PixelU: A U-Shaped Transformer for Efficient End-to-End Pixel Diffusion

Zipeng Guo[1*], Lichen Ma[12*†], Yu He[1*], Xiaolong Fu[1], Jingling Fu[1], Junshi Huang[1‡], and Yan Li[1]

[1]JD.com, [2]Xi'an Jiaotong University
{guozp8888, malichen2020, junshi.huang}@gmail.com

Abstract. End-to-end pixel-space diffusion models bypass the lossy compression of Latent Diffusion Models (LDMs) but struggle to jointly model low-frequency semantics and high-frequency signals in high-dimensional space. Existing works heavily rely on complex pixel decoders to alleviate this issue. In this paper, we challenge this trend by revealing that these decoders primarily compensate for the optimization difficulties inherent to velocity prediction (v-prediction). Under the clean data paradigm (x-prediction), they are redundant. Motivated by this insight, we advocate for simplicity over complexity and introduce PixelU, a minimalist, single-stage U-shaped Diffusion Transformer tailored for pixel space. PixelU abandons auxiliary decoders in favor of zero-cost skip connections, which provide an "information highway" that directly routes uncorrupted high-frequency spatial details from shallow to deep layers. To further enable the backbone to focus exclusively on modeling low-frequency semantics, we introduce a constant-channel spatial down-sampling mechanism as a natural low-pass filter, which compresses deep features into a compact, low-frequency semantic manifold. Extensive experiments demonstrate that this decoupling of frequencies could outperform the strong baseline (JiT-G) with only about 1/3 of its computation cost. On ImageNet 256×256 and 512×512, PixelU achieves FID of 1.63 and 1.92 respectively, surpassing recent pixel-space methods and establishing a simple yet powerful new paradigm for end-to-end diffusion models. Code will be available at https://github.com/gzp6688/PixelU.

Keywords: Image Generation · Diffusion Model · Pixel Diffusion

1 Introduction

The remarkable success of contemporary image generation is largely driven by Latent Diffusion Models (LDMs) [7, 24, 28, 29]. By compressing raw pixels into a latent space via pre-trained VAEs, LDMs effectively filter out high-frequency spatial details to simplify the generation process. However, this reliance on external

* Equal Contribution.
† Project Lead.
‡ Corresponding Author.

VAEs inherently limits fine-grained image fidelity and restricts the architecture to a two-stage pipeline. Consequently, there is a surging interest in end-to-end pixel-space diffusion models, which bypass VAEs to synthesize images directly in the high-dimensional raw pixel space.

Unlike LDMs, pixel-space diffusion [3,6,12,15,38] presents a severe optimization challenge: a single diffusion transformer (DiT) [28] struggles to simultaneously model low-frequency global semantics and high-frequency signals (e.g., high-dimensional noise and fine image details like edges and textures). To alleviate this, a series of recent state-of-the-art works [5, 25, 43, 48] have attempted to preserve high-frequency spatial details by adding architecturally complex pixel decoders. While effective, these specialized modules incur significant computational costs that degrade the overall training and inference efficiency. In stark contrast, recent JiT [18] demonstrate that remarkable generation quality can be achieved on a plain DiT simply by shifting the prediction objective to clean data (x-prediction) rather than velocity (v-prediction).

This leads us to rethink the prevailing architectural trend and raises a critical question: *Are such heavyweight decoders truly necessary for high-frequency modeling in pixel space?* To answer this, we conduct a systematic evaluation of these decoders from the perspective of prediction objectives. Our empirical analysis reveals a crucial insight: *complex pixel decoders primarily compensate for the optimization difficulties inherent to v-prediction.* Under the x-prediction paradigm, which naturally anchors the generative target on a low-dimensional image manifold, these decoders become largely redundant, offering only marginal performance gains at massive computational costs.

Motivated by this insight, we advocate for simplicity over complexity. Rather than relying on complex pixel decoders to recover fine details, we revert to a classic, essentially zero-cost architectural design: **Skip Connections**. Since x-prediction explicitly targets the clean image, skip connections naturally provide the exact "information highway" needed. They directly route uncorrupted high-frequency spatial details from the shallow encoder straight to the decoder, completely replacing the need for complex pixel-wise modeling at the end of the network. Having delegated the preservation of high-frequency details to the skip connections, the backbone network is liberated to focus exclusively on low-frequency global semantics. By visualizing the intermediate predictions $\hat{x}_0$ across the diffusion trajectory, we observe that the generative process intrinsically prioritizes macroscopic low-frequency structures in its early stages. This trajectory demonstrates that before resolving complex high-dimensional details, the core structure of an image is first constructed on a compact, *endogenous semantic manifold* governed by low-frequency signals. To explicitly compel the network to capture this manifold, we introduce **Spatial Down-sampling**. Functioning as a natural, physical low-pass filter, down-sampling systematically attenuates high-frequency noise [40, 45] and compresses entangled features into compact, semantically rich representations, thereby accelerating convergence.

Integrating these insights, we introduce **PixelU**, a minimalist, single-stage U-shaped Diffusion Transformer tailored for pixel space. Unlike traditional U-

Net architectures [9, 29, 30], PixelU performs only a single stage of spatial down- and up-sampling and maintains a constant channel dimension, achieving massive computational savings while yielding outstanding generative quality.

Extensive experiments show that PixelU generates high-quality images when trained end-to-end on pixel space without any autoencoders. On ImageNet 256×256 and 512×512, PixelU achieves FID of 1.63 and 1.92 respectively, outperforming recent pixel-space models by a large margin. Notably, it surpasses the strong JiT-G baseline [18] while requiring merely 1/3 of the computational cost. We would like to highlight that our objective is not to reinvent the U-Net, but to offer a timely reminder to the community. Amidst the recent trend toward flat DiTs [24, 28] and complex pixel decoders [5, 25, 43, 48], the foundational design of skip connections and spatial down-sampling has been largely overlooked. We demonstrate that under the x-prediction objective, these two classic components synergistically decouple high-frequency details from low-frequency semantics, establishing an elegant and compute-friendly new paradigm for pixel diffusion.

