SCALE AS A MODALITY: MULTI-SCALE PIXEL-SPACE DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent diffusion models rely on pretrained autoencoders that introduce lossy compression, decouple representation learning from generative modeling, and constrain the spatial granularity at which images are represented. Pixel-space diffusion removes this bottleneck and, importantly, enables flexible tokenization of the same image at multiple spatial scales. Existing pixel-space diffusion transformers, however, largely adopt two-level coarse-to-fine architectures with limited cross-scale interaction. We introduce MS-pDiT, a multi-scale pixel-space diffusion transformer that treats spatial scale as a modality. MS-pDiT tokenizes the same noisy image at multiple patch sizes, assigns each scale a dedicated transformer stream, and couples these streams through compacted MM-DiT-style joint attention. This design enables bidirectional information exchange across scales while keeping the attention cost independent of the finest patch size. It further establishes the number of scales as a new architectural scaling dimension. On ImageNet 256×256, increasing the number of scales consistently improves FID from 4.39 with two scales to 2.31, 2.26, and 2.19 with three, four, and five scales, respectively; under a comparable parameter budget, introducing an intermediate scale substantially outperforms additional depth. Our analysis further shows that bidirectional cross-scale interaction and late introduction of fine-grained streams are important for effective multi-scale modeling. MS-pDiT achieves FID scores of 1.43 and 1.44 on ImageNet at 256×256 and 512×512, respectively, and naturally extends to text-to-image generation by incorporating text as an additional modality, attaining 0.86 on GenEval and 83.9 on DPG-Bench.

1 INTRODUCTION

Latent diffusion models have become the dominant paradigm for image generation (Rombach et al., 2022; Peebles & Xie, 2023; Esser et al., 2024), with pretrained autoencoders compressing images into compact latent representations before generative modeling. This efficiency, however, introduces several inherent limitations. The autoencoder incurs lossy compression, separates representation learning from generative training, and fixes the spatial granularity at which the diffusion model operates (Yao et al., 2025; Zheng et al., 2026; Yu et al., 2026). Pixel-space diffusion eliminates this bottleneck by modeling RGB images directly (Li & He, 2026; Yu et al., 2026; Ma et al., 2026a; Chen et al., 2025b). Beyond avoiding reconstruction error and enabling end-to-end generative modeling, pixel space offers an important architectural advantage: *flexible tokenization*. Unlike latent representations whose spatial granularity is determined by a pretrained autoencoder, raw pixels can be partitioned at arbitrary patch sizes. The same image can therefore be represented simultaneously at multiple granularities, allowing coarse tokens to capture global structure while finer tokens preserve local appearance and detail.

Existing pixel-space diffusion transformers (Yu et al., 2026; Ma et al., 2026a; Chen et al., 2026; Wang et al., 2026a; He et al., 2026) exploit this flexibility only partially, typically employing only two granularities with a coarse pathway capturing global semantics, while a fine pathway reconstructs or refines local details. Moreover, information is primarily propagated in a coarse-to-fine direction, leaving the fine representation unable to influence the evolving coarse representation.

We address these questions with MS-pDiT, a multi-scale pixel-space diffusion transformer that treats *scale as a modality*. Given the same noisy image, MS-pDiT independently tokenizes it at multiple

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

patch sizes and assigns each granularity a dedicated transformer stream. To enable efficient interaction among streams with vastly different token counts, we compact their representations onto a shared spatial grid and couple them through a single MM-DiT-style joint attention operation (Esser et al., 2024). This formulation enables unrestricted bidirectional information exchange across scales while keeping the attention cost independent of the finest patch size. Fine-grained streams are introduced only in later layers, whereas the coarse stream spans the full network to establish global structure before joint refinement. The final prediction is produced directly from the 1×1 stream, without an image autoencoder or decoder.

A central consequence of this design is that the *number of scales becomes a new architectural scaling dimension*. Experiments indicate that scaling across spatial granularities provides a distinct and substantially more effective use of model capacity than conventional depth scaling. Our analysis further reveals two important principles for multi-scale pixel modeling. First, cross-scale interaction should be bidirectional: restricting attention to coarse-to-fine information flow greatly degrades performance, demonstrating that fine-grained representations provide useful feedback to the coarse stream. Second, fine-scale computation is most effective when introduced in later layers, compared to intermediate ones. With this design, MS-pDiT achieves 1.43 FID on ImageNet 256×256 and 1.44 FID at 512×512, establishing strong performance among pixel-space diffusion models. The same formulation naturally extends to text-to-image generation by treating text as an additional modality, achieving 0.86 on GenEval and 83.9 on DPG-Bench.

Our contributions are summarized as follows:

- We identify flexible tokenization as a fundamental architectural advantage of pixel-space diffusion and propose MS-pDiT, a multi-scale pixel-space diffusion transformer that treats spatial scales as parallel modalities and enables bidirectional cross-scale interaction through compacted joint attention.
- We establish the number of scales as a new architectural scaling dimension. Controlled studies show that introducing intermediate granularities is substantially more effective than allocating comparable capacity to additional depth, and reveal bidirectional cross-scale interaction and late fine-stream computation as key design principles.
- Extensive experiments demonstrate strong performance across both class-conditional and text-to-image generation. MS-pDiT achieves FID scores of 1.43 and 1.44 on ImageNet at 256×256 and 512×512, respectively, and naturally extends to text conditioning, attaining 0.86 on GenEval and 83.9 on DPG-Bench.

2 Related Work

2.1 Latent and Pixel-Space Diffusion

Diffusion transformers are predominantly trained on the latents of a pretrained autoencoder (Rombach et al., 2022; Peebles & Xie, 2023; Ma et al., 2024; Esser et al., 2024). We instead remove the autoencoder entirely. Pixel-space diffusion first reached high resolutions through cascaded models (Ho et al., 2022; Saharia et al., 2022) or U-Nets with aggressive downsampling (Hoogeboom et al., 2023; 2025), and single-stage pixel transformers have recently become competitive through x-prediction, self-supervised pretraining, and perceptual losses (Li & He, 2026; Ma et al., 2026b; Chen et al., 2025b). Most closely related to our work are PixelDiT (Yu et al., 2026), DeCo (Ma et al., 2026a), DiP (Chen et al., 2026), PixNerd (Wang et al., 2026a), and HyperDiT (He et al., 2026), which divide the network into two granularities with strictly coarse-to-fine conditioning, realized through a fine-level decoder, prediction head, or cross-attention branch. PixelU replaces the decoder with skip connections but retains a two-resolution hierarchy (Guo et al., 2026). In contrast, MS-pDiT assigns a complete transformer stream to each of several granularities and couples them bidirectionally. See Appendix A for an extended discussion.

2.2 Multi-scale Generation

Generative models typically exploit scale *sequentially*: they first generate a low-resolution image and subsequently refine it through cascades, nested resolutions, or next-scale prediction (Ho et al., 2022; Teng et al., 2024; Gu et al., 2024; Tian et al., 2024). In MS-pDiT, all streams observe the same image at the same noise level within a single forward pass; scale thus determines *how the signal is tokenized* rather than *when it is generated*.

