Next Patch Prediction for AutoRegressive Visual Generation

Yatian Pang[1,3], **Peng Jin**[1], **Shuo Yang**[1], **Bin Zhu**[1], **Bin Lin**[1], **Chaoran Feng**[1], **Zhenyu Tang**[1], **Liuhan Chen**[1], **Francis E. H. Tay**[3], **Ser-Nam Lim**[5, 6], **Harry Yang**[4, 6], **Li Yuan**[1,2,†]

[1]Peking University, Shenzhen Graduate School
[2]PengCheng Laboratory
[3]National University of Singapore
[4]Hong Kong University of Science and Technology
[5]University of Central Florida
[6]Everlyn

Abstract

Autoregressive models, built based on the Next Token Prediction (NTP) paradigm, show great potential in developing a unified framework that integrates both language and vision tasks. Pioneering works introduce NTP to autoregressive visual generation tasks. In this work, we rethink the NTP for autoregressive image generation and extend it to a novel **Next Patch Prediction** (NPP) paradigm. Our key idea is to group and aggregate image tokens into patch tokens with higher information density. By using patch tokens as a more compact input sequence, the autoregressive model is trained to predict the next patch, significantly reducing computational costs. To further exploit the natural hierarchical structure of image data, we propose a multi-scale coarse-to-fine patch grouping strategy. With this strategy, the training process begins with a large patch size and ends with vanilla NTP where the patch size is 1×1, thus maintaining the original inference process without modifications. Extensive experiments across a diverse range of model sizes demonstrate that NPP could reduce the training cost to $\sim 0.6\times$ while improving image generation quality by up to 1.0 FID score on the ImageNet 256×256 generation benchmark. Notably, our method retains the original autoregressive model architecture without introducing additional trainable parameters or specifically designing a custom image tokenizer, offering a flexible and plug-and-play solution for enhancing autoregressive visual generation.

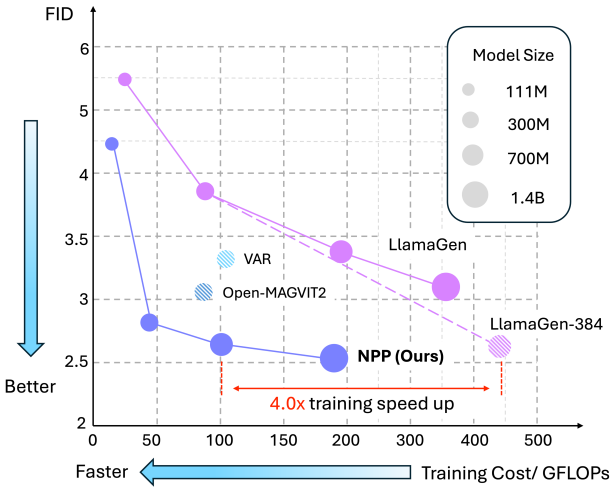

Figure 1: Comparison of NPP and baseline methods. Our method on a diverse range of models achieves higher FID scores with significantly less training cost on the ImageNet 256×256 benchmark.

Introduction

Autoregressive models, foundational to large language models (LLMs) (Vaswani et al. 2017; Devlin et al. 2018; Zhang et al. 2022), generate content through the prediction of subsequent tokens in a sequence. This Next Token Prediction (NTP) paradigm enables LLMs to excel in a variety of natural language processing tasks, exhibiting human-like conversational abilities (Ouyang et al. 2022; Touvron et al. 2023; Bai et al. 2023; Yang et al. 2023; Team 2023; Bi et al. 2024) and demonstrating remarkable scalability (Kaplan et al. 2020a; Henighan et al. 2020; Hoffmann et al. 2022; Wei et al. 2022; Alabdulmohsin, Neyshabur, and Zhai 2022; Chowdhery et al. 2023; Anil et al. 2023). Such advancements illustrate the potential for achieving general-purpose artificial intelligence systems. Inspired by the success of autoregressive models in the language domain, their applications for image generation have been widely explored. Notable approaches, including VQVAE (Van Den Oord, Vinyals et al. 2017), VQGAN (Esser, Rombach, and Ommer 2021) introduce image tokenizers that convert continuous images into discrete tokens, employing autoregressive models to sequentially generate these tokens, thereby achieving image generation. In parallel, diffusion models (Ho, Jain, and Abbeel 2020) emerge as a distinct and rapidly evolving paradigm in image generation. However, the fundamental differences in the underlying methodologies of autoregressive and diffusion models pose significant challenges for developing a unified framework that integrates both language and vision tasks.

More recently, a pioneering work LlamaGen (Sun et al.

