Diffusion Image Generation with Explicit Modeling of Data Manifold Geometry

Duoduo Xue[1], Zhiyu Zhu[1,2]*, Junhui Hou[1]*

[1] Department of Computer Science, City University of Hong Kong, China
[2] City University of Hong Kong (Dongguan), China
duoduxue@cityu.edu.hk, zhiyu.zhu@cityu-dg.edu.cn, jh.hou@cityu.edu.hk

Abstract. Image generative models aim to sample data points from the underlying data manifold, a task that requires learning and decoding a dense, low-dimensional, and compact parameterization space. To achieve this, we propose the Data Manifold-aware Image diffusioN moDel (MIND), a novel framework that explicitly models manifold geometry by integrating discrete patch tokenization into the score function of a continuous diffusion model. This approach successfully leverages both the structural quantification capabilities of discrete tokens and the parallel generation flexibility of continuous diffusion. Moreover, we enable end-to-end differentiable training via a novel soft top-k aggregation mechanism and introduce dual-branch high-frequency feature embedding layers to alleviate the spectral bias of transformer backbones on low-dimensional inputs. Furthermore, for inference, we design a multi-stage transition sampling scheme that dynamically adjusts the sampling scheme based on timestep. Extensive experiments on ImageNet 256×256 demonstrate the effectiveness of MIND. After 80-epoch training, our base model achieves an FID of 22.73 without guidance, nearly halving the 43.47 FID of the vanilla DiT-B/2 baseline. The proposed method reduces FID by 15.95 and 9.06 on average compared with the baselines DiT and SiT, respectively. For image generation on ImageNet 256×256 with guidance, the proposed MIND-B with only 130M parameters achieves an FID of 2.06, surpassing LlamaGen-3B with 3.1B parameters. Our MIND-XL with 715M parameters further reduces the FID to 1.95. Our MIND introduces a fresh perspective on diffusion-based image generation, paving the way for future research and innovation in this community. The code is available here: `https://github.com/xddgit/MIND`.

Keywords: Image generation, Discrete diffusion model, Data manifold, Image tokenizer

1 Introduction

Deep generative models, encompassing generative adversarial networks, variational autoencoders, and normalizing flows, have fundamentally revolutionized

* Corresponding authors.

image synthesis [16, 25, 26]. In recent years, continuous diffusion models and score-based generative models [3,12,17,23,38,48,50], particularly latent diffusion models [4, 42] and Diffusion Transformers (DiT) [35, 40], have become the standard for high-fidelity image generation. Despite their impressive performance, these models typically operate within unbounded Euclidean spaces with infinite continuous states. While theoretical analyses heavily rely on the low-dimensional manifold hypothesis—which posits that high-dimensional natural images reside on a low-dimensional topological structure [2,7,9,33,41,52,54], existing continuous diffusion algorithms rarely model this geometric prior explicitly, aside from a few specialized Riemannian diffusion formulations [10, 27]. Consequently, image generation with standard continuous diffusion models is efficient but struggles to leverage the data manifold geometry optimally.

Parallel to continuous diffusion models, discrete generation paradigms, e.g., next-token prediction [32, 51] or masked diffusion models [6, 58], compress images into discrete tokens via vector quantization [13, 39, 46, 56, 65]. Both of the methods (intentionally or unintentionally) set a decoding order onto the image generation process, e.g., the 1D generation order for the next-token prediction or random order for the discrete diffusion, which is constructed based on the assumption of causality among different patches. However, different from the language data, the success of vision diffusion models [6, 17, 40, 64] has demonstrated that high-fidelity of generation relies on progressively refining each token in parallel, which also applies to consistency-like models [49]. Furthermore, image generation models under the scheme of next-token prediction cannot utilize global bidirectional information during early sampling stages due to the inherent sequential generation scheme [56].

Explicitly constraining continuous diffusion processes to a low-dimensional geometric manifold is highly desirable but fundamentally challenging. Projecting continuous latent features directly onto a strict topological bottleneck (e.g., a low-dimensional hypersphere surface) leads to catastrophic representation collapse. Because continuous autoencoders rely heavily on vector magnitude to encode structural information, the strict ℓ_2-normalization inherent to hyperspherical manifolds destroys this information, resulting in severe perceptual distortion. In contrast, discrete token indices encode information categorically. When mapped to a low-dimensional manifold, discrete tokens rely entirely on angular separation rather than magnitude, demonstrating an innate immunity to bottleneck degradation. See more analysis in Section 3.1.

Motivated by this insight, we propose a novel and general generative framework named data Manifold-aware Image diffusioN moDel (MIND), which explicitly models the data manifold by leveraging the representational robustness of discrete tokenization with the sampling flexibility of continuous diffusion. Specifically, we first discretize images using a pre-trained tokenizer. The discrete tokens are then mapped onto a continuous low-dimensional hypersphere surface, upon which the forward and reverse diffusion processes are formulated. To enable end-to-end differentiable training between the unnormalized network logits output and the hyperspherical continuous latent, we introduce a soft top-k aggregation

and projection mechanism. Furthermore, to alleviate the spectral bias when processing low-dimensional inputs, we introduce a dual-branch feature projection module. This module injects Random Fourier Features to capture high-frequency details, stabilized by a zero-initialized deep residual sinusoidal pathway. During inference, we propose a multi-stage transition sampling scheme that dynamically adjusts the sampling method based on timestep.

We evaluate the proposed method on the ImageNet 256×256 benchmark. Under a restricted computational budget of 80 training epochs, our MIND-B (130M parameters) achieves an FID of 22.73, nearly halving the 43.47 FID of the vanilla DiT-B/2 baseline. In summary, our main contributions are three-fold:

- We analyze the representation collapse phenomenon when projecting continuous latent onto a low-dimensional manifold, demonstrating the fundamental superiority of discrete tokens for low-dimensional manifold modeling.
- We propose a general data manifold-aware image diffusion model explicitly constrained to a hypersphere manifold. It features a differentiable soft top-k bridge for training and multi-stage transition sampling for inference.
- Extensive experiments demonstrate that MIND significantly improves generation quality, achieving average FID reductions of 15.95 and 9.06 over the DiT and SiT baselines, respectively, establishing it as a highly efficient generative paradigm. For image generation on ImageNet 256×256 with guidance, the proposed MIND-B with only 130M parameters achieves an FID of 2.06, surpassing LlamaGen-3B with 3.1B parameters. Our MIND-XL with 715M parameters further reduces the FID to 1.95.

