EAI Endorsed Transactions
on AI and Robotics

Review Article EAI.EU

A Survey of Data-Driven 2D Diffusion Models for Generating Images from Text

Shun Fang[1,*]

[1] Peking University, Lumverse Inc, Beijing, China

Abstract

This paper explores recent advances in generative modeling, focusing on DDPMs, HighLDM, and Imagen. DDPMs utilize denoising score matching and iterative refinement to reverse diffusion processes, enhancing likelihood estimation and lossless compression capabilities. HighLDM breaks new ground with high-res image synthesis by conditioning latent diffusion on efficient autoencoders, excelling in tasks through latent space denoising with cross-attention for adaptability to diverse conditions. Imagen combines transformer-based language models with HD diffusion for cutting-edge text-to-image generation. It uses pre-trained language encoders to generate highly realistic and semantically coherent images, surpassing competitors based on FID scores and human evaluations in DrawBench and similar benchmarks. The review critically examines each model's methods, contributions, performance, and limitations, providing a comprehensive comparison of their theoretical underpinnings and practical implications. The aim is to inform future generative modeling research across various applications.

Keywords: 2D Diffusion Model, DDPM, HighLDM, Imagen

Received on 18 March 2024, accepted on 21 April 2024, published on 22 April 2024

doi: 10.4108/airo.5453

1. Introduction

The field of generative modeling has seen remarkable advancements in recent years, particularly with the advent of sophisticated deep learning architectures and techniques that enable the synthesis of high-quality, intricate data samples. The diffusion models [1], constructed using a tiered arrangement of denoising autoencoder architectures, have demonstrated remarkable outcomes not only in image synthesis [2,3] but also beyond [4,5,6,7], thus establishing the contemporary pinnacle in class-conditioned synthesis [8,9] and high-resolution [10] enhancement tasks. This review paper aims to provide a comprehensive analysis and comparative discussion of three pioneering models at the forefront of this revolution: Denoising Diffusion Probabilistic Models (DDPM) [11], High-Resolution Latent Diffusion Models (HighLDM) [12], and Imagen [13].

Denoising Diffusion Probabilistic Models have emerged as a breakthrough innovation in the generative landscape, leveraging principles from denoising score matching and Langevin dynamics to iteratively refine noise into realistic data points. DDPMs introduce novel parameterizations that ensure uniform input consistency during the reverse process, effectively computing accurate log likelihoods for both continuous and discrete data. These models not only enhance the understanding of diffusion processes but also demonstrate potential for lossless compression through the integration of autoregressive components and Variational Autoencoders (VAEs).

HighLDM represents a significant leap in synthesizing high-resolution images by strategically employing latent diffusion models conditioned on highly efficient pretrained autoencoders. By decomposing the image generation task into a series of denoising steps and diffusion processes, HighLDM achieves state-of-the-art performance in tasks such as image generation, inpainting, and super-resolution,

[*] Corresponding author. Email: fangshun@pku.org.cn

all while minimizing computational requirements. Its unique strength lies in training diffusion models within a latent space, striking a balance between complexity reduction and detail retention, further augmented by cross-attention mechanisms that increase adaptability to diverse conditioning inputs like text or bounding boxes.

On the other hand, the Imagen framework introduces an innovative approach to text-to-image synthesis by integrating transformer-based language models with high-definition diffusion models. A key novelty in Imagen is its efficacious use of pre-trained language models to encode textual information, leading to outstanding sample authenticity and alignment between generated images and their corresponding descriptions. With its large-scale language model components, Imagen attains impressive Frechet Inception Distance (FID) scores on datasets like COCO without direct fine-tuning, thereby demonstrating its capacity to create exceptionally detailed, photorealistic images tightly coupled with linguistic inputs. Human evaluations on the DrawBench benchmark confirm Imagen's superiority over contemporaneous methods in terms of visual quality and semantic correspondence.

Throughout this review, we will examine these models' underlying methodologies, their contributions to the generative modeling domain, and the empirical results they yield. We will critically assess their strengths, limitations, and implications for future research directions, aiming to provide insights into how these cutting-edge techniques can inform and advance the development of generative models across various applications. The subsequent sections will thus delve deeply into the theoretical underpinnings, experimental outcomes, and comparative analyses of DDPM, HighLDM, and Imagen.

2. Methods

We will delve into the fundamental principles and analyze the experimental outcomes of three pivotal generative models: DDPM, HighLDM, and Imagen.

2.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPM) [11] introduce an innovative methodology for training a reverse process mean function estimator that aims at estimating either ˜µt or ϵ, with experimental evidence suggesting superior outcomes when predicting ϵ in the early stages. In terms of data preprocessing, the method involves linear scaling of image information to guarantee uniform input consistency for the neural network's reverse phase, incorporating a discrete decoder to compute precise discrete log likelihoods. The research borrows from principles found in Langevin dynamics and denoising score matching, presenting a streamlined parameterization which simplifies the variational bound inherent in diffusion models. Moreover, the study explores leveraging autoregressive architectures alongside VAE decoders to guarantee lossless compression of discrete datasets. Additionally, the DDPM acknowledges and relates to pertinent literature within the field, such as those concentrating on energy-based frameworks, generative modeling paradigms, and deep learning networks. To summarize, the DDPM significantly advances generative modeling methodologies by proposing a fresh parameterization strategy and examining its efficacy within the context of both diffusion models and variational constraints.

The DDPM methodology introduces the concept of diffusion probabilistic modeling, also referred to as diffusion frameworks, which are parameterized Markov chains trained via variational inference mechanisms. These models are designed to synthesize samples that closely align with the target data distribution following a certain duration of processing time. A notable benefit of these diffusion models lies in their training efficiency and ease of definition, rendering them highly amenable to straightforward implementation and optimization efforts.

The research showcases the capacity of diffusion models to generate high-quality output samples, occasionally surpassing the performance of alternative generative model architectures. Furthermore, a particular configuration of diffusion models establishes an intriguing connection with denoising score matching techniques and annealed Langevin dynamics, thereby contributing to enhanced sample fidelity. Despite not attaining log likelihood scores on par with other likelihood-centric models, diffusion models exhibit considerable potential in generating superior quality samples.

Of particular interest is the sampling process within diffusion models, which can be viewed as a form of gradual decoding—a characteristic that amplifies their generative prowess beyond conventional autoregressive methodologies. In summary, the method's effectiveness in training, its capability to consistently produce high-quality samples, and its innovative sampling procedure collectively constitute significant contributions and strengths of this proposed approach within the realm of generative modeling research.