2 Related Work

2.1 Latent Space Diffusion

Latent diffusion models (LDMs) compress raw images into a compact latent space via pre-trained Variational Autoencoders (VAEs) and subsequently train the diffusion models within this space [29]. Early LDMs primarily relied on CNN-based U-Net architectures [6, 9, 36, 37], whereas the pioneering work of DiT [28] introduced a flat, isotropic Transformer architecture to replace the U-Net.

However, recent studies have begun to re-evaluate the utility of the classic U-Net macro-architecture in the Transformer era. For instance, U-ViT [1] designs a ViT-based U-shaped architecture that assists diffusion generation by incorporating long skip connections between shallow and deep layers. U-DiT [41] further identifies potential redundancies in purely isotropic DiTs, introducing token downsampling within the self-attention mechanism of a U-shaped diffusion Transformer. To further enhance the accuracy of generated semantics and accelerate convergence, studies such as REPA [47] leverage pre-trained large vision models to align intermediate representations. Addressing the spatial dimension inconsistencies caused by downsampling in U-Net architectures, U-REPA [39] proposes a specialized alignment paradigm, successfully extending REPA to U-Nets. DDT [44] explores single-scale frequency decoupling in a compressed latent space, showing that frequency decoupling remains important even in compressed space. Despite these advancements, the two-stage paradigm is inherently bottlenecked by VAEs. Their lossy compression inevitably introduces decoding artifacts and obliterates high-frequency details, imposing a strict upper bound on generation quality and image fidelity [2, 29].

2.2 Pixel Space Diffusion

To overcome the numerous limitations imposed by VAEs, pixel-space diffusion models advocate for direct, end-to-end modeling in the raw pixel space. Early

diffusion models performed denoising directly on pixels, but constrained by the enormous dimensionality, they faced immense challenges regarding computational burden and optimization difficulty. To reduce the computational overhead of high-resolution image generation, previous research explored multi-scale Cascaded Diffusion [10] or Relay Diffusion [38]. These approaches divide the generation process into multiple resolution stages or utilize multi-scale architectures (e.g., PixelFlow [3]). However, they often incur higher training costs and introduce complex inference scheduling.

Recent works are dedicated to advancing pixel-space generation by improving architectures and prediction targets. In terms of architectural innovation, FractalGen [19] constructs a fractal generative model with long-range structures, TarFlow [49] and FARMER [51] integrate Transformers with normalizing flows to directly model pixels, while PixNerd [43] utilizes lightweight neural field layers for pixel-space diffusion. Regarding improvements in prediction targets, JiT [18] demonstrates that shifting the training objective from velocity prediction to directly predicting the clean image (x-prediction) can effectively simplify the learning process in high-dimensional spaces and anchor the generation on a low-dimensional image manifold.

Furthermore, several recent works [5, 8, 23, 25, 43, 48] have introduced architectural frequency decoupling or pixel-wise refinement modules. Specifically, DeCo [25] leverages an additional pixel decoder to generate high-frequency details conditioned on the main DiT. Similarly, DiP [5] appends a co-trained Patch Detailer Head to the global backbone to restore fine-grained local textures. PixelDiT [48] adopts a dual-level architecture, utilizing a dedicated pixel-level DiT specifically for texture refinement. Unlike these approaches, our proposed PixelU achieves frequency decoupling through architectural simplicity. By integrating the x-prediction paradigm with simple skip connections and spatial downsampling, PixelU entirely eliminates the need for heavyweight decoders. This design significantly reduces computational overhead while maintaining superior generation quality, offering an efficient paradigm for pixel-space diffusion.

3 Methodology

3.1 Preliminaries: Pixel Diffusion with x-prediction

Flow Matching (FM) [21] learns a velocity field that maps a prior distribution to the data distribution. Let x be a clean image sampled from the data distribution and $\epsilon \sim \mathcal{N}(0, I)$ be standard Gaussian noise. We can define a noisy state z_t at a continuous timestep $t \in [0, 1]$ using a linear interpolation schedule [7, 22]:

$$z_t = tx + (1 - t)\epsilon. \tag{1}$$

A neural network net_θ can be trained to predict various targets, such as the noise (ϵ), velocity (v), or clean image (x). JiT [18] reveals that directly predicting the clean image (x-prediction) offers superior performance for high-dimensional pixel-space generation compared to ϵ- or v-prediction.

While the model learns to reconstruct x, the training objective is still grounded in the flow matching paradigm. The true velocity is given by the time derivative of the interpolation, yielding $v = x - \epsilon$. To compute the flow matching loss, the network's output x_θ is converted to a predicted velocity v_θ:

$$v_\theta = \frac{x_\theta - z_t}{1 - t}, \tag{2}$$

The optimization objective minimizes the mean squared error between these velocities, which implicitly acts as a dynamically reweighted x-prediction loss:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,x,\epsilon} \left\| v_\theta - v \right\|^2 = \mathbb{E}_{t,x,\epsilon} \left\| \frac{x_\theta - x}{1 - t} \right\|^2. \tag{3}$$

This design effectively benefits from both the optimization stability of x-prediction and the favorable sampling advantages of flow matching.