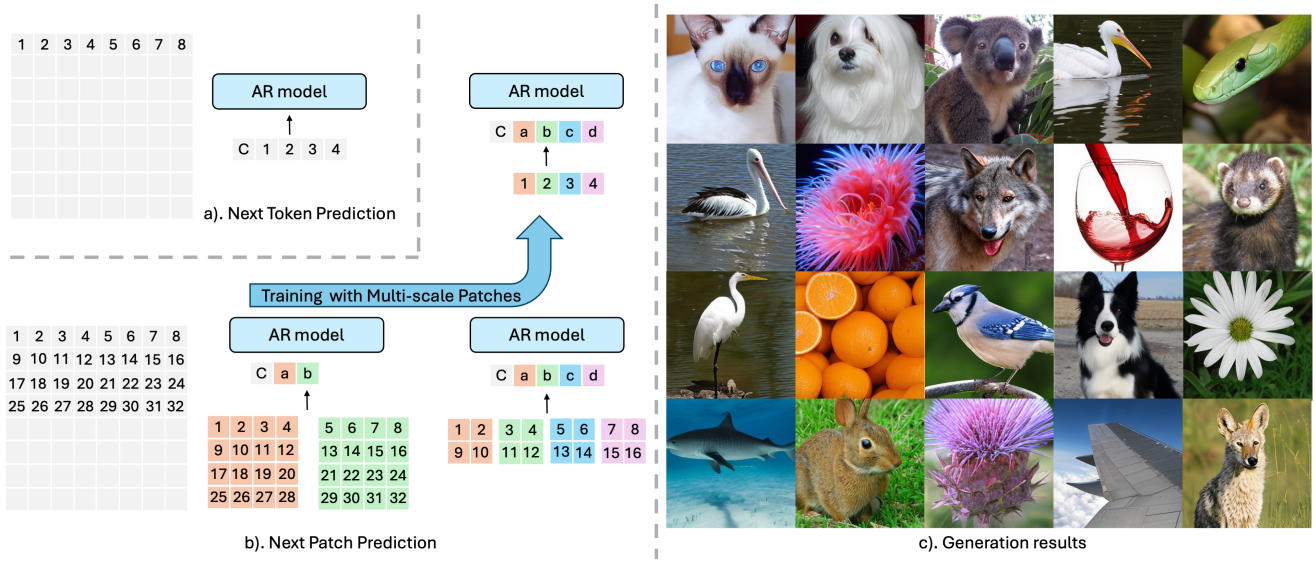

Figure 2: Motivation of the next patch prediction. a). Illustration of next token prediction. b). Demonstration of the proposed next patch prediction. c). Generation results on the ImageNet benchmark. Please zoom in to view.

2024) achieves the next token prediction paradigm for image generation with a vanilla autoregressive model, Llama, bringing the field one step closer to building a unified model between language and vision. However, directly applying NTP from the language domain to the image domain may lead to suboptimal performance due to the distinct properties of the two different modalities.

In this work, we follow the NTP paradigm as shown in Figure 2 a) for autoregressive image generation and rethink the modeling of the NTP paradigm in the following aspects.

(I). The NTP paradigm, widely successful in large language models, leverages the high information density of text tokens. However, image tokens typically exhibit lower information density due to the inherently redundant nature of image data (He et al. 2022). Our key insight is to aggregate multiple image tokens into high information density units referred to as patches[1], which can potentially enhance the performance of autoregressive image generation.

(II). Transformer-based autoregressive models incur substantial computational costs during training, with the total cost approximately scaling as $C \approx 6WN$ (Kaplan et al. 2020b), where W represents the number of model parameters and N denotes the input sequence length. While maintaining the model architecture, we could manage to reduce the input sequence length of image tokens, thus improving training efficiency.

(III). Unlike language data, image modality inherently exhibits hierarchical property in both understanding and generation tasks. This observation suggests that autoregressive image generation could benefit from a multi-scale, coarse-to-fine modeling strategy, which has the potential to improve generation quality and training efficiency.

[1]Here, we define a patch as containing multiple image tokens originally encoded by the VQVAE encoder.

Building on these insights, we introduce Next Patch Prediction (NPP) as shown in Figure 2 b), a simple yet effective method for autoregressive visual generation. Specifically, the input image tokens are grouped and aggregated into patch tokens with higher information density through an intra-patch average operation. With the resulting patch tokens as a shorter input sequence, the autoregressive model is trained to predict the next patch, thus significantly reducing the computational cost. To further exploit the hierarchical nature of images, we propose a multi-scale patch grouping strategy that progressively refines predictions in a coarse-to-fine manner, seamlessly extending the vanilla NTP paradigm to our novel NPP paradigm. Specifically, the training process starts with a large patch size and ends with vanilla NTP where the patch size is 1×1, thus preserving the original inference stage without requiring modifications. Extensive experiments show that our method not only enhances training efficiency but also improves the generation quality. As shown in Figure 1, experiments on a diverse range of models from 100M to 1.4B parameters demonstrate that the NPP paradigm could reduce the training cost to $\sim 0.6\times$ while improving image generation quality by up to 1.0 FID score on the ImageNet 256×256 generation benchmark. Some of the generation results are shown in Figure 2 c). We highlight that our method retains the original autoregressive model architecture without introducing additional trainable parameters or specifically designing a custom image tokenizer. This ensures flexibility for seamless adaptation to various autoregressive models addressing visual generation tasks.

