Journal of Computer Science & Technology
Indexed in: SCIE, EI, INSPEC, CBST, DBLP, etc.
Sponsored by: ICT, CAS & CCF

JCST Papers

Jiang R, Zheng GC, Li T *et al*. A survey of multimodal controllable diffusion models. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(3): 509-541 May 2024. DOI: 10.1007/s11390-024-3814-0

A Survey of Multimodal Controllable Diffusion Models

Rui Jiang[1, †] (江　锐), Guang-Cong Zheng[1, †] (郑光聪), Teng Li[1] (李　藤), Tian-Rui Yang[2] (杨天瑞)
Jing-Dong Wang[3] (王井东), *Fellow, IEEE, Distinguished Member, CCF*
and Xi Li[1, *] (李　玺), *Senior Member, CCF, IEEE*

[1] *College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China*

[2] *Department of Mathematics, Nanjing University, Nanjing 210023, China*

[3] *Baidu Visual Technology Department, Baidu Inc., Beijing 100085, China*

E-mail: jrss@zju.edu.cn; guangcongzheng@zju.edu.cn; tengli19@zju.edu.cn; 201840050@smail.nju.edu.cn; wangjingdong@baidu.com; xilizju@zju.edu.cn

Received September 28, 2023; accepted March 19, 2024.

Abstract Diffusion models have recently emerged as powerful generative models, producing high-fidelity samples across domains. Despite this, they have two key challenges, including improving the time-consuming iterative generation process and controlling and steering the generation process. Existing surveys provide broad overviews of diffusion model advancements. However, they lack comprehensive coverage specifically centered on techniques for controllable generation. This survey seeks to address this gap by providing a comprehensive and coherent review on controllable generation in diffusion models. We provide a detailed taxonomy defining controlled generation for diffusion models. Controllable generation is categorized based on the formulation, methodologies, and evaluation metrics. By enumerating the range of methods researchers have developed for enhanced control, we aim to establish controllable diffusion generation as a distinct subfield warranting dedicated focus. With this survey, we contextualize recent results, provide the dedicated treatment of controllable diffusion model generation, and outline limitations and future directions. To demonstrate applicability, we highlight controllable diffusion techniques for major computer vision tasks application. By consolidating methods and applications for controllable diffusion models, we hope to catalyze further innovations in reliable and scalable controllable generation.

Keywords diffusion model, controllable generation, application, personalization

1 Introduction

In recent years, the realm of artificial intelligence has experienced noteworthy advancements across various domains, encompassing computer vision, natural language processing, and reinforcement learning. And the area of generative models has undergone significant progress, where the primary objective is to produce samples of high fidelity and diversity from intricate data distributions. During the initial stages of generative models, conventional methods such as texture composition[1] and texture mapping[2] were employed. However, more sophisticated techniques like generative adversarial networks (GANs)[3, 4], variational Autoencoders (VAEs)[5], and normalizing flows[6] have risen to prominence as dominant approaches for generation with the passage of time.

More recently, the landscape of generative models has witnessed a paradigm shift, marked by the emergence of diffusion models[7]. This novel family of deep generative models has brought forth a comprehensible parameterization for probabilistic modeling, a sta-

---

Survey

This work is supported in part by the National Science Foundation for Distinguished Young Scholars of China under Grant No. 62225605, the National Natural Science Foundation of China under Grant No. U20A20222, the Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020016, and the Ng Teng Fong Charitable Foundation in the form of ZJU-SUTD IDEA under Grant No. 188170-11102.

†Equally Contributed (Rui Jiang was responsible for the theoretical underpinnings and comprehensive literature review within the survey. Guang-Cong Zheng was responsible for revising and improving the overall article structure.)

*Corresponding Author

ble training procedure supported by theoretical foundations, and a unified loss function characterized by its simplicity.

The structural components of diffusion models revolve around three key elements: a forward process, a reverse process, and a denoising procedure for sampling. The forward process is designed to convert the data distribution into random noise. The reverse process employs a learnable neural network to estimate the transformation kernel step by step to undo the forward process, as outlined by [8]. The sampling procedure obtains random noise and employs the optimized network to generate data. The difference between the sampling procedure and reverse process is that the network used during sampling is already optimized and exclusively employed for inference. These three components can be implemented in either discrete[9, 10] or continuous[11, 12] manners.

Nevertheless, it is crucial to recognize that diffusion models inherently involve a more time-consuming sampling procedure[13] when compared with GANs or VAEs. This extended duration can be attributed to the iterative transformation from the prior distribution into more complex data distributions through ODE, SDE[14–17], or Markov processes, which mandates numerous function evaluations in the process. Additional challenges include the control and steering of the generation.

In response, researchers have proactively proposed a range of solutions to address challenges associated with diffusion models. Advanced solvers on either ODE or SDE[14–17] and model distillation techniques[18] are introduced to expedite the sampling process. Guidance mechanisms are explored to correct the unconditional direction[19] given guiding conditions, reducing the discrepancy between the desired[20] and reference conditional distributions[21]. Such conditions can be of diverse modalities[22, 23], including images[24], texts[25], or 2D poses[26, 27].

Although there are several surveys[28–30] delving into various aspects of diffusion models, many fall short of offering a comprehensive investigation into controllable generation. And certain surveys[31–33] prioritize the application side, providing valuable insights into practical applications but offering limited coverage of controllable techniques.

This survey bridges the gap in the literature by offering a comprehensive and cohesive review of controllable generation. Specifically, we present a taxonomy encompassing various forms of control in the context of diffusion-based image synthesis, providing a succinct summary of diverse techniques and strategies, as illustrated in Fig.1. We also explore different application scenarios where controllable generations are successfully applied. Through a careful examination of these examples, our aim is to provide valuable insights into the potential of controllable diffusion models and to inspire new directions for future research in this dynamic and evolving field.

We will explore the foundational theories and

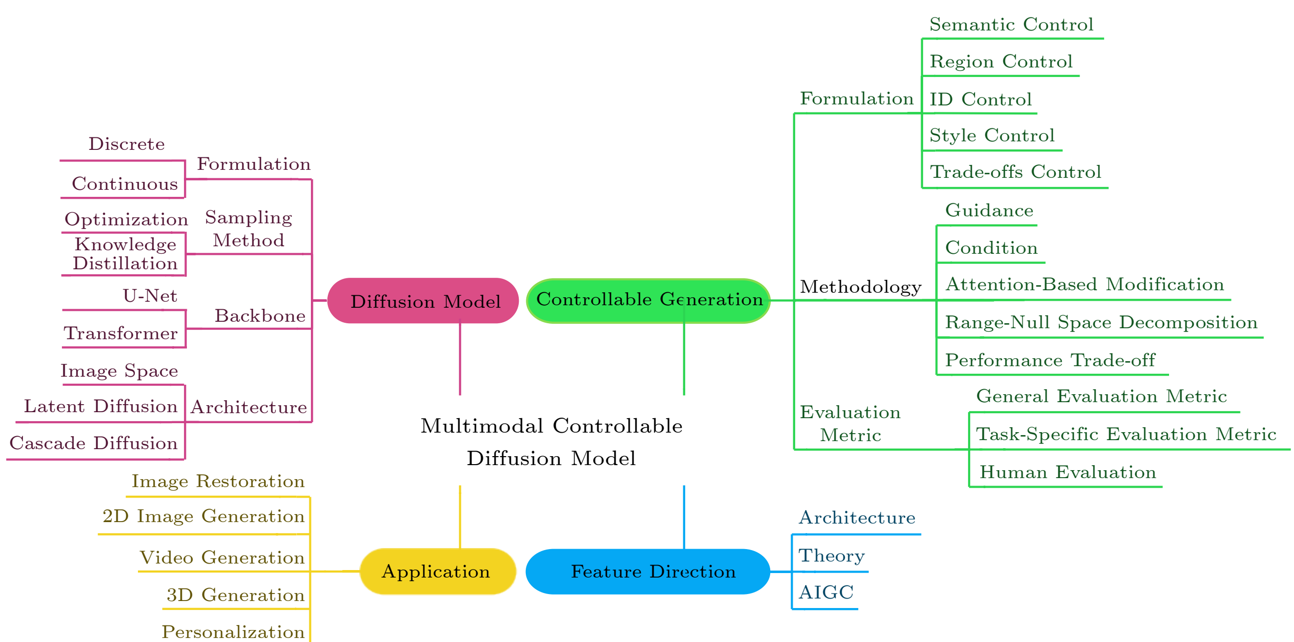


Fig.1. Overview of multimodal controllable diffusion models.

components of diffusion in Section 2. In Section 3, we will discuss several forms of controllable generation, and review the current solutions that have been developed to achieve this. In Section 4, we will explore the diverse applications of controllable generation. Finally, we will conclude with a discussion on the potential research trends and future directions for diffusion-based technologies in Section 5. Finally, we conclude the paper in Section 6.

2 Diffusion Model

2.1 Discrete Form

2.1.1 DDPM

The Denoising Diffusion Probabilistic Model (DDPM)[10] leverages two Markov chains, as well-known as the forward process and reverse process, to generate images of high fidelity. The comprehensive workflow of the diffusion model is illustrated in Fig.2. Diffusion models are a class of latent variable models characterized by the expression: $p_\theta(\boldsymbol{x}_0) := \int p_\theta(\boldsymbol{x}_{0:T})\mathrm{d}\boldsymbol{x}_{1:T}$. In this formulation, the latent variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ possess the same dimensionality as the observed data $\boldsymbol{x}_0$, which is distributed according to $q(\boldsymbol{x}_0)$. In the forward process, noises sampled from a prior distribution, typically standard Gaussian, are applied iteratively to corrupt a clean image $\boldsymbol{x}_0$. This transformation can be achieved by using Markov transition kernels, of which coefficients are denoted sequentially by $\beta_1, \beta_2, \beta_3, \dots, \beta_T$:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\boldsymbol{I}),$$

where $\boldsymbol{I}$ denotes the identity matrix.

Given the addition property of Gaussian, the transition kernel can be reformulated to avoid repetitive steps, making possible direct calculation from $\boldsymbol{x}_0$:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}),$$

where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. At the end of the forward process, $\boldsymbol{x}_T$ will theoretically follow the Gaussian distribution, as $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) \approx \mathcal{N}(0, \boldsymbol{I})$.

The reverse process parameterizes its transition kernel as neural networks and is capable of turning Gaussian noise $\boldsymbol{x}_T$ back to a clean image at timestamp 0:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\big(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\big),$$

where $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ denote the mean and variance of Gaussian, respectively. By rule of thumb, $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ is fixed to a constant β_t in practice.

