NESTED DIFFUSION MODELS USING HIERARCHICAL LATENT PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce nested diffusion models, an efficient and powerful hierarchical generative framework that substantially enhances the generation quality of diffusion models, particularly for images of complex scenes. Our approach employs a series of diffusion models to progressively generate latent variables at different semantic levels. Each model in this series is conditioned on the output of the preceding higher-level model, culminating in image generation. Hierarchical latent variables guide the generation process along predefined semantic pathways, allowing our approach to capture intricate structural details while significantly improving image quality. To construct these latent variables, we leverage a pre-trained visual encoder, which learns strong semantic visual representations, and apply a series of compression techniques, including spatial pooling, channel reduction, and noise injection, in order to control the information capacity at each level of the hierarchy. Across multiple benchmarks, including class-conditioned generation on ImageNet-1k and text-conditioned generation on the COCO dataset, our system demonstrates notable improvements in image quality, as reflected by FID scores. These improvements incur only slight additional computational cost, as more abstract levels of our hierarchy operate on lower-dimensional representations. Our method also enhances unconditional generation, narrowing the performance gap between conditional generation and unconditional generation that leverages neither text nor class labels.

Figure 1: Our proposed nested diffusion models generate images by employing a series of diffusion models to estimate hierarchical semantic representations. We illustrate this process using a 3-level hierarchical system, where images in each row are generated based on the representations of images outlined with red borders from the previous levels, along with image labels. As the hierarchy progresses, the similarity between generated images evolves from abstract semantic similarities to lower-level visual feature similarities.

1 INTRODUCTION

Generative modeling is an unsupervised technique that learns to approximate the distribution of data and can generate novel samples draws from a simple prior distribution. Significant advances have been made in generative models, including GANs (Goodfellow et al., 2014), VAEs (Kingma, 2013; Sønderby et al., 2016; Vahdat & Kautz, 2020; Pervez & Gavves, 2020; Luhman & Luhman, 2022), diffusion models (Gu et al., 2022; 2023; Zhang et al., 2023; Song et al., 2020), and normalizing flows (Papamakarios et al., 2021; Abdal et al., 2021; Wang et al., 2022), which have been proven to be

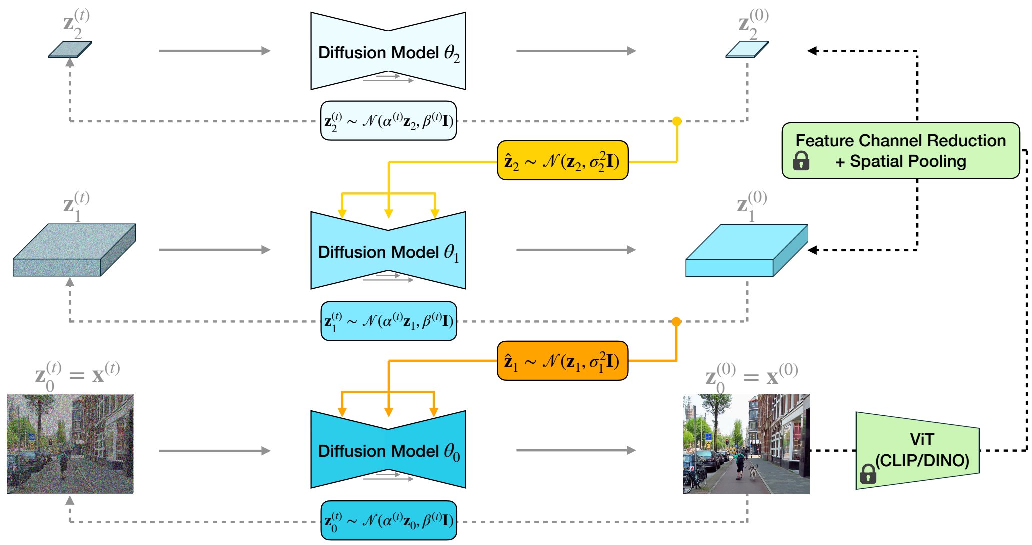


Figure 2: The diagram presents our proposed nested diffusion model, which constructs a hierarchical generative model by sequentially utilizing a series of diffusion models to produce target latent representations, ultimately the generation of final images. In the diagram, direction of arrows with solid gray lines corresponds to generative / backward process, while dotted lines correspond to how we generate training signals for different levels of the hierarchy. These hierarchical targets are obtained from visual features that are extracted using a pre-trained, frozen visual encoder. The features are then post-processed by compressing the representations via spatial pooling, reducing feature channels through singular value decomposition (SVD), and further compressing the information by parameterizing the latent features as a Gaussian distribution.

capable of modeling complex real-world images, videos, and language data (Bao et al., 2023; Nichol et al., 2021; Liu et al., 2024). These models can serve as general-purpose tools for various downstream applications (Regier et al., 2015; Smith et al., 2022; Lanusse et al., 2021; Zhao & Murphy, 2007; Osokin et al., 2017; Lopez et al., 2020).

Recent research highlights another promising aspect: the performance of these models can be enhanced by scaling up the number of model parameters, inspiring subsequent works [] that focus on building ever-larger models. However, we argue that simply increasing model parameters is not an effective solution due to the substantial gap between the data distribution and the prior distribution, as well as the complex, multimodal, and hierarchical nature of real-world data structures, which requires proper structural model design.