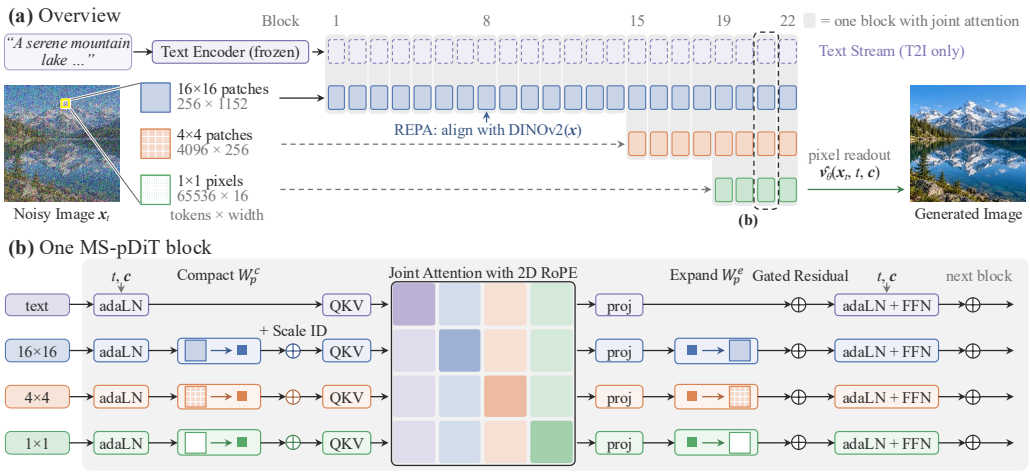

Figure 1: Overview of MS-pDiT. (a) The same noisy image is tokenized at multiple spatial scales and processed by scale-specific transformer streams, with finer streams introduced only in later layers. An optional text stream enables text-to-image generation, while the final prediction is read out directly from the 1×1 pixel stream. (b) In each MS-pDiT block, active image streams are compacted onto a shared spatial grid, augmented with scale identities, and coupled through MM-DiT-style joint attention with shared 2D RoPE, enabling bidirectional information exchange across scales. The attended features are then expanded back to their native resolutions and processed by stream-specific residual and feed-forward layers.

3 Method

In this section, we introduce MS-pDiT, a multi-scale pixel-space diffusion transformer that models the same noisy image at multiple spatial granularities. As illustrated in Figure 1, we first construct scale-specific pixel streams, then couple them through compacted joint attention to enable efficient bidirectional cross-scale interaction, and finally decode the prediction directly from the finest pixel stream. We further describe how the same formulation naturally incorporates text as an additional modality for text-to-image generation.

3.1 Preliminaries

We model RGB pixels directly using linear interpolation (rectified-flow) probability path (Liu et al., 2023; Lipman et al., 2023; Albergo et al., 2025). Let $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ denote a clean image scaled to $[-1, 1]$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ a noise sample of the same shape, and $\boldsymbol{c}$ the condition. For $t \in [0, 1]$, the noisy image and its velocity are

$$\boldsymbol{x}_t = t\,\boldsymbol{x} + (1 - t)\,\boldsymbol{\varepsilon}, \qquad \boldsymbol{v} = \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{x} - \boldsymbol{\varepsilon}, \tag{1}$$

so that $\boldsymbol{x}_0 = \boldsymbol{\varepsilon}$ and $\boldsymbol{x}_1 = \boldsymbol{x}$. A network $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ is trained to regress this velocity with the flow-matching objective

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{c}),\, \boldsymbol{\varepsilon},\, t}\Big[\big\|\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - (\boldsymbol{x} - \boldsymbol{\varepsilon})\big\|_2^2\Big]. \tag{2}$$

The minimizer of Equation 2 is the conditional expectation $\boldsymbol{v}^\star(\boldsymbol{x}_t, t, \boldsymbol{c}) = \mathbb{E}[\boldsymbol{x} - \boldsymbol{\varepsilon} \mid \boldsymbol{x}_t, t, \boldsymbol{c}]$, and integrating the ODE $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t = \boldsymbol{v}^\star(\boldsymbol{x}_t, t, \boldsymbol{c})$ from $t = 0$ to $t = 1$ transports $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to the data distribution $p(\boldsymbol{x} \mid \boldsymbol{c})$ (Lipman et al., 2023; Liu et al., 2023). In contrast to JiT (Li & He, 2026), which predicts $\boldsymbol{x}$ and converts the prediction into a velocity, we regress $\boldsymbol{v}$ directly with a time-independent loss weight, following (Yu et al., 2026).

3.2 Multi-Scale Pixel Streams

Let $\mathcal{P}$ denote the set of patch sizes. For each $p \in \mathcal{P}$, the same noisy image $\boldsymbol{x}_t$ is partitioned into $p \times p$ patches and embedded using a scale-specific linear projection, yielding $N_p = HW/p^2$ tokens $\boldsymbol{X}_p \in \mathbb{R}^{N_p \times d_p}$. Each stream has its own normalization layers, attention projections, and

feed-forward network. Importantly, these streams form parallel representations of the same image rather than successive stages of a coarse-to-fine hierarchy: each is embedded directly from $\boldsymbol{x}_t$, with no stream derived from another. Consequently, fine streams preserve local structure, that is lost under coarse tokenization and cannot be recovered from the coarse stream alone. The predicted velocity is read out directly from the 1×1 stream, eliminating the need for patch-level upsampling at the output. Each stream has its own embedding dimension d_p, with finer streams kept narrower to accommodate their substantially larger token counts. Our default class-conditional model uses $\mathcal{P} = \{1, 4, 16\}$ with $(d_1, d_4, d_{16}) = (16, 256, 1152)$. Complete width configurations for all models are provided in Appendix B.

Streams need not span the full network depth. Let L_p denote the depth of stream p and $L = \max_p L_p$ the total network depth. We right-align the streams so that stream p is active only in the final L_p blocks, with each block instantiating parameters only for its active streams. Most of the network depth is allocated to the coarse stream, which captures global semantics, consistent with prior pixel-space architectures that concentrate capacity in the patch-level pathway (Yu et al., 2026; Ma et al., 2026a). Our default model uses $(L_1, L_4, L_{16}) = (4, 8, 22)$: the 16×16 stream spans all 22 blocks, while the 4×4 and 1×1 streams are introduced only in the final eight and four blocks, respectively. Section 5.4 compares this late anchoring strategy with an alternative that places the fine-stream blocks at intermediate depths (He et al., 2026).

3.3 Compacted Multi-Scale Joint Attention

Direct attention over fine-grained streams is prohibitively expensive. We therefore perform attention on a shared coarser grid. Let the attention patch size a be a common multiple of all $p \in \mathcal{P}$, with $a = 16$ by default. Similar to Yu et al. (2026), for each stream p, the $(a/p)^2$ tokens within each $a \times a$ region are concatenated along the channel dimension and projected to a shared width D:

$$\widetilde{\boldsymbol{X}}_p = \mathrm{concat}_{(a/p)^2}(\boldsymbol{X}_p)\, \boldsymbol{W}_p^{\mathrm{c}} \in \mathbb{R}^{M \times D}, \qquad M = \frac{HW}{a^2}, \qquad \boldsymbol{W}_p^{\mathrm{c}} \in \mathbb{R}^{(a/p)^2 d_p \times D}. \tag{3}$$

Thus, each stream contributes exactly M compact tokens, with tokens at the same spatial position corresponding to the same $a \times a$ image region. After attention, a scale-specific expansion projection $\boldsymbol{W}_p^{\mathrm{e}} \in \mathbb{R}^{D \times (a/p)^2 d_p}$ maps each compact token back to $(a/p)^2$ native-resolution tokens. As a result, attention operates on a sequence length determined by M and $|\mathcal{P}|$, independent of the finest patch size, while each stream maintains its native-resolution representation for normalization and feed-forward computation.

All image streams share the same 2D rotary position embeddings (Su et al., 2024) over the $H/a \times W/a$ attention grid. This shared coordinate system establishes explicit cross-scale correspondence: compact tokens from different streams that represent the same spatial region receive identical rotations, thereby preserving relative positional structure both within and across streams. However, positional information alone cannot distinguish tokens from different streams, as the $|\mathcal{P}|$ tokens associated with the same spatial location are otherwise indistinguishable to the attention operator. We therefore augment each compact token with a learned scale-identity embedding, specific to each block and stream, before attention. Ablating this embedding degrades FID from 2.31 to 3.29 (Appendix C.1), demonstrating the importance of explicitly encoding stream identity.