Here KL divergence is introduced to minimize the distance between the learnable transition kernel and the Bayesian posterior of the forward process derived as $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t\boldsymbol{I})$, by optimizing variational lower bound (VLB), i.e. evidence lower bound (ELBO), on their negative log likelihood:

$$\mathbb{E}_q\Bigg[\underbrace{-\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)}_{L_0} + \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{L_T} + \sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))}_{L_{t-1}}\Bigg],$$

where L_0 and L_T can be ignored for simplicity, that is, with respect to $\boldsymbol{x}_0$:

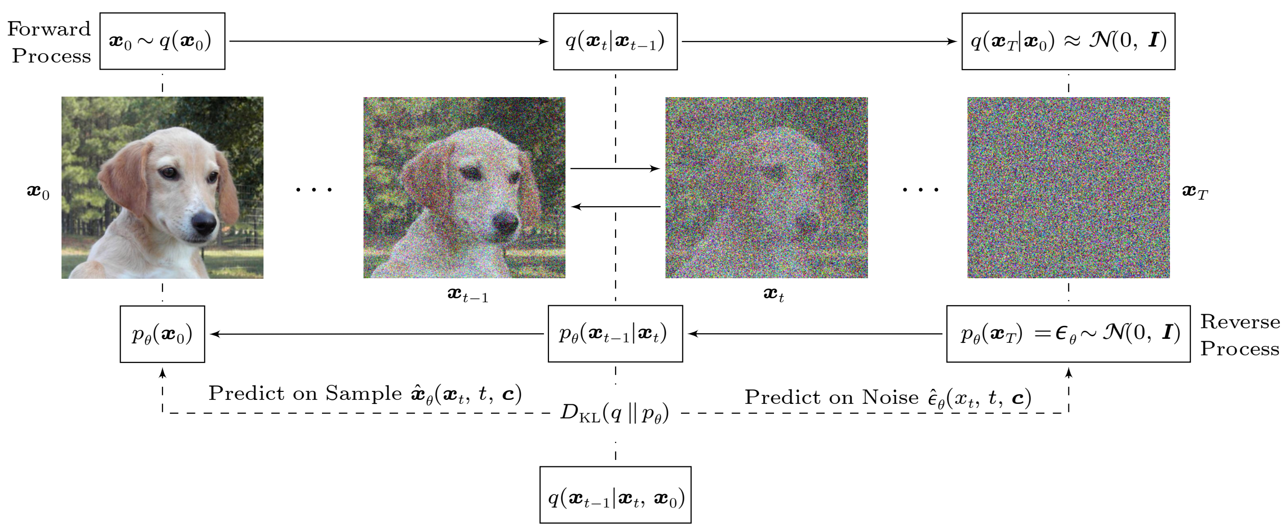


Fig.2. Diffusion models alter the data by adding noise to it, and then generate new data from the noise through the inverse process. In the reverse process, each denoising step requires estimating the transition kernel.

$$L_{t-1}=\mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0)-\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t\right)\right\|^2\right].$$

The reparameterization trick of noise prediction regards $\boldsymbol{x}_t$ as $\boldsymbol{x}_t(\boldsymbol{x}_0,\boldsymbol{\epsilon})=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, thus the loss function can be further simplified as:

$$L_{t-1}=\mathbb{E}_{t,\,\boldsymbol{x}_0,\,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right].$$

On an optimized neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)$, the sampling procedure can be achieved iteratively with random noises $\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I})$ by:

$$\boldsymbol{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)\right)+\sigma_t\boldsymbol{z}.$$

2.1.2 SMLD

Denoising Score Matching with Langevin Dynamics (SMLD) is a method that employs the estimation of scores, representing the gradient of the log probability density with respect to data, at varying noise scales. Score perspective models employ a maximum likelihood-based estimation approach, utilizing the score function of the log-likelihood of the data to estimate the parameters of the diffusion process. The score function $(\nabla x\log p(x))$ of a given data distribution $p(x)$ is estimated through score matching by training a shared neural network s_θ parameterized by θ, which approximates the score of $p(x)s_\theta(x)\approx\nabla x\log p(x)$, achieved by minimizing the corresponding objective:

$$\mathbb{E}_{x\sim p(x)}\left\|s_\theta(x)-\nabla_x\log p(x)\right\|_2^2.$$

However, the computational complexity associated with calculating the gradient of the log density $\nabla x\log p(x)$ hampers the scalability of score matching to deep networks and high-dimensional data. To address this challenge, Song *et al.*[9] proposed the utilization of denoising score matching and sliced score matching techniques. The authors further proposed training a single noise-conditioned score network (NCSN) to estimate scores corresponding to all noise levels. They derive $\nabla_x\log(p_{\sigma_t}(x))$ as $\nabla_{x_t}\log p_{\sigma_t(x_t|x)}=-(x_t-x)/\sigma_t$, given that:

$$\begin{aligned}p_{\sigma_t}(x_t|x)&=\mathcal{N}(x_t,x,\sigma_t^2\boldsymbol{I})\\&=\frac{1}{\sigma_t\sqrt{2\pi}}\times\exp\frac{-(x_t-x)^2}{2\sigma_t^2},\end{aligned}$$

where x_t represents a noised version of x. The process of inference is carried out through the utilization of an iterative technique known as Langevin dynamics. Langevin dynamics employ a Markov Chain Monte Carlo (MCMC) approach to generate samples from a distribution $p(x)$ solely based on its score function, $\nabla x\log p(x)$. To transform from an initial random sample x_0 towards samples from $p(x)$, the algorithm iteratively performs the following steps:

$$x_t^i=x_{t-1}+\frac{\gamma}{2}\nabla_x\log p(x)+\sqrt{\gamma}\times\omega_i,\ i\in[0,N],$$

where ω_i is drawn from a standard normal distribution, and γ represents the friction coefficient of the environment where the particle resides.

2.2 Continuous Form

DDPMs and SMLD can be further generalized to the case of infinite time steps or noise levels, where the perturbation and denoising processes are solutions to stochastic differential equations (SDEs). This formulation is called Score SDE[11], as it leverages SDEs for noise perturbation and sample generation, and the denoising process requires estimating score functions of noise data distributions:

$$\mathrm{d}\boldsymbol{x}=f(\boldsymbol{x},t)\mathrm{d}t+g(t)\mathrm{d}\boldsymbol{w},$$

where $t\in[0,T]$, $f(\cdot,\cdot)$ and $g(\cdot)$ are the drift and diffusion coefficients, respectively, and $\{\boldsymbol{w}_t\}_{t\in[0,\,T]}$ denotes the standard Brownian motion. The forward process in DDPM is a discretization of SDE. For DDPMs, its corresponding SDEs transition kernels are:

$$\begin{aligned}f(\boldsymbol{x},t)&=-\frac{1}{2}\beta(t)\boldsymbol{x},\\g(t)&=\sqrt{\beta(t)}.\end{aligned}$$

For SMLD[9], its corresponding SDEs transition kernels are:

$$\begin{aligned}f(\boldsymbol{x},t)&=0,\\g(t)&=\sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}}.\end{aligned}$$

The trajectories of the reverse SDE share the same marginal density as those of the forward SDE, with the only difference being that they evolve in the opposite time direction.

Moreover, Anderson's work[34] is of considerable importance in the study of diffusion processes, as he showed that the diffusion process can be reversed by solving a time-reverse SDE:

$$\mathrm{d}\boldsymbol{x} = \left[f(\boldsymbol{x},t) - g(t)^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right] \mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}.$$

Song *et al.*[11] found a property that the trajectory of a new type of ordinary differential equation (ODE) called the probabilistic flow ODE has the same marginal density as the trajectory of the time-reverse SDE:

$$\mathrm{d}\boldsymbol{x} = \left[f(\boldsymbol{x},t) - \frac{1}{2}g(t)^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right] \mathrm{d}t. \quad (1)$$

Due to the lack of randomness, ODEs can be solved using larger step sizes, thus speeding up convergence and reducing computational costs. Some work such as DPM-Solver[35] and DPM-Solver++[17] obtains faster sampling speed based on acceleration techniques of ODE. The training objective is defined as:

$$L = \mathbb{E}_t \left[\lambda(t) \mathbb{E}_{x_0} \mathbb{E}_{q(x_t|x_0)} \left\| s_\theta(x_t,t) - \nabla_{x_t} \log p(x_t|x_0) \right\|_2^2 \right],$$

where $\lambda(t)$ is the positive weighting function, $s_\theta(x_t,t)$ is the output of the denoising network at time t.

2.3 Sampling Method

2.3.1 Optimization

A popular approach for optimization sampling in diffusion models centers on directly solving the probability flow ODE (1) from the continuous-time perspective. Denoising diffusion implicit models (DDIM)[15] accelerated sampling by adopting a deterministic process aligned with the probability flow ODE. In subsequent studies[35, 36], DDIM has been interpreted as the result of applying an exponential integrator to the ODE governing variance preserving (VP) diffusion[11]. This interpretation sheds light on the underlying mechanisms of DDIM and its relationship to VP diffusion. Moreover, recent advancements in the field have seen the utilization of advanced ODE solvers in various methodologies, including PNDM[37], EDM[12], DEIS[36], gDDIM[38], and DPM-Solver[35]. For instance, EDM employs efficient Heun's[39] second order ODE solvers to tackle the computational challenges inherent in diffusion models. DPM-Solver[35] proposes improved higher-order ODE solvers tailored for generative modeling, leveraging semi-linear structure and approximating solutions to reduce error. Extensions like DPM-Solver++[17] incorporate data-conditional constraints during ODE integration to improve sample quality and stability.

Other methods based on KL-divergence optimization set the reverse mean and covariance using the Monte Carlo method. Although these methods, such as Analytic-DPM[16] and extended Analytic-DPM[40], provide optimal reverse solutions while accounting for correction at each state, they are restricted in their applicability to specific distributions due to their pre-assumptions.

2.3.2 Knowledge Distillation

Knowledge distillation was originally proposed as a model compression technique, where a smaller "student" network is trained to mimic the outputs of a larger "teacher" model. The key idea is that the student learns an efficient representation that matches the teacher's performance. Recent work has adapted knowledge distillation to compress the sampling procedures of diffusion models[18, 41, 42]. The original sampling process serves as the teacher, while a student with fewer steps is trained to match its outputs using distillation objectives. This allows reducing sampling complexity and cost[43].

Denoising Student[44] and DSNO[45] focus on optimizing the distillation process for maximum speedup and efficiency. This requires a large and costly dataset[46] of teacher samples for distillation. Progressive distillation[18] addresses this by gradually merging pairs of teacher steps into the student. After compressing two steps, the student becomes the teacher for the next round[47]. However, more rounds of progressive distillation can compound errors and degrade sample quality. Managing this trade-off remains an open challenge, with work on new distillation architectures and objectives to allow deeper compression[48].

2.4 Backbone

2.4.1 U-Net

U-Net[49] is implemented with an overlap-tile strategy and mirroring extrapolation to segment images of arbitrary size. U-Net's combination of effective feature localization, skip connections, and computational efficiency has contributed to its widespread adoption. Several architecture modifications are made to adapt U-Net as the backbone of diffusion, including replacing weight normalization[50] with group normalization[51] for learning efficiency, adding dense connections between two groups to help in the vanishing gradient problem, incorporating attention block[52] for higher capacity, and exploring normalization layers as conditions in diffusion models[53].