To sum up, this work contributes in the following ways:

- We propose a simple yet effective method to aggregate image tokens into high information density patch tokens. Meanwhile, with patch tokens as a shorter input sequence, our approach enables the autoregressive model

to efficiently process and predict the next patch tokens, significantly lowering computational costs.

- Leveraging the hierarchical property of image modality, we further introduce a multi-scale patch strategy to seamlessly extend the next token prediction paradigm to our novel next patch prediction paradigm.
- Experiments on a diverse range of models demonstrate that our method could reduce the training cost to $\sim 0.6\times$ while improving image generation quality by up to 1.0 FID score on the ImageNet generation benchmark.

Related Works

Visual Generation

Generative adversarial networks (GANs) (Goodfellow et al. 2014) are the pioneering method for visual generation in the deep learning era, focusing on learning to generate realistic images through adversarial training. Inspired by language model architectures, BERT-style models (Chang et al. 2022, 2023; Yu et al. 2023) emerge, using masked-prediction techniques to learn to predict missing parts of images, much like how BERT predicts masked words in text. Diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Peebles and Xie 2023; Podell et al. 2023; Esser et al. 2024) introduce a novel approach, treating visual generation as a reverse diffusion process, where images are gradually denoised from Gaussian noise through a series of steps. Autoregressive models (Esser, Rombach, and Ommer 2021; Ramesh et al. 2021; Yu et al. 2022), inspired by GPT, predict the next token in a sequence. These methods often involve an image tokenization step (Kingma and Welling 2013; Van Den Oord, Vinyals et al. 2017), converting pixel space into a more semantically meaningful representation and training the autoregressive model with encoded tokens. Some works (Han et al. 2024; Li et al. 2024b) focus on image tokenizer for better compression and reconstruction of image data, which is also crucial for the image generation quality.

Recently, a pioneering work LlamaGen (Sun et al. 2024) introduced the next token prediction paradigm for image generation with a vanilla autoregressive model. VAR (Tian et al. 2024) proposes a novel next scale prediction, however, requiring a specialized multi-scale tokenizer and incurring longer input token sequences. In this work, we follow LlamaGen for autoregressive visual generation and extend the next token prediction paradigm to our novel next patch prediction. Concurrently, a series of works (Pang et al. 2024; Yu et al. 2024; He et al. 2024; Wang et al. 2024b) explore different novel modeling strategies for autoregressive visual generations, including next random token prediction, and parallelized tokens prediction. However, these works do not focus on training efficiency and largely modify the autoregressive property, inevitably introducing additional complexity to the model. In contrast, our method focuses on training efficiency and preserves the original autoregressive model architecture without introducing additional trainable parameters or specifically designing a custom image tokenizer.

Multimodal Foundation Models

Recent advancements in large language models and vision-and-language models (Liu et al. 2024) have demonstrated impressive capabilities in various language and vision tasks. However, unifying the understanding and generation tasks in multimodal large language models is still being explored. Most existing approaches (Sun et al. 2023; Li et al. 2024a) focus on integrating diffusion models with other existing pre-trained models, rather than adopting a unified next-token prediction paradigm. These methods often require complex designs to link two distinct training paradigms, which makes scaling up more challenging and inevitably disconnects visual token sampling from the multimodal large language models. Some pioneering efforts (Lu et al. 2022; Team 2024; Wang et al. 2024a) explore incorporating image generation into large language models using an autoregressive approach, achieving promising results. However, most of them directly adopt the next token prediction paradigm without exploring novel autoregressive visual generation approaches. In this work, our method does not introduce additional trainable parameters or specifically design a custom image tokenizer, ensuring flexibility for seamless adaptation to various autoregressive image generation tasks, including unified vision-language models for understanding and generation tasks.

Method

Preliminaries

We outline the vanilla NTP as shown in Figure 3 b). An input image is first encoded into a sequence of discrete tokens $\mathbf{x} = [x_1, x_2, ..., x_K]$ by a pre-trained VQVAE encoder. The autoregressive model is trained to model the probability distribution of a sequence based on a forward autoregressive factorization. Specifically, the training objective is to maximize the joint probability of predicting the current token x_k given the condition token c and all preceding tokens $[x_1, x_2, ..., x_{k-1}]$:

$$\max_{\theta} p_\theta(\mathbf{x}) = \prod_{k=1}^{K} p_\theta(x_k | c, x_1, x_2, \cdots, x_{k-1}), \quad (1)$$

where p_θ represents the token distribution predictor with an autoregressive model parameterized by θ. The model utilizes a stack of transformer layers with causal attention, commonly known as a decoder-only transformer. During the inference stage, the model takes a class token as the condition and generates the following image tokens in an autoregressive manner. In this work, we focus on exploring the modeling method for input token sequence and retain the original autoregressive model architecture without introducing additional trainable parameters or specifically signing a custom image tokenizer.

