Boosting Resolution Generalization of Diffusion Transformers with Randomized Positional Encodings

Liang Hou[1*], **Cong Liu**[2*†], **Mingwu Zheng**[1], **Xin Tao**[1‡], **Pengfei Wan**[1], **Di Zhang**[1], **Kun Gai**[1]

[1]Kling Team, Kuaishou Technology
[2]Southeast University
{lianghou96, jiangsutx}@gmail.com

Abstract

Resolution generalization in image generation tasks enables the production of higher-resolution images with lower training resolution overhead. However, a key obstacle for diffusion transformers in addressing this problem is the mismatch between positional encodings seen at inference and those used during training. Existing strategies such as positional encodings interpolation, extrapolation, or hybrids, do not fully resolve this mismatch. In this paper, we propose a novel two-dimensional randomized positional encodings, namely RPE-2D, that prioritizes the order of image patches rather than their absolute distances, enabling seamless high- and low-resolution generation without training on multiple resolutions. Concretely, RPE-2D independently samples positions along the horizontal and vertical axes over an expanded range during training, ensuring that the encodings used at inference lie within the training distribution and thereby improving resolution generalization. We further introduce a simple random resize-and-crop augmentation to strengthen order modeling and add micro-conditioning to indicate the applied cropping pattern. On the ImageNet dataset, RPE-2D achieves state-of-the-art resolution generalization performance, outperforming competitive methods when trained at 256^2 and evaluated at 384^2 and 512^2, and when trained at 512^2 and evaluated at 768^2 and 1024^2. RPE-2D also exhibits outstanding capabilities in low-resolution image generation, multi-stage training acceleration, and multi-resolution inheritance.

Introduction

Diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Song, Meng, and Ermon 2020; Song et al. 2020) have effectively replaced traditional generative models such as variational autoencoders (Kingma 2013) and generative adversarial networks (Goodfellow et al. 2014) as the predominant paradigm in the field of image generation due to their strong generative performance (Dhariwal and Nichol 2021; Rombach et al. 2022). Diffusion Transformers (DiTs) (Peebles and Xie 2023) further demonstrate that Transformers (Vaswani 2017) can be effectively scaled within diffusion frameworks, making Transformer-based diffusion architectures one of the central focuses of modern diffusion model research (Lu et al. 2024; Ma et al. 2024; Chen et al. 2023a, 2024b). However, existing image generation models are typically trained at a specific resolution to produce high-quality images only at that resolution. Scaling these models directly to higher resolutions usually incurs a multiplicative increase in training cost, which becomes prohibitive when computation and data resources are limited. This situation calls for models with genuine cross-resolution generalization ability, such that, even when trained solely on low-resolution images, they can still generate high-quality images at higher resolutions, thereby avoiding the substantial costs associated with conventional high-resolution training.

A number of approaches have been proposed to address resolution generalization of image generation. The first line of work (He et al. 2023; Du et al. 2024; Lu et al. 2024; Teterwak et al. 2019; Yang et al. 2019) focuses on enhancing network architectures, but often leads to complex designs that are tightly coupled to specific frameworks or training pipelines. A second line of work (Jin et al. 2023) improves extrapolation by modifying the attention mechanism to account for changes in attention entropy. However, it largely overlooks a key bottleneck: the one-to-one correspondence introduced by positional encodings (PEs), which enables Transformers to perceive positional information but simultaneously constrains the resolution generalization capacity of DiTs. A third line of methods (Zhuo et al. 2024; Lu et al. 2024; Peng et al. 2023; NTK 2024) explicitly targets the limitations that PEs impose on generalization, proposing interpolation-, extrapolation-, or hybrid-based schemes. Yet these methods remain bounded by the intrinsic extrapolation limits of the underlying PEs and do not fully close the PE gap between training and inference.

In this work, we revisit resolution generalization in image generation from the perspective of PEs. We argue that the fundamental reason existing methods perform poorly at resolution extrapolation is that many PEs required at test time have never "truly" appeared during training, leading to a systematic distributional mismatch of PEs between training and inference. To fundamentally alleviate this issue, we posit that all PEs used at test time should, in a statistical sense, be "covered" by the sampling process during training. Guided by this principle, and inspired by the success of one-

[*]These authors contributed equally.
[†]This work was conducted during the author's internship at Kling Team, Kuaishou Technology.
[‡]Corresponding author.

dimensional randomized positional encodings (RPE-1D) in handling length extrapolation in large language models (Ruoss et al. 2023), we propose RPE-2D, a two-dimensional, training-based randomized positional encoding framework tailored for resolution generalization in image generation.