The neural architecture within the DDPM framework involves training a reverse process mean function approximator, denoted by µθ, which targets the prediction of either ˜µt or ϵ, and optionally can be adapted to estimate x0 as well. In terms of data normalization, it is ensured that image data, originally encoded as integers spanning from 0 to 255, undergoes linear transformation to a consistent input range. For attaining precise discrete log-likelihoods, a Gaussian-inspired discrete decoder is incorporated at the terminal stage of the reverse procedure, thus enabling lossless encoding of discrete information without necessitating supplementary noise injection or Jacobian corrections.

$$\boldsymbol{L_{t-1}} = \mathbb{E}_q\left[\frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\sigma}_t^2} \parallel \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, \boldsymbol{t}) \parallel^2\right] + \boldsymbol{C} \quad (\mathbf{1})$$

This equation represents a term in the variational bound of the diffusion model. Here, $\boldsymbol{L_{t-1}}$ is a part of the overall loss function that the model aims to minimize during training.

The expectation $\mathbb{E}_q$ is taken with respect to the approximate posterior distribution $\boldsymbol{q}$. The term inside the expectation represents the mean squared error between the forward process posterior mean $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$ and the reverse process mean $\boldsymbol{\mu}_\theta(\mathbf{x}_t, \boldsymbol{t})$, scaled by the inverse of twice the variance $\boldsymbol{\sigma}_t^2$. The constant $\boldsymbol{C}$ is a term that does not depend on the model parameters $\boldsymbol{\theta}$ and can be ignored during optimization.

$$\boldsymbol{L}_{t-1} - \boldsymbol{C} = \mathbb{E}_{\mathbf{x}_0,\epsilon}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\epsilon) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\epsilon), \boldsymbol{t})\right\|^2\right] \quad (2)$$

Eq. (2) expands Eq. (1) by incorporating the reparameterization of the forward process. Here, $\boldsymbol{\epsilon}$ is a noise term sampled from a standard normal distribution, and$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon})$ represents the application of the forward process starting from $\mathbf{x}_0$ with noise $\boldsymbol{\epsilon}$. The term $\frac{1}{\sqrt{\alpha_t}}$ and $\sqrt{1-\bar{\alpha}_t}$ are factors derived from the diffusion process's time-dependent parameters. This equation shows that the loss function can be expressed in terms of the data density gradient, which is a key concept in denoising score matching.

The significance of Eq. (1) and Eq. (2) lies in their connection to denoising score matching, a technique used in energy-based models. By optimizing an objective similar to denoising score matching, the authors are able to train a diffusion model that can reverse the noise diffusion process and generate high-quality samples. This is achieved by learning the parameters $\boldsymbol{\theta}$ of the reverse process mean function to predict the noise term $\boldsymbol{\epsilon}$ given the current state $\mathbf{x}_t$ of the diffusion process. The model effectively learns to estimate the gradient of the data distribution, which is a powerful inductive bias for generative modeling.

The learning algorithm engages gradient descent steps that are guided by the inferred gradient of the underlying data density, emulating the principles of Langevin dynamics in its approach. The sampling scheme comprises calculating xt−1 based on the predicted ϵ and stochastic perturbations, effectively implementing a multi-scale version of denoising score matching. This holistic methodology amalgamates core concepts from diffusion models, denoising autoencoder methodologies, and variational inference techniques to drive advancements in generative modeling practices.

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, \boldsymbol{t}) = \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t)\right)\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, \boldsymbol{t})\right) \quad \textbf{(3)}$$

Eq. (3) defines the mean function of the reverse process, which is a key component of the diffusion model. The function $\boldsymbol{\mu}_\theta$ is the model's parameterized mean, which is designed to predict the denoised version of the data given the current state $\mathbf{x}_t$ and time step $\boldsymbol{t}$. The term $\tilde{\boldsymbol{\mu}}_t$ represents the mean of the forward process posterior, and $\boldsymbol{\epsilon}_\theta$ is a function approximator that predicts the noise $\boldsymbol{\theta}$ added during the diffusion process. The parameters $\bar{\boldsymbol{\alpha}}_t$ and $\boldsymbol{\beta}_t$ are related to the noise schedule of the diffusion process. This equation essentially captures the idea that the model should learn to reverse the noise accumulation process to recover the original data distribution.

$$\mathbb{E}_{\mathbf{x}_0,\epsilon}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \boldsymbol{t}\right)\right\|^2\right] \quad (4)$$

Eq. (4) further refines the loss function by incorporating the noise term $\boldsymbol{\epsilon}$ and the parameters of the diffusion process. It represents the expected squared difference between the actual noise and the model's prediction of the noise, scaled by the variance $\boldsymbol{\sigma}_t^2$ and the noise schedule parameters $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\alpha}}_t$. The term $\boldsymbol{\alpha}_t$ is a function of the noise schedule that determines the rate at which noise is added to the data. This equation is similar to the denoising score matching objective, where the goal is to minimize the difference between the predicted noise and the actual noise that was added to the data. By optimizing this objective, the model learns to reverse engineer the noise diffusion process, thereby learning to generate data that matches the original distribution.

$$\boldsymbol{L}_{\text{simple}}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \boldsymbol{t}\right)\right\|^2\right] \quad (5)$$

Eq. (5) represents a simplified version of the variational bound used for training the diffusion model, referred to as $\boldsymbol{L}_{\text{simple}}$. It is a key component of the model's objective function, which the authors aim to minimize during the training process.

Here's a breakdown of the components within the equation:

- $L_{\text{simple}}(\theta)$: This is the simplified loss function for the diffusion model, dependent on the model parameters $\boldsymbol{\theta}$.
- $\mathbb{E}_{t,\mathbf{x}_0,\epsilon}$: This denotes the expectation taken over the random variables $\boldsymbol{t}$ (time step), $\mathbf{x}_0$ (the initial data point), and $\boldsymbol{\epsilon}$ (noise).
- $\boldsymbol{\epsilon}$: This is the noise term from the diffusion process.
- $\boldsymbol{\epsilon}_\theta$: This is the model's parameterized prediction of the noise term $\boldsymbol{\epsilon}$, given the data point $\mathbf{x}_0$ and the noise $\boldsymbol{\epsilon}$ at time step $\boldsymbol{t}$.
- $\sqrt{\bar{\boldsymbol{\alpha}}_t}$: This term is related to the noise schedule of the diffusion process and scales the initial data point $\mathbf{x}_0$.
- $\sqrt{1-\bar{\boldsymbol{\alpha}}_t}$: This term also relates to the noise schedule and scales the noise $\boldsymbol{\epsilon}$.
- $\boldsymbol{t}$: This is the time step in the diffusion process, and the expectation is taken over a uniform distribution between 1 and **T**, where **T** is the total number of time steps in the diffusion process.

The essence of Eq. (5) is to minimize the mean squared error between the actual noise $\boldsymbol{\epsilon}$ added to the data during the diffusion process and the noise predicted by the model $\boldsymbol{\epsilon}_\theta$. By optimizing this loss function, the model learns to reverse the diffusion process effectively, which means it learns to generate high-quality samples that match the original data distribution.