3.2 Rethinking Decoders in Pixel Diffusion

Concurrent modeling of low-frequency semantics and high-frequency details in the high-dimensional pixel space poses a fundamental challenge for standard DiT backbones. To tackle this challenge, recent works such as DeCo [25] and PixelDiT [48] rely on pixel decoders to retain high-frequency signals. Unfortunately, these auxiliary modules introduce substantial computational burdens, significantly compromising both training and inference efficiency. In contrast, JiT [18] demonstrates that remarkable efficacy can be achieved on a plain DiT simply by shifting the prediction objective. This compelling contrast leads us to rethink: *Are such complex pixel decoders truly necessary?*

Table 1: Performance comparison of different decoders and prediction targets on ImageNet 256×256. * denotes our reproduced results as the official code is not publicly available. Compared to complex decoders, simple skip connections achieve the best FID under x-prediction without significantly increasing GFLOPs.

ImageNet 256×256, 160 epochs, cfg=1			
Model	**GFLOPs**	**FID**	
		x-pred	**v-pred**
Baseline (JiT-B/16)	21.99	43.65	190.11
+ DDT decoder [44]	25.35	43.48	189.50
+ DeCo decoder [25]	25.91	41.20	**42.07**
+ PixelDiT decoder* [48]	33.78	39.20	44.85
+ Skip Connections (Ours)	23.20	**38.53**	185.54

To investigate the actual gains of these pixel decoders under different prediction targets, we conduct a series of controlled experiments on the ImageNet 256×256 dataset. Building upon the recent effective pixel diffusion baseline

JiT [18], we evaluate the impact of various decoder architectures on generation quality (FID and IS) under both velocity prediction (v-prediction) and data prediction (x-prediction) objectives. Observing the empirical results in Tab. 1, we draw two key conclusions:

- **Pixel decoders primarily serve v-prediction:** The vanilla JiT performs exceptionally poorly under v-prediction (FID 190.11). However, the introduction of DeCo or PixelDiT decoders yields a dramatic performance leap (dropping to 42.07 and 44.85, respectively). This indicates that such heavy decoders mainly serve to alleviate the optimization difficulties caused by v-prediction in the presence of high-frequency noise.
- **Pixel decoders offer limited improvements for x-prediction:** Conversely, when employing the x-prediction paradigm, which anchors the generative target on clean images, the baseline model already demonstrates stable performance (FID 43.65). In this context, introducing DeCo or PixelDiT decoders at the cost of substantial GFLOPs brings only marginal performance gains (dropping to merely 41.20 and 39.20).

3.3 Simplicity over Complexity: Skip Connections

Under the x-prediction paradigm, the network directly targets the clean image rather than the complex high-dimensional noise of v-prediction, making heavyweight pixel decoders essentially redundant. However, a critical challenge remains: the inherent high-frequency signals of clean images (e.g., fine textures and sharp edges) are still prone to being washed out during deep feature extraction. Other studies [4, 27, 34] also show that transformers tend to capture low-frequency semantics well but struggle with high-frequency signals. Based on these findings, we raise a fundamental question: *under the x-prediction paradigm, can we bypass heavyweight decoders and utilize computationally efficient operations to preserve and recover these high-frequency details?* To this end, we discard all auxiliary decoder modules and adopt a classic, computationally lightweight design—**Skip Connections**. As shown in the last row of Tab. 1, merely introducing skip connections between the encoder and decoder allows the model to achieve an FID of 38.53 under x-prediction, outperforming all variants equipped with complex decoders while introducing negligible computational overhead. Since x-prediction explicitly requires the network to output clean pixels, skip connections provide an "information highway" that directly routes pristine high-frequency spatial details from shallow layers to deep layers, perfectly compensating for the spatial information loss in deep features. Notably, this simple design completely fails under v-prediction (FID 185.54), further corroborating that the synergy between x-prediction and skip connections is the key to our breakthrough in pixel diffusion.

3.4 Down-Sampling for Low-Frequency Semantics

By delegating high-frequency details to skip connections, the backbone network is liberated to focus exclusively on modeling low-frequency global semantics. To

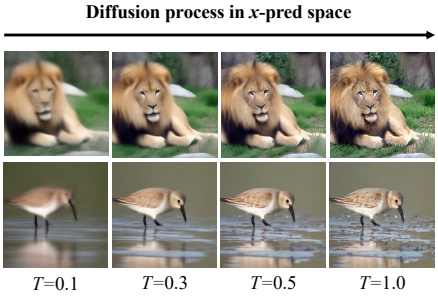


Fig. 1: Visualization of the diffusion process in x-prediction space via intermediate $\hat{x}_0$ at various timesteps T. Diffusion model establishes macroscopic low-frequency structures (e.g., basic shapes and colors) in the early stages, before progressively refining high-frequency microscopic details.

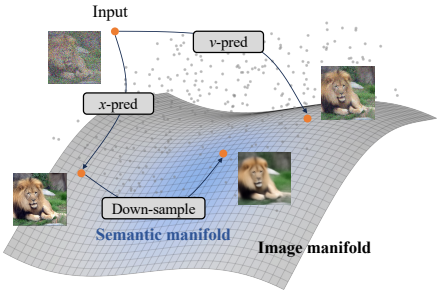


Fig. 2: Unlike v-prediction which points toward the high-dimensional noise space, x-prediction maps the noisy input directly onto the low-dimensional image manifold. Spatial down-sampling further explicitly compresses this representation further into an endogenous, low-frequency semantic manifold (highlighted in blue).

further reveal the generative process of diffusion models, we visualize the intermediate $\hat{x}_0$ at various timesteps across the x-prediction trajectory. As shown in Fig. 1, the generative process intrinsically prioritizes macroscopic, low-frequency structures during its early stages (e.g., $T = 0.1$), followed by a progressive refinement of microscopic high-frequency details. The overall layout and semantic composition of the image are largely established during these initial phases.

While previous work [18] reveals that natural images inherently reside on a low-dimensional manifold, our generative trajectory analysis implies the existence of a fundamental sub-space within it. As illustrated in Fig. 2, this broader image manifold contains a highly compact, endogenous semantic sub-manifold (highlighted in blue) that is predominantly governed by low-frequency signals.