In contrast to conventional approaches that attempt to extend positions along fixed coordinate axes, RPE-2D performs random sampling over a larger two-dimensional grid while only enforcing consistency of order along the horizontal and vertical axes. As a result, all PEs required during high-resolution inference can be regarded as statistically covered by the random sampling process at training time. This reframes an out-of-domain extrapolation task as an in-domain interpolation problem and models them in a unified manner via random selection, thereby avoiding any additional training overhead. Conceptually, each image can be regarded as a cropped, resized, or geometrically transformed view of a larger latent canvas. This is fundamentally different from the one-dimensional textual sequences processed by language models, where it is natural to assume a uniform step size between adjacent tokens along the sequence and to use equally spaced positional encodings to represent their order. In contrast, in two-dimensional visual settings, different views correspond to different regions and scales of the same underlying canvas. From this perspective, using exactly the same, equally spaced positional encodings to model all such views introduces unnecessary constraints on positional modeling. The design of RPE-2D is precisely motivated by this observation: by assigning randomized two-dimensional positional encodings, it aims to weaken the model's reliance on specific positional intervals and instead encourage it to exploit positional order, which is an essential factor that has often been overlooked in prior work.

Concretely, RPE-2D performs without-replacement random sampling along the horizontal and vertical axes of a predefined maximal grid, followed by sorting the sampled indices in ascending order to construct a two-dimensional set of random positions. At test time, we instead adopt a deterministic, equidistant sampling strategy to achieve better generalization in expectation. To further enhance the model's ability to capture positional order, we introduce a data augmentation strategy that combines random resizing and cropping, and employ micro-conditioning to explicitly inject the corresponding cropping and resizing information. This allows the model to preserve the topological structure of images while relying more on positional order than on precise distances. In addition, we incorporate attention scaling and timestep shifting strategies during inference to alleviate performance degradation caused by changes in attention entropy and signal-to-noise ratio when sampling at high resolutions.

We empirically validate RPE-2D on ImageNet at both 256^2 and 512^2 training resolutions. When trained at 256^2 and evaluated at 384^2 and 512^2, as well as trained at 512^2 and evaluated at 768^2 and 1024^2, RPE-2D consistently outperforms strong positional-encoding extrapolation baselines, demonstrating state-of-the-art resolution generalization performance under all evaluation settings. Moreover, integrating RPE-2D with different PE families maintains or improves in-distribution image quality at the training resolution, indicating that RPE-2D is broadly compatible with existing DiT architectures. Beyond upward resolution extrapolation, RPE-2D also supports downward resolution generation, accelerates multi-stage training when fine-tuning to higher resolutions, and enables flexible multi-resolution inheritance, highlighting its practical value for scalable diffusion transformers.

Related Work

Length Generalization in Languages Models

A significant stride in extrapolation has been achieved with ALIBI (Press, Smith, and Lewis 2021), a method that employs local attention to reinforce the model's ability to capture local dependencies within the data. This is crucial as it allows the model to maintain a more refined understanding of the data's structure, thereby improving the quality of extrapolation.Another notable approach is the NTK (NTK 2024), which adjusts the frequency components of the position encodings. This method is designed to preserve the high-frequency information during the extrapolation process, ensuring a more accurate representation of the data's characteristics. YaRN (Peng et al. 2023) is an innovative approach that extends the context window of large language models efficiently. It does so by modifying the attention mechanism to handle longer sequences without the need for fine-tuning, thus maintaining a consistent level of performance across various lengths of input data. The concept of random position encoding (Ruoss et al. 2023) has also gained traction, offering a more natural and elegant solution to the challenge of handling longer sequences during prediction. This method has been shown to be effective not only in language models but also in non-language models, where the generation of images or other data types requires a broader context understanding. Attention Masking is another strategy that has proven effective in language models, which are inherently local in nature. By "forcing" the model to focus on a limited number of tokens, it can effectively manage the increased complexity during prediction. However, its applicability to non-language models is still under exploration .