Each stream applies its own query, key, value, and output projections to its compact tokens. The resulting queries, keys, and values are concatenated along the sequence dimension and processed jointly by a single attention operator:

$$\boldsymbol{Q} = [\boldsymbol{Q}_p]_{p \in \mathcal{P}}, \quad \boldsymbol{K} = [\boldsymbol{K}_p]_{p \in \mathcal{P}}, \quad \boldsymbol{V} = [\boldsymbol{V}_p]_{p \in \mathcal{P}}, \qquad \boldsymbol{O} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_h}}\right)\boldsymbol{V}, \tag{4}$$

using D/d_h heads over a sequence of length $|\mathcal{P}|M$. The output is then split by stream, passed through the corresponding output projection, expanded via $\boldsymbol{W}_p^{\mathrm{e}}$, and added to each stream through a zero-initialized gate. Crucially, no cross-scale attention mask is imposed. In contrast to prior pixel-space architectures that restrict information flow from coarse to fine (Yu et al., 2026; Ma et al., 2026a; Chen et al., 2026; Wang et al., 2026a; He et al., 2026), joint attention permits unrestricted bidirectional interaction among all active scales. A single softmax over the concatenated sequence therefore allows the model to adaptively determine the extent and direction of cross-scale information exchange at each layer. Section 5.3 shows that enforcing scale-causal attention severely degrades performance.

Table 1: Class-conditional generation on ImageNet 256×256. All metrics are computed from 50K samples generated with classifier-free guidance. "Params" denotes the inference parameter count (generator + VAE/RAE decoder), and NFE denotes the number of function evaluations. † Results reported in the original paper. ‡ Results reproduced with the official implementation, using post-modulation for improved training stability. Bold denotes the best result within each group.

Method	Params	Epochs	NFE	FID↓	IS↑	sFID↓	Prec.↑	Rec.↑
Latent-space diffusion								
DiT-XL/2 (Peebles & Xie, 2023)	675+49M	1400	250×2	2.27	278.2	4.60	**0.83**	0.57
SiT-XL/2 (Ma et al., 2024)	675+49M	1400	250×2	2.06	277.5	**4.50**	0.80	0.64
REPA-XL/2 (Yu et al., 2025)	675+49M	800	250×2	1.42	305.7	4.70	0.80	0.64
DDT-XL/2 (Wang et al., 2026b)	675+49M	400	–	1.26	**310.6**	–	0.78	**0.67**
RAE-XL/2 (Zheng et al., 2026)	839+415M	800	–	**1.13**	262.6	–	0.78	**0.67**
Pixel-space diffusion								
ADM-G (Dhariwal & Nichol, 2021)	554M	400	250	4.59	186.7	5.25	**0.82**	0.52
RIN (Jabri et al., 2023)	410M	480	–	3.42	182.0	–	–	–
VDM++ (Kingma & Gao, 2023)	2B	–	–	2.12	278.1	–	–	–
PixelFlow-XL/4 (Chen et al., 2025b)	677M	320	120×2	1.98	282.1	5.83	0.81	0.60
PixNerd-XL/16 (Wang et al., 2026a)	700M	320	100×2	1.95	300.0	4.54	0.80	0.60
JiT-G/16 (Li & He, 2026)	2B	600	100×2	1.82	292.6	–	–	–
PixelGen-XL/16 (Ma et al., 2026b)	676M	160	100×2	1.83	293.6	–	0.79	0.63
DiP-XL/16 (Chen et al., 2026)	631M	600	100×2	1.79	281.9	4.59	0.80	0.63
DeCo-XL/16 (Ma et al., 2026a)	682M	600	100×2	1.69	304.0	4.59	0.79	0.63
PixelDiT-XL† (Yu et al., 2026)	797M	320	100×2	1.61	292.7	4.68	0.78	0.64
PixelDiT-XL‡	797M	320	100×2	1.66	292.7	4.89	0.78	0.64
PixelU-H/16 (Guo et al., 2026)	1.2B	600	100×2	1.63	305.9	5.04	0.79	0.64
MS-pDiT-L	788M	600	100×2	1.55	305.9	4.42	0.78	**0.65**
MS-pDiT-XL	963M	600	100×2	1.50	**308.7**	4.31	0.79	**0.65**
MS-pDiT-H	1.4B	360	100×2	**1.43**	304.6	**4.13**	0.79	**0.65**

3.4 Architectural Details

Viewing scales as distinct modalities naturally extends the architecture to text conditioning. We project precomputed text features to width D, assign them 1D rotary positional embeddings, and concatenate them with the compact image tokens when constructing $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ in Equation 4. The text stream is active in every block and, analogous to each image stream, maintains its own attention projections, adaptive normalization, and feed-forward network. Image generation remains entirely in pixel space: the output is decoded from the 1×1 stream, without an image autoencoder.

Following prior work (Yu et al., 2025), we add a REPA loss to Equation 2, aligning the coarsest-stream tokens with frozen DINOv2 features (Oquab et al., 2024) of the clean image, with weight $\lambda = 0.5$, at block 8 for class-conditional models and block 6 for text-to-image models. The remaining architectural components follow standard practice. Queries and keys are normalized with RMSNorm (Zhang & Sennrich, 2019; Dehghani et al., 2023), and feed-forward networks use SwiGLU (Shazeer, 2020). Each stream maintains its own modulation parameters, applied at the native stream resolution before compaction and following the attention residual.

4 Experiments

We evaluate MS-pDiT on class-conditional ImageNet generation (Deng et al., 2009) and on text-to-image generation. On ImageNet, we report both system-level comparisons with prior work and controlled architectural ablations. Following standard practice, we report FID, sFID, Inception Score (IS), precision, and recall; lower values are better for FID and sFID, and higher values are better for the remaining metrics. For text-to-image generation, we assess compositional and prompt-following ability at 512×512 resolution using GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024).

4.1 Experimental Setup

We train directly on RGB pixels, without a VAE or learned image tokenizer. Class- and text-conditional models share the optimizer, initialization, and classifier-free guidance dropout, and dif-

Table 2: Class-conditional generation on ImageNet 512×512. All metrics are computed from 50K samples generated with classifier-free guidance. Bold denotes the best pixel-space result.

Method	FID↓	sFID↓	IS↑	Precision↑	Recall↑
Latent-space diffusion					
DiT-XL (Peebles & Xie, 2023)	3.04	5.02	240.8	0.84	0.54
SiT-XL (Ma et al., 2024)	2.62	4.18	252.2	0.84	0.57
MaskDiT (Zheng et al., 2024)	2.50	5.10	256.2	0.83	0.56
U-ViT-H (Bao et al., 2023)	4.05	–	263.8	0.84	0.48
REPA (Yu et al., 2025)	2.08	4.19	274.6	0.83	0.58
RAE-XL (Zheng et al., 2026)	1.13	–	259.6	0.80	0.63
Pixel-space diffusion					
ADM (Dhariwal & Nichol, 2021)	3.85	5.86	221.7	**0.84**	0.53
RIN (Jabri et al., 2023)	3.95	–	216.0	–	–
Simple Diffusion (Hoogeboom et al., 2023)	3.54	–	205.0	–	–
VDM++ (Kingma & Gao, 2023)	2.65	–	278.1	–	–
PixNerd-XL (Wang et al., 2026a)	2.84	5.95	245.6	0.80	0.59
EPG-L/32 (Liu et al., 2026)	2.35	–	295.4	0.82	0.57
DeCo-XL/16 (Ma et al., 2026a)	2.22	**4.67**	290.0	0.80	0.60
JiT-H (Li & He, 2026)	1.94	–	309.1	–	–
PixelU-H/32 (Guo et al., 2026)	1.92	–	**322.1**	–	–
PixelDiT (Yu et al., 2026)	1.81	5.61	278.6	0.78	**0.67**
MS-pDiT	**1.44**	4.76	319.2	0.80	0.64