The research delves into data rescaling, applying it such that image data is consistently scaled for neural network processing. This involves using a discrete decoder designed from a Gaussian distribution to accurately derive discrete log likelihoods and ensure lossless encoding lengths for discrete data items. The learning process is governed by gradient descent steps informed by the estimated gradient of the data density, echoing the iterative nature of Langevin dynamics.

$$\mathbf{x_0} \approx \hat{\mathbf{x}}_\mathbf{0} = \frac{\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t)\right)}{\sqrt{\bar{\alpha}_t}} \tag{6}$$

Eq. (6) describes the progressive estimation of the original data point $\hat{\mathbf{x}}_\mathbf{0}$ in the context of the reverse process of the diffusion model. This equation is used to reconstruct the data point $\mathbf{x_0}$ from the sequence of noisy points $\mathbf{x}_t$ generated by the model during the reverse process. The reconstruction is done progressively, starting from the final point $\mathbf{x}_T$ and working backwards towards the original point $\mathbf{x_0}$.

- $\hat{\mathbf{x}}_\mathbf{0}$: This is the estimated or reconstructed version of the original data point.
- $\mathbf{x}_t$: This represents the noisy version of the data point at each time step t in the reverse process.
- $\sqrt{1-\bar{\alpha}_t}$: This term is related to the noise schedule of the diffusion process and scales the function approximator's output $\epsilon_\theta(\mathbf{x}_t)$.
- $\epsilon_\theta(\mathbf{x}_t)$: This is the model's parameterized prediction of the noise term at each time step t.
- $\sqrt{\bar{\alpha}_t}$: This term scales the reconstructed data point $\hat{\mathbf{x}}_\mathbf{0}$ by the inverse of the square root of $\bar{\alpha}_t$, which is a function of the noise schedule.

Eq. (6) effectively calculates a running total of the noisy data points, with each point being denoised as it is included in the sum. This progressive approach allows for the generation of data that starts from a highly noisy state and gradually becomes clearer as more information from the reverse process is incorporated.

$$L = D_{\mathrm{KL}}\left(q(\mathbf{x}_T) \parallel p(\mathbf{x}_T)\right) + \mathbb{E}_q\left[\sum_{t\geq 1} D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1}|\mathbf{x}_t) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\right)\right] + H(\mathbf{x_0}) \tag{7}$$

- $D_{\mathrm{KL}}\left(q(\mathbf{x}_T) \parallel p(\mathbf{x}_T)\right)$: This is the Kullback-Leibler (KL) divergence between the final distribution $q(\mathbf{x}_T)$ and the true data distribution $p(\mathbf{x}_T)$. This term measures the difference between the noise distribution at the end of the diffusion process and the actual noise distribution.
- $\mathbb{E}_q\left[\sum_{t\geq 1} D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1}|\mathbf{x}_t) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\right)\right]$: This is the expected value, taken under the distribution q, of the sum of KL divergences between the approximate posterior $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and the reverse process distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ at each time step. This term encourages the model to learn the reverse process that can effectively undo the noise added by the diffusion process.
- $H(\mathbf{x_0})$: This is the entropy of the original data distribution, which is a measure of the uncertainty in the data.

The goal during training is to minimize this variational bound, which effectively trains the model to generate data that is similar to the true data distribution. By optimizing this objective, the model learns to reverse the diffusion process by learning to predict the noise that was added to the data at each time step and gradually removing it to recover the original, noise-free data.

The sampling methodology consists of calculating xt−1 based on the predicted ϵ and stochastic noise variables, mimicking a multi-scale application of denoising score matching principles. Ultimately, the DDPM consolidates critical aspects from diffusion models, denoising autoencoders, and variational inference methodologies to propel advancements in generative modeling techniques.

Table 1. this table presents a comparison of the FID scores for different models on the LSUN dataset, which is a large-scale image dataset designed for testing the performance of generative models.

Model	LSUN Church	LSUN Cat	LSUN Bedroom
DDPM	7.89	19,75	6.36
DDPM(Large)	-	-	4.90
StyleGAN [14]	4.21	8.53	2.65
StyleGAN2 [15]	3.86	6.93	-
ProgressiveGAN [16]	6.42	37.52	8.34

The header of Table 1 includes Model, LSUN Bedroom, LSUN Church, and LSUN Cat. The “Model” column lists the names of the different generative models under comparison. The columns for “LSUN Bedroom”, “LSUN Church” and “LSUN Cat” represent the FID scores achieved by the respective models on three different subsets of the LSUN dataset: Bedroom, Church, and Cat. The FID score is a measure of the quality of the generated images; a lower FID score indicates better performance, i.e., the generated images are more similar to the real images in the dataset.

Table 1 includes the following models and their corresponding FID scores:

ProgressiveGAN: A GAN-based model that generates images progressively, starting from a low-resolution version and refining it step by step. The FID scores for ProgressiveGAN on the Bedroom, Church, and Cat subsets are 8.34, 6.42, and 37.52, respectively.

StyleGAN: A variant of GAN that uses a style-based generator architecture. The FID scores for StyleGAN are not provided in the table, but the paper refers to the scores reported in the StyleGAN paper as baselines (4.21 for Church and 8.53 for Cat).

StyleGAN2: An improved version of StyleGAN with better image quality and stability. The FID scores for StyleGAN2 on the Bedroom and Church subsets are 3.86 and 6.93, respectively.

DDPM: This refers to the diffusion model proposed in the paper using the simplified training objective. The FID scores for this model on the Bedroom, Church, and Cat subsets are 6.36, 7.89, and 19.75, respectively.

DDPM, large: This is a larger variant of the diffusion model with more parameters, trained specifically on the Bedroom subset. The FID score for this model is 4.90, indicating improved performance over the smaller version.

The table demonstrates that the proposed diffusion model achieves competitive or superior FID scores compared to existing models like ProgressiveGAN and StyleGAN2, particularly on the Bedroom and Church datasets. This suggests that the diffusion model is effective in generating high-quality images that closely resemble the real images in the LSUN dataset.

The hyperparameter optimization routine entailed fine-tuning for sample quality on CIFAR10 and subsequently transferring these optimized configurations to alternative datasets. A variety of hyperparameters underwent calibration, such as the βt schedule, dropout ratio, data enhancement strategies like random flipping along the horizontal axis, and the selection of the optimization technique – Adam being the choice made. In most experimental settings, the learning rate was fixed at 2 × 10^-4, while for experiments dealing with 256 by 256 pixel images, a lower rate of 2 × 10^-5 was adopted.

2.2 High-Resolution Latent Diffusion Models

High-Resolution Latent Diffusion Models (HighLDM) [12] introduce an innovative technique for image synthesis through the strategic utilization of Latent Diffusion Models (LDMs) which have been conditioned on highly effective pretrained autoencoders. By fractionating the process of image construction into denoising autoencoder components and diffusion modeling, this HighLDM has accomplished unprecedented results in generating high-resolution images, inpainting, super-resolution tasks, among others, concurrently reducing computational needs significantly. The pivotal novelty is encapsulated in training the diffusion models within a latent space, thereby facilitating an equilibrium between complexity minimization and intricate detail retention.