Classical approaches to tackle this problem are hierarchical generative modeling within the variational Autoencoders (VAEs) framework (Vahdat & Kautz, 2020; Pervez & Gavves, 2020; Takida et al., 2023), which progressively refines the prior distribution through multiple nested generation steps, enhancing the model's ability to capture complex target distributions. The key to designing such models lies in constructing progressive hierarchical levels of abstraction to guide the generation process effectively. While diffusion and autoregressive models (Yu et al., 2022) operate within this hierarchical framework, their latent variables are typically simple linear transformations of the input data, limiting their ability to generate sufficient abstraction and preserve semantic structures at output.

Conditional generative models, which integrate supplementary inputs like text, class labels, audio, or segmentation maps, demonstrate enhanced generation quality and control compared to their unconditional counterparts with no external context. The conditional input serves a similar role to the upper layers in a two-level generative system, offering high-level guidance to the lower-level generator. However, the scalability of these methods is limited by the availability of such conditional inputs during training. One example of a two-level system is Latent Diffusion (Rombach et al., 2022), which transitions the generation process from pixel space to the bottleneck representations of a VAE (Kingma, 2013), demonstrating improved generation quality through the use of more compact representations. Given that visual data naturally encompasses representations at multiple scales, it is reasonable to extend these models beyond two hierarchical levels to better handle the complexities of real-world data.

In this work, we propose a hierarchical model that employs a series of diffusion models to sequentially generate latent representations at different semantic levels, ultimately producing the final output data. We use pretrained visual encoders, such as CLIP or DINO (Caron et al., 2021), to extract feature maps that capture semantic visual representations. The dimensions of these representations are then reduced using techniques like singular value decomposition (SVD) and spatial average pooling to construct hierarchical representations along both spatial and feature channels. Since we reduce the feature dimensions at higher hierarchy levels, our hierarchical model introduces only a limited computational overhead compared to single-level variants. Throughout our experiments, we find that an effective compression scheme is critical for maintaining strong generative performance. Compared to recent works that build hierarchical diffusion models with VAE latent spaces that encode restricted semantic representations, our method demonstrates significant improvements in generation quality through the use of semantic representation. Furthermore, we quantitatively evaluate our model across various image generation tasks, demonstrating that our proposed approach significantly advances the baseline methods, especially in complex scenarios. Additionally, in text-to-image generation tasks, where text conditions offer rich semantic guidance, our method substantially enhances the overall generation quality.

2 RELATED WORKS

Hierarchical Generative Model: A hierarchical generative model has been proposed to improve generation quality by progressively refining the prior through multiple nested generation steps. In this line of research, hierarchical variational autoencoders (HVAE) (Vahdat & Kautz, 2020; Zhao et al., 2017; Child, 2020; Takida et al., 2023), which extend the latent space of VAEs (Kingma, 2013) to include multiple latent variables, demonstrate improved generation quality. However, HVAE is known to suffer from high variance and collapsed representations, where the top-level variables may be ignored (Vahdat & Kautz, 2020; Child, 2020). To address this issue, Luhman & Luhman (2022) introduced a layer-wise scheduler and network regularization to enhance stability, while Hazami et al. (2022) proposed a simplified architecture.

Recent work has sought to build hierarchical generative systems by freezing the latent variables and leveraging powerful generative models such as diffusion models and autoregressive models. For example, Ho et al. (2022); Gu et al. (2023); Liu et al. (2024) trained a set of diffusion models to handle images at different resolutions, and Tian et al. (2024) trained a hierarchical autoregressive model to predict the residuals between tokenized representations at adjacent resolutions. However, none of these approaches involve training semantic hierarchical representations.

Conditional generation: A conditional diffusion model aims to parameterize the prior as a complex joint distribution conditioned on an input, rather than using a simple Gaussian prior, which significantly enhances the model's capacity to capture intricate data patterns. For images with complex scenes, generation conditioned on image captions Gu et al. (2022); Kang et al. (2023); Reed et al. (2016) has shown notable improvements in both quality and controllability. Zhang et al. (2023); Rombach et al. (2022) extended this conditioning approach to multi-modality, incorporating inputs such as segmentation maps, depth maps, and human joint positions. Another direction in this field is learning the conditional variable itself. Models like DiffAE (Preechakul et al., 2022), SODA (Hudson et al., 2024), and Abstreiter et al. (2021) train an encoder to produce low-dimensional latent variables to assist the generation process, and these works also demonstrate that the encoder can learn meaningful image representations.

Generation with semantic visual representation: State-of-the-art generative models, such as diffusion models and autoregressive models, can be viewed as denoising autoencoders that inherently learn meaningful data representations. Research by Yang & Wang (2023); Tang et al. (2023); Zhang et al. (2024) demonstrates that diffusion models capture semantic visual representations, which can be directly applied to various downstream tasks (Baranchuk et al., 2021; Karazija et al., 2023). Additionally, Zhang & Maire (2023) highlights that the discriminator in GANs also learns strong image representations. Studies like Li et al. (2023a); Jiang et al. (2024) show that incorporating representation learning objectives into the generative framework can further enhance generation quality. Furthermore, Li et al. (2023b); Hu et al. (2023); Wang et al. (2024) leverage semantic representations learned by the encoder to improve generation quality even more.

3 METHODS

Our method employs a structured approach to capture hierarchical semantic representations for image generation. Here, we review diffusion models, one essential component of our system.