2 Related Work

Diffusion Models. Diffusion models have beaten GAN and achieved great success in image generation recently [12, 50]. Start from image generation in pixel space such as DDPM [17] and DDIM [48], latent diffusion models train diffusion models in latent space with much lower dimension by encoding images with an autoencoder to get continuous latent and enable high-resolution generation, the neural backbone is implemented with a time-dependent UNet in [42] and a transformer in DiT [40]. A series of works is built on the pioneering work of DiT. Scalable Interpolant Transformers (SiT) [35] improves DiT with an interpolant framework. REPresentation Alignment (REPA) [61] aligns the diffusion model representation with the pretrained visual representation. REPA-E [30] proposes end-to-end training of the variational autoencoder (VAE) and diffusion model based on representation-alignment loss. Representation autoencoders [64] propose to replace the VAE in DiT with the well-developed and pretrained representation encoders to obtain a semantically rich latent space. Riemannian flow matching with Jacobi regularization [27] proposes to perform diffusion in the feature space of representation learning and constrain the generative process to the manifold geodesics. Deco [37] proposes to decouple the generation as generating high-frequency details and low-frequency semantics in pixel and latent spaces, respectively. To improve the efficiency of diffusion models, several works explore

one-step or few-step generation under the frameworks of consistency model [49], flow-matching [14, 15, 20], and distillation [44].

Conditional diffusion models on class labels are crucial to further improve the generation quality. Class information is incorporated into adaptive group normalization layers together with time step in [12]. Dhariwal *et al.* [12] proposed to train a classifier on noisy images and then use the gradient of the trained classifier to guide the generation model. The widely used classifier-free guidance [18] is proposed to avoid an extra classifier, which trains a single network to parameterize both the unconditional and conditional models and samples with the linear combination of the conditional and unconditional score estimations. Rombach *et al.* [42] proposed flexible conditional image generation by mapping the embedded class label to the intermediate layers of UNet with a cross-attention layer. Karras *et al.* proposed autoguidance that guides the model with an inferior version of itself [24] rather than an unconditional model as in [18], which performs better but needs to train an additional guiding model. However, the diffusion models operator in continuous space with infinite states and cannot incorporate the prior information of the data manifold.

Image generation with Language Models. Parallel to diffusion models, popular language models, such as autoregressive language models, have been extended to image generation since the great success of large language models in natural language generation. Images are firstly tokenized to discrete tokens $\boldsymbol{k}$ based on a well-developed visual tokenizer [60]. The autoregressive language models [51] predict the next token $\boldsymbol{k}_i$ using previous tokens $\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_{i-1}$ and conditional information $\boldsymbol{c}$. The training stage learns the categorical distribution $p_{\boldsymbol{\theta}}(\boldsymbol{k}_i|\boldsymbol{k}_{i-1}, c)$ by minimizing the negative log-likelihood $\mathcal{L}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{k}_i|\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_{i-1}; \boldsymbol{c})$, where $\boldsymbol{\theta}$ is the model parameters. The inference stage generates discrete tokens sequentially based on next-token prediction. The masked language models [6] learns the distribution $p_{\boldsymbol{\theta}}(\boldsymbol{k}_i|\boldsymbol{k}_j, c)$ in the training stage, where mask $\boldsymbol{m} \in \{0, 1\}, m_j = 1, \forall j, m_i = 0, \forall i$. Starts with a fully masked sequence, the inference stage alternatively samples the whole sequence with the given non-masked tokens and remasks the tokens with the lowest probability in decreasing mask ratio. eMIGM (effective and efficient Masked Image Generation Models) [58] unifies the masked diffusion models [34, 47] and masked image generation [6], which systematically explores the design space of masked image generation models. More recently, the diffusion language models have combined the advantages of diffusion models and language generation, which can be generalized to image generation directly [1, 43] by tokenizing images into discrete tokens. However, these language models restrict the whole generation process to a finite discrete space and limit the generation ability.

Several works try to combine the advantages of diffusion models and language models. The Riemannian diffusion language model (RDLM) [21] proposes a continuous diffusion model for language modeling and performs the diffusion process on the hypersphere surface, which is extended by RJF [27] to image generation in the feature space of representation learning. Our MIND is fundamentally different from RDLM [21] and RJF [27]. Specifically, in terms of *motivation*, RJF

aids convergence in high-dimensional representation spaces, whereas we embed a low-dimensional manifold prior. Regarding the *input space*, RJF is purely continuous, while we process both discrete tokens and continuous latents. Finally, concerning the *trajectory*, RDLM and RJF require complex, approximated formulations to strictly constrain trajectories to the hypersphere, whereas we allow off-sphere trajectories, only requiring data points on the surface.

Image Tokenization is the first step for image generation with autoregressive and masked language models. Images can be represented as discrete variables at the pixel level, but this is redundant and has a quadratically increasing computational cost. Image tokenizers compress images into discrete tokens. The general pipeline maps the image to a latent feature map and then quantizes the continuous features as discrete codes. A representative work is VQ-GAN [13] that constructs a learned codebook and finds the nearest codebook entry for each spatial feature vector. VQ-GAN is built on the grounding study named VQ-VAE [39] and introduces discriminator loss into the training objective. A series of works have been proposed to improve the initialization rate and size of the codebook. VQGAN-LC [65] initializes the codebook entry with the feature of the pretrained vision encoder, then keeps the static codebook and optimizes a projector. The Index Backpropagation Quantization (IBQ) updates all codebook embeddings leveraging a soft one-hot categorical distribution [46]. The scale-based tokenizer named VAR proposes a multi-scale VQ autoencoder that quantizes residuals of feature maps recursively [53]. Recently, visual language models have also been incorporated into image tokenizers for extracting semantically rich features [63]. A complementary survey of image tokenizers can be referred to [56].

3 Proposed Method

The low-dimensional manifold hypothesis is widely adopted in the theoretical analysis of image generation with diffusion models [2,7,33,41,52,54], but has not been explicitly utilized in modern generative models. In this paper, we leverage the image tokenization method as the quantification measurement of image patch manifold structure and anchor those discrete points onto a simple and tackleable geometry, as illustrated in Fig. 1.

3.1 Effectively Parameterizing the Image Manifold Geometry

Mapping images or their continuous latent representations onto a compact topological manifold presents a fundamental challenge, as capturing diverse visual information requires highly complex features. Furthermore, facilitating an effective generation process necessitates that real data samples densely populate the parameterization space. This strict high-density requirement severely exacerbates the difficulty of the learning objective.