Moreover, the integration of cross-attention mechanisms amplifies the adaptability of LDMs when subjected to various conditioning inputs such as textual descriptions or bounding box constraints, culminating in strong competitive outcomes across different synthesis scenarios. Not only does the HighLDM delve deeply into the intricacies of model architectural design and the underlying learning strategy, but it also furnishes pretrained instances that extend its applicability beyond the confines of mere image synthesis problems.

The presented methodology introduces LDMs that capitalize on the latent dimensions of robustly pretrained autoencoders to generate high-quality images with minimized computational expenses (Figure 1). By partitioning the image generation process and integrating cross-attention modules within its design, LDMs facilitate meticulous control over synthesis without necessitating retraining, thereby achieving a near-optimal equilibrium between complexity minimization and fine detail retention. This strategy substantially enhances visual authenticity, thus attaining top-tier performance across applications like image restoration, class-controlled image creation, text-guided image synthesis, unconditional image generation, and super-resolution enhancement.

Moreover, LDMs exhibit strong competitiveness across multiple datasets while diminishing both training and inference costs in comparison to pixel-wise diffusion models. Additionally, the adoption of a versatile conditioning mechanism rooted in cross-attention enables multi-modal learning for diverse tasks including class-conditioned, text-driven, and layout-based image synthesis, further highlighting the flexibility and efficacy of this proposed technique.

The proposed technique introduces LDMs that capitalize on the latent dimensions of advanced pretrained autoencoders to facilitate high-resolution image synthesis while minimizing computational demands. By disassembling the image construction process and incorporating cross-attention mechanisms within its structural design, LDMs allow for meticulous control over image generation without necessitating retraining, thereby achieving a nearly ideal equilibrium between complexity minimization and fine detail conservation. This methodology substantially bolsters visual accuracy, thus enabling leading-edge performance in applications such as image completion, class-guided image synthesis, text-to-image translation, unconditional image creation, and super-resolution enhancement.

Furthermore, LDMs exhibit strong competitiveness across numerous datasets while simultaneously reducing both the training overhead and inference expenses when compared with pixel-level diffusion models. Additionally, the deployment of a broadly applicable conditioning system based on cross-attention supports multi-modal learning scenarios in tasks like conditional image synthesis based on classes, texts, or layouts, further underscoring the adaptability and robustness of this presented method.

LDMs introduce a novel technique for synthesizing high-resolution images by harnessing diffusion processes within the latent dimensions of pre-trained autoencoders, thus striking a balance between complexity minimization and maintaining fine details. This methodology is characterized by a dual-phase training regimen: initially, an autoencoder is trained to create a reduced-dimensional yet perceptually congruent representation, followed by the training of diffusion models within this space, which are then referred to as LDMs.

Diffusion Models (DMs) are probabilistic frameworks that progressively denoise a Gaussian-distributed variable in order to learn a given data distribution, with successful iterations often relying on cascades of denoising autoencoders. The LDMs employ a time-variant UNet architecture at their core to facilitate efficient image generation directly from the latent domain. Conditional control mechanisms are integrated into these models, allowing them to respond to inputs like text descriptions or semantic maps, thereby enhancing the versatility of LDMs as conditional image generators.

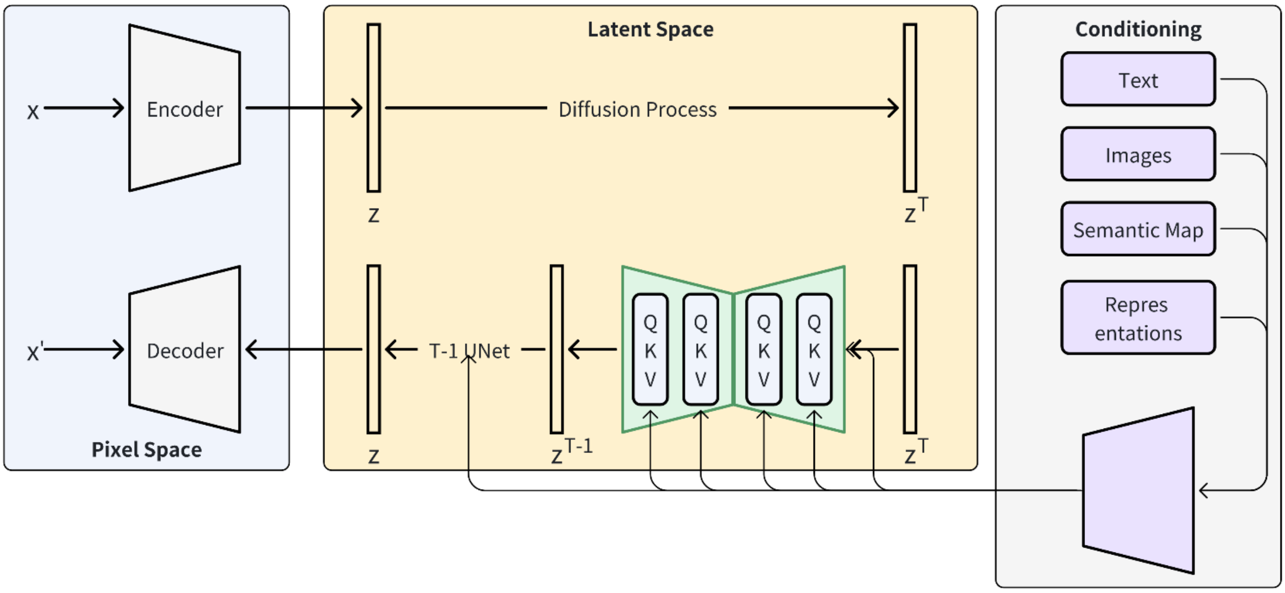

Figure 1. Overview of HighLDM [12]

The HighLDM approach centers around the strategic training of LDMs within a latent space, aiming to optimize computational efficiency and maintain fidelity to fine details. It particularly underscores the application of various conditioning mechanisms to guide the conditional synthesis of images, showcasing the enhanced capabilities of LDMs in this context.

The strategy involves training diffusion models within the latent realm of pretrained autoencoders, striking a balance between complexity minimization and maintaining intricate details. The process commences with teaching an autoencoder to construct a dimensionally reduced yet perceptually equivalent domain, succeeded by the education of diffusion models within this reduced space. LDMs utilize a time-dependent UNet framework as their core structure for generating images effectively from the latent level, incorporating conditioning mechanisms that regulate the synthesis procedure according to inputs such as textual descriptions or semantic layouts.