Diffusion models: A diffusion model, as a generative framework, consists of both a forward (diffusion) process and a backward processes, each spanning a total of taking place over T steps. Let $\mathbf{x} \in \mathbb{R}^d$ denote the original data sample. The forward process defines a sequence of latent variables $\{\mathbf{x}^{(t)}\}_{t=1}^{T}$ obtained by sampling from a Markrov process $q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)$, which is usually parameterized as Gaussian distribution, allowing us to sample $q\left(\mathbf{x}^{(t)}|\mathbf{x}\right) = \prod_{s=1}^{t} q(\mathbf{x}^{(s)}|\mathbf{x}^{(s-1)}) = \mathcal{N}(\mathbf{x}^{(t)}; \alpha^{(t)}\mathbf{x}, \beta^{(t)}\mathbf{I})$ in single step, where $\alpha^{(t)}$ and $\beta^{(t)}$ are hyperparameters of a noise scheduler, ensuring that the signal-to-noise ratio (SNR) decreases as t increases.

In the backward process, the model $\boldsymbol{D}_\theta$ is tasked with estimating the transition probability $p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ and generating data through the process $\prod_{t=1}^{T} p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})p(\mathbf{x}^{(T)})$, where $p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ represents the transition probability estimated by $\boldsymbol{D}_\theta$. It is trained by maximizing the Variational Lower Bound (VLB).

$$\mathcal{L}_{\text{VLB}} = -\sum_{t=1}^{T} D_{\text{KL}}\left(q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}\right) \middle\| p_\theta\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)\right). \tag{1}$$

where $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}\right)$ could be derived using Bayes' rule: $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}\right) = q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \mathbf{x}\right) q\left(\mathbf{x}^{(t-1)}|\mathbf{x}\right) / q\left(\mathbf{x}^{(t)}|\mathbf{x}\right)$. Maximizing RHS of Eqn.1 can be simplified as the training $\boldsymbol{D}_\theta$ to estimate the noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ (Ho et al., 2020):

$$\mathcal{L}_{\text{diffusion}} = \mathbf{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I}),t} \|\boldsymbol{D}_\theta(\alpha^{(t)}\mathbf{x}_0 + \beta^{(t)}\boldsymbol{\epsilon}_t, t) - \boldsymbol{\epsilon}_t\|_2. \tag{2}$$

3.1 NESTED DIFFUSION MODELS

Our proposed nested diffusion models can be seen as a hierarchical generative framework comprising L levels, each employing a diffusion model $\boldsymbol{D}_{\theta_l}$. As illustrated in Figure 2, the model at each level l is responsible for generating its corresponding latent variables $\mathbf{z}_l$. Here $\mathbf{z}_l \in \mathbb{R}^{d_l}$ and $d_l \leq d_{l+1}$, indicating decreasing amount of information when l increases. At the shallowest level of the hierarchy, level 0, the latent variables correspond directly to the data samples, that is, $\mathbf{z}_0 = \mathbf{x}$.

Diffusion with semantic hierarchy: Our design explicitly directs the generation process to follow a semantic hierarchy, where top-level (larger l) corresponds to increasing levels of semantic abstraction, while the bottom level (smaller l) correspond to fine-grained detailed information. This is essential for preserving image semantic structures and producing realistic samples in generative models. In contrast, the latent variable in standard diffusion models, $\mathbf{x}^{(t)}$, is a linear transformation of the input data $\mathbf{x}$ with added Gaussian noise. This means that information abstraction in standard diffusion models occurs at the raw pixel level, through the addition of noise to images, making it challenging for the diffusion models to maintain semantic structure in the generated output.

Markovian generation: At each hierarchical level l, we follow the diffusion model framework and task $\boldsymbol{D}_{\theta_l}$ to estimate the transition probability $p(\mathbf{z}_l^{(t-1)}|\mathbf{z}_l^{(t)}, \mathbf{z}_{l+1})$. At layer l, we assume Markovian generation that $\boldsymbol{D}_{\theta_l}$ only depends on the latent variable $\mathbf{z}_{l+1}$ estimated from the preceding hierarchy. To train our nested diffusion model, we update $\{\boldsymbol{D}_{\theta_l}\}_{l=1}^{L}$ by minimizing the objectives across all L levels and diffusion steps:

$$\begin{aligned} &\sum_{l=0}^{L-2} \sum_{t=1}^{T} D_{\text{KL}}\left(q\left(\mathbf{z}_l^{(t-1)}|\mathbf{z}_l^{(t)}, \mathbf{z}_l, \mathbf{x}\right) \middle\| p_\theta\left(\mathbf{z}_l^{(t-1)}|\mathbf{z}_l^{(t)}, \mathbf{z}_{l+1}\right)\right) \\ &- \sum_{t=1}^{T} D_{\text{KL}}\left(q\left(\mathbf{z}_l^{(t-1)}|\mathbf{z}_l^{(t)}, \mathbf{z}_L, \mathbf{x}\right) \middle\| p_\theta\left(\mathbf{z}_l^{(t-1)}|\mathbf{z}_l^{(t)}\right)\right). \end{aligned} \tag{3}$$