Next Patch Prediction

We introduce the Next Patch Prediction paradigm in Figure 3 a). The input image is initially encoded into image token indexes, which are then mapped to token embeddings

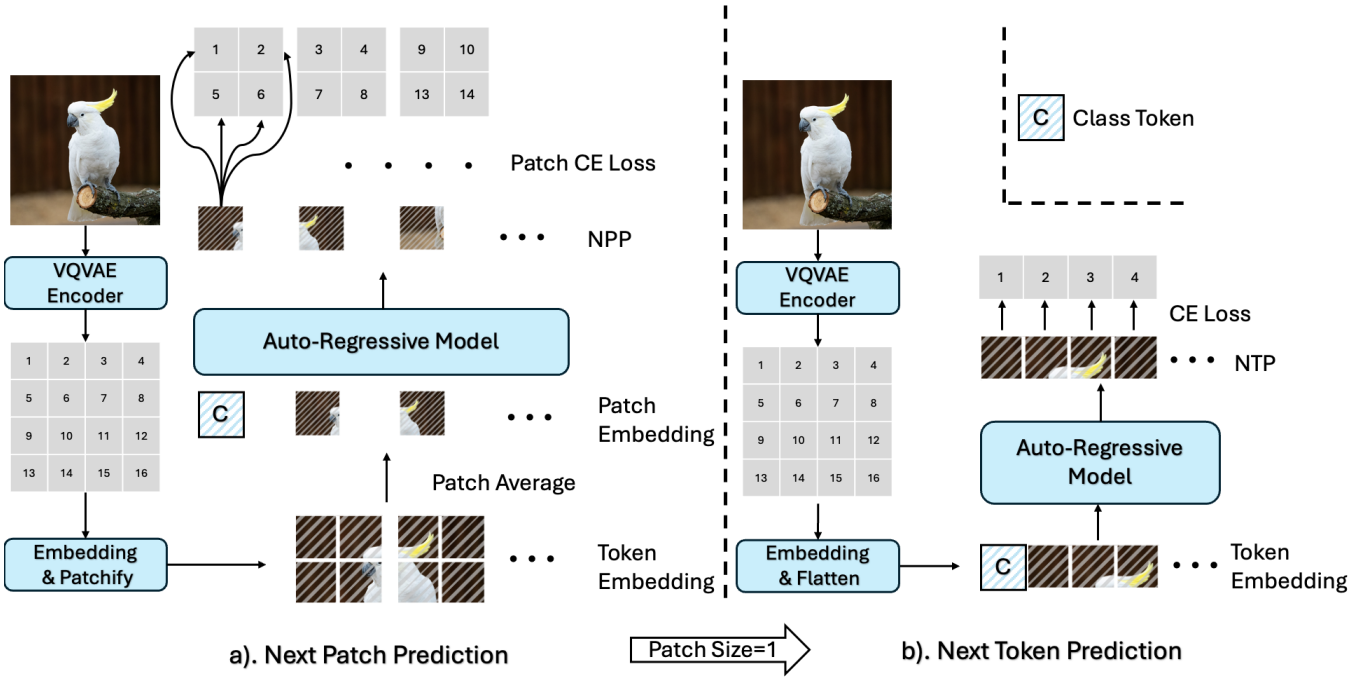

Figure 3: Next Patch Prediction.

of sequence length N. Considering the naturally low information density of image data, our key idea is to aggregate multiple tokens into groups of units containing higher information density. Specifically, we group tokens into non-overlapping patches and generate a sequence of patch embeddings with length $\frac{N}{K}$, where K is the number of tokens associated with each patch. To avoid introducing extra parameters during this compression process, we simply adopt an intra-patch average operation to compute the patch embeddings. Formally, given the embedding function E, for the i-th patch p_i associated with K image tokens x_k^i in the input sequence, the patch embedding is formulated as,

$$E(p_i) = \frac{1}{K}\sum_{k=1}^{K} E(x_k^i). \tag{2}$$

In this way, the original input token embeddings of sequence length N are aggregated into patch embeddings of sequence length $\frac{N}{K}$. With the resulting patch embeddings as input, the autoregressive model is trained to predict the next patch. However, directly maximizing the joint probability as in Equation 1 is difficult due to the absence of an explicit ground truth (GT) index for a patch token. To address this issue, we maintain the original prediction head and propose a patch-wise Cross-Entropy (CE) loss that supervises the model using the associated K image token GT indexes $Index_k^i$ in the next patch p_i. Specifically, given the next patch predictions by the share-weighted prediction head as $Pred_i = P(p_i|c, p_{<i})$, and recalling the patch sequence