Resolution Generalization in Diffusion Models

In the realm of computer vision, extrapolation techniques have been pivotal in advancing the capabilities of models to generate images and predict video sequences beyond the limits of their training data. The development of FiT (Lu et al. 2024) and LuminaNext (Zhuo et al. 2024) has showcased the potential of local attention mechanisms in enhancing the performance of image generation models. Local attention focuses on specific regions within an image, allowing for more detailed and accurate generation of high-resolution images. In addition to these, there are modifications to the network structure, such as attention scale (Jin et al. 2023), neighborhood attention (Hassani et al. 2023), and KV-compression (Chen et al. 2024a). In summary, while current methods have made limited improvements in extrapolation capabilities, they still fail to address the fundamental issue of the position encoding gap between training and prediction.

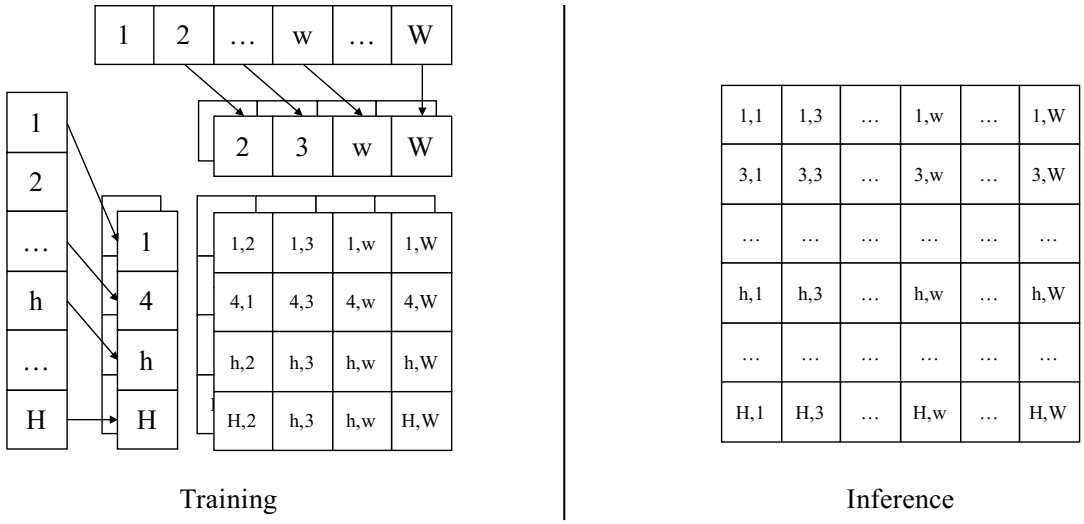

Figure 1: Illustration of RPE-2D for training and inference. During training (left), row and column indices are randomly sampled without replacement from the maximal grid $H \times W$ and sorted to form a set of 2D positions matching the training resolution. During inference (right), a deterministic, approximately equidistant grid matching the inference resolution is used.

Preliminary

Positional Encodings

Sinusoidal PE Positional encodings (PEs) (Vaswani 2017) play a significant role in Transformer-based sequence modeling, as they inject positional information into token representations to compensate for the order-agnostic nature of self-attention. A widely used choice is the sinusoidal PE, which adds to each token embedding $\mathbf{x}_m \in \mathbb{R}^d$ at position $m \in \{1, 2, \ldots, L\}$ a positional vector $\mathrm{PE}(m) := \mathbf{p}_m \in \mathbb{R}^d$, where $d \in \mathbb{N}^+$ is the embedding dimension. Its components are defined as

$$\mathrm{PE}(m, 2i) := p_{m,2i} = \sin(m\theta_i), \tag{1}$$

$$\mathrm{PE}(m, 2i+1) := p_{m,2i+1} = \cos(m\theta_i), \tag{2}$$

where $i \in \{0, 1, \ldots, d/2 - 1\}$ and $\theta_i = b^{-2i/d}$ is the frequency associated with the i-th pair of dimensions, with base $b \in \mathbb{R}^+$.

RoPE Rotary positional encoding (RoPE) (Su et al. 2024) is a form of relative PE that has shown strong length generalization and has become a preferred choice in both modern LLMs and DiTs. Instead of adding a positional vector, RoPE applies a position-dependent rotation to the query and key vectors in self-attention. Let $\mathbf{q}_m \in \mathbb{R}^d$ and $\mathbf{k}_n \in \mathbb{R}^d$ denote the query and key at positions m and n, respectively, and let f denote the attention function. RoPE modifies f as

$$\begin{aligned} f(\mathbf{q}_m, \mathbf{k}_n, m, n) &= (\mathbf{R}_m \mathbf{q}_m)^\top (\mathbf{R}_n \mathbf{k}_n) \\ &= \mathbf{q}_m^\top \mathbf{R}_m^\top \mathbf{R}_n \mathbf{k}_n = \mathbf{q}_m^\top \mathbf{R}_{n-m} \mathbf{k}_n, \end{aligned} \tag{3}$$