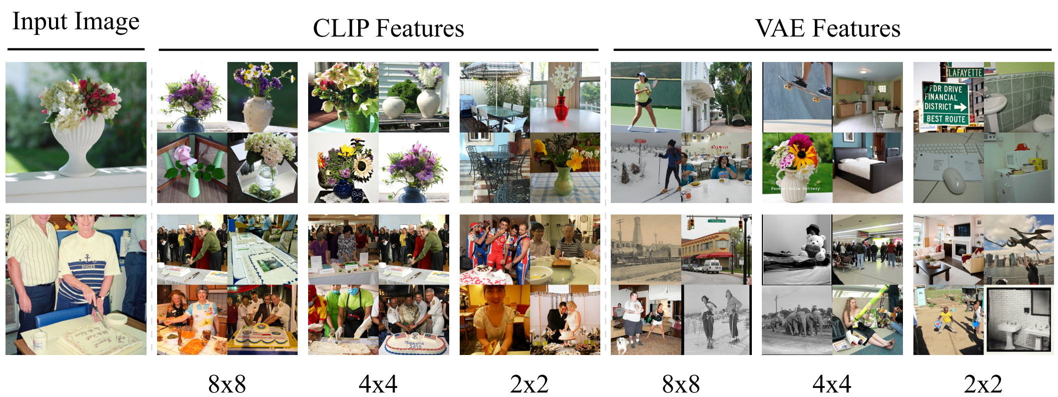


Figure 3: **Visualization of K-Nearest Neighbors (KNN) constructed using latent features**.For each input image, we display its nearest neighbors (KNNs) using features extracted from various hierarchical levels, with the respective spatial dimensions (Height × Width) indicated at the bottom. This is done across two types of visual representations: the CLIP representations and VAE bottlenecks. Unlike the VAE, CLIP learns semantic visual representations, resulting in more meaningful nearest neighbor images. While the VAE features does not produce meaningful neighbors. Using semantic representations to construct features for generation yields meanful

Drawing inspiration from hierarchical VAEs which also includes hierarchical latent variables $\{\mathbf{z}_l\}_{l=1}^{L}$, we enhance its sampling capability by integrating the diffusion model and introducing an additional set of latent variables $\{\mathbf{z}_l^t\}_{t=0}^{T}$ for each level l. This modification allows for multiple sampling steps, as opposed to the single forward pass used in hierarchical VAEs, leading to a more accurate prior estimation. This improvement is vital in hierarchical generative systems, where mismatches between the posterior and prior distributions can compound across levels, potentially degrading the quality of the generated output.

3.2 Hierarchical latent variables via progressive compression

In hierarchical VAEs, both posterior and prior distributions are represented by neural networks, and all latent variables, $\{\mathbf{z}_l\}_{l=1}^{L-1}$, are jointly optimized. This often leads to high variance, particularly in models with more hierarchical levels, as noted in previous studies (Pervez & Gavves, 2020; Vahdat & Kautz, 2020; Child, 2020). The high variance in $\{\mathbf{z}_l\}_{l=1}^{L-1}$ makes diffusion training especially challenging. The diffusion model trains to estimate the entire reverse process $\mathbf{z}_l^{(T)} \to \mathbf{z}_l^{(0)} = \mathbf{z}_l$, using intermediate variable samples $\mathbf{z}_l^{(t)}$. If $\{\mathbf{z}_l\}_L$ changes drastically, both $\mathbf{z}_l$ and $\mathbf{z}_l^{(t)}$ vary significantly during training, complicating the process.

Extraction of features: We initialize $\{\mathbf{z}_l\}_{l=1}^{L-1}$ using features from a pre-trained encoder and freeze them during training. Specifically, we use features from pre-trained models like DINO or CLIP because they learn strong semantic representations and these representations have been shown to significantly enhance the quality of generative models, including GANs (Casanova et al., 2021) and diffusion models (Hu et al., 2023; Li et al., 2023b). Alternatively, other recent methods propose to construct hierarchical diffusion models using VAE bottleneck representations, which offer highly compressed feature maps. In our experiments, we observed a substantial improvement in generation quality when using semantically rich features.

Hierarchical compression: A challenge with using DINO or CLIP features described above is that they often result in highly redundant feature maps. For example, DINO's VIT-B model produces a 14x14x768 feature map, which has the same spatial dimensions as the input image (224x224x3). Such overcomplete representations force the generative model to capture unnecessary correlations, degrading the quality of generated samples. Moreover, this redundancy can disrupt the hierarchical system. If $\mathbf{z}_l$ contain sufficient information to perfectly reconstruct the original data $\mathbf{x}$, then the lower-level latent variables $\{\mathbf{z}_{l'}\}_{l'<l}$ would be meaningless because they do not provide additional information for $\mathbf{x}$.

Therefore, designing an effective progressive compression scheme is critical for managing high-dimensional features and constructing meaningful hierarchical latent variables. Our compression routine involves three key steps:

1. Spatial dimensionality reduction via average pooling: We begin by reducing the spatial dimensions of the feature map through average pooling. This strategy has been used in previous hierarchical models based on original images and VAE bottleneck representations. However, we find that spatial pooling alone is insufficient, as it does not address redundancy in the feature channels.

2. Feature channel reduction via singular value decomposition (SVD): To tackle redundancy in the feature channels, we apply SVD along the feature dimension and retain only the top components as hierarchical features. SVD orders the feature channels by importance based on their singular values, allowing us to conveniently form hierarchical representations by trimming less important channels. To prevent the model from neglecting the trailing channels, we standardize the features to have zero mean and unit variance.

