For the training process*, FAR randomly selects a frequency level i and filters the input image x into the intermediate frequency level x_i. FAR then adopts the mask ratio $[r_i, 1]$ to the token sequence. The output z_i of the FAR model is then conditioned on the diffusion MLP. The diffusion loss models the distribution of each token with the frequency-aware dynamic loss weight w_i. *For the inference process*, we take the intermediate step i as example. FAR takes the masked x_i as input and outputs z_i. With z_i as condition, the diffusion model samples the groundtruth token distribution x. Then, we can filter x to get the next frequency level x_{i+1}.

D Compatibility of VAR and Continuous Tokenizer: Visual Results

In Figure 8, we depict the visual results of combining VAR and continuous tokenizer, corresponding to Table 4 of the main manuscript. The direct combination of VAR and continuous tokenizer demonstrates poor compatibility, generating images with obvious artifacts and poor architectures. On the right part, removing the Residual Quantization (RQ) successfully reduces the artifacts as expected. While, the generation is still inferior due to the challenging distribution modeling of diffusion loss in this case. These visual results exactly match our analyses in the manuscript.

E More Visual Results of Text-to-Image Generation

In this section, we present more text-to-image generation results of our method in Figure 9.

F Visual Results of Ablations on the Mask Mechanism

In Figure 10, we depict the visual results of the ablations on the mask mechanism, corresponding to Table 5 of the main manuscript. In the left part, our FAR can generate high-quality images after employing the diffusion loss distribution modeling simplification strategy. While, the generation diversity is limited. On the right part, we further adopt the mask mechanism. Mask introduces randomness, improving the generation diversity.

G Visual Results as Model and Inference Step Scaling

In Figure 11, we verify the scaling capacity of FAR: including model size and inference step. Scaling up these two factors can consistently improve the generation performance.

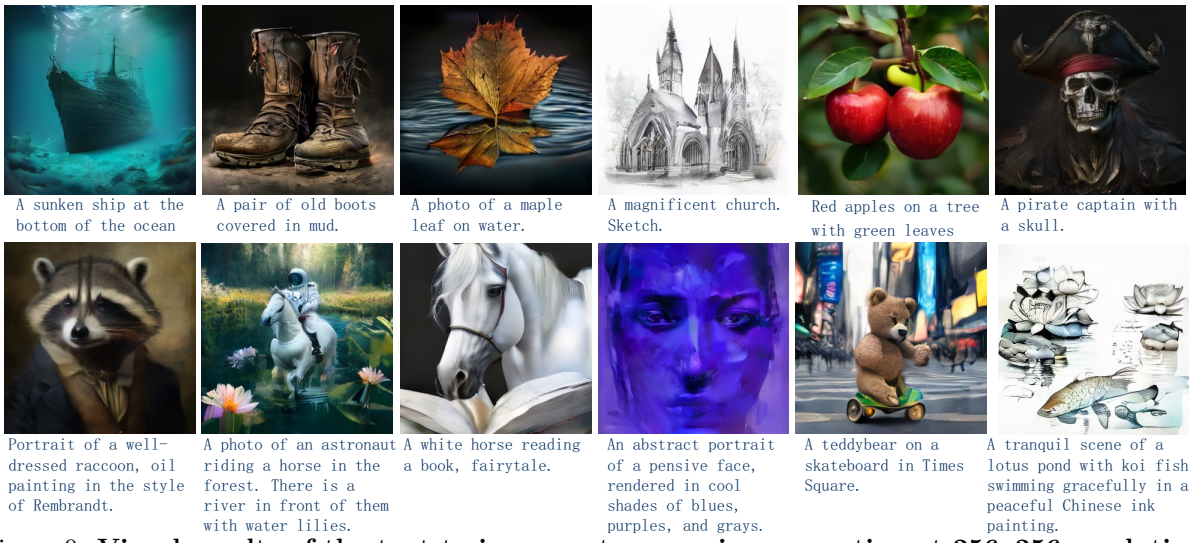

Figure 9: **Visual results of the text-to-image autoregressive generation at 256x256 resolution**.

Figure 10: **The visual results of ablations on the mask mechanism.** Each column corresponds to one class label. Our FAR can generate high-quality images without mask mechanism, while suffers from low diversity within each class. The mask strategy can effectively improves the generation diversity.

H Visual Results at Intermediate Steps

In Figure 12, we present the intermediate generation results along the autoregressive generation process. In the early steps, FAR generates the overall color and structural information, and then refines the details.

Note that several mid-to-late steps appear visually similar, and the final step introduces a much more significant increase in detail, appearing qualitatively different from the preceding steps. This visual discrepancy arises from the mapping between the latent space and pixel space, and we verify this with real image filtering visualization.

As detailed in Sec 4.2, the filtering process operates in the latent space of resolution 16x16. Blurry latent brings visualization difference when decoded back to the pixel space, as the VAE encoder/decoder is trained with clean latent instead of blurry latent. This also explains why the intermediate results present ring artifacts. However, this doesn't affect the generation process as well as the final visual quality, as we only need to decode the final clean latent back to pixel space.