Conditioning signals are integrated into the UNet through a cross-attention mechanism, enabling multimodal learning and supporting tasks like class-specific image generation, text-to-image conversion, and layout-to-image transfiguration. The UNet's foundational architecture is primarily composed of two-dimensional convolutional layers, focusing on the most perceptually significant features using a reweighted boundary constraint. Furthermore, it encompasses a shallow transformer layering self-attention modules, position-wise multilayer perceptrons, and cross-attention layers specifically designed for conditioning purposes. This model design facilitates both efficient image creation and versatile conditioning methodologies for diverse tasks, demonstrating competitive performance across multiple image synthesis challenges while significantly curtailing computational expenses.

The optimization criterion incorporates the training of diffusion models within the latent domain of pretrained autoencoders for high-resolution image generation. The strategy targets computational efficiency by dividing the learning process into distinct compressive and generative stages. The perceptual compression module relies on an autoencoder that has been tutored using a hybrid loss function composed of perceptual metrics and patch-wise adversarial goals, ensuring lifelike reconstructions and mitigating blurriness.

$$\boldsymbol{L_{DM}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{1}),\boldsymbol{t}}\left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}(\boldsymbol{x_t},\boldsymbol{t}) \|_2^2\right] \qquad \mathbf{(8)}$$

Eq. (8) represents the loss function for the latent diffusion model (LDM). It is an expectation over the noise variable $\boldsymbol{\epsilon}$ drawn from a standard normal distribution, the denoised sample $\boldsymbol{\epsilon_\theta}(\boldsymbol{x_t},\boldsymbol{t})$ at time step $\boldsymbol{t}$ and the initial data sample $\boldsymbol{x}$. The loss is calculated as the mean squared error between the noise $\boldsymbol{\epsilon}$ and the denoised sample $\boldsymbol{\epsilon_\theta}(\boldsymbol{x_t},\boldsymbol{t})$. This loss function is used to train the model to reverse the diffusion process and synthesize high-resolution images. The variable T denotes the total number of time steps in the diffusion process.

$$\boldsymbol{L_{LDM}} \coloneqq \mathbb{E}_{\mathcal{E}(\boldsymbol{x}),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{1}),\boldsymbol{t}}\left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}(\boldsymbol{z_t},\boldsymbol{t}) \|_2^2\right] \qquad \mathbf{(9)}$$

Eq. (9) is a variant of Formula 1, where the loss is computed in the latent space after the encoder $\mathbb{E}$ has been applied to the data. Here, $\boldsymbol{z_t}$ represents the latent representation of the data at time step $\boldsymbol{t}$. The encoder $\mathbb{E}$ compresses the high-dimensional image data into a lower-dimensional latent representation, which is then used by the diffusion model $\boldsymbol{\epsilon_\theta}$ to generate the denoised samples. This formulation allows the model to focus on the semantic and structural information in the data, rather than the high-frequency details that may not be perceptually significant.

$$\boldsymbol{L_{LDM}} \coloneqq \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}\left(\boldsymbol{z_t}, \boldsymbol{t}, \boldsymbol{\tau_\theta}(\boldsymbol{y})\right) \|_2^2\right] \quad \textbf{(10)}$$

Eq. (10) extends the loss function to incorporate conditioning on an input $\boldsymbol{y}$, which can be text, semantic maps, or other forms of guidance. The term $\boldsymbol{\tau_\theta}(\boldsymbol{y})$ represents a conditioning mechanism that maps the input $\boldsymbol{y}$ to an intermediate representation that is used by the diffusion model $\boldsymbol{\epsilon_\theta}$. This allows the LDM to generate images that are not only high-resolution and perceptually realistic but also aligned with the specific context or content specified by $\boldsymbol{y}$. This conditional loss function enables the model to perform tasks such as text-to-image synthesis, where the generated images are guided by textual descriptions.

The autoencoder networks are adversarially fine-tuned with constraints imposed on the latent manifold to preserve low variance and yield high-quality reconstructions. When it comes to training diffusion models in the latent space, two distinctive scenarios are identified based on different regularization techniques: one involving KL-divergence regularization in the latent realm and another leveraging Vector Quantization (VQ) regularization. In the scenario where KL-regularization is employed, sampling takes place according to the mean and standard deviation within the latent field; conversely, in the VQ-regularized case, the quantization step is integrated directly into the decoding phase.

$$\boldsymbol{L}_{\text{Autoencoder}}\& = \min_{\mathcal{E},\mathcal{D}} \max_{\psi} \left(\boldsymbol{L_{rec}}\left(\boldsymbol{x}, \boldsymbol{\mathcal{D}}\left(\boldsymbol{\mathcal{E}}(\boldsymbol{x})\right)\right) - \boldsymbol{L_{adv}}\left(\boldsymbol{\mathcal{D}}\left(\boldsymbol{\mathcal{E}}(\boldsymbol{x})\right)\right) + \textbf{log } \boldsymbol{D_\psi}(\boldsymbol{x}) + \boldsymbol{L_{reg}}(\boldsymbol{x}; \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{D}}) \right) \quad (11)$$

Eq. (11) represents the full objective function for training the autoencoder model, which is a critical component in the proposed LDMs. The autoencoder consists of an encoder $\boldsymbol{\mathcal{E}}$ and a decoder $\boldsymbol{\mathcal{D}}$.

- $\boldsymbol{L_{rec}}\left(\boldsymbol{x}, \boldsymbol{\mathcal{D}}\left(\boldsymbol{\mathcal{E}}(\boldsymbol{x})\right)\right)$: This term represents the reconstruction loss, which measures the difference between the original input image $\boldsymbol{x}$ and the reconstructed image $\boldsymbol{\mathcal{D}}\left(\boldsymbol{\mathcal{E}}(\boldsymbol{x})\right)$. The reconstruction loss encourages the decoder $\boldsymbol{\mathcal{D}}$ to produce outputs that are as close as possible to the original images.
- $\boldsymbol{L_{adv}}\left(\boldsymbol{\mathcal{D}}\left(\boldsymbol{\mathcal{E}}(\boldsymbol{x})\right)\right)$: This term is the adversarial loss, where the decoder $\boldsymbol{\mathcal{D}}$ is encouraged to generate images that can fool a discriminator $\boldsymbol{\psi}$. The discriminator is trained to distinguish between real images and the reconstructed images produced by the decoder. This adversarial training helps to improve the quality and realism of the reconstructed images.
- $\textbf{log } \boldsymbol{D_\psi}(\boldsymbol{x})$: This term is the log-likelihood of the discriminator's output when given a real image $\boldsymbol{x}$. It encourages the discriminator $\boldsymbol{\psi}$ to output high probability for real images, which is a standard component in adversarial training.
- $\boldsymbol{L_{reg}}(\boldsymbol{x}; \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{D}})$: This term is the regularization loss, which is used to control the variance of the latent space. The regularization can be achieved through different methods, such as a small weighted KL divergence from the latent distribution to a standard normal distribution (as in a variational autoencoder) or through the use of a vector quantization layer. The regularization helps to ensure that the latent space is meaningful and does not contain excessively high-variance representations that could lead to overfitting or poor image quality.