To explicitly compel the backbone network to capture this endogenous semantic manifold, we incorporate **Spatial Down-sampling** modules within the network. Inherently functioning as a natural low-pass filter, down-sampling systematically attenuates high-frequency noise and compresses the entangled features. As depicted by the mapping trajectory in Fig. 2, this operation effectively anchors the representation from the broader image manifold directly into the compact, semantically rich sub-manifold. This aligns with existing findings [45] that down-sampling aids diffusion denoisers by automatically discarding higher-frequency subspaces dominated by noise.

3.5 PixelU Architecture

Based on the preceding insight, we introduce **PixelU**, a minimalist U-shaped Diffusion Transformer designed specifically for pixel-space generation by integrating skip connections and spatial down-sampling. As illustrated in Fig.

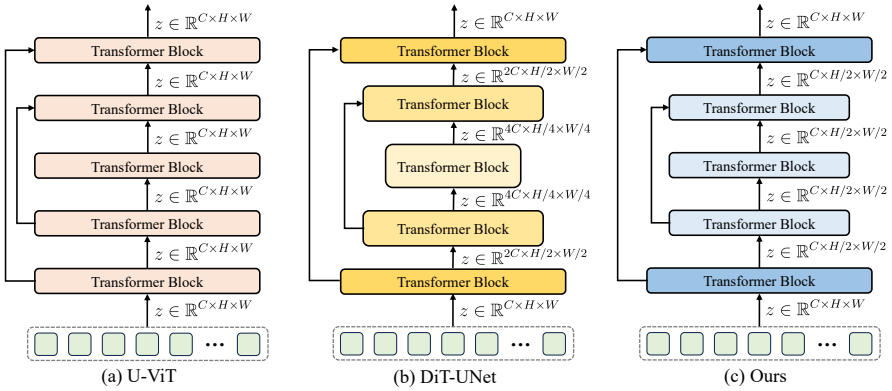


Fig. 3: Architectural comparison of Diffusion Transformers. (a) U-ViT maintains a flat, full-resolution sequence ($H \times W$). (b) DiT-UNet uses multi-stage down-sampling with expanding channel dimensions (e.g., $2C, 4C$). (c) PixelU employs a minimalist design with a single stage of down-sampling ($H/2 \times W/2$) and strictly maintains a constant channel dimension throughout the network to minimize computational overhead.

3, while PixelU shares a macroscopic U-shaped topology with previous models [6, 9, 36, 37], its internal dimensional routing diverges significantly. First, unlike U-ViT [1] (Fig. 3a), which maintains a computationally expensive flat, full-resolution sequence ($H \times W$) and lacks an explicit low-frequency bottleneck, PixelU incorporates spatial down-sampling. This operation explicitly filters high-frequency noise and compresses the spatial resolution ($H/2 \times W/2$), forcing the deep layers to focus on the endogenous semantic manifold. Second, rather than adopting the traditional U-Net paradigm [12, 41] (Fig. 3b) that employs multi-stage down-sampling coupled with continuous channel expansion (e.g., $2C$, $4C$), PixelU employs a drastically simplified design. It performs only a *single stage* of down/up-sampling and strictly maintains a *constant channel dimension* (C) across all intermediate layers. By avoiding multi-stage complexities and channel expansion, this streamlined design significantly mitigates the computational burden, yielding substantial reductions in GFLOPs. Ultimately, this architecture establishes a minimalist and efficient paradigm for end-to-end pixel diffusion.

4 Experiments

4.1 Experiment Setup

Model Settings. To empirically validate the efficacy and scalability of PixelU, we conduct extensive experiments on ImageNet using three different model scales: Base (B), Large (L), and Huge (H). The detailed configurations for each variant are detailed in Tab. 2.

Implementation Details. For class-to-image generation, our implementation follows the training setup of [18] on the ImageNet-1K dataset [31] at 256×256

Fig. 4: Selected 256 × 256 and 512 × 512 resolution samples from PixelU-H. We use a classifier-free guidance scale cfg = 4.0.

Table 2: Configurations of PixelU architecture at different model sizes.

Model	Params	Channel	Heads	Depth
PixelU-B/16	163.7M	768	12	[4,4,4]
PixelU-B/32	165.8M	768	12	[4,4,4]
PixelU-L/16	522.8M	1024	16	[8,8,8]
PixelU-L/32	525.5M	1024	16	[8,8,8]
PixelU-H/16	1168.3M	1152	16	[8,20,8]
PixelU-H/32	1171.8M	1152	16	[8,20,8]

Table 3: Comparing PixelU against JiT on ImageNet 256×256 with CFG.

Model	GFLOPs	FID↓	IS↑
JiT-B/16	25	3.66	275.1
PixelU-B/16	21.8	3.40	276.01
JiT-L/16	88	2.36	298.5
PixelU-L/16	73.1	2.17	295.88
JiT-H/16	182	1.86	303.4
JiT-G/16	383	1.82	292.6
PixelU-H/16	136.9	1.63	305.88

and 512×512 resolutions. We use a global batch size of 1024 and the AdamW optimizer with $\beta = (0.9, 0.95)$ and a constant learning rate of 2×10^{-4} on 8×B200 GPUs. Throughout the diffusion training, we employ logit-normal sampling, where $\text{logit}(t) \sim \mathcal{N}(-0.8, 0.8^2)$, and Exponential Moving Average (EMA) with a decay of 0.9999. We employ REPA [35, 47] with a loss weight of $\lambda_{\text{repa}} = 0.01$ and apply the alignment at the middle block of the network. For evaluation, we report FID, sFID, Inception Score, and Precision–Recall on 50K samples following ADM [6]. We use 50 Heun sampling steps with Classifier-Free Guidance (CFG) [11] and the CFG interval [16].