2.4.2 Transformer

The Transformer architecture[54] has become a focus for incorporation into diffusion models for generative modeling[55]. Transformers offer useful abilities for modeling long-range dependencies in image and sequence data[56]. Recent work by Peebles and Xie[57] proposed the Transformer-based diffusion model DiTs, showing improved sample quality on image modeling tasks. Follow-up work U-Vit[58] and MDT[59] has continued modifying the Transformer architecture design for diffusion. They prove that the inclusion of long skip connections is crucial for diffusion-based image modeling, while down/up connections play a key role. Despite promising results, several challenges remain. Modeling long sequences is still costly, with quadratic memory and compute requirements. More work is needed to scale up Transformers to handle high-resolution multimodal data.

2.5 Architecture

2.5.1 Image Space Diffusion Model

Image space diffusion models, exemplified by the seminal model DDPM[10], function by directly diffusing and sampling within the pixel domain, as depicted in Fig.3(a). This approach offers conceptual simplicity and facilitates direct optimization of the data distribution[60]. By leveraging neural networks, image space diffusion models effectively capture both local and global image features, resulting in the generation of high-quality and visually-coherent samples. Moreover, the image space optimization allows for the integration of image-specific techniques, such as perceptual loss functions[61], to improve the alignment between generated samples and the target distribution.

However, it is important to note that generating high-dimensional data, like images, through pixel-level sampling can pose computational challenges and often necessitates significant computational resources compared with those in the latent space[62]. Additionally, image space diffusion may occasionally produce pixel values outside the valid range, resulting in noticeable clipping artifacts[63].

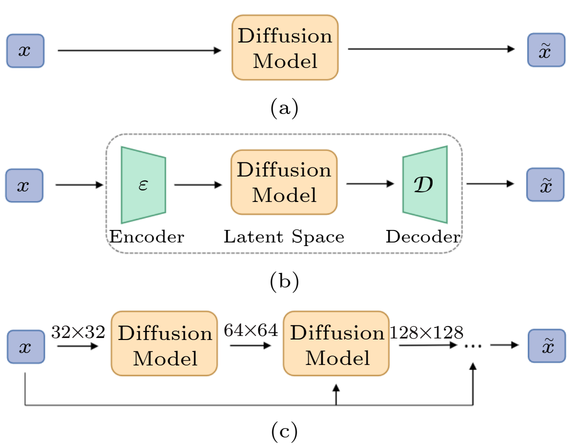


Fig.3. Architecture of (a) Image Space Diffusion Model (DM), (b) Latent Diffusion Model (LDM), and (c) Cascade Diffusion Model (CDM).

2.5.2 Latent Diffusion Model

Latent space diffusion models have emerged as a powerful generative modeling approach for images and other modalities[64]. Unlike typical generative models that directly output pixels or waveforms, latent diffusion operates in a learned compact latent space, as depicted in Fig.3(b). Specifically, the model encodes the data into this lower-dimensional latent space, then performs diffusion and sampling followed by decoding to the output[63, 65].

Working in the latent space provides several advantages. Firstly, sampling complex high-dimensional data like images is more stable and efficient in the compressed latent representation[66]. Secondly, the decoder acts as a strong prior to convert sampled latent codes into realistic outputs[67, 68]. Finally, manipulating the latent space gives fine-grained control over attributes of generated samples[69, 70]. Notable latent diffusion models such as DALL-E 2[71], Audioldm[72], and SLD[73] have demonstrated state-of-the-art sample quality and training stability. Meanwhile, the latent-based Diffusion method is also shining in the field of video generation[74, 75].

Latent space diffusion models, although promising for generative modeling, are not without their limitations and challenges. One drawback pertains to the loss of pixel-level granularity in the generated samples, stemming from their operation within a compressed latent space. Furthermore, the interpretability of the latent space poses a concern, as unraveling the semantic correspondence between latent dimensions and resulting image transformations remains an ongoing challenge. The reliance on the learned prior distribution represents an additional limitation, as it may result in the generation of samples exhibiting a pronounced dependence on the prior, potentially leading to a lack of diversity or deviation from the desired distribution.

2.5.3 *Cascade Diffusion Model*

Cascading refers to a multi-stage generative modeling approach for producing high-resolution images, introduced by Saharia *et al.*[76]. It involves training a pipeline of separate models at progressively increasing resolutions, as depicted in Fig.3(c). Any type of generative model could be used in a cascading pipeline[77].

This cascading strategy confers several benefits. By initially sampling at low resolutions, it is more computationally efficient and stable[78, 79]. The super-resolution models can then focus on adding high-frequency details on top of the low-frequency structure[80]. Cascading also allows combining different specialized models for each resolution[76]. After training a single model and progressively splitting it into specialized models for different synthesis stages, the eDiff-I[81] ensemble outperforms previous large-scale text-to-image diffusion models on the standard benchmark and allows for the exploitation of a variety of embeddings for conditioning.

Cascade diffusion models are not exempt from certain limitations and challenges. A notable drawback resides in the inherent sequentiality of the diffusion process. This protracted sequence engenders sluggish convergence and augmented computational complexity, thereby rendering cascade diffusion models computationally burdensome and time-intensive. Moreover, the sequential nature of diffusion engenders the potential for accumulated errors at each step, thereby jeopardizing the preservation of fine-grained details and veracity in the generated samples.

3 Controllable Generation

3.1 Formulation

Controllable generation can take various forms, depending on the specific domain. Here are some common forms of controllable generation in Fig.4.

3.1.1 *Semantic Control*

Semantic-controlled image generation refers to the ability to precisely manipulate salient image attributes or characteristics during image generation. This precise controllability allows for fine-grained adjustments to the generated images. Its applications range widely from class-to-image[7, 78] and text-to-image[79, 82–84] generation to synthetic data augmentation[85, 86]. The main challenges are endowing generative models with semantic understanding so they can represent image attributes disentangled from other factors and respond precisely to semantic controls. This precision control during generation results in images with user-specified semantic characteristics.

3.1.2 *Spatial Control*

Spatial-controlled image generation refers to fine-grained controls on the contents in specific regions of the generated images. Layout- or segmentation-guided approaches[83, 87–91] perform generation spatially conditioned on bounding boxes or segmentation maps. Sketch- or edge-guided approaches[22, 84, 92–96] synthesize images by completion from either user scribbles or detected edges of the reference image. Depth-guided approaches[22, 84, 94–97] constrain the process by depth priors, which can be estimated in the monocular manner for practice. Skeleton-guided approaches[22, 27, 84, 93, 95, 96] calibrate human poses in the synthesis using keypoints generated by pre-trained OpenPose[98].

Recent efforts[99] have focused on combining spatial coordinates alongside natural language descriptions to achieve precise region control in text-to-image generation. The pioneering work of ControlNet[22]

Fig.4. Example of semantic control, spatial control, ID control, and style control.

and FreeControl[100] successfully instills spatial information from multi-modal guiding maps, like sketches, depth maps, or human poses into a trainable copy of denoising U-Net, which is affiliated to the original frozen model via zero convolution and is capable of visually-compelling and textually-coherent synthesis. In addition, the authors of ControlNet designed Fooocus① with many optimizations and quality improvements built in and automated, turning manual settings on other pages into automatic configuration.

3.1.3 ID Control

ID-controlled image generation refers to conditioning image synthesis on user-specified identity information to generate images of specific individuals. ID-conditioned image generation was first introduced in StyleGAN[4] to control stochastic variation in GANs. Unique IDs were mapped to seeds that control the latent space sampling. StyleCLIP[101] and StyleSpace[102] extend ID-conditioning by introducing text-conditional control through CLIP.

In the field of diffusion, the concept of ID control has been further expanded to object customization, allowing users to have fine-grained control over the generation process to tailor outputs according to their individual preferences and specific requirements. These diffusion methodologies can be broadly categorized into optimization-based techniques[103] and encoder-based approaches[104, 105]. Optimization-based methods exhibit the potential to preserve identity with fidelity; however, they often suffer from time-consuming computations and may occasionally lead to overfitting. Conversely, contemporary encoder-based approaches offer the advantage of zero-shot performance, but they may sacrifice identity preservation or generate outcomes of copy-pasting.

3.1.4 Style Control

While diffusion models can generate remarkable photorealistic images, controlling specific attributes like visual style remains difficult when conditioned solely on text prompts or example images. This limitation constrains the full creative potential of generative art. Recent work has begun tackling finer-grained control through techniques like style-based guidance[106–109], where separate style and content latent codes are decoded to isolate stylistic factors. Some approaches explore directional style transfer via weighted interpolation in the latent space[109]. Energy-guided methods[110] draw inspiration from classifier guidance[7], utilizing estimated loss gradients to guide the generation process at each sampling step. These methods employ carefully designed energy functions to assess the discrepancy between the generated output and the target style. To improve efficiency, coarse-grained predictions are often used instead of directly utilizing the output of the diffusion model.

Moreover, it is noteworthy that style transfer can be effectively achieved through the process of fine-tuning a pre-existing model. Personalization-based methodologies encompass the practice of refining a pre-trained model using sophisticated techniques such as Textual Inversion[111], Dreambooth[112], or LoRA[113]. Subsequently, the fine-tuned model is employed to decode the latent codes of inverted content images. This approach shares resemblances with the GAN Adaptation method. However, most control mechanisms remain discrete rather than continuous. An open research direction is enabling fluid, granular manipulation of attributes like color, texture, lighting, etc. This could be achieved by mapping generative parameters to an intuitive creative interface[114]. If generative models could smoothly interpolate between granular artistic attributes based on interactive human guidance, it would greatly empower creative expression.

3.1.5 Controllability Trade-Off

Fidelity-Diversity Trade-Off. Balancing diversity with fidelity to the user preference is a key aspect of controllability in generative models. The fidelity-diversity trade-off is delineated in Fig.5(a). Models that adhere too strictly to conditional inputs may suffer from outputs lack of variety, while models that introduce too much randomness can deviate from user intent. Recent work has aimed to improve trade-offs through technical advances. For example, DALL-E 2[71] uses a context-conditioned variation module that maintains fidelity to the text prompt while still allowing for diversity by sampling different latent codes. Similarly, Parti[115] separates the text embedding into a content code for fidelity and a style code for diversity.

Faithfulness-Realism Trade-Off. The trade-off between faithfulness and realism pertains to finding a

①https://github.com/lllyasviel/Fooocus, May 2024.