To mitigate these challenges, we employ discrete tokenization to quantize the manifold structure, an approach that significantly reduces the reparameterization burden associated with continuous latent spaces. To validate this advan-

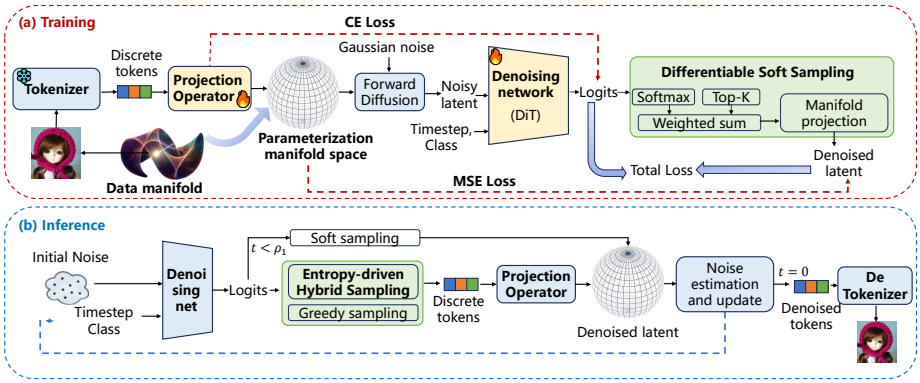

Fig. 1: Illustration of the proposed **MIND**, an image generation framework with explicit modeling of data manifold geometry.

tage, we empirically contrast two distinct methodologies: a *continuous projection* that maps patchified VAE latents directly onto a 6-dimensional hypersphere (where the exceptionally low dimensionality enforces a high-density space), and a *discrete approach* that projects VQ token indices into the identical manifold. For a rigorous and fair comparison, both methods are restricted to linear/MLP layers with the same parameter level and trained on the same randomly selected image, with carefully controlled optimization and architectural conditions (detailed in the *Supplementary material*). By decoding the reconstructed bottleneck features back into the RGB domain, we measure the Learned Perceptual Image Patch Similarity (LPIPS) [62], as illustrated in Fig. 2. Our results indicate that the continuous latent formulation suffers from representation collapse due to its heavy reliance on feature magnitude to encode structural information. In contrast, the discrete token path encodes information categorically and relies entirely on angular separation. This grants it an innate robustness against the magnitude-destructive bottleneck, consistently yielding significantly lower perceptual distortion and demonstrating the inherent superiority of discrete tokens for manifold parameterization. Experiments with higher

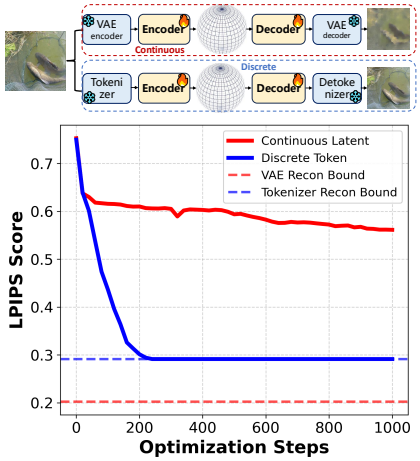

Fig. 2: Continuous (top) and discrete projection (bottom). The plot shows the LPIPS distance between the reconstructed and original images during optimization. Solid lines represent the models with manifold constraints, whereas dashed lines indicate the unconstrained reconstruction.

dimensions are shown in Fig. S-5 of the *Supplementary material*.

3.2 Diffusion with Explicit Modeling of Data Manifold

Formulation of Diffusion Process. Given an image $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W}$ sampled from data distribution p_{data}, it is tokenized as discrete token $\boldsymbol{k} \in \mathbb{R}^N$ using a tokenizer with vocabulary size V. The token $\boldsymbol{k} \in \mathbb{R}^N$ is then mapped to the continuous latent feature by a projection operator $\mathcal{P}_{\boldsymbol{\theta}}(\cdot)$ parameterized by $\boldsymbol{\theta}$ to get a continuous representation of dimension L $\boldsymbol{x}_0 \in \mathbb{R}^{N\times L}$ on the hypersphere surface satisfying $\sum_{l=0}^{L-1} \boldsymbol{x}_0^2(n,l) = R^2$ for all $n = 0, \cdots, N-1$.

The forward and reverse diffusion processes are performed based on the hypersphere surface with radius R. The forward diffusion process starts from the continuous representation $\boldsymbol{x}_0 \in \mathbb{R}^{N\times L}$ and perturbs it to get noisy latent,

$$\boldsymbol{x}_t = c_1\sqrt{1-t}\cdot\boldsymbol{w} + c_2\sqrt{t}\cdot\boldsymbol{x}_0, \tag{1}$$

where $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$,$t \in [0,1]$, the end of forward diffusion process is the Gaussian noise $\boldsymbol{x}_1 = c_1\boldsymbol{w}$. The reverse diffusion process starts from the sampled Gaussian noise $\boldsymbol{x}_1 \sim \mathcal{N}(\mathbf{0}, c_1^2\mathbf{I})$, the network $\boldsymbol{s}_{\boldsymbol{\phi}}(\cdot, t)$ parametrized by $\boldsymbol{\phi}$ gets noisy latent and outputs denoised logits $\tilde{\boldsymbol{x}}_{0t} = \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t) \in \mathbb{R}^{N\times V}$. The denoised latent $\dot{\boldsymbol{x}}_{0t}$ is predicted by sampling discrete tokens from $\tilde{\boldsymbol{x}}_{0t}$ and then mapping the discrete tokens to continuous latent.

Training Pipeline. During the training stage, to maintain the differentiability of the discrete token sampling process while ensuring the output strictly resides on the target hypersphere manifold, we introduce a soft sampling mechanism. Let $\tilde{\boldsymbol{x}}_{0t}(n) \in \mathbb{R}^V$ denote the predicted logits of the n-th token over the vocabulary of size V. We first extract the set of indices corresponding to the top-k logit values, denoted as $\mathcal{I}_k$. The normalized soft weights α_i are computed via the softmax function over these selected elements:

$$\alpha_i = \exp(x_i) / \sum_{j\in\mathcal{I}_k} \exp(x_j), \forall i \in \mathcal{I}_k, \tag{2}$$

where x_i is the i-th element of $\tilde{\boldsymbol{x}}_{0t}(n)$. Subsequently, we aggregate the corresponding continuous latent using the weights α_i. Since the convex combination of points on a hypersphere falls into the interior of the sphere, we explicitly project the aggregated feature onto the hypersphere surface via ℓ_2-normalization:

$$\dot{\boldsymbol{x}}_{0t}(n) = \frac{\sum_{i\in\mathcal{I}_k} \alpha_i \mathcal{P}_{\boldsymbol{\theta}}(i)}{\left\|\sum_{i\in\mathcal{I}_k} \alpha_i \mathcal{P}_{\boldsymbol{\theta}}(i)\right\|_2}, \tag{3}$$

where $\dot{\boldsymbol{x}}_{0t} \in \mathbb{R}^{N\times L}$ is the denoised continuous latent at timestep t, seamlessly bridging the discrete token space and the continuous diffusion manifold.