Figure 2: Class-conditional samples from MS-pDiT-XL on ImageNet 256×256 (CFG scale 4.0).

fer only in the condition stream, data, and resolution schedule. On ImageNet, ablations use the $(1, 4, 16)$ streams and are evaluated after 80 epochs; MS-pDiT-L and MS-pDiT-XL are trained for 600 epochs and MS-pDiT-H for 360. Metrics are computed on 50K samples drawn with a flow DPM solver. The text-to-image model follows the three-stage recipe and data of DeCo (Ma et al., 2026a). Architectures, learning rates, guidance intervals, and the full sampling protocol are given in Appendix B.

4.2 Class-Conditional Generation

Results at 256×256. Table 1 compares our models with latent- and pixel-space generators. Among pixel-space methods, MS-pDiT-L improves upon the reproduced PixelDiT-XL baseline from 1.66 to 1.55 FID at a comparable inference parameter budget. It also attains higher recall (0.65 vs. 0.64) and substantially lower sFID (4.42 vs. 4.89), suggesting more faithful spatial structure rather than merely sharper textures. For MS-pDiT-XL, increasing the depth further reduces FID to 1.50 and sFID to 4.31, with an IS of 308.7. MS-pDiT-H extends the model to all five scales $(1, 2, 4, 8, 16)$ and achieves 1.43 FID and 4.13 sFID after only 360 epochs. This result mirrors the scale-wise scaling law, with additional scales continuing to improve the model. Relative to the strongest pixel-space baselines, MS-pDiT-H improves FID by 0.18 over PixelDiT-XL and by 0.20 over PixelU-H/16.



Figure 3: Class-conditional samples from MS-pDiT on ImageNet 512×512 (CFG scale 4.0).



Figure 4: Text-to-image samples from MS-pDiT at 512×512. Prompts are drawn from DrawBench and PartiPrompts and are listed, from left to right, in Table 14.

Results at 512×512**.** Table 2 summarizes the comparison at higher resolution. The advantage of multi-scale streams increases with resolution. We hypothesize that higher resolutions require modeling structure over a wider range of spatial scales, which a sequence of intermediate granularities captures more effectively than a two-level hierarchy. MS-pDiT achieves 1.44 FID and 319.2 IS, outperforming PixelDiT (1.81), PixelU-H/32 (1.92), and JiT-H (1.94) with a large margin. Figures 2 and 3 present samples at both resolutions, and additional samples from other classes are provided in Appendix D.

4.3 Text-to-Image Generation

The same multi-scale interface accommodates a text stream without any change to the pixel readout. As shown in Table 3, MS-pDiT attains a GenEval score of 0.86 at 512×512, matching the best pixel-space result, reported by DeCo-XXL/16, and outperforming PixelDiT-T2I (0.78) and PixNerd-XXL/16 (0.73). On DPG-Bench, it achieves 83.9, exceeding DeCo by 2.5 points and PixelDiT-T2I by 0.2 points despite a shorter training schedule. These results indicate that treating text as an additional modality alongside several image granularities is a viable approach to pixel-space text-to-image generation. Figure 4 presents 512×512 samples, and the corresponding prompts are listed in Appendix D.

5 Analysis

Unless otherwise stated, all analysis experiments use the 788M-parameter $(1, 4, 16)$ model trained for 80 epochs on ImageNet 256×256, with the same sampling and evaluation protocol as the main comparison.

5.1 Scale-Wise Scalability

Table 4 examines whether introducing additional scale-specific computation benefits a model that already contains a coarse 16×16 stream and a pixel-level 1×1 stream. Adding an eight-block 4×4 stream reduces FID from 4.39 to 2.31 and increases IS from 241.5 to 302.1. By comparison, allocating a similar parameter budget to additional coarse-stream depth yields a substantially smaller improvement (Section 5.2). These results support our central empirical finding: the number of represented granularities provides a distinct and effective axis for scaling model capacity. Across the evaluated configurations, FID generally improves as more scales are introduced: from 4.39 with two scales to 2.31 with three, 2.26 with four $(1, 2, 4, 16)$, and 2.19 with all five $(1, 2, 4, 8, 16)$.

Table 3: Comparison of pixel-space text-to-image models at 512×512. We report the per-category GenEval scores, the overall GenEval score, and the average DPG-Bench score.

		GenEval↑							DPG↑
Method	**Params**	**Single**	**Two**	**Count.**	**Colors**	**Pos.**	**Attr.**	**Overall**	**Avg.**
PixelFlow	0.9B	–	–	–	–	–	–	0.60	77.9
PixNerd	1.2B	0.97	0.86	0.44	0.83	0.71	0.53	0.73	80.9
PixelGen	1.1B	0.99	0.88	0.59	0.90	0.70	0.70	0.79	–
PixelDiT	1.3B	**1.00**	**0.94**	0.70	0.90	0.53	0.65	0.78	83.7
DeCo	1.1B	**1.00**	0.92	**0.72**	0.91	**0.80**	**0.79**	**0.86**	81.4
MS-pDiT	1.4B	0.99	**0.94**	**0.72**	**0.94**	0.78	0.77	**0.86**	**83.9**

Table 4: Scale-wise model scaling on ImageNet 256×256. Each scale column gives the number of transformer blocks assigned to the corresponding patch stream, and a dash indicates that the stream is absent. FID improves as scales are added, and the five-scale model performs best.

Stream Depth					Training Settings		Quality				
16×16	8×8	4×4	2×2	1×1	Params	Epochs	FID↓	IS↑	sFID↓	Prec.↑	Rec.↑
22	-	-	-	4	651M	80	4.39	241.5	6.23	0.73	0.59
22	-	8	-	4	788M	80	2.31	302.1	4.77	0.81	0.58
22	-	8	8	4	910M	80	2.26	298.5	4.66	0.81	0.58
22	8	8	-	4	914M	80	2.35	300.0	4.67	0.81	0.59
22	8	8	8	4	1.0B	80	2.19	296.4	4.73	0.81	0.59

Performance therefore does not saturate within the range of scales considered, suggesting further headroom along the scale dimension at the current training budget.

5.2 Choice of Scales

The preceding comparison varies both the number of scales and the parameter count. Table 5 therefore controls for model size: all five configurations contain 758–788M parameters and share the same 1×1 output stream. Allocating the additional capacity solely to the coarse stream, 26 blocks at 16×16 together with the pixel stream achieves an FID of 3.63, while distributing depth uniformly across three scales (13 blocks per stream) performs slightly worse, with an FID of 3.74. In contrast, every multi-scale configuration with an asymmetric depth allocation substantially outperforms both baselines. These results indicate that the gain observed in Table 4 arises primarily from introducing an intermediate granularity rather than merely increasing parameter count. Among the candidate intermediate scales, 4×4 performs best, achieving 2.31 FID and 4.77 sFID, compared with 2.42 FID for 8×8 and 2.54 for 2×2.

5.3 Cross-Scale Information Flow

Our default joint attention permits bidirectional information flow across scales, marking a key departure from prior pixel-space diffusion architectures. In PixelDiT, DeCo, DiP, PixNerd, and HyperDiT, information flows predominantly from a coarse semantic pathway to a finer pathway, without a corresponding fine-to-coarse interaction. If such a strictly top-down inductive bias were sufficient, explicitly enforcing it should preserve or improve performance.