where $\mathbf{R}_m$ and $\mathbf{R}_n$ are rotation matrices that depend on the absolute positions, and $\mathbf{R}_{n-m} := \mathbf{R}_m^\top \mathbf{R}_n$ depends only on the relative offset $(n - m)$. For a single 2D subspace (a pair of channels), the relative rotation matrix takes the form

$$\mathbf{R}_{n-m} = \begin{pmatrix} \cos((n-m)\theta_i) & -\sin((n-m)\theta_i) \\ \sin((n-m)\theta_i) & \cos((n-m)\theta_i) \end{pmatrix}, \tag{4}$$

The full d-dimensional rotation matrix is block-diagonal, composed of such 2×2 rotation blocks.

2D Positional Encodings

For image-like data with a two-dimensional structure, PEs are typically extended to 2D by composing two independent 1D PEs along the horizontal and vertical axes. Taking 2D RoPE as an example, consider the query $\mathbf{q}_{x_1,y_1}$ at spatial position (x_1, y_1) and the key $\mathbf{k}_{x_2,y_2}$ at position (x_2, y_2). The corresponding 2D rotary matrix can be written as

$$\mathbf{R}_{x_2-x_1,\, y_2-y_1} = \begin{pmatrix} \mathbf{R}_{x_2-x_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{y_2-y_1} \end{pmatrix}, \tag{5}$$

where $\mathbf{R}_{x_2-x_1}$ and $\mathbf{R}_{y_2-y_1}$ are 1D RoPE rotation matrices along the horizontal and vertical directions, respectively, and the full 2D rotation is realized as a block-diagonal composition of the two. This construction naturally adapts RoPE to 2D grids while preserving its relative-position property along each axis.

Method

2D Randomized Positional Encodings

We consider resolution generalization in image generation, where a model is trained only at a low resolution due to computational constraints but is expected to generate images at higher resolutions at test time. Let $h_{\text{train}}, w_{\text{train}} \in \mathbb{N}^+$ denote the spatial size of the training images (or VAE latents), and $h_{\text{test}}, w_{\text{test}} \in \mathbb{N}^+$ that of the test images, with $h_{\text{test}} > h_{\text{train}}$ and $w_{\text{test}} > w_{\text{train}}$. Under such resolution extrapolation, many positional encodings required at test time inevitably lie outside the range seen during training.

NTK (NTK 2024) and YaRN (Peng et al. 2023) extend the usable context range by combining interpolation and extrapolation, but they do not resolve a fundamental issue: the

positional encoding associated with each token differs between training and inference. Inspired by one-dimensional randomized positional encodings (RPE-1D) (Ruoss et al. 2023) in LLMs, we reinterpret resolution extrapolation in image generation as an interpolation problem and propose *2D Randomized Positional Encodings* (RPE-2D). The core idea is to ensure that all positional encodings used at test time are statistically covered by the training-time sampling process. By randomly assigning positions to image patches in a structured manner, every test-time position lies within the training distribution, thereby improving robustness to positional shifts.

As illustrated in Fig. 1, RPE-2D extends RPE-1D, originally designed for text, to a two-dimensional setting suitable for images. A naive extension would be to flatten the $h_{\text{train}} \times w_{\text{train}}$ patches into a 1D sequence and sample positions from a longer 1D range of length HW, where $H > h_{\text{test}} > h_{\text{train}}$ and $W > w_{\text{test}} > w_{\text{train}}$ are hyperparameters. However, such flattening ignores the inherent 2D structure of images and entangles horizontal and vertical neighbors in an unnatural way, leading to distorted distances along the two axes. For 2D image data, the horizontal and vertical axes are naturally decoupled. RPE-2D therefore performs independent randomized position sampling along each axis. Formally, at each training step we sample, without replacement, index sets $\mathcal{X} \subset \{1, 2, \ldots, H\}$ and $\mathcal{Y} \subset \{1, 2, \ldots, W\}$ such that $|\mathcal{X}| = h_{\text{train}}$ and $|\mathcal{Y}| = w_{\text{train}}$. We then sort them in ascending order, $\mathcal{X} = \{x_1, \ldots, x_{h_{\text{train}}}\}$ with $x_1 < x_2 < \cdots < x_{h_{\text{train}}}$ and $\mathcal{Y} = \{y_1, \ldots, y_{w_{\text{train}}}\}$ with $y_1 < y_2 < \cdots < y_{w_{\text{train}}}$. The 2D random position set is constructed via the Cartesian product