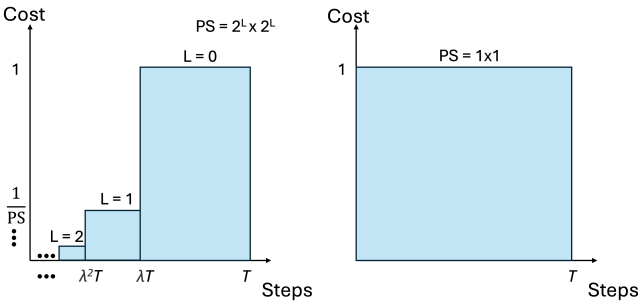

Figure 4: Multi-scale Next Patch Prediction. The patch grouping function begins with a large patch size, resulting in a short sequence length. As training progresses, the patch size is gradually reduced to 1×1.

length $\frac{N}{K}$, the loss function is formulated as:

$$L = -\frac{1}{N}\sum_{i=1}^{\frac{N}{K}}\sum_{k=1}^{K}\log(Pred_i). \tag{3}$$

However, simply training with this objective leads to the issue that all tokens in a patch are predicted to be the same during the inference stage. To address this issue and seamlessly extend the next token prediction paradigm to our novel next patch prediction paradigm, we propose a multi-scale, coarse-to-fine patch grouping strategy that leverages the natural hierarchical structure of image data as illustrated in Fig-

Type	Model	#Para.	FID↓	IS↑	Precision↑	Recall↑
GAN	BigGAN (Brock, Donahue, and Simonyan 2018)	112M	6.95	224.5	0.89	0.38
	GigaGAN (Kang et al. 2023)	569M	3.45	225.5	0.84	0.61
	StyleGan-XL (Sauer, Schwarz, and Geiger 2022)	166M	2.30	265.1	0.78	0.53
Diffusion	ADM (Dhariwal and Nichol 2021)	554M	10.94	101.0	0.69	0.63
	CDM (Ho et al. 2022)	–	4.88	158.7	–	–
	LDM-4 (Rombach et al. 2022)	400M	3.60	247.7	–	–
	DiT-L/2 (Peebles and Xie 2023)	458M	5.02	167.2	0.75	0.57
Mask.	MaskGIT (Chang et al. 2022)	227M	6.18	182.1	0.80	0.51
	MaskGIT-re (Chang et al. 2022)	227M	4.02	355.6	–	–
VAR	VAR-$d16$ (Tian et al. 2024)	310M	3.30	274.40	0.84	0.51
	VAR-$d20$ (Tian et al. 2024)	600M	2.57	302.60	0.83	0.56
	VAR-$d24$ (Tian et al. 2024)	1.0B	2.09	312.90	0.82	0.59
AR	VQGAN (Esser, Rombach, and Ommer 2021)	227M	18.65	80.4	0.78	0.26
	VQGAN (Esser, Rombach, and Ommer 2021)	1.4B	15.78	74.3	–	–
	VQGAN-re (Esser, Rombach, and Ommer 2021)	1.4B	5.20	280.3	–	–
	ViT-VQGAN (Yu et al. 2021)	1.7B	4.17	175.1	–	–
	ViT-VQGAN-re (Yu et al. 2021)	1.7B	3.04	227.4	–	–
	RQTran. (Lee et al. 2022)	3.8B	7.55	134.0	–	–
	RQTran.-re (Lee et al. 2022)	3.8B	3.80	323.7	–	–
	GPT2-re (Esser, Rombach, and Ommer 2021)	1.4B	5.20	280.3	–	–
	Open-MAGVIT2-B (Luo et al. 2024)	343M	3.08	258.3	0.85	0.51
AR	LlamaGen-B (Sun et al. 2024)	111M	5.46	193.61	0.83	0.45
	LlamaGen-L (Sun et al. 2024)	343M	3.80	248.28	0.83	0.52
	LlamaGen-L-384† (Sun et al. 2024)	343M	3.07	256.06	0.83	0.52
	LlamaGen-XL (Sun et al. 2024)	775M	3.39	227.08	0.81	0.54
	LlamaGen-XL-384† (Sun et al. 2024)	775M	2.62	244.08	0.80	0.57
	LlamaGen-XXL (Sun et al. 2024)	1.4B	3.10	253.61	0.83	0.53
Ours	NPP-B	111M	4.47	229.25	0.86	0.46
	NPP-L	343M	2.76	266.34	0.83	0.56
	NPP-XL	775M	2.65	281.03	0.83	0.57
	NPP-XXL	1.4B	2.54	286.13	0.84	0.56
	NPP-2B	2B	2.28	290.22	0.85	0.57