Table 6 tests this hypothesis using two cross-scale attention masks. The scale-causal mask allows a query at scale p to attend only to keys at scales $p' \geq p$, permitting fine streams to access coarser streams while preventing the reverse interaction. The strict variant additionally removes same-scale attention from all but the coarsest stream, yielding a purely top-down information pathway. With the default late anchoring, the scale-causal mask degrades FID from 2.31 to 3.47, while the strict mask yields 2.55. Under middle-layer anchoring, the corresponding FIDs are 2.50 and 3.24. In both settings, unrestricted bidirectional attention achieves the best performance.

Table 5: Controlled comparison of intermediate patch scales on ImageNet 256×256. Models have comparable parameter counts and share a 1×1 output stream and a 16×16 backbone stream. The 4×4 intermediate stream gives the best FID, IS, and sFID.

Stream Depth					Training Settings		Quality				
16×16	8×8	4×4	2×2	1×1	Params	Epochs	FID↓	IS↑	sFID↓	Prec.↑	Rec.↑
22	-	8	-	4	788M	80	**2.31**	**302.1**	**4.77**	**0.81**	0.58
13	-	13	-	13	768M	80	3.74	237.8	5.63	0.76	**0.59**
26	-	-	-	4	758M	80	3.63	254.6	5.72	0.76	0.58
22	-	-	8	4	773M	80	2.54	292.5	5.26	0.80	0.58
22	8	-	-	4	777M	80	2.42	298.2	5.12	0.80	**0.59**

Table 6: Effect of cross-scale information flow and fine-stream anchor location on ImageNet 256×256. "Last Layers" denotes the default right-aligned schedule, and "Middle Layers" places the fine-stream blocks in the middle of the coarse backbone. The scale-causal mask allows a query to attend only to its own or coarser scales, and the strict variant additionally removes same-scale attention except in the coarsest stream.

Method	**Epochs**	**FID↓**	**IS↑**	**sFID↓**	**Prec.↑**	**Rec.↑**
Last Layers (Ours)	80	**2.31**	**302.1**	**4.77**	**0.81**	0.58
+ Scale Causal Mask	80	3.47	262.2	5.58	0.76	0.58
+ Strict Scale Causal Mask	80	2.55	299.4	5.18	0.79	**0.59**
Middle Layers	80	**2.31**	277.4	4.99	0.80	**0.59**
+ Scale Causal Mask	80	2.50	282.6	5.10	0.80	0.58
+ Strict Scale Causal Mask	80	3.24	265.2	5.70	0.78	0.58

These results suggest that fine-to-coarse communication provides useful information beyond what is available through a purely top-down pathway. Because coarse tokenization discards local spatial detail, fine streams can expose information unavailable to the 16×16 representation alone, allowing coarse-scale features to incorporate finer-grained evidence during joint attention. Such interactions cannot be represented by a strictly coarse-to-fine architecture. We therefore retain fully bidirectional joint attention across active scales.

5.4 Where to Anchor the Fine Streams

We next examine where the limited fine-stream depth should be allocated relative to the coarse backbone. Right-aligning the fine streams in the final layers achieves 2.31 FID and 302.1 IS. Moving the same number of fine-stream blocks to intermediate depths, with individual fine-stream layers interleaved with several coarse-backbone layers as in HyperDiT, yields a similar FID but lowers IS to 277.4 and worsens sFID to 4.99. These results favor late anchoring as the default design. Under this configuration, the coarse stream first develops a global representation, after which the fine streams are introduced to jointly refine local appearance and produce the final pixel-level output.

6 Conclusion

We introduced MS-pDiT, a multi-scale pixel-space diffusion transformer that exploits flexible pixel tokenization by treating spatial scale as a modality. By representing the same noisy image at multiple granularities and coupling scale-specific streams through compacted joint attention, MS-pDiT enables efficient and bidirectional information exchange across scales while generating images directly in pixel space. Our experiments establish the number of scales as an effective architectural scaling dimension, with additional granularities providing substantial gains beyond simply increasing network depth. Further analysis highlights the importance of bidirectional cross-scale communication and late fine-stream computation, suggesting that effective pixel-space generation benefits from first establishing global structure and subsequently integrating fine-grained information through joint refinement. Across ImageNet class-conditional and text-to-image generation, MS-pDiT demonstrates strong performance and scales naturally to higher resolutions and additional conditioning modalities. These results suggest that multi-scale representation provides a promising architectural direction for scalable pixel-space generative modeling.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experiment-related coding, as well as drafting, editing, and polishing the manuscript. Generative AI tools were not used to generate the core research ideas, hypotheses, or methodological contributions of this work. All AI-assisted code was reviewed, tested, and validated by the authors. The authors take full responsibility for the final content of this work, including all text, claims, code, experimental results, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix B specify the model configurations, optimizer settings, timestep and guidance schedules, and evaluation protocol for every reported result, including the exact stream widths, per-stream depths, attention patch size, and REPA settings used in each ablation. All class-conditional results are obtained on the public ImageNet-1K training set and evaluated against the standard ADM reference statistics for FID, sFID, IS, precision, and recall; they can therefore be reproduced without access to proprietary data. We will release the training and evaluation code, the configuration file for each table entry, and the model checkpoints.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679. IEEE, 2023.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025b.

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. Dip: Taming diffusion models in pixel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36136–36146, 2026.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

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

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

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 30234–30262, 2024.

Zipeng Guo, Lichen Ma, Yu He, Xiaolong Fu, Jingling Fu, Junshi Huang, and Yan Li. Pixelu: A u-shaped transformer for efficient end-to-end pixel diffusion. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part LX*, volume 17060 of *Lecture Notes in Computer Science*, pp. 425–442. Springer, 2026. doi: 10.1007/978-3-032-37092-1_24. URL https://doi.org/10.1007/978-3-032-37092-1_24.

Yu He, Lichen Ma, Zipeng Guo, Xinyuan Shan, Jingling Fu, Dong Chen, Junshi Huang, and Yan Li. Hyperdit: Hyper-connected transformers for high-fidelity pixel-space diffusion. *arXiv preprint arXiv:2605.15741*, 2026.

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 2023. URL https://proceedings.mlr.press/v202/hoogeboom23a.html.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18062–18071. IEEE, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Allan Jabri, David J. Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14569–14589. PMLR, 2023. URL https://proceedings.mlr.press/v202/jabri23a.html.

Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36:65484–65516, 2023.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 936–944. Ieee, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Keli Liu, Julius Berner, Y HoiM, Hongkai Zheng, Jiahong Wu, Xiangxiang Chu, et al. There is no vae: End-to-end pixel-space generative modeling via self-supervised pre-training. In *International Conference on Learning Representations*, volume 2026, pp. 102818–102839, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43600–43610, 2026a.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. Pixelgen: Improving pixel diffusion with perceptual supervision. *arXiv preprint arXiv:2602.02493*, 2026b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=qTlcbLSm4p.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. In *International Conference on Learning Representations*, volume 2026, pp. 43559–43580, 2026a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 40633–40642, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=vTBjBtGioE.