Based on the diffusion process formulated with the explicitly modeled hypersphere manifold prior, the neural network $\boldsymbol{s}_{\boldsymbol{\phi}}$ is optimized by minimizing the

Algorithm 1 Training Process of MIND

Input: Dataset p_{data}, pre-trained tokenizer, initialized networks $\boldsymbol{s}_{\boldsymbol{\phi}}$ and $\mathcal{P}_{\boldsymbol{\theta}}$, noise schedules c_1, c_2, timestep range $[t_{min}, t_{max}]$.

repeat
  Sample image $\boldsymbol{I} \sim p_{data}$;
  Extract discrete tokens: $\boldsymbol{k} = \text{Tokenizer}(\boldsymbol{I})$;
  Project to hypersphere manifold: $\boldsymbol{x}_0 = \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{k})$;
  Sample timestep $t \sim \mathcal{U}[t_{min}, t_{max}]$ and Gaussian noise $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  Compute noisy latent: $\boldsymbol{x}_t = c_1\sqrt{1-t} \cdot \boldsymbol{w} + c_2\sqrt{t} \cdot \boldsymbol{x}_0$;
  Predict unnormalized logits: $\tilde{\boldsymbol{x}}_{0t} = \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t)$;
  Obtain denoised latent $\dot{\boldsymbol{x}}_{0t}$ via soft sampling over $\tilde{\boldsymbol{x}}_{0t}$;
  Compute loss: $\mathcal{L}(\boldsymbol{s}_{\boldsymbol{\phi}}, \mathcal{P}_{\boldsymbol{\theta}}) = \text{CE}(\tilde{\boldsymbol{x}}_{0t}, \boldsymbol{k}) + \text{MSE}(\dot{\boldsymbol{x}}_{0t}, \boldsymbol{x}_0)$;
  Take gradient descent step on $\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\mathcal{L}$ and update $\boldsymbol{\theta}, \boldsymbol{\phi}$;
until network converges

cross-entropy (CE) loss between denoised logits $\tilde{\boldsymbol{x}}_{0t}$ and the discrete tokens $\boldsymbol{k}$ and the mean squared error (MSE) loss between $\boldsymbol{x}_0$ and the denoised latent $\dot{\boldsymbol{x}}_{0t}$:

$$\mathcal{L}(\boldsymbol{s}_{\boldsymbol{\phi}}, \mathcal{P}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{k}\sim p_{data}, t\in[0,1]}\{\text{CE}(\tilde{\boldsymbol{x}}_{0t}, \boldsymbol{k}) + \lambda \cdot \text{MSE}(\dot{\boldsymbol{x}}_{0t}, \boldsymbol{x}_0)\}, \tag{4}$$

where $\tilde{\boldsymbol{x}}_{0t} = \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t)$, $\boldsymbol{x}_t = c_1\sqrt{1-t} \cdot \boldsymbol{w} + c_2\sqrt{t} \cdot \boldsymbol{x}_0$, $\boldsymbol{x}_0 = \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{k})$, λ controls the strength of MSE loss. Algorithm 1 demonstrates the training pipeline of the proposed method.

Inference Pipeline. During the inference stage, the denoised logits are sampled to obtain the denoised latent $\dot{\boldsymbol{x}}_{0t} = \mathcal{P}_{\boldsymbol{\theta}}(\text{S}(\tilde{\boldsymbol{x}}_{0t}))$. To ensure that the generated token sequences remain in the high-probability manifold of the codebook space while preserving diversity in the early stage of the generation, we propose a parameterized multi-stage transition sampling strategy with thresholds ρ_1 and ρ_2. In the initial phase ($t < \rho_1$), the operator $\mathcal{P}(\text{S}(\cdot))$ inherits the soft sampling mechanism from the training stage in Eq. (3). During the middle phase ($\rho_1 < t < \rho_2$), the operator $\text{S}(\cdot)$ is implemented by an entropy-aware discrete hybrid-filtering scheme. For the given distribution $\boldsymbol{p} = \text{Softmax}(\tilde{\boldsymbol{x}}_{0t})$, we compute the Shannon entropy H. The sampling temperature is scaled adaptively: $\tau_{adj} = \tau \cdot (2.5e^{-H/3} + 0.6)$ following [36], forcing the model to perform confident sampling when the predictive entropy is high. We then apply a hybrid filter combining Top-k and Nucleus (Top-p) constraints, drawing the discrete token from the renormalized distribution. In the terminal phase ($t > \rho_2$), the operator $\text{S}(\cdot)$ collapses to greedy sampling to eliminate stochasticity and ensure maximal local consistency. The detailed sampling scheme is summarized in Algorithm 3. The denoised latent $\dot{\boldsymbol{x}}_{0t} \in \mathbb{R}^{N\times L}$ is perturbed to get the input of the next timestep,

$$\boldsymbol{x}_{t+\Delta t} = c_2\sqrt{t+\Delta t} \cdot \dot{\boldsymbol{x}}_{0t} + c_1\sqrt{1-(t+\Delta t)} \cdot (\eta \boldsymbol{w}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t) + (1-\eta)\boldsymbol{w}), \tag{5}$$

where $\boldsymbol{w}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t) = (\boldsymbol{x}_t - c_2\sqrt{t} \cdot \dot{\boldsymbol{x}}_{0t})/(c_1\sqrt{1-t})$, $\eta \in [0, 1]$. Algorithm 2 demonstrates the inference pipeline of the proposed method.

Network Architecture. The projector operator $\mathcal{P}_{\boldsymbol{\theta}}$ extracts the continuous latent $\boldsymbol{x}_0 \in \mathbb{R}^{N\times L}$ via an embedding layer, where each L-dimensional vector is

Algorithm 2 Inference Process of MIND

Input: Trained networks $\boldsymbol{s_\phi}$ and $\mathcal{P}_{\boldsymbol{\theta}}$, noise schedules c_1, c_2, variance control coefficient $\eta \in [0,1]$, sampling timestep schedule, thresholds ρ_1, ρ_2.

Sample initial noise: $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, c_1^2\mathbf{I})$;
for t=0 **to** 1 **do**
  Predict unnormalized logits: $\tilde{\boldsymbol{x}}_{0t} = \boldsymbol{s_\phi}(\boldsymbol{x}_t, t)$;
  if $t < \rho_1$ **then**
    $\dot{\boldsymbol{x}}_{0t} \leftarrow$ Soft sampling via Eq. (3);
  else
    Sample discrete tokens with Algorithm 3: $\hat{\boldsymbol{k}}_t = \mathrm{S}(\tilde{\boldsymbol{x}}_{0t})$;
    Project back to hypersphere manifold: $\dot{\boldsymbol{x}}_{0t} = \mathcal{P}_{\boldsymbol{\theta}}(\hat{\boldsymbol{k}}_t)$;
  end if
  Estimate noise component: $\boldsymbol{w_\phi}(\boldsymbol{x}_t, t) = (\boldsymbol{x}_t - c_2\sqrt{t}\cdot\dot{\boldsymbol{x}}_{0t})/(c_1\sqrt{1-t})$;
  Sample stochastic noise $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  Compute latent for the next timestep:
    $\boldsymbol{x}_{t+\Delta t} = c_2\sqrt{t+\Delta t}\cdot\dot{\boldsymbol{x}}_{0t} + c_1\sqrt{1-(t+\Delta t)}\cdot\big(\eta\boldsymbol{w_\phi}(\boldsymbol{x}_t, t) + (1-\eta)\boldsymbol{w}\big)$;
end for
Return DeTokenizer($\hat{\boldsymbol{k}}_1$).