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) \mid x \in \mathcal{X},\, y \in \mathcal{Y}\}. \tag{6}$$

For the patch at training index (i, j), where $1 \le i \le h_{\text{train}}$ and $1 \le j \le w_{\text{train}}$, its randomized positional encoding is defined as

$$\text{RPE}(i, j) := \text{PE}(x_i, y_j) \in \mathbb{R}^d,$$

with $(x_i, y_j) \in \mathcal{X} \times \mathcal{Y}$ and $\text{PE}(\cdot, \cdot)$ denoting any 2D positional encoding function (e.g., SinPE or RoPE). This construction preserves the monotonic order along each axis and induces a consistent 2D grid structure: along any fixed row, vertical coordinates are aligned, and along any fixed column, horizontal coordinates are aligned, while the actual intervals between sampled positions vary across training steps and thus prevent the model from memorizing specific lengths.

At test time, RPE-2D uses deterministic and approximately equidistant positions. Given a maximal grid of size $H \times W$, we choose $x_1 = 1$, $x_{h_{\text{test}}} = H$ and $y_1 = 1$, $y_{w_{\text{test}}} = W$, with spacings $x_{i+1} - x_i = \lfloor H/h_{\text{test}} \rfloor$ and $y_{j+1} - y_j = \lfloor W/w_{\text{test}} \rfloor$. In this way, all test-time positions lie within the support of the randomized training positions while covering the full spatial extent of the maximal grid, effectively turning resolution extrapolation into interpolation over a shared 2D positional range. Our RPE-2D training paradigm is orthogonal to the specific choice of positional encoding and can be applied on top of both sinusoidal PEs and RoPE; we empirically validate this compatibility in our experiments (see Table 3).

Data Augmentation and Micro-Conditioning

To further enhance the model's ability to perceive the order of image patches, we jointly apply resize and crop operations to transform the "collected" high-resolution images into low-resolution inputs suitable for training. The resize operation encourages the model to capture global structure, while the crop operation prompts it to attend to local details. Importantly, the low-resolution images produced by these two operations are kept at the same spatial resolution. To address the issue of image incompleteness introduced by cropping, we design a micro-conditioning mechanism. We first upsample each low-resolution image in the training set to a high resolution (if necessary) and record its base resolution as $\mathbf{c}_{\text{original}} = (h_{\text{original}}, w_{\text{original}})$. During each training iteration, we then randomly select start and end coordinates from a set of cropping options (including the no-crop case corresponding to global resizing) and crop the base image accordingly, yielding crop coordinates $\mathbf{c}_{\text{crop}} = (c_{\text{top}}, c_{\text{left}}, c_{\text{down}}, c_{\text{right}})$. The cropped region is subsequently resized to a target resolution $\mathbf{c}_{\text{resize}} = (h_{\text{target}}, w_{\text{target}})$, where $h_{\text{target}} \times w_{\text{target}}$ matches the desired training resolution. These three types of conditioning information are injected into the model via adaLN (Xu et al. 2019). Concretely, each component is independently embedded using Fourier feature encoding (Tancik et al. 2020), and the resulting embeddings are concatenated into a single vector. We then add this vector to the DiT (Peebles and Xie 2023) timestep embedding, thereby providing the model with explicit information about the original resolution, cropping pattern, and final resize configuration.

Training-Free Sampling Strategy

Attention Scale In addition to the changes in PEs, resolution extrapolation inevitably leads to an increase in the number of image patches, creating another inconsistency between testing and training. Since attention is scale-dependent, this dependency arises from the fact that the entropy of attention changes as the number of patches increases (Jin et al. 2023). We also attempt to use the proposed scaling factor to mitigate the variations in attention entropy,

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\log_n m}{\sqrt{d}}\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}, \tag{7}$$

where $m = h_{\text{test}} \times w_{\text{test}}$ and $n = h_{\text{train}} \times w_{\text{train}}$ represent the number of patches during testing and training, respectively.