The overall objective function is minimized over the encoder and decoder parameters, while simultaneously maximizing the adversarial loss (hence the min-max formulation). This results in an autoencoder model that can effectively compress and reconstruct images, with the additional benefit of being robust to adversarial examples. The autoencoder is a key part of the LDM framework, as it provides the latent space in which the diffusion process takes place, leading to high-resolution and high-fidelity image synthesis.

In summary, the loss function integrates reconstruction error terms, adversarial losses, regularization components, and tailored sampling methodologies that align with the specific characteristics of the latent space, all contributing to efficient and effective image synthesis processes.

Table 2. Quantitative comparison of the layout-to-image models on the COCO and OpenImages datasets

Method	Open-Images $\mathbf{256^2}$	Open-Images $\mathbf{512^2}$	COCO $\mathbf{256^2}$
	FID	FID	FID
LDM-8(100 steps)	-	-	42.06
LDM-4(200 steps)	32.02	35.80	40.91
VQGAN+T [17]	45.33	48.11	56.58
SPADE [18]	-	-	41.11
OC-GAN [19]	-	-	41.65
LostGAN-V2 [20]	-	-	42.55

Table 2 presents a quantitative comparison of layout-to-image models on the COCO and OpenImages datasets. The table provides a detailed analysis of the performance of different models in generating images from layout inputs, which typically consist of bounding boxes and categories that describe the content and arrangement of objects within an image.

Method: This column lists the different models that have been evaluated for layout-to-image synthesis. The models include LostGAN-V2, OC-GAN, SPADE, VQGAN+T, and the proposed LDM-8 and LDM-4 from the paper, along with their respective configurations (e.g., the number of steps for training and whether they were fine-tuned from OpenImages or trained from scratch on COCO).

FID: This column shows the FID (Fréchet Inception Distance) scores for each model. FID is a metric used to evaluate the quality of generated images by comparing them to real images using the Inception network. A lower FID score indicates better performance, as it suggests the generated images are more similar to real images.

IS: This column displays the Inception Score (IS) for each model. IS is another metric that measures the quality of generative models by assessing the diversity and faithfulness of the generated images. A higher IS score indicates better performance.

Precision: This column presents the Precision scores for the models. Precision measures the proportion of generated images that are relevant to the input layout, out of all the images generated by the model.

Recall: This column shows the Recall scores for the models. Recall measures the proportion of relevant images that were generated, out of all the images that should have been generated based on the input layout.

Nparams: This column indicates the number of trainable parameters in millions for each model.

Table 2 highlights the performance of the proposed LDM models (LDM-8 and LDM-4) in comparison to other state-of-the-art models. It shows that the LDM models are able to achieve competitive performance in layout-to-image synthesis, with LDM-8 and LDM-4 outperforming or matching the results of other models like LostGAN-V2, OC-GAN, SPADE, and VQGAN+T. This demonstrates the effectiveness of the LDM approach in generating high-quality and diverse images from layout descriptions.

2.3 Imagen

The Imagen [13] framework introduces an innovative text-to-image generative model, which harnesses the power of expansive transformer-based language models in tandem with high-definition diffusion models to attain photorealistic image synthesis deeply rooted in linguistic comprehension. The pivotal novelty of Imagen is encapsulated by its efficient utilization of pre-trained language models for encoding textual inputs aimed at image creation, thus demonstrating superior capabilities in both sample authenticity and the alignment between images and their corresponding texts.

By scaling up the dimensions of its language model components, Imagen attains unprecedented FID scores on the COCO dataset without undergoing direct training on it, thereby illustrating its outstanding capacity to generate images of exceptional quality that are tightly coupled with the given text descriptions. In the DrawBench benchmark—a holistic assessment tool for evaluating text-to-image generation models—Imagen surpasses several recent methodologies in terms of human-judged sample excellence and the congruence between generated images and input text.

A critical breakthrough in Imagen's design is the strategic and effective application of large-scale, pretrained language architectures like T5 for text representation during image synthesis, which demonstrates superior results with respect to sample authenticity and the harmony between images and their corresponding textual descriptions. Moreover, the study underscores the importance of adaptive thresholding techniques and judicious architectural selections in the U-Net structure, which are instrumental in fostering the generation of more realistic and intricate images. In essence, Imagen's core strengths can be attributed to its adept employment of sizeable frozen language models as powerful text transformers, the incorporation of dynamic threshold controls for elevated image fidelity, and the establishment of a robust evaluation framework for assessing text-to-image generative models, all of which vividly demonstrate its innovative strides and advancements within the discipline of text-guided image synthesis.

In summary, Imagen underscores the critical role played by large-scale, frozen pre-trained language models as highly effective text encoders in the realm of text-to-image synthesis. It emphasizes the significance of scaling such language models to enhance overall performance and serves as a catalyst for further investigation along this research pathway.

The proposed approach outlines several core components and stages in the Imagen text-to-image diffusion framework. To begin with, Imagen employs a sequential process that starts with a foundational 64×64 model followed by two progressive, text-conditioned super-resolution diffusion models, capable of incrementally upgrading image resolution to dimensions such as 256×256 and 1024×1024. This is accomplished by leveraging cascaded diffusion architectures augmented with noise conditioning to generate high-quality images.

In terms of its neural network design, Imagen adopts an adjusted U-Net structure that conditions on text embeddings using a pooled embedding vector and cross-attention mechanisms applied across multiple scales of text embeddings. Layer Normalization is integrated into both the attention and pooling layers for processing these text embeddings. An innovative feature incorporated within the method is dynamic thresholding, which adjusts pixel values during sampling to prevent saturation and thereby enhance the photorealism and alignment between generated images and their corresponding texts, particularly when employing large guidance weights.

Furthermore, Imagen introduces a tailored variant of the U-Net architecture, branded Efficient U-Net, specifically designed for its super-resolution modules. This variant incorporates modifications like relocating model parameters to low-resolution blocks, resizing skip connections, and reversing the order of downscaling/upscaling operations to optimize computational speed and memory efficiency.

In summary, the Imagen methodology combines these inventive features to deliver unprecedented levels of high-fidelity text-to-image synthesis, characterized by remarkable photorealism and tight correspondence with input text.

The loss function implemented within the Imagen methodology revolves around training the diffusion model to transform noisy instances into data points by optimizing a weighted quadratic error metric. More precisely, this entails teaching the diffusion model to progressively remove noise from $\boldsymbol{z_t}$until it approximates x through minimizing the squared difference between the forecasted output $\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\boldsymbol{z_t}, \boldsymbol{\lambda_t}, \boldsymbol{c})$ and the actual data sample x, where the magnitude of the loss is modulated by a weighting function $\mathbf{w}(\boldsymbol{\lambda_t})$, which significantly impacts the quality of generated samples. This iterative denoising procedure serves as a cornerstone for creating high-fidelity images by meticulously refining noisy inputs towards realistic data representations.