Algorithm 3 Multi-Stage Transition Sampling Operator S($\cdot$)

Input: Logits $\tilde{\boldsymbol{x}}_{0t}$, $t \in [0,1]$, thresholds ρ_1, ρ_2, base temperature τ, params p, k.
Output: Sampled token $\hat{\boldsymbol{k}}_t$.
if $\rho_1 \le t < \rho_2$ **then**
  {**Entropy-Driven Hybrid Sampling**}
  Predict entropy: $\boldsymbol{p} \leftarrow \mathrm{Softmax}(\tilde{\boldsymbol{x}}_{0t})$, $H \leftarrow -\sum \boldsymbol{p}\log\boldsymbol{p}$;
  Adaptive temperature: $\tau_{adj} \leftarrow \tau\cdot(2.5\cdot e^{-H/3} + 0.6)$;
  $\mathcal{V} \leftarrow \mathrm{TopK}(\tilde{\boldsymbol{x}}_{0t}/\tau_{adj}, k) \cap \mathrm{Nucleus}(\boldsymbol{p}, p)$;
  $\hat{\boldsymbol{k}}_t \sim \mathrm{Multinomial}\,(\mathrm{Softmax}(\tilde{\boldsymbol{x}}_{0t}/\tau_{adj}|\mathcal{V}))$;
else if $t \ge \rho_2$ **then**
  Greedy sampling: $\hat{\boldsymbol{k}}_t \leftarrow \arg\max(\tilde{\boldsymbol{x}}_{0t})$;
end if
return $\hat{\boldsymbol{k}}_t$.

factored into d_{sub}-dimensional hyperspherical subspaces through sub-vector normalization. The geometrically constrained $\boldsymbol{x}_0$ is then perturbed to yield the noisy latent $\boldsymbol{x}_t$. The denoising network $\boldsymbol{s_\phi}$ is implemented based on the DiT. Before feeding into the DiT backbone, the continuous latent is processed through two branches named the base high-frequency mapper and the residual sinusoidal projection module. The high-frequency mapper maps the input $\boldsymbol{x}_t$ using a learnable projection matrix $\boldsymbol{B} \in \mathbb{R}^{L\times d_{map}}$, which is initialized from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with a large variance to amplify high-frequency signals. The projected features are then transformed by sine and cosine functions,

$$\boldsymbol{f}_{base} = [\sin(2\pi\boldsymbol{x}_t\boldsymbol{B}), \cos(2\pi\boldsymbol{x}_t\boldsymbol{B})]\,. \tag{6}$$

The concatenated features $\boldsymbol{f}_{base} \in \mathbb{R}^{2d_{map}}$ are subsequently passed through a shallow Multi-Layer Perceptron (MLP) with Layer Normalization (LN) and

GELU activations to yield the base hidden states $\boldsymbol{h}_{base} \in \mathbb{R}^{d_{hidden}}$, matching the hidden dimension of the DiT backbone. The residual sinusoidal projection module expands the continuous latent $\boldsymbol{x}_t$ using a deterministic sinusoidal positional encoding to get a high-dimensional vector $\boldsymbol{f}_{freq}$, which is passed through an MLP equipped with LN and SiLU activations. Crucially, to ensure stable gradient flow and prevent catastrophic divergence at initialization, the weight matrix of the final linear layer in this residual MLP is strictly initialized to zero. Finally, the outputs from both branches are aggregated via a residual addition $\boldsymbol{h}_{input} = \boldsymbol{h}_{base} + \boldsymbol{h}_{res}$ before interacting with the timestep and class conditions. The backbone network utilizes the transformer-based diffusion architecture in DiT, which scales across multiple model capacities. Conditioning is handled via an adaptive LayerNorm that injects combined timestep and class label representations into each Transformer block, while spatial dependencies are modeled using rotary positional embeddings.

4 Experiments

4.1 Experimental Settings

To evaluate our proposed method, we followed the setup of DiT [40] and SiT [35] to perform experiments on the standard ImageNet dataset [11]. Each image was processed to the resolution of 256×256 and encoded into discrete token sequences of length 256 using the pre-trained tokenizers Index Backpropagation Quantization (IBQ) [46] or GigaTok [57] with a vocabulary size of V. These discrete tokens are mapped to continuous latent features on the hypersphere surface. The model was trained within the continuous feature space using the forward noising formulation in Eq. (1) and the hybrid loss combining cross-entropy and MSE defined in Eq. (4). To enable classifier-free guidance during inference, we randomly replaced class labels with an unconditional token with a probability of $p_{drop} = 0.1$ during training. All models were optimized using the AdamW optimizer, utilizing chunked gradient checkpointing to optimize memory consumption. Detailed training hyperparameters, including specific learning rates and batch sizes for each experimental setting, are comprehensively summarized in Table S-8 of the *Supplementary material*. During the inference phase, we employed a custom SDE solver in Algorithm 2 with 250 denoising steps to iteratively map Gaussian noise back to the data manifold. At each timestep t, the backbone network $\boldsymbol{s}_{\boldsymbol{\phi}}$ processed the continuous latent state to predict the unnormalized logit distribution over the discrete token vocabulary. We applied classifier-free guidance with scale cfg for condition alignment.

4.2 Performance Evaluation

We report standard generative metrics: Fréchet Inception Distance (FID), Inception Score (IS), Precision, and Recall on 50,000 samples that are generated with class-balanced sampling following [64].

Table 1: Class-conditional image generation without guidance on Imagenet 256× 256. Bold values indicate the best performance. "N/A" denotes that the metric was not reported in the original publication. All models are trained for 80 epochs.