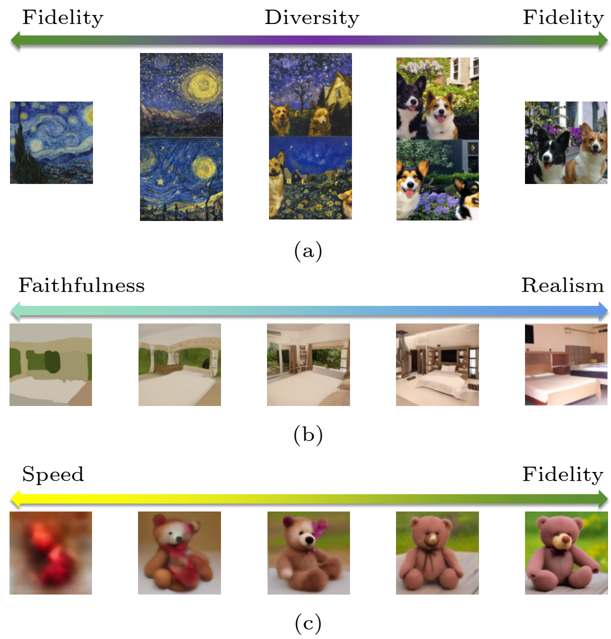


Fig.5. Example of trade-offs control. (a) Fidelity-diversity trade-off. (b) Faithfulness-realism trade-off. (c) Speed-fidelity trade-off.

balance where the generated images closely adhere to the prompt (faithfulness) while also exhibiting a natural and realistic appearance (realism). The faithfulness-realism trade-off is delineated in Fig.5(b). By introducing additional Gaussian noise along with stochastic diffusion, the synthesized images are more realistic but less faithful[116]. The optimal balance produces images that both fulfill the user's intent and visualize the request in a realistic style.

Speed-Fidelity Trade-Off. There is an inherent trade-off between speed and fidelity (image quality). Using more diffusion steps results in higher quality images but takes longer to generate each sample. The speed-fidelity trade-off is delineated in Fig.5(c). Using fewer steps speeds up sampling but can reduce image quality. One way to adjust this trade-off is by changing the number of diffusion steps. More steps improve fidelity at the cost of speed. Fewer steps increase speed but may introduce artifacts or reduce image coherence.

3.2 Methodologies

3.2.1 *Guidance*

This category of work utilizes a frozen pre-trained diffusion model as a foundation model but introduces modifications to the sampling method, incorporating feedback from the guidance function to guide the image generation process. For instance, Dhariwal and Nichol[7] proposed classifier-guidance, where a classifier was trained on images of different noise scales to serve as the guidance function. For the generation conditioned on y, the classifier-guidance method entails the replacement of $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ with $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})$. The condition $\boldsymbol{y}$ can be of various forms, such as text[117], class[78], image-based[22], and multi-modal condition[118].

For the sake of simplicity and without sacrificing generality, we will discuss guidance using the output in the form of scores. The objective is to learn the score of the conditional model, represented as $\nabla \log p\left(\boldsymbol{x}_t \mid \boldsymbol{y}\right)$, at a noise level t. To simplify the notation, we use ∇ as shorthand for $\nabla_{\boldsymbol{x}_t}$. Applying Bayes' rule, we can derive the following equation:

$$
\begin{aligned}
& \nabla \log p\left(\boldsymbol{x}_t \mid \boldsymbol{y}\right) \\
& =\nabla \log \left(\frac{p\left(\boldsymbol{x}_t\right) p\left(\boldsymbol{y} \mid \boldsymbol{x}_t\right)}{p(\boldsymbol{y})}\right) \\
& =\nabla \log p\left(\boldsymbol{x}_t\right)+\nabla \log p\left(\boldsymbol{y} \mid \boldsymbol{x}_t\right)-\nabla \log p(\boldsymbol{y}) \\
& =\underbrace{\nabla \log p\left(\boldsymbol{x}_t\right)}_{\text{unconditional score}}+\underbrace{\nabla \log p\left(\boldsymbol{y} \mid \boldsymbol{x}_t\right)}_{\text{adversarial gradient}} .
\end{aligned} \quad (2)
$$

Note that in the forward process, $\boldsymbol{x}_t$ is obtained from $\boldsymbol{x}_{t-1}$ by adding a noise, which will not contribute to the classification, so the gradient of $\log p(\boldsymbol{y})$ with respect to $\boldsymbol{x}_t$ is zero. The final result involves learning an unconditional score function combined with the adversarial gradient of a classifier, $p(\boldsymbol{y}|\boldsymbol{x}_t)$.

Classifier guidance[7, 119, 120] involves training the score of an unconditional diffusion model and a classifier simultaneously. The classifier takes in noisy input $\boldsymbol{x}_t$ and predicts the conditional information $\boldsymbol{y}$. During the sampling process, the overall conditional score function for annealed Langevin dynamics[121] is computed by adding the unconditional score function to the adversarial gradient of the classifier. To control the influence of conditioning information, classifier guidance introduces a hyperparameter γ to scale the adversarial gradient. Therefore, the learned score function under classifier guidance can be summarized as follows:

$$
\nabla \log p\left(\boldsymbol{x}_t \mid \boldsymbol{y}\right)=\nabla \log p\left(\boldsymbol{x}_t\right)+\gamma \nabla \log p\left(\boldsymbol{y} \mid \boldsymbol{x}_t\right) . \quad (3)
$$

Intuitively, by setting $\gamma=0$, the conditional diffusion model learns to disregard the conditioning information completely. On the other hand, when γ takes on a larger value, the model becomes more inclined to generate samples that closely align with the conditioning information. However, this emphasis on adherence to conditioning information comes at the expense of sample diversity[122].

Classifier guidance diffusion incorporates gradient information towards the target category in each step of the reverse process to achieve targeted image generation. This process bears similarities to optimization-based image generation algorithms, where a fixed network directly optimizes the image itself. Consequently, previous optimization-based image generation algorithms can be adapted to the diffusion model by modifying the condition type in guided diffusion. For example, semantic guide diffusion (SGD)[120] introduces two forms of category guidance: reference-based and text-based guidance. By designing corresponding gradient items, SGD achieves the desired guidance effect and produces high-quality results.

However, learning a classifier may come with extra costs and training instability[123], as it requires training on data with scheduled noise levels. This instability is further compounded by the fact that training on noisy data can be difficult due to the destruction of the data structure caused by the addition of more and larger noise according to the noise schedule[124]. Furthermore, generating images via gradients can lead to adversarial attack effects[125], where imperceptible details fool classifiers and are not actually generated conditionally, raising concerns about the reliability of the generated images.

3.2.2 Condition

Methods in this category involve the training of new diffusion models that incorporate the prompt as an additional input[119, 123, 126]. For instance, the approach proposed in [123] employs classifier-free guidance using class labels as prompts. The diffusion model in this case is trained via linear interpolation between the unconditional and conditional outputs of the denoising networks. In classifier-free guidance (CFG)[123], the authors proposed an alternative approach where a separate classifier model is not trained. Instead, they utilized an unconditional diffusion model and a conditional diffusion model. To obtain the score function under CFG, we can rearrange (2) to demonstrate the following relationship:

$$\nabla \log p\left(\boldsymbol{y} \mid \boldsymbol{x}_t\right)=\nabla \log p\left(\boldsymbol{x}_t \mid \boldsymbol{y}\right)-\nabla \log p\left(\boldsymbol{x}_t\right).$$

By substituting this derived expression into (3), we obtain the following result:

$$\begin{aligned}
&\nabla \log p\left(\boldsymbol{x}_t \mid \boldsymbol{y}\right)\\
&=\nabla \log p\left(\boldsymbol{x}_t\right)+\gamma \nabla \log p\left(\boldsymbol{x}_t \mid \boldsymbol{y}\right)-\gamma \nabla \log p\left(\boldsymbol{x}_t\right)\\
&=\underbrace{\gamma \nabla \log p\left(\boldsymbol{x}_t \mid \boldsymbol{y}\right)}_{\text{conditional score}}+\underbrace{(1-\gamma) \nabla \log p\left(\boldsymbol{x}_t\right)}_{\text{unconditional score}}.
\end{aligned}$$

When γ is set to 0, the conditional model completely disregards the conditioner and learns an unconditional diffusion model. On the other hand, when γ is set to 1, the model learns the vanilla conditional distribution without any additional guidance. When γ is greater than 1, the diffusion model not only prioritizes the conditional score function, but also moves away from the unconditional score function. This means that the model reduces the likelihood of generating samples that do not utilize conditioning information[122], favoring samples that explicitly incorporate it. However, this comes at the expense of reduced sample diversity, as the model becomes more focused on accurately matching the conditioning information.

The study in [126] investigates scenarios where the guidance function takes the form of a known linear degradation operator. A conditional model is then trained to tackle linear inverse problems. In another extension to classifier-free guidance, [119] introduces an approach for text-conditional image generation, using descriptive phrases as prompts. The diffusion model is trained with the objective of maintaining similarity between the CLIP[127] representations of the created images and the text prompts. However, one significant drawback is that the necessity to retrain the diffusion model for each new application makes them computationally intensive and potentially time-consuming.

3.2.3 Attention-Based Modification

Some approaches such as [128–133] utilize cross-attention in U-Net control to enable conditional generation. They discover a significant local similarity in the cross-attention map[134] between word features and objects, which serves as a valuable editing medium. Specifically, let the original text description be $\mathcal{P}$, the diffusion model generation process be $\boldsymbol{x}_T \rightarrow \boldsymbol{x}_0=\boldsymbol{I}$, the edited text is described as $\mathcal{P}^*$, we would like to get the edited image I^*. In a cross-attention layer, the image features $\phi(\boldsymbol{x}_t)$ are linearly mapped to Q. The text embedding is obtained by linear mapping as K and V, the final output:

$$\hat{\phi}\left(\boldsymbol{x}_t\right)=\operatorname{Softmax}\left(\frac{Q K^{\mathrm{T}}}{\sqrt{d}}\right) V.$$

In order to edit the image, we have created an image with both $\mathcal{P}$ and $\mathcal{P}^*$ conditions, then at time step t there will be two attention maps M_t^* and $\hat{M}_t$, which are obtained by a well-designed editing function. The new attention map can be edited by overwriting the original attention map with $\hat{M}_t = \mathrm{Edit}(M_t, M_t^*, t)$. The purpose of editing can be achieved by overwriting the original attention map $\hat{M}_t$. Additionally, [116] enables image translation by adjusting the initial noisy images. There has been some new progress in this field recently[93, 133].