4.2 Class-to-Image Generation

ImageNet 256×256. As summarized in Tab. 4 and Tab. 3, our PixelU-H/16 achieves an exceptional FID of 1.63 and an IS of 305.88 after 600 training epochs, surpassing recent state-of-the-art pixel-space models. Notably, PixelU consistently outperforms the JiT [18] baseline across various model sizes while

Table 4: Quantitative results on ImageNet 256x256 and 512x512 with CFG. PixelU achieves superior performance in end-to-end pixel diffusion and is competitive with two-stage latent diffusion models.

		Params	Epochs	NFE	GFLOPs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
256×256	*Latent-space Diffusion*									
	LDM-4-G [29]	400M	170	-	-	3.60	-	247.6	0.87	0.48
	DiT-XL/2 [28]	675M + 86M	1400	250×2	119	2.27	4.60	278.2	0.83	0.57
	SiT-XL/2 [24]	675M + 86M	1400	250×2	119	2.06	4.50	284.0	0.83	0.59
	MaskDiT [52]	675M	1600	-	-	2.28	5.67	276.5	0.80	0.61
	REPA-XL/2 [47]	675M + 86M	800	250×2	119	1.42	4.70	305.7	0.80	0.64
	LightningDiT [46]	675M	800	-	119	1.35	4.15	295.3	0.79	0.65
	SVG-XL [33]	675M	1400	-	-	1.92	-	264.9	-	-
	DDT-XL [44]	675M	400	-	119	1.26	-	310.6	0.79	0.65
	RAE-XL [50]	839M	800	-	146	1.13	-	262.6	0.78	0.67
	Pixel-space Diffusion									
	StyleGAN-XL [32]	-	-	-	-	2.30	4.02	265.1	0.78	0.53
	ADM-G [6]	554M	400	250	1120	4.59	5.25	186.7	**0.82**	0.52
	RDM [38]	553M + 553M	400	-	-	1.99	**3.99**	260.4	0.81	0.58
	CDM [10]	-	2160	-	-	4.88	-	158.7	-	-
	RIN [14]	410M	480	-	334	3.42	-	182.0	-	-
	VDM++ [15]	2B	-	250×2	555	2.12	-	267.7	-	-
	JetFormer [42]	2.8B	-	-	-	6.64	-	-	0.69	0.56
	Simple Diffusion [12]	2B	-	250×2	555	2.44	-	256.3	-	-
	FractalMAR-H [19]	848M	600	-	-	6.15	-	**348.9**	0.81	0.46
	FARMER [51]	1.9B	320	-	-	3.60	-	269.2	0.81	0.51
	EPG [17]	583M	800	-	-	2.04	-	283.2	0.80	0.56
	PixelFlow-XL/4 [3]	677M	320	120×2	2909	1.98	5.83	282.1	0.81	0.60
	PixNerd-XL/16 [43]	700M	320	100×2	134	1.95	4.54	300	0.80	0.60
	JiT-G [18]	2B	600	100×2	383	1.82	-	292.6	0.79	0.62
	DeCo-XL/16 [25]	682M	600	100×2	-	1.69	4.59	304	0.79	0.63
	PixelGen-XL/16 [26]	676M	160	100×2	-	1.83	4.59	293.6	0.79	0.63
	PixelU-H/16	1168M	320	100×2	137	1.77	5.29	306.86	0.78	0.63
	PixelU-H/16	1168M	600	100×2	137	**1.63**	5.04	305.88	0.79	**0.64**
512×512	*Latent-space Diffusion*									
	DiT-XL/2 [28]	675M + 86M	600	250×2	525	3.04	5.02	240.8	0.84	0.54
	SiT-XL/2 [24]	675M + 86M	600	250×2	525	2.62	4.18	252.2	0.84	0.57
	Pixel-space Diffusion									
	ADM-G [6]	554M	400	250	1983	7.72	6.57	172.7	**0.87**	0.53
	RIN [14]	320M	-	250	415	3.95	-	216.0	-	-
	SimpleDiffusion [12]	2B	800	250×2	555	3.54	-	205.0	-	-
	VDM++ [15]	2B	800	250×2	555	2.65	-	278.1	-	-
	PixNerd-XL/16 [43]	700M	340	100×2	583	2.84	5.95	245.6	0.80	0.59
	JiT-H/32 [18]	682M	600	100×2	183	1.94	-	309.1	-	-
	DeCo-XL/16 [25]	682M	340	100×2	-	2.22	**4.67**	290.0	0.80	**0.60**
	PixelU-H/32	1152M	600	100×2	138	**1.92**	5.98	**322.10**	0.80	0.58

maintaining a substantially lower computational cost. This efficiency advantage scales remarkably to larger models: our PixelU-H/16 drastically outperforms the massive JiT-G (FID 1.82) while utilizing significantly fewer parameters (1.15B vs. 2B) and demanding only about 1/3 of the computational cost (137 GFLOPs vs. 383 GFLOPs). It also surpasses competitive architectures like DeCo-XL/16 [25] (FID 1.69) and completely eclipses earlier models such as ADM [6] and PixelFlow-XL/4 [3]. Furthermore, PixelU demonstrates remarkable convergence efficiency: at just 320 training epochs, it already reaches an FID of 1.76, decisively beating PixNerd-XL/16 [43] (1.95) and PixelFlow-XL/4 (1.98) trained for similar durations. We note that while the concurrent PixelDiT [48] and SiD2 [13] reports a FID of 1.61 and 1.38, its implementation and training pipelines remain closed-source. For fairness, we include its FID for reference