Table 1: Model comparisons on class-conditional ImageNet 256×256 benchmark. "-re" means using rejection sampling. "†" means the model is trained on 384×384 resolution.

ure 4. Specifically, the grouping function begins with a large kernel size, resulting in large patches and a short patch sequence length, allowing the autoregressive model to capture coarse representations. As training progresses, the patch size is gradually reduced to 1×1, enabling the model to learn finer details. This strategy seamlessly extends NTP to NPP, making the NPP inference process identical to the vanilla NTP inference stage. To balance training efficiency and model performance, we introduce a segment scheduling factor λ and set the number of patch levels $\#L$. During the total training steps T, each segment is represented as $\lambda^L T - \lambda^{(L-1)} T$ with a patch size (PS) of $2^L \times 2^L$, where L denotes the current patch level. The computational cost is reduced by a factor of $\frac{1}{PS}$ due to the shorter sequence length at each level.

Experiments

Implementation Details

All the models are trained for 300 epochs following the same setting of LlamaGen (Sun et al. 2024). In the main results, the number of patch levels $\#L$ is set to 2 and λ is 0.5, which means that in the first 150 epochs, the model is trained with a patch size of 2×2 and for the last 150 epochs, the patch size is reduced to 1×1.

Main results

We compare our method with various baseline works on class-conditional ImageNet 256×256 benchmark and show the results in Table 1. Our method achieves state-of-the-art performance on a diverse model size from 100M to 1.4B parameters compared to baseline methods. Specifically, the NPP-L with only 343M parameters achieves a 2.76 FID score, significantly surpassing state-of-the-art AR models with similar parameters including LlamaGen-L-384 (Sun et al. 2024) (FID 3.07), Open-MAGVIT2-B (Luo et al. 2024) (FID 3.08). It also outperforms the widely used VAR-$d16$ (Tian et al. 2024) (FID 3.30) and the diffusion model DiT-L/2 (Peebles and Xie 2023) (FID 5.02). Moreover, compared with LlamaGen-XL and LlamaGen-XXL, our method consistently outperforms the baseline work trained on 256×256 resolution.

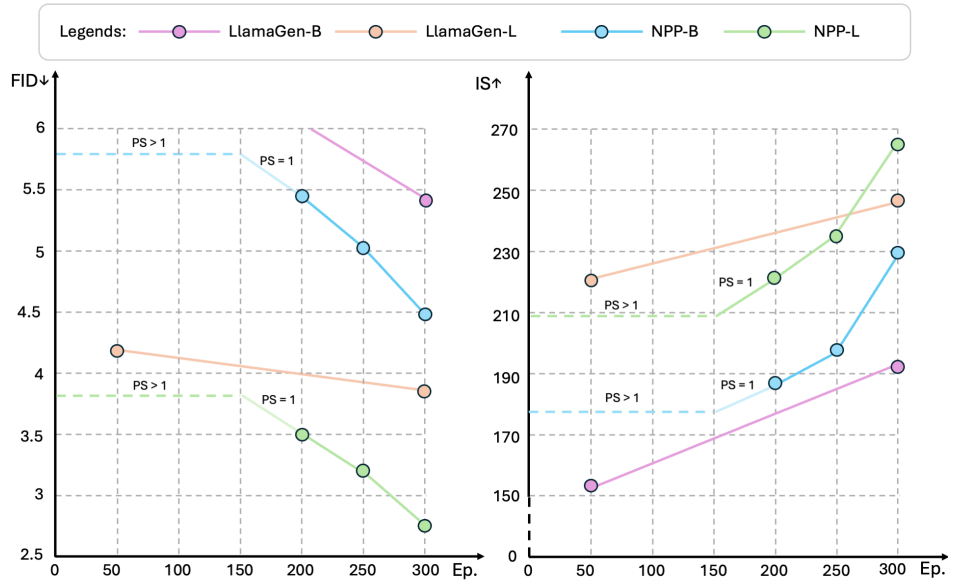

Figure 5: Comparison of NPP and the baseline method LlamaGen.

Model	FID↓	Cost(GFLOPs)↓	Thr.(imgs/sec)↑
LlamaGen-B	5.46	25.06 (1.00×)	∼5888 (1.0×)
NPP-B	4.47	15.70 (0.63×)	∼7625 (1.3×)
VAR-*d*16	3.30	105.70 (1.27×)	∼ 1078 (0.5×)
LlamaGen-L	3.80	83.54 (1.00×)	∼ 2201 (1.0×)
NPP-L	2.76	47.95 (0.57×)	∼ 3469 (1.6×)
VAR-*d*20	2.57	204.40 (1.06×)	∼ 690 (0.7×)
LlamaGen-XL	3.39	193.35 (1.00×)	∼922 (1.0×)
LlamaGen-XL†	2.62	434.11 (2.25×)	∼ 410 (0.5×)
NPP-XL	2.65	102.78 (0.53×)	∼1613 (1.8×)
LlamaGen-XXL	3.10	355.72 (1.00×)	∼448 (1.0×)
LlamaGen-XXL†	2.34	798.64 (2.25×)	∼ 298 (0.7×)
NPP-XXL	2.54	189.11 (0.53×)	∼ 640 (1.4×)

Table 2: Comparisons training cost on class-conditional ImageNet 256×256 benchmark.