3.2.4 Range-Null Space Decomposition

Recent techniques, such as [135–138], directly modify intermediate results to achieve zero-shot image restoration. DDNM[138] elucidates the essence of these methods. DDNM begins by addressing noise-free linear image inverse problems, wherein an image $y = Ax$ is degraded. Here, A represents a linear operator and x denotes the original image. The objective of image restoration is to obtain an estimated result $\hat{\boldsymbol{x}}$ that satisfies two constraints:

$$\mathrm{Consistency}: A\hat{\boldsymbol{x}} \equiv \boldsymbol{y},$$
$$\mathrm{Realness}: \hat{\boldsymbol{x}} \sim q(\boldsymbol{x}),$$

where $q(\boldsymbol{x})$ represents the distribution of the ground truth (GT) images. This problem possesses a general solution that analytically fulfills the consistency constraint:

$$\hat{\boldsymbol{x}} = A^{\dagger}\boldsymbol{y} + (\boldsymbol{I} - A^{\dagger}A)\boldsymbol{x}_r. \tag{4}$$

Here, $A^{\dagger}$ represents the pseudo-inverse of A, satisfying the condition $A^{\dagger}AA \equiv A$, while $\boldsymbol{x}_r$ denotes the unknown null-space variable that needs to be solved. It is worth noting that (4) originates from the range-null space decomposition[138–140]. Furthermore, $(I - A^{\dagger}A)\boldsymbol{x}_r$ is a generalization to $A\boldsymbol{x} = 0$ since $A(I - A^{\dagger}A)\boldsymbol{x}_r \equiv (A - A)\boldsymbol{x}_r \equiv 0$ regardless of $\boldsymbol{x}_r$. A crucial step in employing diffusion models for inverse problems involves considering each estimation $\boldsymbol{x}_{0|t}$ as the null-space variable $\boldsymbol{x}_r$ in (4):

$$\hat{\boldsymbol{x}}_{0|t} = A^{\dagger}\boldsymbol{y} + (\boldsymbol{I} - A^{\dagger}A)\boldsymbol{x}_{0|t}.$$

Subsequently, the obtained consistent result $\hat{\boldsymbol{x}}_{0|t}$ is utilized for subsequent sampling purposes.

3.2.5 Performance Trade-Offs

Truncation. Truncation trick is a technique used in GANs, flow models, and VAEs to trade off diversity for improved sample quality and fidelity. It works by restricting the sampling distribution, for instance by reducing the variance of noise inputs. This yields higher fidelity outputs but with less diversity. For example, BigGAN[141] uses truncated sampling to improve fréchet inception distance (FID)[142] at the cost of reduced inception score (IS)[143]. However, straightforward truncated sampling techniques prove ineffective for diffusion models[7]. Simply limiting noise variance during sampling leads to low-quality, blurry outputs. The sequential sampling process in diffusions requires more sophisticated techniques to restrict diversity and improve fidelity. Recent progress has been made with heuristic guidance[76] and latent space modeling[64].

Timestep Respacing. Timestep respacing is a technique to adjust the spacing between timesteps in the diffusion process, with the goal of improving sample quality. The three main types of respacing schedules are leading, linspace, and trailing. The original DDPM[10] proposes fixed, equally-spaced timesteps, setting the baseline for future work. IDDPM[53] and ADM[7] utilize linspace-style spacing, with denser steps at the start/end. IDDPM demonstrates improved sample quality over linear spacing. ADM learns the spacing adaptively during training to allocate more steps for challenging generations. DDIM[15] and PNDM[37] employ leading-style spacing, with more steps early on. DDIM dynamically adjusts timesteps during sampling, adding steps for high-precision regions. PNDM spaces steps based on a Beta CDF, concentrating them in key areas. DPM-Solver[35] uses trailing-style spacing, with denser steps at the end.

3.3 Evaluation Metrics

Accurate evaluation metrics play a vital role in driving the advancement of research. However, evaluation can be challenging due to the involvement of multiple attributes that contribute to the quality of generated results, making image evaluation subjective in nature. We also list the evaluation metrics and performance of the different methods on different benchmarks for the corresponding application scenarios in the subsequent tables. In Table 1, a compilation of representative work from various domains is presented, alongside corresponding code links.

General Evaluation Metrics. In general image quality evaluation, metrics such as IS[143] and FID[142] are commonly used. IS is a widely used measure of the quality and diversity of generated images scored by the Inception model. However, it has faced criti-

Table 1. Open Resources of Diffusion Models

Application	Diffusion Model	Year & Publication	Open Source Code Link
Image Restoration	RePaint[137]	2021 CVPR	https://github.com/andreas128/RePaint
	IterInpaint[144]	2023 arXiv	https://github.com/j-min/IterInpaint
	DDRM[136]	2022 NeurIPS	https://github.com/bahjat-kawar/ddrm
	SR3[80]	2022 TPAMI	Image-Super-Resolution-via-Iterative-Refinement
	Palette[77]	2022 SIGGRAPH	Palette-Image-to-Image-Diffusion-Models
	SRDiff[145]	2022 Neurocomputing	https://github.com/LeiaLi/SRDiff
	GDP[146]	2023 CVPR	https://github.com/Fayeben/GenerativeDiffusionPrior
Class to Image	ADM-G[7]	2021 NeurIPS	https://github.com/openai/guided-diffusion
	ED-DPM[147]	2022 ECCV	https://github.com/ZGCTroy/ED-DPM
	LDM[64]	2022 CVPR	https://github.com/Stability-AI/stablediffusion
	DiT[57]	2023 CVPR	https://github.com/facebookresearch/DiT
	MDT[59]	2023 ICCV	https://github.com/sail-sg/MDT
	Simple diffusion[60]	2023 arxiv	https://github.com/rkstgr/simple-diffusion
Text to Image	GLIDE[119]	2022 ICML	https://github.com/openai/glide-text2im
	Imagen[76]	2022 NeurIPS	https://github.com/lucidrains/imagen-pytorch
	VQ-Diffusion[13]	2022 CVPR	https://github.com/cientgu/VQ-Diffusion
	Parti[115]	2022 arXiv	https://github.com/lucidrains/parti-pytorch
	Muse[79]	2023 arXiv	https://github.com/lucidrains/muse-maskgit-pytorch
	SDD[82]	2023 arXiv	https://github.com/nannullna/safe-diffusion
	GLIGEN[83]	2023 CVPR	https://github.com/gligen/GLIGEN
Text to Video	RVD[99]	2023 Entropy	https://github.com/buggyyang/rvd
	FDM[148]	2022 NeurIPS	flexible-video-diffusion-modeling
	MCVD[149]	2022 NeurIPS	https://github.com/voletiv/mcvd-pytorch
	Make-A-Video[150]	2023 ICLR	https://github.com/lucidrains/make-a-video-pytorch
	Make-Your-Video[151]	2023 arXiv	https://github.com/AILab-CVC/Make-Your-Video
	Follow-Your-Pose[152]	2024 AAAI	https://github.com/mayuelala/FollowYourPose
	LFDM[74]	2023 CVPR	https://github.com/nihaomiao/CVPR23_LFDM
	VideoComposer[75]	2023 arXiv	https://github.com/ali-vilab/videocomposer
	ControlVideo[153]	2023 arXiv	https://github.com/YBYBZhang/ControlVideo
	VideoFusion[154]	2023 CVPR	text-to-video-synthesis
Text to 3D	DreamFusion[155]	2023 ICLR	https://github.com/chinhsuanwu/dreamfusionacc
	Magic3D[156]	2023 CVPR	https://github.com/chinhsuanwu/dreamfusionacc
	Fantasia3D[157]	2023 ICCV	https://github.com/Gorilla-Lab-SCUT/Fantasia3D
	Zero-1-to-3[158]	2023 arXiv	https://github.com/cvlab-columbia/Zero-1-to-3
	Magic123[159]	2023 arXiv	https://github.com/guochengqian/Magic123
	SyncDreamer[160]	2023 arXiv	https://github.com/liuyuan-pal/SyncDreamer
	LAS-Diffusion[161]	2023 SIGGRAPH	https://github.com/Zhengxinyang/LAS-Diffusion
Personalization	Textual Inversion[111]	2022 ICLR	https://github.com/rinongal/textual_inversion
	DreamBooth[112]	2023 CVPR	https://github.com/Victarry/stable-dreambooth
	Custom Diffusion[25]	2023 CVPR	https://github.com/adobe-research/custom-diffusion
	SVDiff[162]	2023 ICCV	https://github.com/mkshing/svdiff-pytorch
	Perfusion[163]	2023 SIGGRAPH	https://github.com/lucidrains/perfusion-pytorch
	HyperNetworks[164]	2017 ICLR	https://github.com/g1910/HyperNetworks
	LoRA[113]	2021 ICLR	https://github.com/microsoft/LoRA
	ELITE[165]	2023 ICCV	https://github.com/csyxwei/ELITE
	ProFusion[166]	2023 arXiv	https://github.com/drboog/ProFusion
	Mix of Show[167]	2023 NeurIPS	https://github.com/TencentARC/Mix-of-Show

cism for its lack of robustness and sensitivity to noise. FID demonstrates greater robustness compared with IS and provides a better overall assessment of the quality of generated images. However, FID assumes a Gaussian distribution for image features, which may not always hold true in practice. Moreover, there are also evaluation metrics based on reference images. For instance, PSNR is an image quality evaluation indicator based on the difference between corresponding pixel points of two images. SSIM[168] measures image similarity in terms of brightness, contrast, and structure. It has been revealed[169] that PSNR is more sensitive to additive Gaussian noise than SSIM, while the opposite is observed for jpeg compression. To address the problem that traditional metrics (PSNR, SSIM, etc.) disagree with human judgments under some circumstances, Zhang *et al.*[170] proposed perceptually-learned metric called Learned Perceptual Image Patch Similarity (LPIPS), evaluating how well image quality perception models actually correspond to human visual perception.

Task-Specific Evaluation Metrics. Fréchet video distance (FVD)[171] is a new metric for generative models of video based on FID, considering the temporal coherence of the visual content across a sequence of frames as well as its visual presentation at any given point in time. The CLIP score[172] is a metric that captures the semantic relationships between pairs of natural language and image inputs by learning the meaningful associations between them. Sajjadi *et al.*[173] improved the traditional precision and recall by calculating directly from distributions, which was further improved by Kynkäänniemi *et al.*[174] in 2019. Let P_r and P_g denote the probability distributions of the real and generated data, respectively. Recall quantifies the extent to which data generated by P_g matches P_r, while precision measures the proportion of generated images that belong to P_r. Recent work by [175] introduced an enhanced aesthetic prediction model called Improved-Aesthetic-Predictor (LAION-Aesthetics V2), built on LAION-Aesthetics V1[175]. This large-scale aesthetic database allows training a model to predict human-like aesthetic scores for natural images.

Human Evaluation. There has been a trend towards using human evaluation to assess model performance[76, 176, 177], as some commonly used objective evaluation metrics are not sufficient to accurately evaluate the quality of generated images.