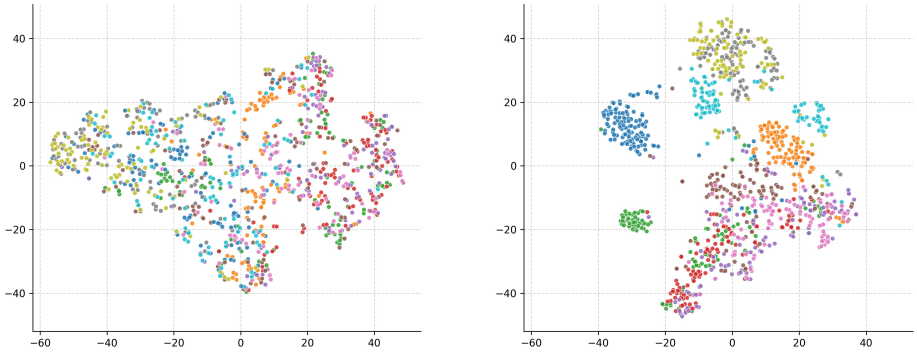
but exclude it from Tab. 4. When compared to the latent-space paradigm, our end-to-end PixelU outperforms widely adopted baselines such as DiT-XL/2 [28] (FID 2.27) and SiT-XL/2 [24] (FID 2.06) using less than half the training epochs (600 vs. 1400 epochs). While advanced latent models equipped with pre-trained autoencoders (e.g., DDT [44], RAE [50]) still hold slightly lower FIDs, PixelU significantly narrows this gap without relying on lossy VAE compressions, proving the immense potential of pure pixel-space modeling.

ImageNet 512×512. Scaling end-to-end pixel diffusion to higher resolutions presents significant challenges due to the quadratic increase in spatial complexity. Despite this, at the 512×512 resolution, our PixelU-H/16 achieves a leading FID of 1.92 and an IS of 322 after 600 epochs. PixelU consistently outperforms existing pixel-based baselines, including DeCo-XL/16 [25] (FID 2.22) and PixNerd-XL/16 [43] (FID 2.84), as well as larger 2B-parameter models like VDM++ [15] (FID 2.65) and SimpleDiffusion [12] (FID 3.54). Furthermore, PixelU demonstrates clear advantages over representative two-stage latent diffusion models. Compared to DiT-XL/2 [28] (FID 3.04) and SiT-XL/2 [24] (FID 2.62) trained for 600 epochs, PixelU delivers significantly better image fidelity.

4.3 Analysis

Feature visualization with t-SNE. To further validate the efficacy of our architectural design in capturing low-frequency semantics, we visualize the intermediate feature spaces of the baseline (Flat JiT) and our PixelU using t-SNE. As illustrated in Fig. 5(a), the feature representations of the flat baseline are scattered and heavily entangled across different classes. This occurs because the full-resolution architecture lacks a structural bottleneck, leaving its feature manifold highly susceptible to interference from high-frequency textures and irrelevant pixel-level noise. Conversely, Fig. 5(b) demonstrates that PixelU generates remarkably compact and well-separated class clusters. This striking contrast visually confirms our hypothesis: the spatial down-sampling mechanism successfully acts as an explicit low-pass filter. By systematically shedding redundant, class-irrelevant high-frequency noise, it compels the bottleneck layers to learn a highly pure and discriminative global semantic representation.

Frequency Energy Distribution. To further validate the superiority of our frequency decoupling mechanism, we present a comparative analysis of the frequency energy distribution between the baseline (flat JiT) and our proposed PixelU (as shown in Figure 6). Specifically, we apply a 2D Fast Fourier Transform (2D-FFT) to the spatial feature maps of different blocks to compute their channel-averaged magnitude spectra. The flat architecture of the baseline JiT suffers from persistent frequency entanglement, where high-frequency noise and low-frequency semantics remain tightly coupled across all layers. Lacking an explicit spatial bottleneck, this isotropic design forces the backbone to model both simultaneously, significantly exacerbating optimization difficulty in pixel space. Conversely, PixelU demonstrates an explicit and highly structured frequency dynamic. While the shallow encoder captures abundant high-frequency spatial details, the middle bottleneck exhibits a precipitous drop in high-frequency energy.



(a) Baseline (Flat JiT) feature space (b) PixelU (Ours) feature space

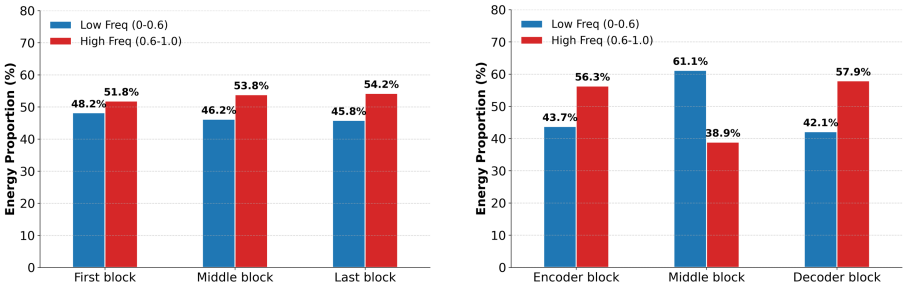
Fig. 5: Feature visualization with t-SNE for 10 ImageNet classes (100 random samples per class), with each class shown in a distinct color. These features are extracted from the intermediate blocks of both the baseline (JiT) and PixelU.

This confirms that our spatial down-sampling acts as an endogenous low-pass filter, effectively isolating a pure, low-dimensional semantic manifold. Subsequently, high-frequency energy sharply rebounds in the decoder block, proving that skip connections function as an information highway, seamlessly routing pristine high-frequency details to the output to compensate for the bottleneck's compression. Collectively, these quantitative observations substantiate that PixelU achieves elegant frequency decoupling through the highly efficient synergy of spatial down-sampling and skip connections.