	#Params	FID ↓	IS ↑	Precision ↑	Recall ↑
DiT-S/2 [40]	33M	68.40	N/A	N/A	N/A
SiT-S/2 [35]	33M	57.64	24.78	0.41	0.60
MIND-S(Our)	∼35M	**40.72**	**31.51**	**0.48**	**0.61**
eMIGM-S [58]	97M	44.47	19.82	0.49	0.57
DiT-B/2 [40]	130M	43.47	N/A	N/A	N/A
SiT-B/2 [35]	130M	33.02	43.71	0.53	**0.63**
MIND-B(Our)	∼ 130M	**22.73**	**56.17**	**0.55**	0.58
DiT-L/2 [40]	458M	23.33	N/A	N/A	N/A

Image Generation without Guidance. We performed a comprehensive system-level evaluation comparing our proposed model against the vanilla DiT baseline. To ensure a fair comparison under a constrained computational budget, all models were evaluated after training for 80 epochs on the ImageNet 256×256 dataset without classifier-free guidance. Table S-8 of the *Supplementary material* summarizes the hyperparameters during inference. As shown in Table 1, the proposed MIND significantly outperforms the vanilla DiT, demonstrating vastly superior performance. Specifically, at the small (S) scale, the proposed MIND-S achieves a remarkable FID of 40.72, yielding a massive absolute improvement of 27.68 over DiT-S/2 (68.40). This trend is further amplified at the base (B) scale, where the proposed MIND-B reaches an FID of 22.73, nearly halving the FID of DiT-B/2 (43.47). Furthermore, compared to more advanced frameworks such as the flow-based SiT and the discrete diffusion-based model eMIGM [58], the proposed MIND establishes state-of-the-art FID while maintaining highly competitive diversity. It should be noted that the proposed method using 130M parameters even outperforms DiT-L with 458M parameters.

Image Generation with Guidance. Table 2 presents the quantitative results for class-conditional image generation with guidance on ImageNet 256×256. In the absence of publicly available 80-epoch performance and checkpoints for DiT, SiT, and eMIGM, we report results by our training from scratch following the default hyperparameter configurations specified in the original works. We sampled with different classifier-free guidance scale cfg and default hyperparameters in the original works. eMIGM was sampled with the recommended higher cfg since it adopts the time interval guidance strategy by default. The proposed MIND demonstrates superior generative performance across different model scales and classifier-free guidance scales. Under the small-scale setting (∼35M parameters), the proposed MIND-S significantly outperforms the baseline models. Specifically, at $cfg = 1.5$, MIND-S improves the FID by 22.09 and 13.29 compared with DiT-S/2 and SiT-S/2, respectively. MIND-S surpasses eMIGM-S with only one-third

Table 2: Class-conditional image generation with guidance on ImageNet 256×256. Bold values indicate the best performance. All models were trained for 80 epochs. Our method achieves a superior FID of 7.97, outperforming the baseline DiT (10.95) and flow-based SiT (9.07).

	#Params	cfg	FID ↓	IS ↑	Precision ↑	Recall ↑
DiT-S/2 [40]	33M	1.5	44.42	34.57	0.47	0.56
SiT-S/2 [35]	33M	1.5	35.62	43.00	0.52	**0.56**
MIND-S(Our)	∼35M	1.5	**22.33**	**60.32**	**0.63**	0.52
DiT-S/2 [40]	33M	2.0	29.29	55.37	0.57	0.51
SiT-S/2 [35]	33M	2.0	22.96	67.81	0.62	**0.51**
MIND-S(Our)	∼35M	2.0	**14.93**	**92.65**	**0.74**	0.45
eMIGM-S [58]	97M	1.3	41.18	21.23	0.50	0.56
eMIGM-S [58]	97M	4.0	23.52	30.83	0.61	0.51
DiT-B/2 [40]	130M	1.5	20.01	73.00	0.65	0.56
SiT-B/2 [35]	130M	1.5	16.88	84.26	0.66	**0.56**
MIND-B(Our)	∼ 130M	1.5	**12.15**	**100.16**	**0.71**	0.51
DiT-B/2 [40]	130M	2.0	10.95	119.12	0.76	0.47
SiT-B/2 [35]	130M	2.0	9.07	137.07	0.77	**0.47**
MIND-B(Our)	∼ 130M	2.0	**7.97**	**154.20**	**0.81**	0.43

of the parameters. Scaling up to the base configuration (∼130M parameters), the proposed MIND-B consistently exhibits strong capabilities. At $cfg = 1.5$, MIND-B reaches the best FID of 12.15 and the highest IS of 100.16. When the guidance scale is increased to 2.0, our model reduces FID by 2.98 and 1.1 compared with DiT-B/2 and SiT-B/2, respectively.

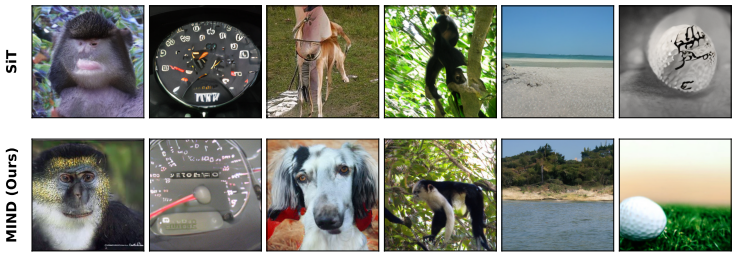

Fig. 3: Visual comparison between our MIND-B and SiT-B/2 with cfg=2.0.

Fig. 3 illustrates a qualitative comparison between SiT-B/2 and the proposed MIND-B, using the same class labels with $cfg = 2.0$. Our method synthesizes significantly more realistic images with precise semantic alignment, effectively avoiding the structural distortions observed in SiT-B/2. These visual results align with our superior quantitative scores in FID, IS, and Precision, which

Table 3: Quantitative comparison of image generation models on ImageNet 256×256. With only ∼130M parameters, our method achieves performance comparable to or exceeding discrete models with billions of parameters and Continuous models with hundreds of millions of parameters. "-re": rejection sampling, "-G": GigaTok, *: taken from VAR [53].