We provide visualizations of how real images look under different frequency levels (i.e., the inputs to each outer-loop step) in Figure 13, as well as its comparison with the intermediate generation process. Both filter types are available, including (a) Spatial down/up-sample (default), and (b) Fourier-domain low-pass filter. Both these two visualizations demonstrate similar visual results as our intermediate generation results, especially type (a), which is the same as training setting.

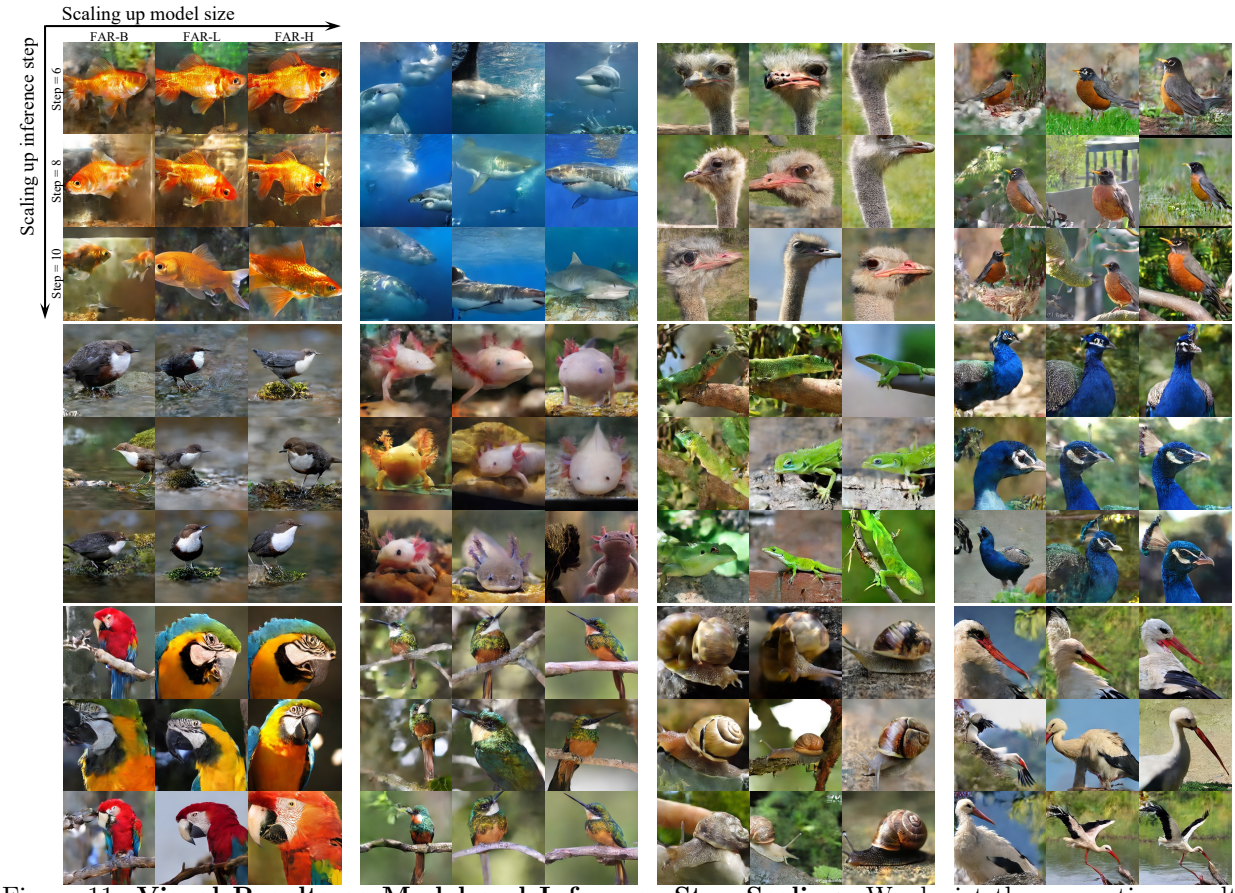

Figure 11: **Visual Results as Model and Inference Step Scaling.** We depict the generation results when increasing the model size and inference step. Scaling up these two factors can consistently improve the generation performance.

I Different Low-Pass Filters

We explore two frequency filtering types: (a) first down-sample then up-sample in the spatial domain, (b) low-pass filter in the Fourier domain. We find that they yield similar performance, as shown in Table 8. Since different frequency filtering methods only slightly differ in the filter. Besides, our method processes the filtered image in the spatial domain, which further narrows the difference of different low-pass filters (Different filter designs indeed similarly convert clean image to images of different blurry level, and these blurry level is visually small in difference, and enjoy similar modeling difficulty.). We thus hold that the frequency filtering methods make small differences to the final performance. By default, we empirically adopt type (a) for simplicity.