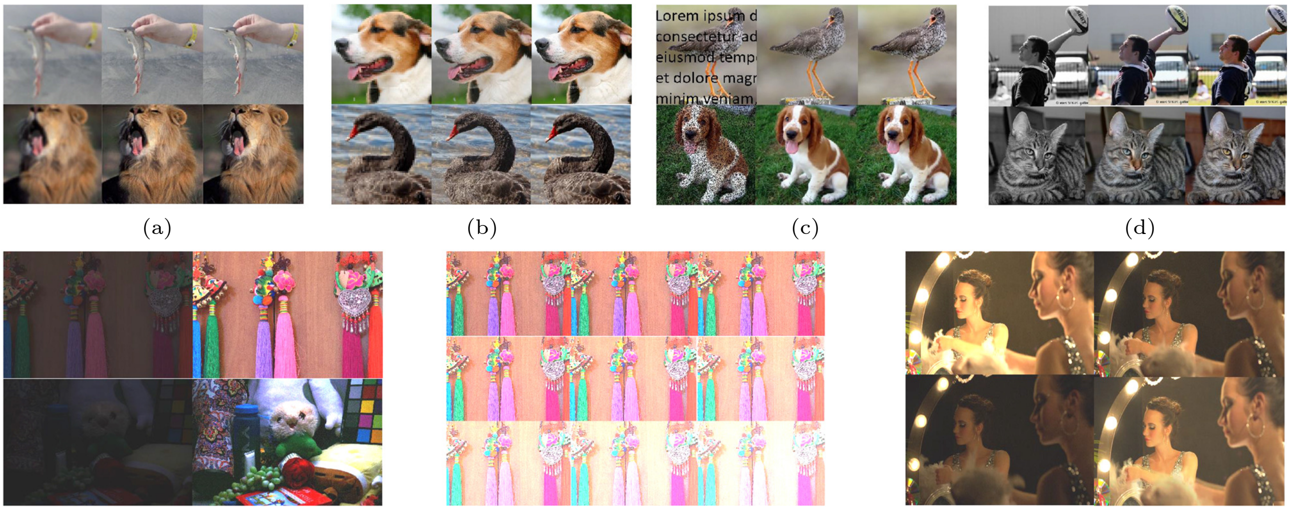
4 Applications

4.1 Image Restoration

Image restoration has been a longstanding fundamental challenge in computer vision, aiming to recover an original image from a degraded version affected by noise or distortion. In recent years, the diffusion model has emerged as a promising approach for image restoration. Its strength lies in effectively handling complex, high-dimensional data and generating high-quality samples from probability distributions. Moreover, many image restoration tasks can be framed as linear inverse problems.

RePaint[137] showcases the generalization capability of unconditionally trained diffusion models for inpainting tasks. By conditioning on available pixels, the model effectively utilizes the strong image prior learned by DDPMs. In the context of masked prediction, DiffMAE[178] introduces a conditional objective that approximates pixel distributions based on visible regions. This formulation allows for efficient extension to video inpainting and recognition tasks. Additionally, IterInpaint[144] was proposed as a novel inpainting baseline, extending the stable diffusion approach for layout-guided inpainting.

Kawar *et al.*[136] introduced Denoising Diffusion Restoration Models (DDRM), an efficient unsupervised posteriori sampling method. Inspired by variational inference, DDRM utilizes a pre-trained denoised diffusion generation model to solve linear inverse problems. The Palette[77] employs conditional diffusion models to establish a unified framework for four distinct image generation tasks: colorization, inpainting, uncropping, and JPEG restoration. Fei *et al.*[146] presented Generative Diffusion Prior (GDP), a method for image restoration. Unlike existing techniques that assume known degradation and require supervised training, GDP models the posterior distributions of natural images through unsupervised sampling. It leverages a pre-trained DDPM to address linear inverse, non-linear, and blind problems. The versatility of GDP is demonstrated on various tasks, including super-resolution, deblurring, denoising, and multi-degradation recovery (see Fig.6). In Tables 2 and 3, we list the comparison of diffusion's performance in six common image recovery domains (inpainting, super-resolution, shadow removal, deblur, colorization, and enlighten), and the best results on different datasets are in bold. The upward and downward arrows indicate that the bigger is better and the smaller is better, respectively.



(a) (b) (c) (d)

(e) (f) (g)

Fig.6. Image restoration results from RePaint[146]. Restoration type: (a) deblur, (b) super-resolution, (c) inpainting, (d) colorization, (e) low-light image enhancement, (f) non-linear enhancement, and (g) multiple-guidance enhancement.

Table 2. Comparison of Diffusion Models' Performance in Mainstream Image Recovery Domains on ImageNet-1k[179]

Restoration	Model	PSNR↑	SSIM↑	FID↓	Cons↓
Inpainting	DGP[180]	27.59	0.82	60.65	414.60
	SNIPS[181]	17.55	0.74	103.50	587.90
	DDRM[136]	34.28	0.95	24.09	**4.08**
	GDP-$\boldsymbol{x}_t$ [146]	31.06	0.93	20.24	8.80
	GDP-$\boldsymbol{x}_0$ [146]	**34.40**	**0.96**	**16.58**	5.29
4x super resolution	DGP[180]	21.65	0.56	152.85	158.74
	SNIPS[181]	22.38	0.66	154.43	21.38
	RED[182]	24.18	0.71	98.30	27.57
	DDRM[136]	**26.53**	**0.78**	40.75	19.39
	GDP-$\boldsymbol{x}_t$ [146]	24.27	0.67	64.67	80.32
	GDP-$\boldsymbol{x}_0$ [146]	24.42	0.68	**38.24**	**6.49**
Deblur	DGP[180]	26.00	0.54	136.53	475.10
	SNIPS[181]	24.73	0.69	17.11	60.11
	RED[182]	21.30	0.58	69.55	63.20
	DDRM[136]	**35.64**	**0.98**	4.78	50.24
	GDP-$\boldsymbol{x}_t$ [146]	25.86	0.75	5.00	54.08
	GDP-$\boldsymbol{x}_0$ [146]	25.98	0.75	**2.44**	**41.27**
Colorization	DGP[180]	18.42	0.71	94.59	305.59
	DDRM[136]	**22.12**	0.91	47.05	37.33
	GDP-$\boldsymbol{x}_t$ [146]	21.30	0.86	66.43	75.24
	GDP-$\boldsymbol{x}_0$ [146]	21.41	**0.92**	**37.60**	**36.92**

4.2 2D Image Generation

4.2.1 Class to Image

DDPM[10] pioneered the use of diffusion probabilistic models for conditional image synthesis. By incorporating class labels and noise into the generative process, DDPM demonstrated the feasibility of utilizing diffusion for controlled image generation. Building upon this work, ADM-G[7] introduces architectural improvements such as classifier guidance, which enhances sample quality by providing conditioning signals during sampling. CDM[78] further advances controllability by employing a cascaded pipeline of diffusion models to synthesize higher resolution images in a step-wise manner. This cascade approach

Table 3. Comparison of Performance of Diffusion Models in Image Recovery Domains

Restoration	Dataset	Model	PSNR↑	SSIM↑	FID↓	LPIPS↓
Inpainting	CelebaHQ[183]	RePaint[137]	–	–	6.98	0.060
		SDM[178]	–	–	4.05	**0.052**
		SDGM[136]	–	–	4.68	0.057
		LDM[64]	–	–	**1.50**	0.137
		LDM(w/o attention)[64]	–	–	2.37	0.146
Shadow removal	ImageNet-1k[179])	DHAN[184]	20.42	0.69	109.35	0.247
		IR-SDE[185]	20.30	0.66	74.35	0.152
		U-Net baseline	20.69	**0.71**	102.10	0.236
		Refusion[186]	**21.88**	0.69	**56.22**	**0.121**
Enlighten	LOL[187]	LightenNet[188]	10.29	0.45	90.91	–
		Retinex-Net[187]	17.24	0.55	129.99	–
		EnlightenGAN[189]	17.44	0.74	82.60	–
		KinD[190]	**17.57**	**0.82**	**74.52**	–
		GDP-$\boldsymbol{x}_t$[146]	7.32	0.57	238.92	–
		GDP-$\boldsymbol{x}_0$[146]	13.93	0.63	75.16	–
Enlighten	VE-LOL-L[191]	LightenNet[188]	13.26	0.57	82.26	–
		Retinex-Net[187]	16.41	0.64	135.20	–
		EnlightenGAN[189]	17.45	0.75	86.51	–
		KinD[190]	**18.07**	**0.78**	80.12	–
		GDP-$\boldsymbol{x}_t$[146]	9.45	0.50	152.68	–
		GDP-$\boldsymbol{x}_0$[146]	13.04	0.55	**78.74**	–

helps mitigate compounding errors. ED-DPM[147] proposes entropy-driven sampling and training schemes to improve conditional image generation with diffusion models. These schemes alleviate vanishing gradient issues during the denoising process. LDM[64] introduces a novel approach by separating the training of autoencoders and diffusion models. This bifurcated process allows each component to focus on its specific capabilities, resulting in performance gains. MDT[59] accelerates training by incorporating masked self-attention, which improves the modeling of spatial relationships in images. This work demonstrates the potential of integrating transformer architectures. Fig.7 showcases the results of class-to-image generation. A performance comparison between some of the methods is listed in Tables 4 and 5.

4.2.2 Text to Image

Text to image generation involves the generation of an image that corresponds to a descriptive text. Two typical problems in text-to-image generation are attribute misbinding and missing objects. Attribute misbinding, where visual characteristics are incorrect-

(a) (b) (c)

Fig.7. Class-to-image results from DiT[57]. Resolution: (a) 512, (b) 256, and (c) 64.

Table 4. Performance Comparison on Class to Image on ImageNet and FFHQ with Resolution 256×256

Dataset	Model	FID↓	IS↑	Precision↑	Recall↑
ImageNet-1k 256×256[179]	BigGAN-deep[141]	6.95	171.40	0.87	0.28
	StyleGAN-XL[192]	2.30	256.12	0.78	0.53
ImageNet-1k 256×256[179]	ADM[7]	10.94	100.98	0.69	0.63
	ADM-U[7]	7.49	127.49	0.72	0.63
	ADM-G[7]	4.59	186.70	0.82	0.52
	CDM[78]	4.88	158.71	–	–
	LDM-8[64]	15.51	79.03	0.65	0.63
	LDM-4[64]	10.56	103.49	0.71	0.62
	LDM-4-G[64]	3.60	247.67	0.87	0.48
	DiT-XL/2[57]	9.62	121.50	0.67	**0.67**
	DiT-XL/2-G[57]	2.27	**278.24**	**0.83**	0.57
	ViT-XL+Min-SNR-5[193]	**2.06**	–	–	–
	Simple diffusion (U-Net)[60]	3.76	171.60	–	–
	Simple diffusion (U-ViT)[60]	2.77	211.80	–	–
FFHQ 256×256[4]	DDPM[10]	8.33	–	–	–
	p2[194]	7.00	–	–	–
	LDM[64]	**4.98**	–	0.73	0.50
	SD[195]	10.50	–	–	–

Note: GAN-based results are included, distinguished from the diffusion-based results by a divider for comprehensiveness.

Table 5. Performance Comparison on Class to Image on CIFAR10 and ImageNet

Dataset	Model	FID↓	IS↑
CIFAR10 32×32[196]	BigGAN[141]	14.70	9.22
	StyleGAN-XL[192]	**1.85**	–
	SDE[11]	2.20	**9.89**
	DDPM[10]	3.17	9.46
	LSGM[197]	2.10	–
	EDM[12]	2.04	9.84
ImageNet-1k 512×512[179]	BigGAN-deep[141]	8.43	177.90
	StyleGAN-XL[192]	**2.41**	**267.75**
	ADM[7]	23.24	58.06
	ADM-U[7]	9.96	121.78
	ADM-G[7]	7.72	172.71
	DiT-XL/2[57]	12.03	105.25
	DiT-XL/2-G[57]	3.04	240.82
	Simple diffusion (U-Net)[60]	4.30	171.00
	Simple diffusion (U-ViT)[60]	3.54	205.30

ly paired with objects, stems from inadequate alignment between modalities[198]. Missing objects occur when models fail to generate portions of an image described in text[199].