4.4 Ablation Study

Contribution of Core Components. Table 5 presents an ablation study on the core components of our PixelU architecture: spatial down-sampling and skip connections. Starting from the full-resolution baseline (FID 43.65, 21.99 GFLOPs), integrating only skip connections improves generation quality (FID 38.53) by directly propagating spatial details across layers. Conversely, applying spatial down-sampling in isolation degrades the generative performance (FID 48.87) despite reducing the computational cost to 18.25 GFLOPs. This degradation occurs because the down-sampling bottleneck inevitably discards high-frequency details. However, combining both components achieves the optimal performance (FID 34.92) while maintaining a lower computational overhead (19.46 GFLOPs) than the baseline. This demonstrates their crucial synergy: down-sampling efficiently extracts low-frequency semantics and reduces computational costs, while skip connections effectively compensate by recovering the filtered high-frequency details. Together, they form a highly efficient paradigm for end-to-end pixel-space generation.

Block Allocation Strategy. Tab. 6 shows that a balanced $[4, 4, 4]$ configuration achieves the optimal FID of 34.92. Allocating insufficient blocks to the middle



(a) Baseline (Flat JiT) energy distribution

(b) PixelU (Ours) energy distribution

Fig. 6: Frequency energy distribution across different blocks. We normalize the radial frequency to $r \in [0, 1]$ and calculate the energy proportions within the low-frequency ($r \leq 0.6$) and high-frequency ($r > 0.6$) bands.

Table 5: Ablation study on core components. The combination of skip connections and spatial down-sampling yields the best generative performance while effectively reducing the computational cost compared to the baseline.

ImageNet 256×256, 160 epochs, cfg=1					
GFLOPs	**Params**	**Metrics**		**Skip Connections**	**Downsampling**
		FID↓	**IS ↑**		
21.99	131.1M	43.65	34.08	✗	✗
18.25	153.6M	48.87	30.29	✗	✓
23.20	135.8M	38.53	38.57	✓	✗
19.46	158.3M	**34.92**	**43.51**	✓	✓

bottleneck (e.g., $[5, 2, 5]$) restricts the network's capacity to explore the low-frequency semantic manifold. Conversely, reducing the encoder/decoder blocks (e.g., $[3, 6, 3]$) compromises high-frequency detail extraction and fusion via skip connections. Thus, a balanced allocation best harmonizes high-frequency and low-frequency modeling.

Number of Down-sampling Stages. Tab. 7 validates our single-stage down-sampling design. A flat architecture (0 stages) struggles with frequency entanglement, yielding a suboptimal FID of 38.53. Conversely, employing 2 down-sampling stages degrades the FID to 42.70. Because PixelU strictly maintains a constant channel dimension, multiple spatial compressions create an overly restrictive bottleneck that irreversibly destroys crucial representations. Consequently, a single down-sampling stage optimally decouples frequencies without catastrophic information loss.

General-purpose Improvements. To demonstrate the compatibility of our architecture with standard diffusion enhancements, we progressively integrate several established techniques into the vanilla PixelU baseline. As shown in Tab. 8, incorporating in-context class tokens [18, 20] and REPA [35, 47] yields steady,

Table 6: Ablation study of the block allocation strategy.

ImageNet 256×256, 160 epochs, cfg=1			
Depth	**GFLOPs**	**FID↓**	**IS↑**
[5, 2, 5]	22.48	35.47	41.56
[4, 4, 4]	19.46	**34.92**	**43.51**
[3, 6, 3]	16.44	38.08	40.25

Table 7: Ablation study of the number of down-sampling (DS) stages.

ImageNet 256×256, 160 epochs, cfg=1			
Depth	**DS Stages**	**FID↓**	**IS↑**
12	0	38.53	38.57
[4, 4, 4]	1	**34.92**	**43.51**
[2, 2, 4, 2, 2]	2	42.70	36.40

Table 8: Ablation study of general-purpose improvements. Progressive integration of established techniques into the vanilla PixelU baseline consistently enhances generation quality across different model scales.

Model Components	**Epoch**	**PixelU-B/16**			**PixelU-H/16**		
		GFLOPs	**FID↓**	**IS↑**	**GFLOPs**	**FID↓**	**IS↑**
PixelU (vanilla)	160	19.46	34.92	43.51	117.86	12.41	95.37
+ In-context class tokens	160	21.42	30.88	51.84	136.32	8.93	124.06
+ REPA	160	21.76	26.86	58.46	136.88	6.95	142.78
+ CFG and CFG interval	160	21.76	4.39	230.41	136.88	2.42	290.26
	600	21.76	**3.40**	**276.01**	136.88	**1.63**	**305.88**

incremental gains in both FID and IS across different model scales. Subsequently, the application of Classifier-Free Guidance (CFG) [11] alongside the CFG interval [16] provides a substantial leap in generative quality, drastically reducing the FID of PixelU-H/16 from 6.95 to 2.42 at 160 epochs. Finally, extending the training schedule to 600 epochs allows the network to fully converge, achieving our optimal performance of 1.63 FID for the PixelU-H/16 model.

5 Conclusion

In this work, we present PixelU, a minimalist, single-stage U-shaped Diffusion Transformer tailored for efficient pixel-space generation. We demonstrate that under the x-prediction paradigm, complex pixel decoders used in prior works become largely redundant. Instead, PixelU advocates for simplicity by leveraging zero-cost skip connections to directly route uncorrupted high-frequency spatial details from shallow to deep layers. Concurrently, a constant-channel spatial down-sampling mechanism acts as a natural low-pass filter, allowing the backbone to focus exclusively on a compact, low-frequency semantic manifold. Extensive experiments show that this elegant frequency decoupling enables PixelU to surpass recent end-to-end pixel-space models at a significantly lower computational cost. While latent diffusion models currently still retain a slight edge, PixelU significantly narrows this gap without the lossy decoding artifacts inherent in autoencoders. Ultimately, our findings highlight that fundamental architectural designs can also establish a powerful and highly efficient new paradigm for end-to-end pixel diffusion.