A Extended Related Work

A.1 Latent Diffusion and Diffusion Transformers

Latent diffusion moves denoising into the compressed space of a pretrained autoencoder (Rombach et al., 2022), and DiT established the transformer as the standard backbone in this space (Peebles & Xie, 2023); it was subsequently extended to flow matching by SiT (Ma et al., 2024) and to text conditioning by MM-DiT (Esser et al., 2024). Much subsequent work targets the autoencoder itself. VA-VAE aligns latents with vision foundation model features (Yao et al., 2025), REPA-E tunes the autoencoder and the diffusion model jointly (Leng et al., 2025), and representation autoencoders replace the variational bottleneck with frozen self-supervised features (Zheng et al., 2026). REPA shows that aligning intermediate denoiser features with DINOv2 substantially accelerates training (Yu et al., 2025; Oquab et al., 2024), and DDT decouples semantic encoding from detail decoding within the latent space (Wang et al., 2026b).

A.2 Pixel-Space Diffusion

Diffusion models were originally formulated in pixel space (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Dhariwal & Nichol, 2021), where high resolutions were first reached

through cascades of separately trained super-resolution stages (Ho et al., 2022; Saharia et al., 2022; Teng et al., 2024) or through architectures that concentrate computation on a small set of latent tokens (Jabri et al., 2023). U-Net and U-ViT variants with aggressive early downsampling and long skip connections remain competitive at high resolution (Hoogeboom et al., 2023; 2025; Kingma & Gao, 2023). The recent generation of single-stage pixel transformers is more directly related to our work. PixelFlow uses a single set of parameters but traverses cascading resolutions during sampling (Chen et al., 2025b). JiT argues that predicting clean data rather than a noised quantity is what makes large pixel patches tractable (Li & He, 2026), and EPG attains strong FID by pretraining the encoder with a self-supervised objective prior to generative fine-tuning (Liu et al., 2026). PixelGen adds LPIPS and DINO perceptual supervision on top of x-prediction (Ma et al., 2026b).

Most closely related to our setting are models that explicitly partition the network by granularity. PixelDiT runs a 16×16 patch-level DiT and a 1×1 pixel-level DiT, connects them through pixel-wise adaptive normalization, and compacts pixel tokens before attention (Yu et al., 2026). DeCo replaces the fine pathway with an attention-free pixel decoder and reweights the flow-matching loss with JPEG-derived frequency priors (Ma et al., 2026a). DiP employs a convolutional patch detailer head (Chen et al., 2026), and PixNerd predicts per-patch neural-field weights from the coarse stream (Wang et al., 2026a). HyperDiT lets a small-patch stream query multi-level semantic anchors of a large-patch stream through cross-attention, using scale-aware rotary embeddings to align the two coordinate systems (He et al., 2026). In each case, there are exactly two granularities, conditioning runs strictly from coarse to fine, and the coarse stream never attends to the fine one. PixelU instead argues that such decoders are unnecessary under x-prediction and replaces them with skip connections around a downsampled bottleneck (Guo et al., 2026); this removes the decoder but retains a single tokenization and a two-resolution hierarchy. In contrast, we employ several granularities, assign each a complete transformer stream, and couple them bidirectionally within a single attention operation.

A.3 Multi-Scale Representations

Multi-scale representations have long been central to computer vision, from feature pyramids to hierarchical vision transformers (Lin et al., 2017; Liu et al., 2021). In generative modeling, scale is also widely exploited, but is typically organized *sequentially*: cascaded diffusion models generate low-resolution images before refining them at higher resolutions (Ho et al., 2022), while visual autoregressive models progressively generate images from coarse to fine scales (Tian et al., 2024). Other approaches, such as Matryoshka diffusion, jointly model nested resolutions within a shared architecture (Gu et al., 2024). In contrast, MS-pDiT uses multiple scales as concurrent representations within a single forward pass. All streams observe the same image at the same resolution and noise level, with scale controlling *how the signal is tokenized* rather than *when it is generated.*

B Detailed Settings and Hyperparameters

B.1 Initialization and Shared Optimization

Patch-embedding and compaction matrices are initialized with Xavier initialization, whereas the adaptive-normalization and output projections are zero-initialized. Consequently, every residual block and every cross-scale gate is initialized as the identity mapping, and the streams become coupled only as training progresses. Both class- and text-conditional models optimize with AdamW at a constant learning rate, clip gradients to unit norm, maintain an exponential moving average (EMA) of the weights for evaluation, and drop the condition with probability 0.1 to enable classifier-free guidance (CFG). They differ only in the condition stream, data pipeline, resolution schedule, and the per-task hyperparameters below.

B.2 Model Configurations

Table 7 lists the configurations used for the main results. All models share a joint-attention patch size of $a = 16$, a per-head dimension of 72 (class-conditional) or 64 (text-conditional), SwiGLU feed-forward networks with an expansion ratio of 4, RMSNorm on queries and keys, and 2D rotary position embeddings on the compact grid.

Table 7: Model configurations. Depths are listed from the finest to the coarsest stream. Because the streams are right-aligned, the total number of blocks equals the largest per-stream depth.

Model	Patch sizes	Depths	Stream widths	D	Heads	Params
MS-pDiT-L	$(1, 4, 16)$	$(4, 8, 22)$	$(16, 256, 1152)$	1152	16	788M
MS-pDiT-XL	$(1, 4, 16)$	$(4, 12, 26)$	$(16, 256, 1152)$	1152	16	963M
MS-pDiT-H	$(1, 2, 4, 8, 16)$	$(4, 8, 12, 20, 26)$	$(16, 64, 256, 512, 1152)$	1152	16	1.4B
MS-pDiT-T2I	$(1, 4, 16)$	$(2, 6, 14)$	$(16, 256, 1536)$	1536	24	1.4B

B.3 Class-Conditional Training and Sampling

Class-conditional models are trained directly on RGB images from the ImageNet-1K training set, without a VAE or any learned image tokenizer. Unless otherwise specified, ablations use the $(1, 4, 16)$ streams with widths $(16, 256, 1152)$ and depths $(4, 8, 22)$, matching MS-pDiT-L in Table 7. Training is performed in bfloat16 with a global batch size of 256 on eight GPUs, using AdamW ($\beta = (0.9, 0.999)$, learning rate 10^{-4}, no weight decay), gradient clipping at 1.0, and an EMA decay of 0.9999. MS-pDiT-L and MS-pDiT-XL are trained for 600 epochs and MS-pDiT-H for 360 epochs, whereas ablation models are evaluated after 80 epochs. Timesteps are sampled as $t = \mathrm{sigmoid}(\epsilon)$ with $\epsilon \sim \mathcal{N}(0, 1)$ and then time-shifted, with a shift of 1.0 at 256×256 and 3.0 at 512×512. Class labels are dropped with probability 0.1 for classifier-free guidance. REPA uses DINOv2 ViT-B/14 features, a three-layer MLP projector, alignment at block 8, and $\lambda = 0.5$.

For evaluation at 256×256, we generate 50K samples with a 100-step flow DPM solver. The 80-epoch ablations use a CFG scale of 3.25 applied over $t \in [0.1, 1.0]$, whereas the main models use a CFG scale of 2.75 applied over $t \in [0.1, 0.9]$. At 512×512, we initialize from the 256×256 checkpoint, fine-tune with a reduced learning rate and a shifted time schedule, and sample with a CFG scale of 3.5. All distributional metrics are computed against the standard ADM ImageNet reference statistics (Dhariwal & Nichol, 2021).

B.4 Text-to-Image Training

The text-conditional model uses image streams $(1, 4, 16)$ with per-stream widths $(16, 256, 1536)$ and depths $(2, 6, 14)$, a joint width of $D = 1536$ with 24 heads, and an attention patch size of 16, identical to that of the class-conditional models. Pixels are read out from the 1×1 stream. Text is encoded by a frozen Qwen3-1.7B model (Yang et al., 2025), projected from 2048 channels to the joint width, and assigned its own RoPE frequencies with $\theta = 10^4$. We use a fixed 128-token context without an attention mask, so padding tokens remain visible to the image streams. Text embeddings are dropped with probability 0.1 for classifier-free guidance.