3. Information reduction through Gaussian distribution parameterization: To enhance the level of feature abstraction, we introduce Gaussian noise to $\mathbf{z}_l$, represented as $\hat{\mathbf{z}}_l \sim \mathcal{N}(\mathbf{z}_l, \sigma_l^2 \mathbf{I})$ for $l = 0, \ldots, L-2$, where $\sigma_l \in \mathbf{R}$ is a fixed value based on the hierarchical level. This process limits the amount of information that can be transmitted, which can be measured by the KL divergence $D_{KL}\left(\mathcal{N}\left(\mathbf{z}_l, \sigma_l^2\right), \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)\right)$. A large variance σ_l^2 substantially limits the information capacity. With this parameterization, the loss function becomes:

$$\begin{aligned}\mathcal{L}_{\text{nested\_diffusion}} = \sum_{l=1}^{L-2} \quad & \mathbf{E}_{\hat{\mathbf{z}}_{l+1} \sim \mathcal{N}\left(\mathbf{z}_{l+1}, \sigma_{l+1}^2 \mathbf{I}\right), \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I}), t} \| \boldsymbol{D}_{\theta_l}(\alpha^{(t)} \mathbf{z}_l + \beta^{(t)} \boldsymbol{\epsilon}_t, \hat{\mathbf{z}}_{l+1}, t) - \boldsymbol{\epsilon}_t \|_2 \\ + \quad & \mathbf{E}_{\boldsymbol{\epsilon}_t \sim \mathcal{N}(0,1), t} \| \boldsymbol{D}_{\theta_{L-1}}(\alpha^{(t)} \mathbf{z}_{L-1} + \beta^{(t)} \boldsymbol{\epsilon}_t, t) - \boldsymbol{\epsilon}_t \|_2 \end{aligned} \tag{4}$$

In our experiments, this parameterization played a vital role in maintaining and improving generation quality as the number of hierarchical levels increased.

4 Experiments

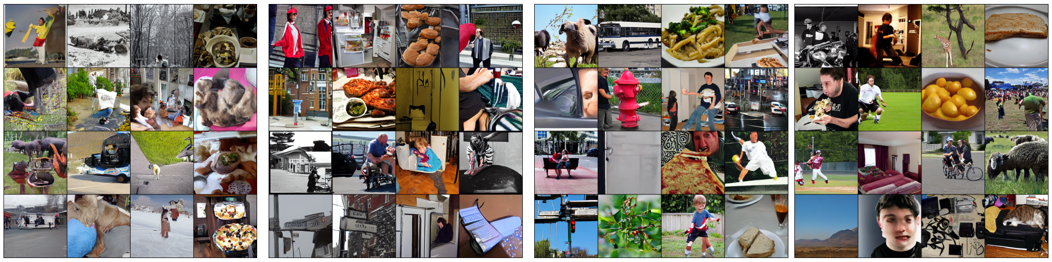
We present the setup and results of our experiments, where we evaluate the performance of our nested diffusion model across various tasks. Our primary focus is to explore the model's effectiveness in both conditional and unconditional image generation scenarios using the COCO-2014(Lin et al., 2014) and ImageNet-100 datasets (Russakovsky et al., 2015), with additional large-scale experiments on ImageNet-1k.

4.1 Experimental Setup

Nested Diffusion Models. We utilize U-ViT (Bao et al., 2023), a ViT-based UNet model with an encoder-decoder architecture, as the foundation of our nested diffusion model. This model employs skip connections and performs diffusion in the latent space of a pre-trained VAE, reducing the input size from 256x256x3 to 32x32x4, which enables efficient handling of high-resolution images. We use the default diffusion scheduler, sampler, and hyperparameters from U-ViT (Bao et al., 2023).

For constructing the nested diffusion model, we instantiate the U-ViT model at each hierarchical level, maintaining consistent configurations across all levels, except for the input data shape $\mathbf{z}^l$ and the conditional feature $\hat{\mathbf{z}}^{l+1}$. The higher hierarchical levels progressively reduce the dimensionality of $\mathbf{z}^l$, resulting in minimal additional computational overhead despite an increase in parameters. We defer further optimizations in parameter efficiency to future work.

To incorporate conditional features $\hat{\mathbf{z}}^{l+1}$, we use deconvolutional layers to upsample them to match the resolution of $\mathbf{z}^l$. These features are then concatenated as tokens every two attention blocks, followed by two fully connected layers. During training, we randomly drop the conditional features with a 50% probability to facilitate classifier-free guidance (CFG) Ho & Salimans (2022) for improving image generation quality. We use model configurations from U-ViT (Bao et al., 2023) and utilize the ViT-small, ViT-medium, and ViT-large configurations for COCO, ImageNet-100, and ImageNet-1k, respectively. Unless stated otherwise, all models are trained for 1000 epochs.



(a) $L = 1$ (b) $L = 2$ (c) $L = 3$ (d) $L = 4$

Figure 4: Unconditional image generation on the COCO dataset is performed across various hierarchical levels. At $L = 1$, it corresponds to standard diffusion models. As more levels are stacked, the generated images exhibit more coherent visual structures and improved overall image quality.

Hierarchical Latent Variables. The hierarchical latent variables $\{\mathbf{z}^l\}_{l=1}^L$ are constructed using a pre-trained visual encoder. For ImageNet experiments, we extract visual features, with the shape 14x14x768, from the final layer of MoCo-v3 (ViT-B/16), a leading self-supervised visual representation learner. For COCO experiments, we use CLIP (ViT-B/16), a multi-modal encoder that aligns visual and textual representations and also use the final visual features as our representations. We apply singular value decomposition (SVD) on the training set and retain the leading channels. Spatial average pooling is used to produce representations at varying resolutions. For COCO experiments, we generate a 5-level hierarchical latent variable structure with progressively smaller spatial and channel dimensions: $\{8 \times 8 \times 64, 6 \times 6 \times 56, 4 \times 4 \times 48, 2 \times 2 \times 40\}$. We utilize fewer levels and feature resolutions for ImageNet compared to COCO, as it's a simpler dataset. The shapes of our latent variables are: $6 \times 6 \times 32, 4 \times 4 \times 24, 2 \times 2 \times 16$

4.2 Unconditional Image Generation

To generate realistic images in an unconditional setting, a generative model must recognize the semantic structures of the images effectively. This is particularly challenging when images during the generation process are heavily corrupted, often by Gaussian noise or random masking. Traditional training objectives, usually based on pixel-wise distance, treat each pixel independently and provide no direct structural guidance in the output space, requiring the model to learn these structures internally in its latent space. If the model struggles to capture these semantic structures, the resulting output is likely to lack coherence. Our proposed approach addresses this challenge by introducing explicit semantic guidance via an external encoder that learns visual semantic representations, thus reducing the complexity of the task of the generative model.