Method	Family	#Params	Epochs	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Discrete Models							
VQGAN [13]	AR	1.4B	N/A	15.78	74.3	N/A	N/A
VQGAN-re [13]	AR	1.4B	N/A	5.20	280.3	N/A	N/A
ViTVQ [59]	AR	1.7B	360	4.17	175.1	N/A	N/A
ViTVQ-re [59]	AR	1.7B	360	3.04	227.4	N/A	N/A
RQTran. [29]	AR	3.8B	N/A	7.55	134.0	N/A	N/A
RQTran.-re [29]	AR	3.8B	N/A	3.80	323.7	N/A	N/A
LlamaGen-B [51]	AR	111M	300	5.46	193.61	0.83	0.45
LlamaGen-L [51]	AR	343M	300	3.07	256.06	0.83	0.52
LlamaGen-XL [51]	AR	775M	300	2.62	244.08	0.80	0.57
LlamaGen-XXL [51]	AR	1.4B	300	2.34	253.90	0.80	0.59
LlamaGen-3B [51]	AR	3.1B	300	2.18	263.33	0.81	0.58
VAR-d16 [53]	VAR	310M	200-350	3.30	274.4	0.84	0.51
VAR-d20 [53]	VAR	600M	200-350	2.57	302.6	0.83	0.56
VAR-d24 [53]	VAR	1.0B	200-350	2.09	312.9	0.82	0.59
VAR-d30 [53]	VAR	2.0B	200-350	1.92	323.1	0.82	0.59
IBQ-B [46]	AR	342M	300	2.88	254.73	0.84	0.51
IBQ-L [46]	AR	649M	350	2.45	267.48	0.83	0.52
IBQ-XL [46]	AR	1.1B	400	2.14	278.99	0.83	0.56
IBQ-XXL [46]	AR	2.1B	450	2.05	286.73	0.83	0.57
MaskGIT [6]	Masked	227M	300	6.18	182.1	0.80	0.51
MaskGIT-re [6]	Masked	227M	300	4.02	355.6	N/A	N/A
MIND-B(Ours)	**Diff.+Dis.**	**∼130M**	896	**2.21**	**260.54**	**0.80**	**0.58**
MIND-B(Ours)	**Diff.+Dis.**	**∼130M**	1000	**2.18**	**258.46**	**0.80**	**0.58**
MIND-B-G(Ours)	**Diff.+Dis.**	**∼130M**	1600	**2.06**	**268.03**	**0.78**	**0.62**
MIND-XL(Ours)	**Diff.+Dis.**	**∼715M**	1000	**1.97**	**297.93**	**0.78**	**0.61**
MIND-XL-G(Ours)	**Diff.+Dis.**	**∼715M**	1600	**1.95**	**293.78**	**0.75**	**0.67**
Continuous Models							
BigGAN [5]	GAN	112M	N/A	6.95	224.5	0.89	0.38
GigaGAN [22]	GAN	569M	124	3.45	225.5	0.84	0.61
StyleGan-XL [45]	GAN	166M	N/A	2.30	265.1	0.78	0.53
ADM-G [12]	Diff.	554M	400	4.59	186.70	0.82	0.52
ADM-G, ADM-U [12]	Diff.	554M	400	3.94	215.84	0.83	0.53
LDM-4-G [42]	Diff.	400M	167	3.60	247.7	0.87	0.48
DiT-L/2* [40]	Diff.	458M	1400	5.02	167.2	0.75	0.57
DiT-XL/2 [40]	Diff.	675M	1400	2.27	278.2	0.83	0.57
SimDiff [19]	Diff.	2B	800	2.77	211.8	N/A	N/A
SiT-XL/2 [35]	Diff.	675M	1400	2.06	270.3	0.82	0.59

Table 3: Quantitative comparison (continued). **: Models used large-scale pre-trained representation network.

Method	Family	#Params	Epochs	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Continuous Models							
eMIGM-XS [58]	Masked	69M	800	3.62	224.91	0.80	0.51
eMIGM-S [58]	Masked	97M	800	2.87	254.48	0.80	0.54
eMIGM-B [58]	Masked	208M	800	2.32	278.97	0.81	0.57
eMIGM-L [58]	Masked	478M	800	1.72	304.16	0.80	0.60
eMIGM-H [58]	Masked	942M	800	1.57	305.99	0.80	0.63
MAR-B [31]	MAR	208M	800	2.31	281.7	0.82	0.57
MAR-L [31]	MAR	479M	800	1.78	296.0	0.81	0.60
MAR-H [31]	MAR	943M	800	1.55	303.7	0.81	0.62
REPA** [61]	Diff.	675M	800	1.29	306.3	0.79	0.64
REPA-E** [30]	Diff.	675M	800	1.12	302.9	0.79	0.66
DDT** [55]	Diff.	675M	400	1.26	310.6	0.79	0.65
RAE** [64]	Diff.	839M	800	1.13	262.6	0.78	0.67
PixelFlow [8]	Flow	677M	320	1.98	282.1	0.81	0.60
MeanFlow-XL/2 [14]	Flow	676M	1000	2.20	N/A	N/A	N/A
IMF-XL/2 [15]	Flow	610M	800	1.54	N/A	N/A	N/A

demonstrate that prioritizing sample fidelity and precision over diversity (Recall) yields more visually compelling and structurally accurate generative outcomes. Fig. S-4 of the *Supplementary material* provides some visual results from our MIND-B with cfg=4.0. We believe that the performance of larger models will be further unlocked by increasing vocabulary size in future work.

4.3 System-level Comparisons with Prior Work

In this section, we present the ultimate performance evaluation of our model against existing state-of-the-art methods. For inference, the classifier-free guidance scale is applied in a limited interval without entropy-driven adaptive temperature as proposed in [28]. The training and inference parameters remain consistent with Table S-8 of the *Supplementary material* except for those summarized in Table S-9. MIND-B-G and MIND-XL(-G) are sampled with NPU-compatible code.

As shown in Table 3, our MIND with about 130M parameters achieves a highly competitive FID of 2.06 (IS: 268.03, Precision: 0.78), successfully outperforming mainstream baselines that are substantially larger across multiple generative paradigms. Specifically, MIND surpasses billion-parameter discrete models like LlamaGen-3B (3.1B, FID 2.18), RQTran. [29] (3.8B, FID 3.80), IBQ-XXL (2.1B, FID 2.05) [46], VAR [53] (1.0B, FID2.09).

Furthermore, our method maintains a distinct advantage over continuous architectures. Specifically, compared with its baseline DiT-XL/2 [40], MIND im-

proves the generation ability from FID=2.27 to 1.95. The proposed MIND, using a stronger tokenizer named GigaTok [22], outperforms the 2B-parameter SimDiff [19] and SiT-XL/2 [35]. Compared with recent masked generative models that rely on continuous tokens, such as eMIGM-B [58] and MAR-B [31], the proposed MIND achieves comparable performance. We anticipate that the proposed framework can achieve even greater performance by introducing a residual term to mitigate the information loss inherent in vector quantization. Ultimately, our approach challenges the conventional reliance on massive parameter scaling, demonstrating that a compact ∼130M model can effectively match or exceed the performance of state-of-the-art architectures ranging from 300M to 3B parameters.

More visual results, ablation studies, and efficiency analysis are provided in the *Supplementary material*.

5 Conclusion and Discussion

In this paper, we introduced a novel generative framework that explicitly models the data manifold by bridging discrete image tokenization with continuous hyperspherical diffusion. To enable this discrete-continuous hybrid system, we proposed a differentiable soft top-k aggregation mechanism for stable training and an entropy-driven hybrid sampling operator for robust inference. Furthermore, our dual-branch high-frequency feature mapper effectively resolves the spectral bias when processing low-dimensional inputs. Extensive experiments on ImageNet demonstrate that our approach significantly outperforms the DiT baseline, flow-based continuous models, and discrete masked diffusion models.