Timestep Shift When generating large images with diffusion models, the increase in resolution leads to an increase in the signal-to-noise ratio (SNR) of the noise schedule used in training (Hoogeboom, Heek, and Salimans 2023). Therefore, it is necessary to adjust the inference timestep spacing during sampling to maintain the SNR as much as possible. Specifically, We follow SD3 (Esser et al. 2024) to map the time tep $t_n \in \{1, 2, \ldots, T\}$ for n patches in training to the timestep $t_m \in \{1, 2, \ldots, T\}$ for m patches in inference to approximate the same level of SNR,

$$t_m = \left\lfloor \frac{\sqrt{\frac{m}{n}} \times \frac{t_n}{T}}{1 + \left(\sqrt{\frac{m}{n}} - 1\right) \times \frac{t_n}{T}} \right\rfloor \times T. \tag{8}$$

Method	ImageNet 256×256									
	384×384					512×512				
	FID↓	sFID↓	IS↑	Precision↑	Recall↑	FID↓	sFID↓	IS↑	Precision↑	Recall↑
PI	18.87	41.59	260.97	0.8312	0.0602	30.64	57.76	159.72	0.676	0.055
Ext	15.95	37.79	374.74	0.8970	0.0636	28.35	54.77	232.41	0.607	0.181
NTK	16.56	35.92	375.28	0.9203	0.0686	27.88	49.8	227.45	0.619	0.177
YaRN	16.97	26.08	264.34	0.7864	0.1050	19.13	34.31	253.16	0.749	0.151
RPE-2D	**15.63**	**14.40**	**385.67**	**0.9631**	**0.1174**	**17.95**	**18.23**	**348.99**	**0.849**	**0.181**

Table 1: Comparison of RPE-2D with different methods on resolution extrapolation trained on ImageNet 256×256.

Method	ImageNet 512×512									
	768×768					1024×1024				
	FID↓	sFID↓	IS↑	Precision↑	Recall↑	FID↓	sFID↓	IS↑	Precision↑	Recall↑
PI	27.24	68.63	150.53	**0.8441**	0.373	38.64	92.49	139.38	**0.7066**	0.361
Ext	20.57	53.65	223.39	0.8184	0.462	45.77	116.54	180.16	0.5698	0.499
NTK	21.58	46.11	225.00	0.7883	0.452	32.90	73.69	**216.12**	0.6304	0.579
YaRN	55.21	75.84	62.93	0.5547	0.513	50.65	79.04	103.38	0.6010	0.461
RPE-2D	**20.45**	**40.46**	**271.05**	0.8271	**0.512**	**25.40**	**47.18**	192.05	0.8109	**0.535**

Table 2: Comparison of RPE-2D with different methods on resolution extrapolation trained on ImageNet 512×512.

Experiments

Experimental Setup

Training Settings We follow DiT (Peebles and Xie 2023) using ImageNet-256^2 and ImageNet-512^2 as training datasets, employing the DiT-XL/2 network architecture while keeping the other training hyper-parameters unchanged[1]. On the ImageNet-256^2, we trained the model from scratch using the proposed random position encoding for 400k iterations, and compared it with the baseline. Subsequently, we apply the weights obtained from training ImageNet256^2 for 400k iterations to ImageNet512^2 for an additional 800k iterations, and compared the resolution extrapolation results with the baseline method.

Evaluation Metrics Following DiT, we use FID (Heusel et al. 2017), sFID (Nash et al. 2021), IS (Salimans et al. 2016), and Precision/Recall (Kynkäänniemi et al. 2019) as the quantitative evaluation metrics in the experiments. We also follow the default value of 4.0 for "cfg-scale" in the `sample.py` file in the DiT official code.

Comparisons

RPE-2D is a training approach for positional encodings rather than a specific encoding form, making it theoretically compatible with any type of positional encoding. We apply RPE-2D to both SinPE and RoPE, and the results in Table 3 show that combining either PE with RPE-2D consistently improves performance.

We then compare RPE-2D with PI (Chen et al. 2023b), extrapolation (Ext), NTK (NTK 2024), and YaRN (Peng et al. 2023) for resolution extrapolation, all built on top of RoPE (Su et al. 2024). It is worth noting that all competitors are implemented with two-dimensional positional encodings: in particular, NTK and YaRN are extended to their 2D RoPE-style versions by applying Eq. (5) following FiT (Lu et al. 2024). Starting from weights trained on ImageNet at 256^2 for 400k iterations, we extrapolate to 384^2 and 512^2. As shown in Table 1, RPE-2D achieves state-of-the-art metrics at both 384^2 and 512^2, reducing the previous best sFID of **34.31** obtained by YaRN to **18.23**, and thus substantially improving resolution extrapolation. The qualitative results in Fig. 2 further indicate that RPE-2D maintains superior visual quality at 512^2, suggesting that our method effectively pushes the practical extrapolation range beyond that of YaRN and NTK.