We initially assessed the performance of the nested diffusion model on unconditional image generation tasks using the COCO-2014 and ImageNet-100 datasets. For COCO-2014, we follow the text-to-image evaluation protocol, calculating the FID between 30K generated images and those from the validation set. For ImageNet-100, where the validation set contains only 5K images - insufficient for reliable FID statistics - we use all 50K training images as a reference and compute FID on 50K generated images. We adopt the default hyperparameters for classifier-free guidance, as outlined in Bao et al. (2023), for conditional generation, substituting the ground truth text or class labels with our generated hierarchical latent variables $\hat{\mathbf{z}}^l$. We report our results in multiple depths of the model L and different conditional noise levels σ_L in Table 1. **Improved performance with more hierarchy levels** L. Compared to the baseline model, our nested diffusion model $\boldsymbol{D}_L$ produces better image quality as we deepen the hierarchy by increasing the depth L. Even though the same model configuration is applied to each level $\boldsymbol{D}_{\theta_l}$, the computational increase, measured in GFlops, remains minimal, particularly with deeper models. It is notable that as we add more hierarchical levels, the performance of unconditional image generation approaches that of conditional generation.

Impact of σ_l. As detailed in our methods section, $sigma_l$ governs the amount of information conveyed by the conditional latent variable and enforces the hierarchical structure. We validate this for $L \leq 4$, where nonzero σ_L significantly improves image quality due to the potential redundancy in $\hat{\mathbf{z}}_L$ at lower levels of the hierarchy. The optimal choice of σ_L for $L = 2$ brings even a significant improvement in image quality despite the fact that $\mathbf{z}^2$ (8x8x64) and $\mathbf{z}^1$ (32x32x4) have the same

Model	Model Config			Fréchet inception distance (FID)↓			
	size of $\mathbf{z}^L$	GFlops Growth	Params Growth	σ_L = 0.0	0.5	1.0	1.5
$L = 1$	$32 \times 32 \times 4$	22.70	44.13M	**32.73**	-	-	-
$L = 2$	$8 \times 8 \times 64$	8.54	58.06M	25.60	16.12	**13.24**	13.32
$L = 3$	$6 \times 6 \times 56$	1.42	59.58M	9.69	**8.29**	8.57	8.78
$L = 4$	$4 \times 4 \times 48$	0.72	59.51M	7.04	**6.86**	7.41	7.86
$L = 5$	$2 \times 2 \times 40$	0.71	59.48M	**6.27**	6.74	7.27	7.45
$L = 1$	text conditional generation			6.30	-	-	-

(a) Unconditional image generation on COCO-2014

Model	Model Config			Fréchet inception distance (FID)↓			
	size of $\mathbf{z}^L$	GFlops Growth	Params Growth	σ_L = 0.0	0.5	1.0	1.5
$L = 1$	$32 \times 32 \times 4$	71.60	130.7M	**44.40**	-	-	-
$L = 2$	$6 \times 6 \times 32$	30.02	100.1M	31.69	17.45	**15.31**	15.40
$L = 3$	$4 \times 4 \times 24$	1.01	100.1M	13.66	11.77	**11.12**	11.34
$L = 4$	$2 \times 2 \times 16$	0.59	100.1M	12.79	11.80	**11.21**	12.09

(b) Unconditional image generation on ImageNet-100

Table 1: Unconditional image generation results on COCO-2014 and ImageNet-100 for nested diffusion models $\boldsymbol{D}_L$. We evaluate image quality across various L (model depths) and σ_L, which determines the information capacity of the final conditional variables $\hat{\mathbf{z}}^L$. For model $\boldsymbol{D}_L$, we select the optimal σ_l values for $l < L$ from earlier levels, highlighted in bold in the table. Image quality improves with increasing model depth, with only a slight increase in computational cost, measured in GFlops, compared to the previous level. Across all L, it's important to add noise to conditional signal, especially for earlier levels of hierarchy, to ensure proper hierarchical dependency.

feature dimension. As we reduce the size of the feature to higher levels, the difference in image quality between $\sigma_L = 0$ and nonzero α_L diminishes, as the $\hat{\mathbf{z}}_L$ has a smaller dimension of the feature that carries less information.

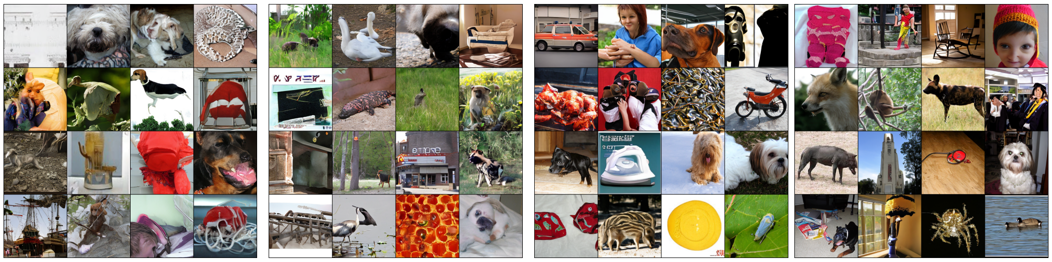
4.3 Conditional Image Generation

We also evaluated our model on conditional generation tasks, including conditional text and class generation. Text, compared to class labels, offers more detailed information, making the generation process easier. However, there are still gaps in the transfer of information, such as the shape and texture of the object, between the conditional input and the generated images. Our approach addresses these gaps through hierarchical generation, leading to improved performance.