References

1. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: 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)
2. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Lu, Y., Han, S.: Deep compression autoencoder for efficient high-resolution diffusion models. In: International Conference on Learning Representations (ICLR) (2025)
3. Chen, S., Ge, C., Zhang, S., Sun, P., Luo, P.: Pixelflow: Pixel-space generative models with flow. arXiv preprint arXiv:2504.07963 (2025)
4. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: International Conference on Learning Representations (ICLR) (2023)
5. Chen, Z., Zhu, J., Chen, X., Zhang, J., Hu, X., Zhao, H., Wang, C., Yang, J., Tai, Y.: Dip: Taming diffusion models in pixel space. arXiv preprint arXiv:2511.18822 (2025)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 8780–8794 (2021)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
8. He, Y., Ma, L., Guo, Z., Shan, X., Fu, J., Chen, D., Huang, J., Li, Y.: Hyperdit: Hyper-connected transformers for high-fidelity pixel-space diffusion. arXiv preprint arXiv:2605.15741 (2026)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Tim, S.: Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research (JMLR) **23**, 1–33 (2022)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. In: International Conference on Machine Learning (ICML). pp. 13213–13232. PMLR (2023)
13. Hoogeboom, E., Mensink, T., Heek, J., Lamerigts, K., Gao, R., Salimans, T.: Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18062–18071 (2025)
14. Jabri, A., Fleet, D., Chen, T.: Scalable adaptive computation for iterative generation. arXiv preprint arXiv:2212.11972 (2022)
15. Kingma, D., Gao, R.: Understanding diffusion objectives as the elbo with simple data augmentation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 65484–65516 (2023)
16. Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., Lehtinen, J.: Applying guidance in a limited interval improves sample and distribution quality in diffusion models. Advances in Neural Information Processing Systems **37**, 122458–122483 (2024)
17. Lei, J., Liu, K., Berner, J., HoiM, Y., Zheng, H., Wu, J., Chu, X.: Advancing end-to-end pixel-space generative modeling via self-supervised pre-training. In: International Conference on Learning Representations (ICLR) (2025)

18. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
19. Li, T., Sun, Q., Fan, L., He, K.: Fractal generative models. arXiv preprint arXiv:2502.17437 (2025)
20. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
22. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
23. Ma, L., Guo, Z., He, Y., Fu, X., Liu, L., Fu, J., Huang, J., Li, Y.: Frequency-booster: Full-frequency modeling for high-fidelity pixel diffusion. arXiv preprint arXiv:2605.17759 (2026)
24. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. arXiv preprint arXiv:2401.08740 (2024)
25. Ma, Z., Wei, L., Wang, S., Zhang, S., Tian, Q.: Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. arXiv preprint arXiv:2511.19365 (2025)
26. Ma, Z., Xu, R., Zhang, S.: Pixelgen: Pixel diffusion beats latent diffusion with perceptual loss. arXiv preprint arXiv:2602.02493 (2026)
27. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (ICLR) (2022)
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (2023)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
31. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
32. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
33. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. arXiv preprint arXiv:2510.15301 (2025)
34. Si, C., Yu, W., Zhou, P., Zhou, Y., Wang, X., Yan, S.: Inception transformer. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 23495–23509 (2022)
35. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? arXiv preprint arXiv:2512.10794 (2025)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2021)

38. Teng, J., Zheng, W., Ding, M., Hong, W., Wangni, J., Yang, Z., Tang, J.: Relay diffusion: Unifying diffusion process across resolutions for image synthesis. arXiv preprint arXiv:2309.03350 (2023)
39. Tian, Y., Chen, H., Zheng, M., Liang, Y., Xu, C., Wang, Y.: U-repa: Aligning diffusion u-nets to vits. arXiv preprint arXiv:2503.18414 (2025)
40. Tian, Y., Tu, Z., Chen, H., Hu, J., Xu, C., Wang, Y.: U-dits: Downsample tokens in u-shaped diffusion transformers. Advances in Neural Information Processing Systems **37**, 51994–52013 (2024)
41. Tian, Y., Tu, Z., Chen, H., Hu, J., Xu, C., Wang, Y.: U-dits: Downsample tokens in u-shaped diffusion transformers. Advances in Neural Information Processing Systems **37**, 51994–52013 (2024)
42. Tschannen, M., Pinto, A.S., Kolesnikov, A.: Jetformer: An autoregressive generative model of raw images and text. arXiv preprint arXiv:2411.19722 (2024)
43. Wang, S., Gao, Z., Zhu, C., Huang, W., Wang, L.: Pixnerd: Pixel neural field diffusion. arXiv preprint arXiv:2507.23268 (2025)
44. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. arXiv preprint arXiv:2504.05741 (2025)
45. Williams, C., Falck, F., Deligiannidis, G., Holmes, C.C., Doucet, A., Syed, S.: A unified framework for u-net design and analysis. Advances in Neural Information Processing Systems **36**, 27745–27782 (2023)
46. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025)
47. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: International Conference on Learning Representations (ICLR) (2025)
48. Yu, Y., Xiong, W., Nie, W., Sheng, Y., Liu, S., Luo, J.: Pixeldit: Pixel diffusion transformers for image generation. arXiv preprint arXiv:2511.20645 (2025)
49. Zhai, S., Zhang, R., Nakkiran, P., Berthelot, D., Gu, J., Zheng, H., Chen, T., Bautista, M.A., Jaitly, N., Susskind, J.: Normalizing flows are capable generative models. arXiv preprint arXiv:2412.06329 (2024)
50. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
51. Zheng, G., Zhao, Q., Yang, T., Xiao, F., Lin, Z., Wu, J., Deng, J., Zhang, Y., Zhu, R.: Farmer: Flow autoregressive transformer over pixels. arXiv preprint arXiv:2510.23588 (2025)
52. Zheng, H., Nie, W., Vahdat, A., Anandkumar, A.: Fast training of diffusion models with masked transformers. arXiv preprint arXiv:2306.09305 (2023)