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,t}\left[\boldsymbol{w_t} \parallel \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\boldsymbol{\alpha_t}\mathbf{x} + \boldsymbol{\sigma_t}\boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x} \parallel_2^2\right] \qquad (\mathbf{12})$$

Eq. (12) represents the denoising objective of the diffusion model. The diffusion model, denoted by $\hat{\mathbf{x}}_{\boldsymbol{\theta}}$, is trained to reverse the forward process of adding noise to the data, starting from the data $\mathbf{x}$ and a conditioning signal $\mathbf{c}$. The process adds Gaussian noise over time, represented by $\boldsymbol{\epsilon}$, following a schedule determined by $\boldsymbol{\alpha_t}$ and $\boldsymbol{\sigma_t}$. The term $\boldsymbol{w_t}$ is a weighting function that emphasizes certain values of $\boldsymbol{t}$, which is sampled uniformly from the interval [0, 1]. The goal is to minimize the squared error between the model's prediction $\hat{\mathbf{x}}_{\boldsymbol{\theta}}$ and the original data $\mathbf{x}$, at each time step $\boldsymbol{t}$, to learn how to effectively denoise the noisy inputs back to the original data distribution.

$$\tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_{\boldsymbol{t}}, \mathbf{c}) = \boldsymbol{w}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_{\boldsymbol{t}}, \mathbf{c}) + (\mathbf{1} - \boldsymbol{w})\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_{\boldsymbol{t}}) \qquad (\mathbf{13})$$

Eq. (13) is part of the classifier-free guidance technique, an alternative method to improve sample quality in conditional diffusion models without relying on a pretrained classifier. The term $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_{\boldsymbol{t}}, \mathbf{c})$ represents the conditional prediction (or noise prediction) given the noisy input $\mathbf{z}_{\boldsymbol{t}}$ and the conditioning signal $\mathbf{c}$. The unconditional prediction $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_{\boldsymbol{t}})$ is used as a baseline when no conditioning signal is provided. The parameter $\boldsymbol{w}$ is the guidance weight, which controls the influence of the conditioning signal on the model's predictions. When $\boldsymbol{w} = \mathbf{1}$, the guidance effect is disabled, while increasing $\boldsymbol{w}$ strengthens the guidance. The model $\overline{\boldsymbol{\theta}}$ is then trained to minimize the difference between the adjusted prediction and the true data, which helps in generating images that align with the conditioning text while maintaining high fidelity.

Both of these formulas are essential components of the Imagen model, with the first focusing on the denoising process and the second on the conditioning of the model during sampling to generate text-conditional images. Together, they enable Imagen to synthesize photorealistic images that closely align with the input text descriptions.

The rationale behind reducing the generative process to a denoising task is substantiated by its optimization as a weighted variational lower bound on the log likelihood of the data under the specified diffusion model framework. Moreover, the model architecture adopts the ϵ-prediction parameterization scheme and computes the squared error loss in the ϵ domain, with the time-step variable t sampled according to a cosine progression. The sampling commences from pure noise at z1 and systematically produces a sequence of points $\mathbf{z}_{t_1}, \ldots, \mathbf{z}_{t_T}$, each one having a decreasing level of noise content, all guided by the underlying diffusion model during the refinement stages.

Table 3. MS-COCO 256×256 FID-30K.

Model	Zero-shot FID-30K	FID-30K
DALL-E [21]	17.89	-
DALL-E2 [22]	10.39	-
LAFITE [23]	26.94	-
GLIDE [24]	12.24	-
Make-A-Scene [25]	-	7.55
LAFITE	-	8.12
XMC-GAN [26]	-	9.33
DM-GAN+CL [27]	-	20.79
DF-GAN [28]	-	21.42
DM-GAN [29]	-	32.64
AttnGAN [30]	-	35.49
Imagen	7.27	-

Table 3 presents a comparison of different models on the MS-COCO 256 × 256 FID-30K benchmark. The table provides a quantitative analysis of the performance of various text-to-image synthesis models. Table 3 is divided into two main sections: Model FID-30K and Zero-shot FID-30K.

Model FID-30K: This section displays the FID scores for each model on the COCO dataset. The FID score is a metric used to measure the similarity between generated images and real images. A lower FID score indicates that the generated images are more realistic and closer to the real images. The table lists several models, including AttnGAN, DM-GAN, DF-GAN, and others, along with their corresponding FID-30K scores. The lower the score, the better the model performs in terms of image fidelity.

Zero-shot FID-30K: This section shows the zero-shot FID-30K scores for the models. Zero-shot learning refers to the ability of a model to generalize its learned knowledge to unseen data without any further training. In this context, the models are evaluated on the COCO dataset without being specifically trained on it. The zero-shot FID-30K score reflects how well the models can generate images that are realistic and aligned with the text descriptions from the COCO dataset without any prior exposure to it. Again, a lower score is better, indicating that the model can effectively understand and generate images corresponding to text descriptions even without direct training on that dataset.

Talbe 3 highlights the performance of Imagen, which achieves a state-of-the-art FID score of 7.27, significantly outperforming other methods such as DALL-E 2, GLIDE, and others. This demonstrates the effectiveness of Imagen's approach to text-to-image synthesis, which leverages large transformer language models and diffusion models to generate photorealistic images with deep language understanding.

3. Comparative Analysis and Future Research

3.1 Comparative Analysis

In this article, we will compare and contrast three significant research papers in the domain of generative models and diffusion models, focusing on their methodologies, applications, and experimental outcomes.

DDPM

Method: This paper introduces a class of latent variable models known as diffusion probabilistic models, which are trained using variational inference to produce high-quality image samples. The authors present a connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, leading to a progressive lossy decompression scheme.

Application: The primary application is high-quality image synthesis, with a focus on generating samples that match or exceed the quality of existing generative models.

Experiments: The authors demonstrate the effectiveness of their models on the CIFAR10 and LSUN datasets, achieving state-of-the-art FID scores and Inception scores. They also discuss the sample quality in relation to the log likelihood of the models.

Imagen

Method: Imagen combines the power of large transformer language models for text understanding with high-fidelity diffusion models for image generation. It utilizes a frozen T5-XXL encoder to map text into embeddings and cascaded diffusion models for image generation.

Application: The main application is text-to-image synthesis, where the model generates photorealistic images based on textual descriptions.

Experiments: Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset without training on it. It also performs well in human evaluations, where generated samples are found to be on par with COCO data in image-text alignment.

HighLDM

Method: This work proposes latent diffusion models that apply diffusion models in the latent space of pretrained autoencoders. It introduces cross-attention layers into the model architecture, making diffusion models flexible for various conditioning inputs.

Application: The paper focuses on high-resolution image synthesis, including tasks like inpainting, super-resolution, and text-to-image synthesis.