Training proceeds in the three stages listed in Table 8, which follow the public recipe of DeCo (Ma et al., 2026a). Per-GPU batch sizes, learning rates, and data are taken unchanged from its released configurations, so that the comparison in Table 3 isolates the effect of the architecture. The high-resolution stage is shorter than that of DeCo (80K vs. 100K steps); all other settings are identical. Stages 1 and 2 train on the BLIP3o pretraining corpus (Chen et al., 2025a), with stage 2 restricted to images whose shorter side is at least 512 pixels, and stage 3 fine-tunes on BLIP3o-60k. Each stage resumes from the previous checkpoint and continues the cumulative step count. Steps are counted as optimizer updates without gradient accumulation.

Table 8: Per-stage text-to-image training settings. The global batch size is the product of the per-GPU batch size, the number of GPUs, and the gradient accumulation factor (always one). Steps denote the number of additional optimizer updates in each stage.

Stage	Res.	Data	Steps	GPUs	Global batch	LR	Shift
1. Pretrain	256^2	BLIP3o	200K	24	1536	2×10^{-4}	2.0
2. High-res.	512^2	BLIP3o ($\geq$512px)	80K	32	512	1×10^{-4}	4.0
3. SFT	512^2	BLIP3o-60k	40K	32	512	1×10^{-5}	4.0

All stages use AdamW without weight decay, a constant learning rate, gradient clipping at 1.0, bfloat16 mixed precision with attention accumulated in fp32, gradient checkpointing, and an EMA decay of 0.9999. The Adam betas are $(0.9, 0.95)$ for the two pretraining stages and $(0.9, 0.999)$ for supervised fine-tuning. Timesteps are sampled from a logit-normal distribution with mean 0 and standard deviation 1 and then time-shifted by the factor given in Table 8. The loss combines the flow-matching objective with representation alignment at block 6 (weight 0.5, DINOv2 ViT-B/14 targets).

Evaluation protocol. All reported text-to-image results are obtained with the final stage-3 EMA weights and a second-order flow DPM solver with 50 sampling steps, a CFG scale of 4.0 applied over the full time interval, and a time shift of 3.0, generating at 512×512 in bfloat16. For GenEval (Ghosh et al., 2023), we sample four images for each of the 553 official prompts with seed 0 and score them with the official detector pipeline; the reported value is the mean over the six task categories. For DPG-Bench (Hu et al., 2024), we sample four images for each of the 1065 prompts with seed 0 and score them with the official ELLA implementation.

C Additional Analysis

C.1 Scale Identity Embedding

Because all streams are compacted onto the same attention grid and share the same rotary coordinates, position alone does not indicate the tokenization from which a token originates. Table 9 shows that the learned scale-identity embedding is necessary: removing it degrades FID from 2.31 to 3.29 and sFID from 4.77 to 6.02. This embedding is the multi-scale counterpart of the modality embeddings used in multimodal transformers.

Table 9: Ablation of the learned scale-identity embedding on ImageNet 256×256. Scale-identity embeddings distinguish streams that share the same compact attention coordinates; removing them degrades FID from 2.31 to 3.29.

Method	Epochs	FID↓	IS↑	sFID↓	Prec.↑	Rec.↑
MS-pDiT	80	**2.31**	**302.1**	**4.77**	**0.81**	0.58
w/o Scale Identity Embedding	80	3.29	265.2	6.02	0.76	**0.59**

C.2 Attention Patch Size

The attention patch size a determines the length of the joint sequence and, consequently, the granularity at which streams exchange information. Doubling a to 32 reduces the number of attention tokens by a factor of four, but it also quadruples the number of native tokens that each compact token must summarize. Table 10 shows that this trade-off is unfavorable: FID degrades from 2.31 to 3.00 and IS from 302.1 to 252.6. We therefore use $a = 16$ throughout, a choice that also renders the compaction of the coarse stream an identity mapping in the spatial dimension.

Table 10: Effect of the joint-attention patch size a on ImageNet 256×256 after 80 epochs. Larger attention patches shorten the joint sequence but coarsen cross-scale correspondence.

Attention Patch Size	Epochs	FID↓	IS↑	sFID↓	Prec.↑	Rec.↑
16 (ours)	80	**2.31**	**302.1**	**4.77**	**0.81**	0.58
32	80	3.00	252.6	5.53	0.77	**0.60**

C.3 Frequency-Aware Flow-Matching Loss

DeCo reweights the flow-matching residual in the DCT domain using JPEG quantization tables, on the grounds that a single network would otherwise expend capacity on perceptually irrelevant high frequencies (Ma et al., 2026a). We evaluated the same loss on top of our model. As shown

in Table 11, it yields no improvement: FID increases from 2.31 to 2.40 and sFID from 4.77 to 4.95, while precision and recall remain unchanged. We interpret this result as mild support for our framing. Frequency reweighting informs a single-scale network which components of the residual are perceptually important; when separate streams already specialize by granularity, such guidance is largely redundant.

Table 11: Effect of the frequency-aware flow-matching (FreqFM) loss on ImageNet 256×256. Adding FreqFM does not improve any of the evaluated metrics after 80 epochs.

Method	**Epochs**	**FID↓**	**IS↑**	**sFID↓**	**Prec.↑**	**Rec.↑**
MS-pDiT	80	**2.31**	**302.1**	**4.77**	**0.81**	**0.58**
+ $\mathcal{L}_{\text{FreqFM}}$	80	2.40	300.0	4.95	**0.81**	**0.58**

C.4 Representation Alignment

Table 12 shows that representation alignment remains important for optimization in pixel space, even with multiple streams. Removing REPA increases FID from 2.31 to 5.30 and reduces IS from 302.1 to 195.3 after 80 epochs. Replacing these spatial tokens with learned register tokens improves IS and sFID but worsens FID to 2.56. We therefore align the spatial tokens of the coarse stream in the main model.

C.5 Computational Cost

Table 13 reports the parameter count and per-forward compute of pixel-space diffusion models at 256×256. MS-pDiT-L requires 186 GFLOPs, which is 30 GFLOPs (19%) more than PixelDiT-XL at a nearly identical parameter count (788M vs. 797M), while improving FID from 1.66 to 1.55 and sFID from 4.89 to 4.42 (Table 1). Relative to JiT-G/16, MS-pDiT-L uses less than half the compute and 40% of the parameters yet attains a lower FID (1.55 vs. 1.82). It is also roughly $16\times$ cheaper than PixelFlow-XL/4 and $6\times$ cheaper than ADM-G, which process long token sequences or dense convolutional feature maps at high resolution.

D Qualitative Examples

Figures 5 and 6 present one sample for each of twenty ImageNet classes at 256×256 (MS-pDiT-XL) and at 512×512 (MS-pDiT), all generated with a CFG scale of 4.0. The classes are disjoint from those shown in Figures 2 and 3. Table 14 lists the prompts for the text-to-image samples in Figure 4. Figure 7 shows sixteen additional 512×512 text-to-image samples, covering photorealistic scenes, portraits, landscapes, painterly styles, and multi-object compositional prompts from PartiPrompts; their prompts are listed in Table 15.

Table 12: Ablation of representation alignment (REPA) on ImageNet 256×256. "Largest Patch" aligns the compact spatial tokens of the 16×16 stream with frozen DINOv2 features, whereas "Registers" aligns learned, non-spatial register tokens instead. All models are evaluated after 80 epochs.