GLIDE[119] draws inspiration from the success of guided diffusion models[7] in generating photorealistic samples, and the capability of text-to-image models to handle free-form prompts[123]. GLIDE employs guided diffusion to address the problem of text-conditional image synthesis. Imagen[76] has presented a text-to-image diffusion model along with a comprehensive benchmark, indicating that Imagen performs better when compared with various approaches such as LDM[64], GLIDE[119], and DALL-E 2[71]. The key discovery behind Imagen is that text embedding from large language models (LLMs) pre-trained on a plain text corpus is very effective for text-to-image synthesis. An example is shown in Fig.8. The work of VQ-Diffusion[13] introduces a novel vector-quantized diffusion model for text-to-image generation. This approach effectively reduces unidirectional bias and circumvents the accumulation of prediction errors. Parti[115] demonstrates the efficacy of scaling autoregressive models to enhance text-to-image generation using a ViT-VQGAN[200] image tokenizer. This approach enables the models to effectively integrate and visually convey world knowledge with a high degree of accuracy. Muse[79] is a novel approach that leverages a masked modeling task in the discrete token space to generate high-fidelity images from text. Specifically, given a text embedding extracted from a pre-trained LLM, Muse is trained to predict randomly masked image tokens. Compared with pixel-space diffusion models such as Imagen[76] and DALL-E 2[71], Muse demonstrates superior efficiency by virtue of its use of discrete tokens and requiring fewer sampling iterations. The performance comparison of various text-to-image methods on the MS-CoCo dataset has been outlined in Table 6.

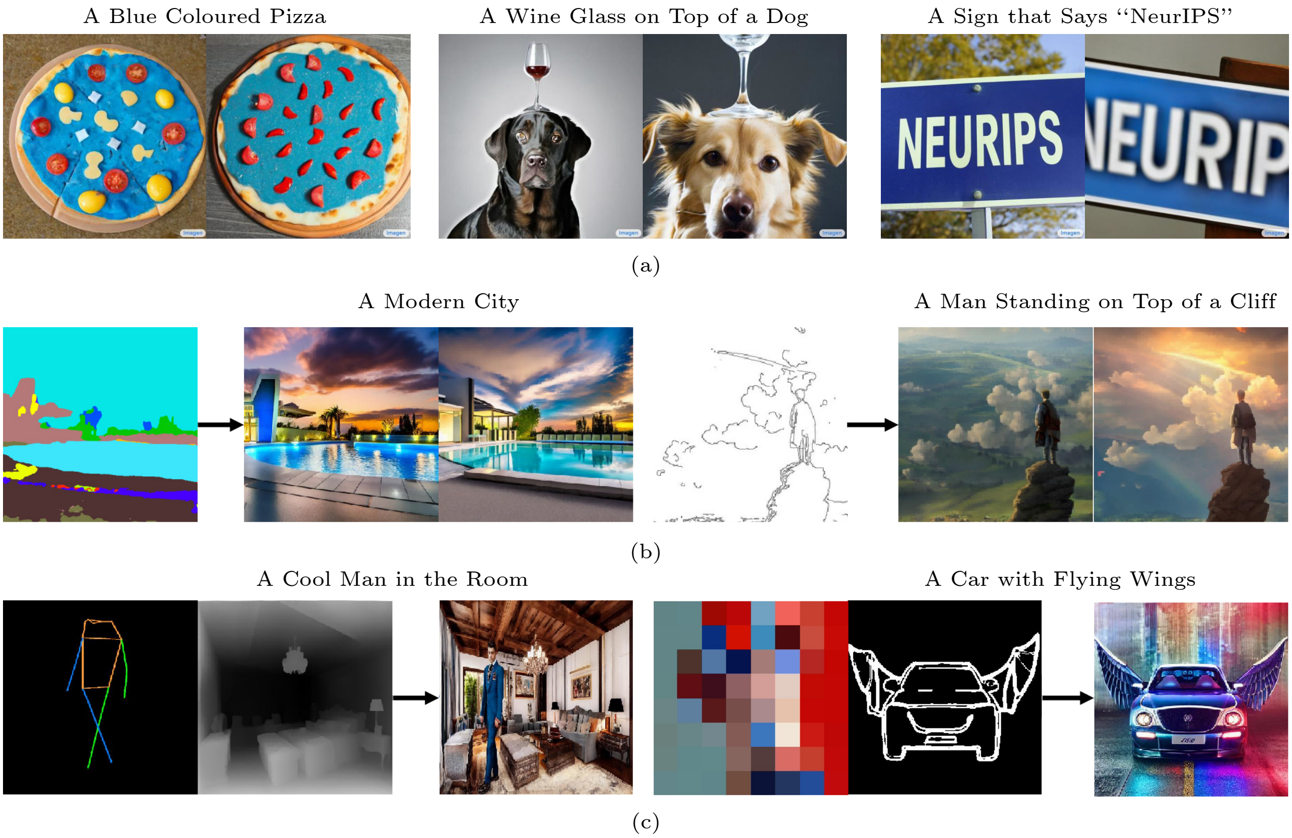


Fig.8. Text-to-Image results from [22, 83]. Condition: (a) text only, (b) text and single condition, and (c) text and multiple conditions.

Table 6. Performance Comparison on Text to Image Generation on the MS-CoCo Dataset[201]

Model	FID ↓
LAFITE[202] (GAN-based)	26.94
CogView[203] (Transformer-based)	27.10
LDM[78]	12.63
VQ-Diffusion[13]	13.86
DALL-E 2[71]	10.39
Parti[115]	7.23
GLIDE[119]	12.24
Muse[79]	7.88
Imagen[7]	7.27
eDiff-I[81]	**6.95**

In order to address the problem that stable diffusion methods may generate images containing harmful information, Kim *et al.*[82] proposed safe Self-Distillation Diffusion (SDD) and employed an exponential moving average teacher to diminish catastrophic forgetting. GLIGEN[83] is a novel approach that extends existing large-scale text-to-image diffusion models by allowing them to be conditioned on grounding inputs, achieving open-world grounded text-to-image generation with caption and bounding box condition inputs. Mou *et al.*[84] proposed using simple and lightweight T2I-Adapters to explicitly control the generation of text-to-image models by aligning internal knowledge with external control signals, achieving rich control and editing effects in the color and structure of the generation results, with attractive properties of practical value such as composability and generalization ability.

4.3 Video Generation

Diffusion-based generative models heavily boost the field of video generation, as first promoted by RVD[99] and followed by subsequent work, making possible significant progress on conditional control, resolution, and temporal consistency. FDM[148] applies diffusion models to improve long-term video prediction. MCVD[149] adapts conditional tasks like future prediction and interpolation. Imagen Video[204] and Make-A-Video[150] each constructs a cascade pipeline to utilize spatial and temporal super-resolution models for high-resolution time-consistent videos. Dreamix[205] fine-tunes a video diffusion model on aligned text and low-resolution frames to improve fidelity. Several work[62, 206, 207] follows the LDM[64] paradigm and successfully transfers generators from the image space to the video space after fine-tuning on video sequences by introducing an extra temporal axis.

There has been a surge of interest in conditional video generation based on pretrained text-to-image or text-to-video models. With fixed spatial weights and learnable temporal weights tuned on video data, Make-Your-Video[151] allows re-rendering of the appearance of source video given extra depth conditions. Follow-Your-Pose[152] uses pose as guidance for the synthesis of human-like character videos. In LFDM[74], the action class is served as condition and is warped in the latent space based on the generated temporally-coherent flow. VideoComposer[75], as an extension to Composer[97], takes multiple kinds of images as conditions, which are fused in the latent space and interact within the U-Net via cross-attention. Control-Video[153] seamlessly incorporates with ControlNet[22], which is tailored into video domain through the augmentation of self-attention with a comprehensive fully cross-frame interaction mechanism. MV-Diffusion[208] improves temporal consistency by explicit motion modeling through global trajectory information and a motion trend attention block. EVDModel[209] reduces computation costs in video synthesis by minimizing convolutional redundancy. VideoFusion[154] addresses the challenges of applying diffusion models to high-dimensional data spaces by employing a decomposed diffusion process involving a shared base noise and varying residual noises along the time axis.

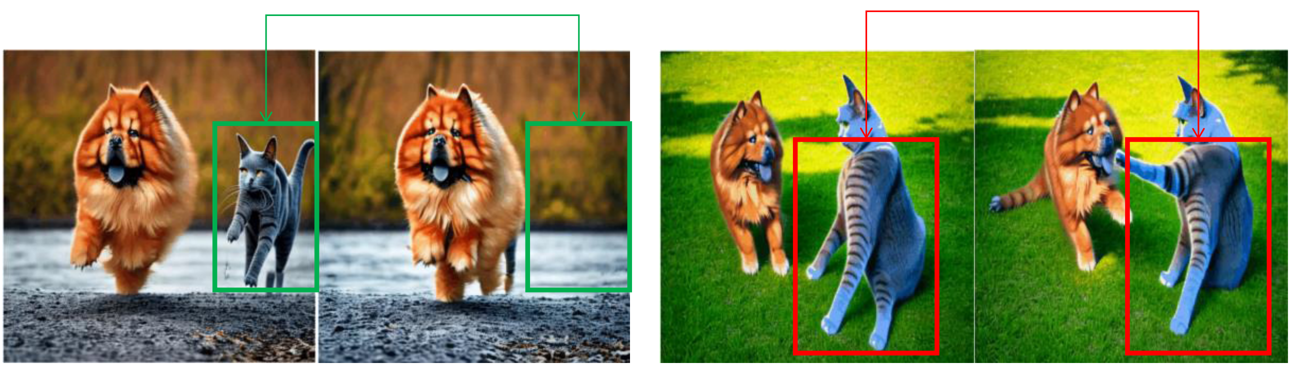
Diffusion-based video generation has witnessed rapid advancements in architecture, conditioning, and temporal modeling, leading to overall improvement. However, certain challenges still persist, such as identity loss, minimizing flicker, and effectively modeling intricate physics across extended timeframes (as shown in Fig.9). The incorporation of robust image priors and the integration of temporal knowledge are expected to have a significant impact on addressing these challenges and shaping the future of diffusion-based video generation. In Tables 7 and 8, performance comparisons within the video generation field are provided, with distinct labeling for zero-shot methods and other approaches.

4.4 3D Generation

3D synthesis presents significant challenges due to the limited availability of large-scale labeled 3D datasets and the absence of efficient architectures for denoising 3D data. The results of the 3D generation are depicted in Fig.10.

To address these challenges, recent research has focused on a research direction known as Score Distillation Sampling (SDS)[155] or Score Jacobian Chaining (SJC)[216] in the field of text-to-3D generation. SDS involves optimizing 3D representations by aligning their rendered images with regions of high probability density conditioned on the accompanying text. This optimization process is supervised using pretrained 2D diffusion models.