Experiments: LDMs achieve state-of-the-art scores for image inpainting and class-conditional image synthesis. They also show competitive performance on unconditional image generation and super-resolution, significantly reducing computational requirements compared to pixel-based diffusion models.

Comparison

Model Methods: DDPM focuses on the connection between diffusion models and score matching, Imagen emphasizes the use of large language models for text understanding, and LDM explores the application of diffusion models in a compressed latent space. Each paper presents a unique approach to improving the quality and efficiency of generative models.

Application Domains: While DDPM and LDM cover a broad range of image synthesis tasks, Imagen specifically targets text-to-image synthesis, showcasing the versatility of diffusion models in various domains.

Experimental Effectiveness: All three papers demonstrate state-of-the-art results in their respective tasks, with Imagen and LDM particularly highlighting the computational efficiency gains over previous methods. Imagen's human evaluation results are particularly noteworthy, indicating a close alignment with human perception of image quality and relevance to text prompts.

In summary, these papers collectively advance the field of generative modeling by introducing novel methods and demonstrating their effectiveness across diverse applications. Each paper contributes to the understanding of how to generate high-quality images, either by improving the generative process itself, integrating language understanding, or reducing computational overheads.

3.2 Future Research

The future research directions in the field of generative models, particularly focusing on diffusion models and their applications, can be expanded along several dimensions:

Improving Sample Quality and Diversity: Future research could focus on enhancing the quality and diversity of samples generated by diffusion models. This may involve developing new training techniques, exploring different model architectures, or incorporating additional modalities such as audio and video to create multimodal generative models.

Reducing Computational Costs: Given the computationally intensive nature of diffusion models, there is a need for more efficient training and inference methods. This could include the development of algorithms that reduce the number of required neural network evaluations, methods for distributed training, and approaches that leverage hardware accelerators more effectively.

Enhancing Text-to-Image Synthesis: The integration of large language models with diffusion models for text-to-image synthesis presents opportunities for research. This includes improving the understanding of textual prompts, generating more contextually relevant images, and exploring cross-modal interactions to refine the alignment between text and images.

Addressing Societal and Ethical Concerns: As generative models become more powerful, it is crucial to address their potential misuse. Research could focus on developing frameworks for responsible AI, including methods for detecting and mitigating deepfakes, ensuring data privacy, and promoting fairness and inclusivity in model training.

Interdisciplinary Applications: The application of diffusion models beyond image synthesis could be explored. This includes fields such as medical imaging, where models could assist in generating patient-specific organ models, or in the creative arts, where they can aid in the design process.

Theoretical Understanding: There is a need for a deeper theoretical understanding of diffusion models. This includes research on the convergence properties of these models, the exploration of different noise schedules and their impact on sample quality, and the development of new theoretical frameworks to explain the inductive biases of diffusion models.

Combining with Other Generative Approaches: Hybrid models that combine the strengths of diffusion models with other generative approaches, such as GANs or autoregressive models, could be an area of future research. Such combinations might lead to models that are more robust and versatile.

Data Efficiency and Transfer Learning: Research could explore how diffusion models can be made more data-efficient, potentially through transfer learning or by leveraging pre-trained models on large datasets. This would allow for the application of these models in scenarios where data is scarce or expensive to obtain.

By pursuing these research directions, the field can continue to push the boundaries of what is possible with generative models, while also addressing the challenges and ethical considerations that come with these powerful tools.

4. Conclusion

DDPM, HighLDM, and Imagen are cutting-edge generative models. DDPM revolutionizes diffusion modeling with novel parametrization for continuous data and discrete likelihoods. HighLDM excels in high-resolution synthesis using latent diffusion and autoencoders, minimizing compute while maintaining detail. Imagen combines transformer-based language understanding with diffusion models to create photorealistic images from text, achieving state-of-the-art semantic alignment and visual fidelity. Each model uniquely advances its domain, showcasing versatility, efficiency, or linguistic grounding.

References

[1] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S., 2015, June. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (pp. 2256-2265). PMLR.

[2] Ho, J., Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, pp.6840-6851.

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

[4] Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M. and Chan, W., 2020. Wavegrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713.

[5] Kingma, D., Salimans, T., Poole, B. and Ho, J., 2021. Variational diffusion models. Advances in neural information processing systems, 34, pp.21696-21707.

[6] Kong, Z., Ping, W., Huang, J., Zhao, K. and Catanzaro, B., 2020. Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761.

[7] Mittal, G., Engel, J., Hawthorne, C. and Simon, I., 2021. Symbolic music generation with diffusion models. arXiv preprint arXiv:2103.16091.

[8] Dhariwal, P. and Nichol, A., 2021. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, pp.8780-8794.

[9] Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M. and Salimans, T., 2022. Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research, 23(47), pp.1-33.

[10] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J. and Norouzi, M., 2022. Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence, 45(4), pp.4713-4726.

[11] Ho, J., Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, pp.6840-6851.

[12] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

[13] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. and Ho, J., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35, pp.36479-36494.

[14] Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[15] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T., 2020. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8110-8119).

[16] Karras, T., Aila, T., Laine, S. and Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

[17] Jahn, M., Rombach, R. and Ommer, B., 2021. High-resolution complex scene synthesis with transformers. arXiv preprint arXiv:2105.06458.

[18] Park, T., Liu, M.Y., Wang, T.C. and Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2337-2346).

[19] Sylvain, T., Zhang, P., Bengio, Y., Hjelm, R.D. and Sharma, S., 2021, May. Object-centric image generation from layouts. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 3, pp. 2647-2655).

[20] Sun, W. and Wu, T., 2021. Learning layout and style reconfigurable gans for controllable image synthesis. IEEE transactions on pattern analysis and machine intelligence, 44(9), pp.5070-5087.

[21] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I., 2021, July. Zero-shot text-to-image generation. In International conference on machine learning (pp. 8821-8831). Pmlr.

[22] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2), p.3.

[23] Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J. and Sun, T., 2021. Lafite: Towards language-free training for text-to-image generation. arxiv 2021. arXiv preprint arXiv:2111.13792, 2.

[24] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. and Chen, M., 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

[25] Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D. and Taigman, Y., 2022, October. Make-a-scene: Scene-based text-to-image generation with human priors. In European Conference on Computer Vision (pp. 89-106). Cham: Springer Nature Switzerland.

[26] Zhang, H., Koh, J.Y., Baldridge, J., Lee, H. and Yang, Y., 2021. Cross-modal contrastive learning for text-to-image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 833-842).

[27] Ye, H., Yang, X., Takac, M., Sunderraman, R. and Ji, S., 2021. Improving text-to-image synthesis using contrastive learning. arXiv preprint arXiv:2107.02423.

[28] Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X.Y., Wu, F. and Bao, B., 2020. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. arXiv preprint arXiv:2008.05865, 2(6).

[29] Zhu, M., Pan, P., Chen, W. and Yang, Y., 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5802-5810).

[30] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X., 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324).