We further fine-tune the models on ImageNet at 512^2 for an additional 800k iterations and extrapolate to 768^2 and 1024^2. As reported in Table 2, RPE-2D continues to obtain the best overall performance at higher resolutions, while PI achieves competitive results on the precision metric. Fig. 2 presents qualitative comparisons between our method and baseline approaches: only RPE-2D consistently preserves both global structure and fine details across resolutions. In contrast, methods such as NTK and YaRN, which combine interpolation and extrapolation, tend to exhibit structural artifacts, whereas PI, as a purely interpolation-based method, often suffers from noticeable detail loss.

Ablation Studies

We conduct ablation studies on the main components of RPE-2D: (i) the random resize-and-crop augmentation with micro-conditioning (Cond-Aug), (ii) attention scaling, and (iii) timestep shifting.

Our Cond-Aug treats the collected low-resolution training

[1]https://github.com/facebookresearch/DiT

Figure 2: Qualitative results of RPE-2D against different positional encoding extrapolation methods at different resolutions.

images as resized or cropped views of larger images, where different sampling intervals and starting points correspond to resizing scale factors and cropping coordinates. As shown in Table 4, Cond-Aug reduces the FID from **20.78** to **19.12** and improves the IS from **293.96** to **325.76**, yielding substantial gains at extrapolated resolutions and indicating that it strengthens the modeling of positional order.

In addition, the carefully designed attention scaling and timestep shifting further improve IS while significantly reducing sFID, as reported in Table 4, demonstrating their effectiveness when combined with RPE-2D for high-resolution sampling.

Applications

Low-Resolution Image Generation As shown in Fig. 3, RPE-2D can not only generate images with higher resolutions than those used for training, but also synthesize images at lower resolutions, e.g., generating 128^2 images when the training resolution is 256^2. This demonstrates that RPE-2D exhibits resolution generalization in both upward and downward directions.

Multi-Stage Training Acceleration Because RPE-2D enables high-quality high-resolution generation from models trained at lower resolutions, a natural application is to facilitate multi-stage, multi-resolution training. We take a model pre-trained on ImageNet at 256^2, fine-tune it at a resolution of 512^2, and compare the loss convergence of standard RoPE against RoPE equipped with our randomized positional encoding scheme. As illustrated in Fig. 4, the model with RPE-2D starts from a lower loss and converges more rapidly, which is beneficial for staged training of large-scale models.

Conclusion

This work investigates the resolution generalization problem in diffusion transformers from the perspective of positional encodings (PEs). Previous approaches have not fully addressed the inconsistency of PEs between training and test-

Method	ImageNet 256 × 256					ImageNet 512 × 512				
	FID↓	sFID↓	IS↑	Precision↑	Recall↑	FID↓	sFID↓	IS↑	Precision↑	Recall↑
SinPE	17.82	11.33	359.32	0.927	0.149	14.75	10.02	277.92	0.8341	0.143
RoPE	17.30	11.07	**366.19**	0.956	**0.176**	14.17	7.16	313.73	**0.8677**	0.176
SinPE w/ RPE-2D	17.33	11.21	358.43	0.931	0.155	14.53	8.72	302.55	0.8324	0.145
RoPE w/ RPE-2D	**16.92**	**11.02**	362.14	**0.959**	0.170	**14.09**	**6.61**	**398.35**	0.8399	**0.225**

Table 3: Comparison of RPE-2D applied on absolute position encoding (SinPE) and relative position encoding (RoPE).

Method	ImageNet 256 × 256									
	256 × 256					512 × 512				
	FID↓	sFID↓	IS↑	Precision↑	Recall↑	FID↓	sFID↓	IS↑	Precision↑	Recall↑
RPE-2D	17.17	11.19	358.17	0.9579	0.1683	20.78	27.29	293.96	0.8155	0.155
+ Cond-Aug	17.09	11.13	359.55	0.9587	0.1687	19.12	23.91	325.76	0.8399	0.163
+ Attention Scale	17.02	11.05	361.73	0.9578	0.1693	18.33	19.17	347.82	0.8469	0.177
+ Timestep Shift	**16.92**	**11.02**	**362.14**	**0.9591**	**0.1703**	**17.95**	**18.23**	**348.99**	**0.8493**	**0.181**

Table 4: Ablation study on components of RPE-2D. Ablation components are progressively integrated in sequence.