For this experiment, we use the same setup as in the unconditional generation tasks, with results presented in Table. 2. Similar to the unconditional generation results, we observe clear performance improvements with hierarchical levels $L = 2$, and the selection of σ_2 remains crucial to overall performance.

However, the additional conditional ground truth input causes the performance gains from increasing model depth to grow more slowly compared to the unconditional task. This can be attributed to the overlap in functionality between the conditional input and the higher levels of the deeper nested diffusion models, both of which capture abstract representations.

Choices of visual representations. We examine the effect of different visual representation sources on constructing the latent variable, with the results shown in Table. 4. Instead of utilizing the encoder's representation, we experimented with using the bottleneck from a VAE. The same procedure and hyperparameters were applied to construct the hierarchical latent variable $\{\mathbf{z}_l\}_l^L$ for $L = 3$. Although VAE learns a compact bottleneck representation, it does not capture strong semantic information. Consequently, when the hierarchical latent variable is constructed by downsampling the feature dimensions, the latent space does not retain coherent semantic structures. As a result, the generation quality with $L = 3$ for VAE-based representations is inferior to our approach using MoCo-v3 in both conditional and unconditional tasks.



(a) $L = 1$ (b) $L = 2$ (c) $L = 3$ (d) $L = 4$

Figure 5: Unconditional image generation on the ImageNet-100 dataset is performed across multiple hierarchical levels. At $L = 1$, it corresponds to traditional diffusion models. As more levels are introduced, the generated images exhibit greater visual coherence and improved quality. It's important to note that this performance enhancement comes with minimal computational cost, as the feature dimensions are reduced at higher levels in the hierarchy.

	Fréchet inception distance (FID)↓			
Model	$\sigma_L = 0.0$	0.5	1.0	1.5
$L = 1$	**6.30**	-	-	-
$L = 2$	9.18	5.43	**5.24**	5.28
$L = 3$	**5.24**	5.26	5.45	5.74

(a) Conditional image generation on COCO-2014

	Fréchet inception distance (FID)↓			
Model	$\sigma_L = 0.0$	0.5	1.0	1.5
$L = 1$	**6.93**	-	-	-
$L = 2$	7.16	**4.88**	5.15	5.41
$L = 3$	**5.16**	5.99	6.47	6.92

(b) Conditional image generation on ImageNet-100

Table 2: We evaluated conditional image generation using nested diffusion models, denoted as $\boldsymbol{D}_L$, on the COCO-2014 and ImageNet-100 datasets. The evaluation focused on image quality across various model depths L and noise levels σ_L, utilizing the same hierarchical setup as in the unconditional generation experiments. Our findings indicate that nested diffusion models improve generation quality. In contrast to the unconditional case, the optimal performance was achieved at $L = 2$, likely due to the redundancy between the conditional input and the highest level of deeper nested models, both offering high-level guidance.

Recent work, RCG (Li et al., 2023b) proposes a two-level hierarchical generative system using the final output from the MoCo-v3 encoder, which is a 256-dimensional vector. Compared to our two-level system where $\mathbf{z}_2 \in \mathbb{R}^{8\times8\times256}$, RCG employs more compact feature representations. However, our approach consistently delivers better generation quality in both conditional and unconditional settings.

Large scale experiments on ImageNet 1K. To examine the performance of applying method to a larger scale dataset, we apply our approaches to ImageNet-1k. We adopt the configurations of U-ViT-L from (Bao et al., 2023) and reproduce the baseline FID as 3.8 despite their official performance is 3.4. We then takes construct $\mathbf{z}_1 \in \mathbb{R}^{6\times6\times32}$. Due to the resources constraint, we were only able to run experiments on a two level system $L = 2$ for conditional image generation and our methods improves the FID from 3.8 to 3.2 .

5 Conclusion

In this work, we introduced the nested diffusion model, a hierarchical generative framework that effectively generates images by following a semantic hierarchy. Our approach builds upon a series of hierarchical latent variables derived from pre-trained visual encoders, followed by feature compression techniques. These latent variables guide the generative process, enabling the model to capture detailed structural information while preserving high image quality. By progressively abstracting and compressing feature representations at multiple levels, we achieve significant improvements in generation performance with minimal computational overhead. Our results demonstrate that this structured, hierarchical design outperforms traditional diffusion models in both conditional and unconditional generation tasks. Rather than solely scaling model parameters, we advocate for a rethinking of generative model design that emphasizes structural organization. Future research

Model	FID	Training Dataset
DALL-E-12B (Ramesh et al., 2021)	28.00	DALL-E (250M)
CogView (Ding et al., 2021)	27.10	Internal data (30M)
GLIDE (Nichol et al., 2021)	12.24	DALL-E (250M)
DALL-E 2 (Ramesh et al., 2022)	10.39	DALL-E (250M)
Imagen (Saharia et al., 2022)	7.27	Internal Data/LAION (860M)
Re-Imagen (Chen et al., 2022)	5.25	KNN-ImageText/COCO(50M)
CM3Leon-7B (Yu et al., 2023)	4.88	Internal Data(350M)
Parti-20B (Yu et al., 2022)	**3.22**	LAION/FIT/JFT/COCO(4.8B)
VQ-Diffusion (Gu et al., 2022)	19.75	COCO(83K)
Friro (Fan et al., 2023)	8.97	COCO(83K)
U-ViT-S (Bao et al., 2023) (Ours $L = 1$)	5.95	COCO(83K)
Ours($L = 2$)	**4.74**	COCO(83K)

Table 3: Results of text conditional image generation on COCO-2014. The upper half shows larger models trained with more data and the bottom half shows the models that are only trained on training split of COCO. When trained only on COCO, our models outperform all the compared methods. It worth noting that we're better than most of the larger models, shown on the top half.