Table 8: Results under different low-pass filters

Filters	FID↓	IS↑	Pre↑	Rec↑
a	4.05	290.2	0.80	0.52
b	4.21	291.3	0.80	0.51

J Failure Case Visualization

We present the generation failure case of FAR in Figure 14. The current version of T2I FAR struggles with very fine repetitive and structural textures (e.g., dense grids, buildings, and bottles). While, note that

Figure 12: **Visual results at intermediate steps.** The intermediate generation results (total step 10) autoregressively refine the details, aligning perfectly with our frequency progression design.

Figure 13: **Real image filtering visualization.** The intermediate filtering results of real image with two types of filter, aligning similarly with our intermediate generation results.

the failure case of FAR may not be relevant to the model design, but more the training data scale and distribution coverage, given that the current T2I version is trained with very limited resources.

K Prompts for Figure 1 in Main Manuscript

The following part presents the prompts for the text-to-image generation results in Figure 1 of the main manuscript:

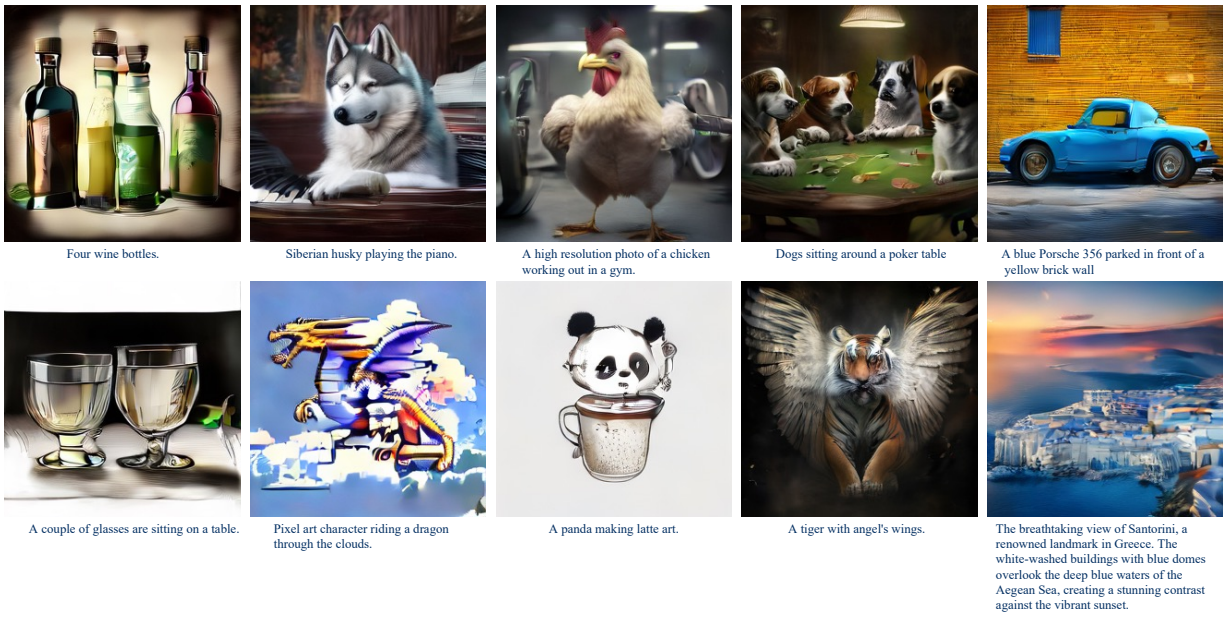

Figure 14: **Failure case visualization.** FAR struggles with very fine repetitive and structural textures (e.g., dense grids, buildings, and bottles). While, note that the failure case may not be relevant to the model design, but more the training data scale and distribution coverage, given that the current T2I version is trained with very limited resources.

- A mountain village built into the cliffs of a canyon, where bridges connect houses carved into rock, and waterfalls flow down into the valley below.
- An otherworldly forest of giant glowing mushrooms under a vibrant night sky filled with distant planets and stars, creating a dreamlike, cosmic landscape.
- A close-up photo of a bright red rose, petals scattered with some water droplets, crystal clear.
- A photo of a palm tree on water.
- A bird made of crystal.
- A tranquil scene of a Japanese garden with a koi pond, painted in delicate brushstrokes and a harmonious blend of warm and cool colors.
- Paper artwork, layered paper, colorful Chinese dragon surrounded by clouds.
- A still life of a vase overflowing with vibrant flowers, painted in bold colors and textured brushstrokes, reminiscent of van Gogh's iconic style.
- A peaceful village nestled at the foot of towering mountains in a tranquil East Asian watercolor scene.
- An enchanted garden where every plant glows softly, and creatures made of light and shadow flit between the trees, with a waterfall flowing in the background.
- A lion teacher wears a suit in the forest.
- A cloud dragon flying over mountains, its body swirling with the wind.