Figure 3: Generated images at different resolutions, including 128×128, 256×256, 512×512, 768×768, and 1024×1024, where the model is trained only at resolutions of 256 × 256 and 512 × 512.

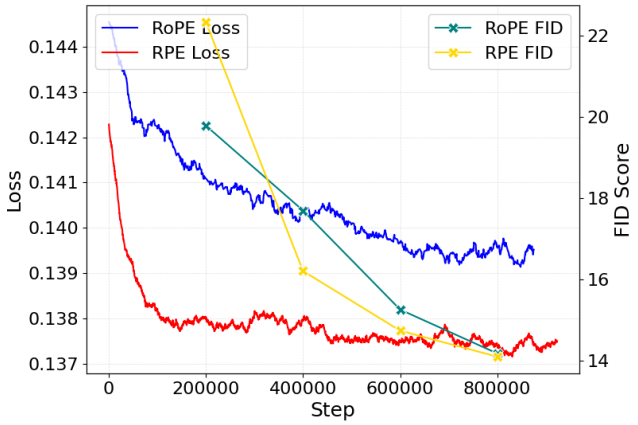

Figure 4: Training loss and FID curves of RoPE and RPE-2D.

ing. We propose RPE-2D, ensuring that the PEs during testing are all trained. By modeling the position orders among image patches rather than their absolute distances, our method bridge the gap between training and testing. Additionally, we propose random data augmentation further enhance the model's ordering modeling while reducing its dependency on the exact number of tokens. To address the potential issue of image incompleteness caused by random data augmentation, we also introduce micro-conditioning, enabling the model to perceive the specific augmentation methods applied. During high-resolution inference, we also employ attention scaling and timestep shifting to address issues related to attention entropy increase and signal-to-noise ratio mismatch. Experimental results on ImageNet-256/512 demonstrate that our proposed method significantly outperforms existing competing approaches in the resolution generalization problem.

References

2024. Ntk-aware Scaled Rope Allows Llama Models to Have Extended (8k+) Context Size Without Any Fine-tuning and Minimal Perplexity Degradation. Accessed: 2024-4-10.

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2024a. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023a. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023b. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Chen, S.; Xu, M.; Ren, J.; Cong, Y.; He, S.; Xie, Y.; Sinha, A.; Luo, P.; Xiang, T.; and Perez-Rua, J.-M. 2024b. GenTron: Diffusion Transformers for Image and Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6441–6451.

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

Du, R.; Chang, D.; Hospedales, T.; Song, Y.-Z.; and Ma, Z. 2024. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6159–6168.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

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.

Hassani, A.; Walton, S.; Li, J.; Li, S.; and Shi, H. 2023. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6185–6194.

He, Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2023. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

Hoogeboom, E.; Heek, J.; and Salimans, T. 2023. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, 13213–13232. PMLR.

Jin, Z.; Shen, X.; Li, B.; and Xue, X. 2023. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36: 70847–70860.

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

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32.

Lu, Z.; Wang, Z.; Huang, D.; Wu, C.; Liu, X.; Ouyang, W.; and Bai, L. 2024. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*.

Ma, N.; Goldstein, M.; Albergo, M. S.; Boffi, N. M.; Vanden-Eijnden, E.; and Xie, S. 2024. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*.

Nash, C.; Menick, J.; Dieleman, S.; and Battaglia, P. W. 2021. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

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

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Press, O.; Smith, N. A.; and Lewis, M. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

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.

Ruoss, A.; Delétang, G.; Genewein, T.; Grau-Moya, J.; Csordás, R.; Bennani, M.; Legg, S.; and Veness, J. 2023. Randomized positional encodings boost length generalization of transformers. *arXiv preprint arXiv:2305.16843*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.;

and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33: 7537–7547.

Teterwak, P.; Sarna, A.; Krishnan, D.; Maschinot, A.; Belanger, D.; Liu, C.; and Freeman, W. T. 2019. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10521–10530.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Xu, J.; Sun, X.; Zhang, Z.; Zhao, G.; and Lin, J. 2019. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32.

Yang, Z.; Dong, J.; Liu, P.; Yang, Y.; and Yan, S. 2019. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10561–10570.

Zhuo, L.; Du, R.; Xiao, H.; Li, Y.; Liu, D.; Huang, R.; Liu, W.; Zhao, L.; Wang, F.-Y.; Ma, Z.; et al. 2024. Lumina-Next: Making Lumina-T2X Stronger and Faster with Next-DiT. *arXiv preprint arXiv:2406.18583*.