Figure 6: Generation results.

To better compare our method with the strong baseline LlamaGen (Sun et al. 2024), we provide a comprehensive study as shown in Figure 5. We report the detailed evaluation metrics FID and IS as training epochs increase. For the base model and large model, our method consistently outperforms LlamaGen during the training process, improving the FID score and inception score. In general, the proposed method outperforms the baseline work LlamaGen by improving the image generation quality up to 1.0 FID scores with significantly higher IS scores.

Training Cost Study

We provide a comprehensive study on the training cost as shown in Table 2. We compare baseline methods, including LlamaGen (Sun et al. 2024) and VAR (Tian et al. 2024) with our method across various model sizes. The average computation cost (GFLOPs per batch) and the actual training throughput (images per second) are presented. For models with 100M-300M parameters, NPP reduces the computation cost to $\sim$ 0.6× and speeds up the training process by $\sim$ 1.3× to 1.6×. Surprisingly, NPP even achieves better generation quality (lower FID scores) with significantly better training efficiency. For large models with 600M-1.4B parameters, our method achieves the best balance between model performance and training efficiency. Specifically, NPP-XL achieves a similar FID score as LlamaGen-XL-384 (2.65 vs 2.62), but with only $\sim$ 0.25× training cost and speeds up the training process by a $\sim$ 4× model throughput.

Generation Results

In Figure 6, we present generation results by NPP on ImageNet 256×256 benchmark. Our NPP is capable of generating high-quality images with diversity and fidelity. More

Model	#para.	FID↓	IS↑	Precision↑	Recall↑	Training Cost↓
LlamaGen-B($PS = 1 \times 1$)	111M	5.46	193.61	0.83	0.45	$1.0\times$
NPP-B ($PS = 2 \times 2$)	111M	4.47	229.25	0.86	0.46	$0.625\times$
NPP-B ($PS = 4 \times 4$)	111M	4.92	222.81	0.86	0.45	$0.531\times$
LlamaGen-L ($PS = 1 \times 1$)	343M	3.80	248.28	0.83	0.52	$1.0\times$
NPP-L ($PS = 2 \times 2$)	343M	2.76	266.34	0.83	0.56	$0.625\times$
NPP-L ($PS = 4 \times 4$)	343M	2.89	262.80	0.83	0.55	$0.531\times$

(a) Comparisons of models trained with different patch sizes.

Model	#para.	FID↓	IS↑	Precision↑	Recall↑	Training Cost↓
LlamaGen-L ($\lambda = 0$)	343M	3.80	248.28	0.83	0.52	$1.0\times$
NPP-L ($\lambda = 1/2$)	343M	2.76	266.34	0.83	0.56	$0.625\times$
NPP-L ($\lambda = 2/3$)	343M	2.79	263.75	0.83	0.55	$0.5\times$
NPP-L ($\lambda = 3/4$)	343M	2.81	262.22	0.83	0.55	$0.43\times$
NPP-L ($\lambda = 4/5$)	343M	2.92	260.68	0.83	0.55	$0.4\times$

(b) Comparisons of models trained with different segment factor λ.

Model	#para.	FID↓	IS↑	Precision↑	Recall↑	Training Cost↓
LlamaGen-B ($\#L = 1$)	111M	5.46	193.61	0.83	0.45	$1.0\times$
NPP-B ($\#L = 2$)	111M	4.47	229.25	0.86	0.46	$0.625\times$
NPP-B ($\#L = 3$)	111M	4.62	231.57	0.86	0.46	$0.578\times$
NPP-B ($\#L = 4$)	111M	4.68	228.31	0.86	0.46	$0.572\times$
LlamaGen-L ($\#L = 1$)	343M	3.80	248.28	0.83	0.52	$1.0\times$
NPP-L ($\#L = 2$)	343M	2.76	266.34	0.83	0.56	$0.625\times$
NPP-L ($\#L = 3$)	343M	2.79	264.30	0.83	0.56	$0.578\times$
NPP-L ($\#L = 4$)	343M	2.84	258.60	0.83	0.56	$0.572\times$

(c) Comparisons of models trained with different numbers of patch level.

Table 3: Ablation studies on key design choices.

generation results are provided in the appendix.