	FID↓	
Visual Representations	Cond	Uncond
None	6.93	44.40
MoCo-v3	5.16	11.12
VAE	7.24	48.23

(a) We study the impact the difference features sources for hierarchical generative model. For VAE, we adopt the same procedure and uses the same parameters to construct the $\{\mathbf{z}_l\}_l^L$

	FID↓	
Methods	Cond	Uncond
RCG	8.04	38.40
Ours (L=2)	4.88	15.31

(b) Comparison to RCG (Li et al., 2023b), a recent hierarchical genreative model with $L = 2$. It utilizes the 256-dimensional output vector from MoCo's final layer to construct latent varible $\mathbf{z}_2$

Table 4: Results on the impact of different visual representations on ImageNet-100 demonstrate that using semantic representations, rather than VAEs which primarily capture low-level features, significantly enhances generation quality. Additionally, compressing information through Gaussian noise controlled by σ_l, as opposed to RCG's use of a fixed 256-dimensional vector, is essential for achieving high-quality outputs. All experiments are running with the same network architecture with only the variation on the conditional features.

will focus on further enhancing the efficiency of these hierarchical models and expanding their applicability to a wider range of generative tasks across diverse domains.

REFERENCES

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

Korbinian Abstreiter, Sarthak Mittal, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. *arXiv preprint arXiv:2105.14257*, 2021.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 579–587, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua M Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficientvdvae: Less is more. *arXiv preprint arXiv:2203.13751*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL `https://arxiv.org/abs/2006.11239`.

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.

Vincent Tao Hu, David W Zhang, Yuki M Asano, Gertjan J Burghouts, and Cees GM Snoek. Self-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18413–18422, 2023.

Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23115–23127, 2024.

Ruoxi Jiang, Peter Y Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors. *Advances in Neural Information Processing Systems*, 36, 2024.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10134, 2023.

Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023.

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

François Lanusse, Rachel Mandelbaum, Siamak Ravanbakhsh, Chun-Liang Li, Peter Freeman, and Barnabás Póczos. Deep generative models for galaxy image simulations. *Monthly Notices of the Royal Astronomical Society*, 504(4):5543–5555, 2021.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023a.

Tianhong Li, Dina Katabi, and Kaiming He. Self-conditioned image generation via generating representations. *arXiv preprint arXiv:2312.03701*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Qihao Liu, Zhanpeng Zeng, Ju He, Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Alleviating distortion in image generation via multi-resolution diffusion models. *arXiv preprint arXiv:2406.09416*, 2024.

Romain Lopez, Adam Gayoso, and Nir Yosef. Enhancing scientific discoveries in molecular biology with deep generative models. *Molecular systems biology*, 16(9):e9198, 2020.

Eric Luhman and Troy Luhman. Optimizing hierarchical image vaes for sample quality. *arXiv preprint arXiv:2210.10205*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Anton Osokin, Anatole Chessel, Rafael E Carazo Salas, and Federico Vaggi. Gans for biological image synthesis. In *Proceedings of the IEEE international conference on computer vision*, pp. 2233–2242, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Adeel Pervez and Efstratios Gavves. Variance reduction in hierarchical variational autoencoders. 2020.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060–1069. PMLR, 2016.

Jeffrey Regier, Andrew Miller, Jon McAuliffe, Ryan Adams, Matt Hoffman, Dustin Lang, David Schlegel, and Mr Prabhat. Celeste: Variational inference for a generative model of astronomical images. In *International Conference on Machine Learning*, pp. 2095–2103. PMLR, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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.

Michael J Smith, James E Geach, Ryan A Jackson, Nikhil Arora, Connor Stone, and Stéphane Courteau. Realistic galaxy image simulation via score-based generative models. *Monthly Notices of the Royal Astronomical Society*, 511(2):1808–1818, 2022.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yuhta Takida, Yukara Ikemiya, Takashi Shibuya, Kazuki Shimada, Woosung Choi, Chieh-Hsin Lai, Naoki Murata, Toshimitsu Uesaka, Kengo Uchida, Wei-Hsiang Liao, et al. Hq-vae: Hierarchical discrete representation learning with variational bayes. *arXiv preprint arXiv:2401.00365*, 2023.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6232–6242, 2024.

Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2604–2612, 2022.

Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18938–18949, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Xiao Zhang and Michael Maire. Structural adversarial objectives for self-supervised representation learning. *arXiv preprint arXiv:2310.00357*, 2023.

Xiao Zhang, David Yunis, and Michael Maire. Deciphering'what'and'where'visual pathways from spectral clustering of layer-distributed neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4165–4175, 2024.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from generative models. *arXiv preprint arXiv:1702.08396*, 2017.

Ting Zhao and Robert F Murphy. Automated learning of generative models for subcellular location: building blocks for systems biology. *Cytometry part A*, 71(12):978–990, 2007.