Method	**Epochs**	**FID↓**	**IS↑**	**sFID↓**	**Prec.↑**	**Rec.↑**
Largest Patch (Ours)	80	**2.31**	302.1	4.77	0.81	**0.58**
No REPA	80	5.30	195.3	5.63	**0.83**	0.47
Registers	80	2.56	**312.5**	**4.43**	**0.83**	0.56

Table 13: Parameters and compute of pixel-space diffusion models on ImageNet 256×256. GFLOPs are measured for a single forward pass without classifier-free guidance, counting one multiply–add as one FLOP.

Method	Params	GFLOPs
ADM-G (Dhariwal & Nichol, 2021)	554M	1120
RIN (Jabri et al., 2023)	410M	334
VDM++ (Kingma & Gao, 2023)	2B	555
PixelFlow-XL/4 (Chen et al., 2025b)	677M	2909
PixNerd-XL/16 (Wang et al., 2026a)	700M	134
JiT-G/16 (Li & He, 2026)	2B	383
PixelDiT-XL (Yu et al., 2026)	797M	156
MS-pDiT-L	788M	186



Figure 5: Class-conditional samples from MS-pDiT-XL on ImageNet 256×256. Classes from left to right, top to bottom: robin (15), sulphur-crested cockatoo (89), hornbill (93), hummingbird (94), white stork (127), flamingo (130), common iguana (39), African chameleon (47), starfish (327), anemone fish (393), rhinoceros beetle (306), chickadee (19), lion (291), Pembroke corgi (263), Persian cat (283), snow leopard (289), brown bear (294), meerkat (299), fox squirrel (335), and gorilla (366).

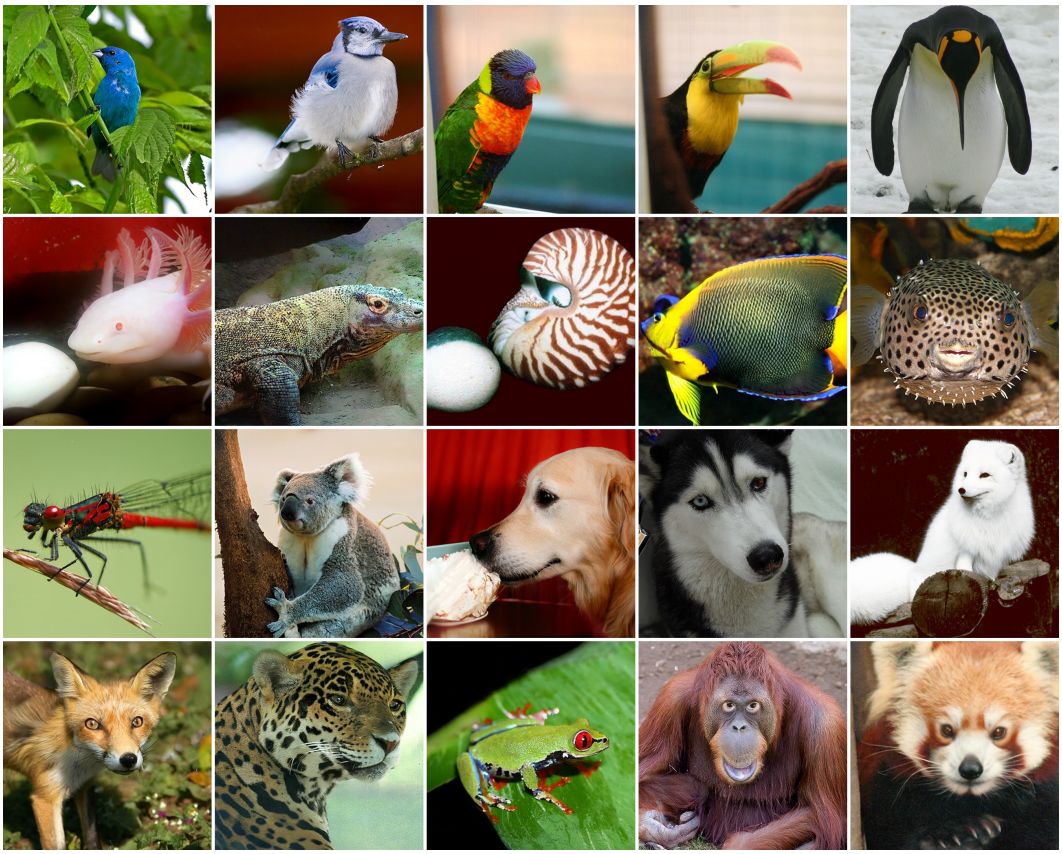

Figure 6: Class-conditional samples from MS-pDiT on ImageNet 512×512. Classes from left to right, top to bottom: indigo bunting (14), jay (17), lorikeet (90), toucan (96), king penguin (145), axolotl (29), Komodo dragon (48), chambered nautilus (117), rock beauty (392), puffer (397), damselfly (320), koala (105), golden retriever (207), Siberian husky (250), Arctic fox (279), red fox (277), jaguar (290), tree frog (31), orangutan (365), and lesser panda (387).

Table 14: Prompts for Figure 4, ordered from left to right.

#	Prompt
1	a painting of a fox in the style of starry night.
2	a close-up of the eyes of an owl.
3	A fluffy baby sloth with a knitted hat trying to figure out a laptop, close up, highly detailed, studio lighting, screen reflecting in its eyes.
4	an old man with a long grey beard and green eyes.
5	A photo of a light bulb in outer space traveling the galaxy with a sailing boat inside the light bulb.
6	a sport car melting into a clock, surrealist painting in the style of Salvador Dali.
7	A photograph of the inside of a subway train. There are red pandas sitting on the seats. One of them is reading a newspaper. The window shows the jungle in the background.
8	An aerial view of Ha Long Bay.
9	A richly textured oil painting of a young badger delicately sniffing a yellow rose next to a tree trunk. A small waterfall can be seen in the background.
10	a portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. A full moon over the city of Los Angeles is in the background at night.



Figure 7: Additional text-to-image samples from MS-pDiT at 512×512. Prompts are taken from PartiPrompts and are listed, from left to right and top to bottom, in Table 15.

Table 15: Prompts for Figure 7, ordered from left to right and top to bottom.

#	Prompt
1	A smiling sloth is wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The sloth is holding a quarterstaff and a big book. The sloth is standing on grass a few feet in front of a shiny VW van with flowers painted on it. wide-angle lens from below.
2	a painting of street in Paris
3	A teddy bear wearing a motorcycle helmet and cape is standing in front of Loch Awe with Kilchurn Castle behind him
4	a portrait of young girl
5	A punk rock squirrel in a studded leather jacket shouting into a microphone while standing on a stump and holding a beer on dark stage.
6	Snow mountain and tree reflection in the lake
7	A wombat sits in a yellow beach chair, while sipping a martini that is on his laptop keyboard. The wombat is wearing a white panama hat and a floral Hawaiian shirt. Out-of-focus palm trees in the background.
8	a blue cow is standing next to a tree with red leaves and yellow fruit. the cow is standing in a field with white flowers. impressionistic painting
9	a photograph of an ostrich wearing a fedora and singing soulfully into a microphone
10	A train ride in the monsoon rain in Kerala. With a Koala bear wearing a hat looking out of the window. There is a lot of coconut trees out of the window.
11	a whale breaching near a mountain
12	A raccoon wearing formal clothes, wearing a tophat and holding a cane. The raccoon is holding a garbage bag. Oil painting in the style of Hokusai.
13	A photo of an astronaut riding a horse in the forest. There is a river in front of them with water lilies.
14	A high resolution photo of a large bowl of ramen. There are several origami boats in the ramen of different colors.
15	a light shining on a giraffe in a street
16	A cozy living room with a painting of a corgi on the wall above a couch and a round coffee table in front of a couch and a vase of flowers on a coffee table