Ablation Studies

Effect of Patch Size. We study the effect of different patch sizes and present the results in Table 3a. In this experiment, we modify the multi-scale grouping strategy to skip intermediate patch size and set the segment scheduling factor $\lambda = 1/2$. The models are ablated with different patch sizes adopted in the first $1/2$ number of training epochs. We observe NPP with different patch sizes consistently outperforms LlamaGen. However, with a larger patch size such as $PS = 4 \times 4$, the learned knowledge cannot be smoothly transferred to the case with $PS = 1 \times 1$, leading to a slight performance drop where FID scores were reduced by 0.45 for NPP-B and 0.13 for NPP-L. Therefore, we choose patch size $PS = 2 \times 2$ as the default setting.

Effect of Segment Scheduling Factor λ. We provide a study on the effect of segment scheduling factors adopted in the proposed multi-scale patch grouping strategy, as shown in Table 3b. In this study, the multi-scale patch grouping strategy is disabled and the patch size is set to $PS = 2 \times 2$. λ factors are scanned from 1/2 to 4/5. We observe that a larger λ factor results in lower training computational cost but with slight performance degradation that FID scores are increased from 2.76 to 2.92. Hence, to balance training efficiency and model performance, we set $\lambda = 1/2$ by default.

Effect of Multi-scale Patch Grouping Strategy. We present a study on the effect of the multi-scale patch grouping strategy as shown in Table 3c. In this experiment, we set $\lambda = 1/2$ and compare different numbers of patch levels $\#L$. Experiments show this strategy makes a trade-off between training computational cost and image generation quality. Moreover, with this strategy, the training process ends with vanilla NTP where the patch size is 1×1, thus preserving the original inference stage without modifications.

Conclusion

In this work, we introduce a novel Next Patch Prediction paradigm that improves autoregressive image generation quality and efficiency by grouping and aggregating image tokens into high-density patch tokens. We further introduce a multi-scale patch strategy to seamlessly bridge the Next Patch Prediction with the vanilla next token prediction paradigm. Our approach reduces the computational cost to $\sim 0.6\times$ while improving image generation quality by up to 1.0 FID score on the ImageNet benchmark. We highlight that our method retains the original autoregressive model architecture without introducing additional trainable parameters or custom image tokenizers, thereby making the next patch prediction paradigm seamlessly adapted to various autoregressive models addressing image generation tasks.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (No. 62332002, 62202014, 62425101) and Shenzhen Science and Technology Program (KQTD20240729102051063).

References

Alabdulmohsin, I. M.; Neyshabur, B.; and Zhai, X. 2022. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35: 22300–22312.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11315–11325.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; Lacey, K.; Goodwin, A.; Marek, Y.; and Rombach, R. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206.

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

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2024. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. arXiv:2412.04431.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, Y.; Chen, F.; He, Y.; He, S.; Zhou, H.; Zhang, K.; and Zhuang, B. 2024. ZipAR: Accelerating Autoregressive Image Generation through Spatial Locality. *arXiv preprint arXiv:2412.04062*.

Henighan, T.; Kaplan, J.; Katz, M.; Chen, M.; Hesse, C.; Jackson, J.; Jun, H.; Brown, T. B.; Dhariwal, P.; Gray, S.; et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1): 2249–2281.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10124–10134.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020a. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020b. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11523–11532.

Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024a. Autoregressive Image Generation without Vector Quantization. *arXiv preprint arXiv:2406.11838*.

Li, X.; Qiu, K.; Chen, H.; Kuen, J.; Gu, J.; Raj, B.; and Lin, Z. 2024b. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; and Kembhavi, A. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Luo, Z.; Shi, F.; Ge, Y.; Yang, Y.; Wang, L.; and Shan, Y. 2024. Open-MAGVIT2: An Open-Source Project Toward Democratizing Auto-regressive Visual Generation. *arXiv preprint arXiv:2409.04410*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pang, Z.; Zhang, T.; Luan, F.; Man, Y.; Tan, H.; Zhang, K.; Freeman, W. T.; and Wang, Y.-X. 2024. RandAR: Decoder-only Autoregressive Visual Generation in Random Orders. *arXiv preprint arXiv:2412.01827*.

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

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv preprint arXiv:2406.06525*.

Sun, Q.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; Wang, Y.; Gao, H.; Liu, J.; Huang, T.; and Wang, X. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Team, C. 2024. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv preprint arXiv:2405.09818*.

Team, I. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. *arXiv preprint arXiv:2404.02905*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. 2024a. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Wang, Y.; Ren, S.; Lin, Z.; Han, Y.; Guo, H.; Yang, Z.; Zou, D.; Feng, J.; and Liu, X. 2024b. Parallelized Autoregressive Visual Generation. *arXiv preprint arXiv:2412.15119*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yu, J.; Li, X.; Koh, J. Y.; Zhang, H.; Pang, R.; Qin, J.; Ku, A.; Xu, Y.; Baldridge, J.; and Wu, Y. 2021. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10459–10469.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2024. Randomized Autoregressive Visual Generation.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.