Note that while we instantiated our method on the basic DiT architecture to highlight its fundamental superiority, our significant gains over DiT show its immense potential for image generation. We will integrate advanced trajectory formulations, such as mean flow and consistency models, to enhance sampling efficiency and enable high-fidelity, few-step, or single-step generation. Additionally, incorporating self-supervised representation learning could further enrich feature expression and accelerate training convergence to get better performance.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62422118, and in part by the Hong Kong Research Grants Council under Grants N_CityU1114/25 and 11219324.

References

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: NeurIPS. pp. 17981–17993 (2021)

2. Azangulov, I., Deligiannidis, G., Rousseau, J.: Convergence of diffusion models under the manifold hypothesis in high-dimensions. arXiv preprint arXiv:2409.18804 (2024)
3. Bao, F., Li, C., Zhu, J., Zhang, B.: Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In: ICLR (2022)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR. pp. 22563–22575 (2023)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
6. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: CVPR. pp. 11305–11315 (2022)
7. Chen, M., Huang, K., Zhao, T., Wang, M.: Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In: ICML. pp. 4672–4712 (2023)
8. Chen, S., Ge, C., Zhang, S., Sun, P., Luo, P.: Pixelflow: Pixel-space generative models with flow. arXiv preprint arXiv:2504.07963 (2025)
9. Cui, H., Pehlevan, C., Lu, Y.M.: A solvable model of learning generative diffusion: theory and insights. In: NeurIPS. pp. 5253–5296 (2025)
10. De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y.W., Doucet, A.: Riemannian score-based generative modelling. In: NeurIPS. pp. 2406–2422 (2022)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. pp. 8780–8794 (2021)
13. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR. pp. 12873–12883 (2021)
14. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. In: NeurIPS. pp. 75460–75482 (2025)
15. Geng, Z., Lu, Y., Wu, Z., Shechtman, E., Kolter, J.Z., He, K.: Improved mean flows: On the challenges of fastforward generative models. In: CVPR. pp. 30467–30476 (2026)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020)
18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
19. Hoogeboom, E., Heek, J., Salimans, T.: Simple diffusion: End-to-end diffusion for high resolution images. In: ICML. pp. 13213–13232 (2023)
20. Huang, Y., Wang, S.H., Bertozzi, A.L., Wang, B.: RMFlow: Refined mean flow by a noise-injection step for multimodal generation. In: ICLR (2026)
21. Jo, J., Hwang, S.J.: Continuous diffusion model for language modeling. In: NeurIPS. pp. 97394–97430 (2025)
22. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: CVPR. pp. 10124–10134 (2023)
23. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS. pp. 26565–26577 (2022)

24. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. In: NeurIPS. pp. 52996–53021 (2024)
25. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4396–4405 (2019)
26. Kobyzev, I., Prince, S.J., Brubaker, M.A.: Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(11), 3964–3979 (2021)
27. Kumar, A., Patel, V.M.: Learning on the manifold: Unlocking standard diffusion transformers with representation encoders. arXiv preprint arXiv:2602.10099 (2026)
28. 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. In: NeurIPS. pp. 122458–122483 (2024)
29. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: CVPR. pp. 11523–11532 (2022)
30. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. arXiv preprint arXiv:2504.10483 (2025)
31. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. In: NeurIPS. vol. 37, pp. 56424–56445 (2024)
32. Liu, Y., Qu, L., Zhang, H., Wang, X., Jiang, Y., Gao, Y., Ye, H., Li, X., Wang, S., Du, D.K., et al.: Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction. arXiv preprint arXiv:2505.21473 (2025)
33. Loaiza-Ganem, G., Ross, B.L., Hosseinzadeh, R., Caterini, A.L., Cresswell, J.C.: Deep generative models through the lens of the manifold hypothesis: A survey and new connections. Transactions on Machine Learning Research (2024)
34. Lou, A., Meng, C., Ermon, S.: Discrete diffusion modeling by estimating the ratios of the data distribution. In: ICML. pp. 32819–32848 (2024)
35. 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. In: ECCV. pp. 23–40 (2024)
36. Ma, X., Zhao, F., Ling, P., Qiu, H., Wei, Z., Yu, H., Huang, J., Zeng, Z., Ma, L.: Towards better & faster autoregressive image generation: From the perspective of entropy. In: NeurIPS. pp. 31466–31497 (2025)
37. Ma, Z., Wei, L., Wang, S., Zhang, S., Tian, Q.: Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In: CVPR. pp. 43600–43610 (2026)
38. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171 (2021)
39. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NeurIPS. pp. 6309–6318 (2017)
40. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
41. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., Goldstein, T.: The intrinsic dimension of images and its impact on learning. In: ICLR (2021)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
43. Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., Kuleshov, V.: Simple and effective masked diffusion language models. In: NeurIPS. pp. 130136–130184 (2024)
44. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: ICLR (2022)

45. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 Conference Proceedings (2022)
46. Shi, F., Luo, Z., Ge, Y., Yang, Y., Shan, Y., Wang, L.: Scalable image tokenization with index backpropagation quantization. In: ICCV. pp. 16037–16046 (2025)
47. Shi, J., Han, K., Wang, Z., Doucet, A., Titsias, M.K.: Simplified and generalized masked diffusion for discrete data. In: NeurIPS. pp. 103131–103167 (2024)
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
49. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML. pp. 32211–32252 (2023)
50. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS. p. 11895–11907 (2019)
51. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
52. Tang, R., Yang, Y.: Adaptivity of diffusion models to manifold structures. In: Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. pp. 1648–1656 (2024)
53. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: scalable image generation via next-scale prediction. In: NeurIPS. pp. 84839–84865 (2024)
54. Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., Qu, Q.: Diffusion models learn low-dimensional distributions via subspace clustering. In: ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy (2025)
55. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. In: CVPR. pp. 40633–40642 (2026)
56. Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., Yao, Y., Shen, H., Wan, Z., Huang, J., et al.: Autoregressive models in vision: A survey. Transactions on Machine Learning Research (2025)
57. Xiong, T., Liew, J.H., Huang, Z., Feng, J., Liu, X.: Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In: ICCV. pp. 18770–18780 (2025)
58. You, Z., Ou, J., Zhang, X., Hu, J., ZHOU, J., Li, C.: Effective and efficient masked image generation models. In: ICML. pp. 72730–72746 (2025)
59. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved VQGAN. In: ICLR (2022)
60. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., Gong, B., Yang, M.H., Essa, I., Ross, D.A., Jiang, L.: Language model beats diffusion - tokenizer is key to visual generation. In: ICLR (2024)
61. 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: ICLR (2025)
62. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
63. Zheng, A., Wen, X., Zhang, X., Ma, C., Wang, T., YU, G., Zhang, X., QI, X.: Vision foundation models as effective visual tokenizers for autoregressive generation. In: NeurIPS. pp. 62656–62675 (2025)
64. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. In: ICLR (2026)

65. Zhu, L., Wei, F., Lu, Y., Chen, D.: Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. In: NeurIPS. pp. 12612–12635 (2024)