One notable application of SDS is DreamFusion[155], which utilizes the noise residual to optimize Neural Radiance Fields (NeRF) and has been extended by much later work. For example, Magic3D[156] introduces a two-stage coarse-to-fine optimization framework that incorporates sparse grids and differentiable rendering, leading to accelerated optimization and improved fidelity. Dream3D[215] initializes the neural field by a 3D shape prior extracted from the text-to-shape phase and is capable of generating high-quality 3D contents after optimized in a



(a) (b)

Fig.9. Problems with video generation between consecutive frames. (a) ID loss. (b) Temporal inconsistency.

Table 7. Performance Comparison on Text to Video on Dataset UCF-101[210]

Diffusion-Based	Model	Zero-Shot	IS ↑	FVD ↓
Yes	CogVideo[211]	✓	23.55	751.34
No	MagicVideo[207]	✓	–	699.00
	Make-A-Video[150]	✓	33.00	**367.23**
	Make-Your-Video[151]	✓	–	330.49
	Video LDM[62]	✓	**33.45**	550.61
	VideoFusion[154]	✓	17.49	639.90
	Video-LDM[62]	✓	33.45	550.61

Note: Methods listed above the horizontal line in the table are not based on diffusion, whereas those below the line are diffusion-based.

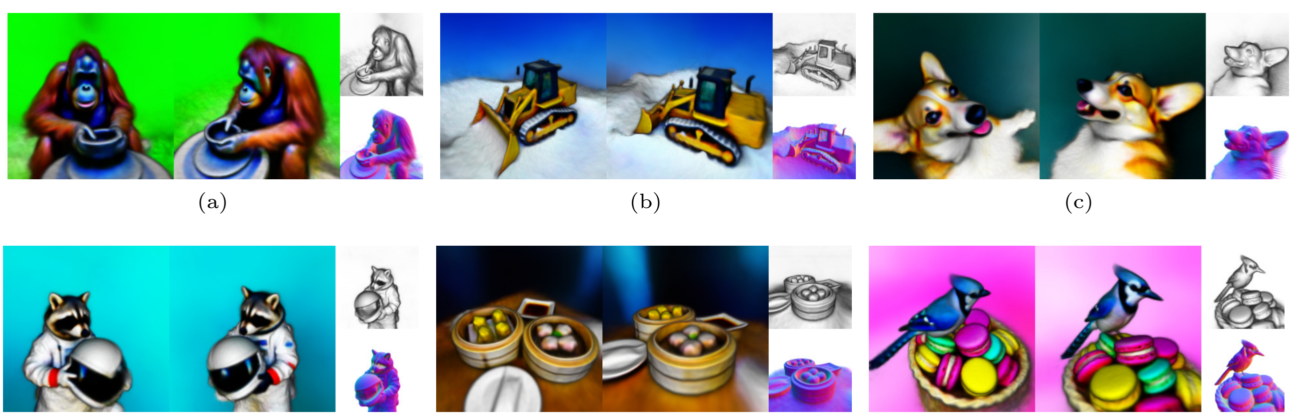
Table 8. Performance Comparison on Text to Video on Dataset MSR-VTT[212]

Diffusion-Based	Model	Zero-Shot	CLIPSIM ↑
Yes	CogVideo[211]	✓	0.261 4
No	GODIVA[213]	–	0.240 2
	NUWA[214]	–	0.243 9
	Make-A-Video[150]	✓	**0.304 9**
	Video LDM[62]	✓	0.292 9
	VideoFusion[154]	✓	0.279 5
	VideoComposer[75]	✓	0.293 2

CLIP-guided manner. Zero-1-to-3[158] unveils the viewpoint-aware ability of pre-trained diffusion model by fine-tuning on camera extrinsics as condition for novel view synthesis, yet followed by an SJC-based optimization on neural fields to further enable 3D reconstruction. Magic123[159] incorporates both 2D priors from SD and 3D priors from Zero-1-to-3 in SDS loss, with an extra hyperparameter to trade off exploration against exploitation of the generated geometry. Fantasia3D[157] disentangles appearance learning from geometry modeling under normal map supervision and introduces fully Bidirectional Reflectance Distribution Function (BRDF) into text-to-3D tasks, thus enables photorealistic rendering of material surfaces.

Efforts have also been made to enhance multiview consistency and local controllability in text-to-3D synthesis. SyncDreamer[160] generates multiview-consistent images from a single-view image by synchronizing intermediate states using a 3D-aware feature attention mechanism. By jointly training the model on multi-view images (from 3D assets) and 2D image-text pairs, they proposed multi-view diffusion models, which can be used as a multi-view 3D prior agnostic to 3D representations. Wonder3D[217] proposes a cross-domain diffusion model that generates multiview normal maps and the corresponding color images, achieving high-quality reconstruction results, robust generalization, and good efficiency compared with prior work. MVDream[218] believes that large-scale 2D data is crucial to generalizable 3D generation. Rodin[219] utilizes latent conditioning and 3D-aware convolution to create high-fidelity 3D avatars from a single portrait or text prompt, allowing for text-based semantic manipulation of the avatars. LAS-Diffusion[161] addresses challenges related to quality, local controllability, and generalizability by employing signed distance function (SDF) representation and a view-aware local attention mechanism.

In summary, the combination of neural rendering, multimodal representations, and diffusion modeling has shown promise for high-fidelity 3D synthesis. The advancements in SDS, such as DreamFusion[155] and its extensions, have improved the optimization process. Additionally, the development of methods has enhanced the overall quality, multi-view consistency, and local controllability of the generated 3D outputs. However, challenges remain in scaling synthesis, reducing optimization costs, and improving coherence.



(a) (b) (c) (d) (e) (f)

Fig.10. Text-to-3D results from Dream3D[215]. Text: (a) an orangutan making a clay bowl on a throwing wheel, (b) a bulldozer clearing away a pile of snow, (c) a corgi taking a selfie, (d) a raccoon astronaut holding his helmet, (e) a table with dim sum on it, and (f) a jay standing on a basket of macarons.

Future progress in 3D synthesis will rely on leveraging 3D priors and shape representations to overcome these challenges and achieve even higher levels of fidelity. Performance comparisons for the 3D generation domain are detailed in Table 9.

Table 9. Performance Comparison on Text to 3D Generation on Dataset GSO[220]

Model	Chamfer Dist ↓	Volume IoU ↑
Realfusion[221]	0.081 9	0.274 1
Magic123[159]	0.051 6	0.452 8
One-2-3-45[222]	0.062 9	0.452 8
Shape-E[223]	0.043 6	0.358 4
Zero-1-to-3[158]	0.033 9	0.503 5
SyncDreamer[160]	**0.026 1**	**0.542 1**

4.5 Personalization

The personalization involves the generation of images with specific and unique concepts, modifications of their appearance, and compositions of new characters and scenes. In essence, personalization allows users to communicate with a generative model and specify their desired output with greater precision and flexibility. Fig.11 illustrates four prevalent approaches to personalization generation.

Embedding Tuning. Textual Inversion[111] is noteworthy in the field of embedding tuning. It generates images with a similar style to the training images using a limited set of images and defining new keywords. To achieve this, a novel keyword needs to be defined, one that is not currently present in the existing model. This keyword is assigned a distinct numerical value, similar to other tokens in the tokenizer. The keyword is then transformed into an embedding, and the text transformer maps it to the most suitable embedding vector for the newly provided keyword. Improving upon Textual Inversion, $\mathcal{P}+$[224] introduces an inversion space that encompasses multiple textual conditions corresponding to each layer of the denoising U-Net in the diffusion model. This enhancement offers better disentanglement and control over image synthesis.

Embedding-Weight Tuning. Compared with the textual inversion method, DreamBooth[112] employs a rare word instead of a new word to prevent overfitting. Additionally, DreamBooth fine-tunes the entire model, whereas textual inversion only adjusts the text embedding component. Custom Diffusion[25] introduces a method for co-training multiple concepts or constrained optimization of several existing concept models. SVDiff[162] fine-tunes the singular values of weight matrices, reduces the risk of overfitting and language-drifting, and introduces a Cut-Mix-Unmix data-augmentation technique to enhance multi-subject image generation. Perfusion[163] is a personaliza-

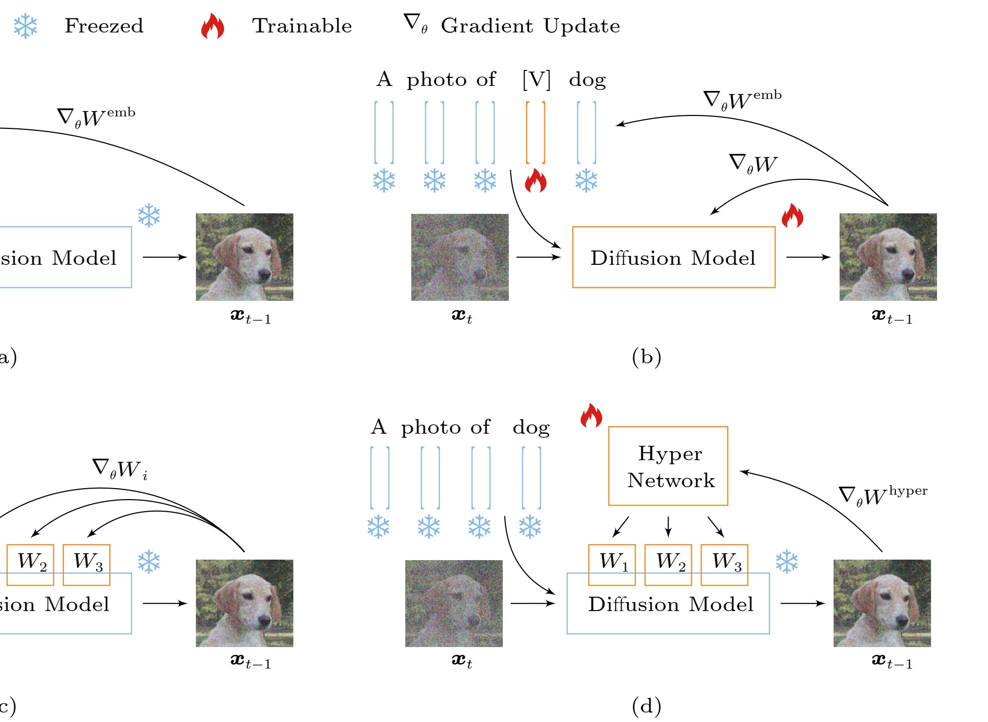


Fig.11. Methods for personalization. (a) Textual inversion[111]. (b) Dreambooth[112]. (c) LoRA[113]. (d) HyperNetwork[164]. W^{emb}: learnable text encoding. W: model parameters. W_i: lora parameters for layer i. W^{hyper}: hypernetwork parameters.

tion method that uses dynamic rank-1 updates and a mechanism that "locks" new concepts' cross-attention keys to their superordinate category to balance visual-fidelity and textual-alignment, allowing run-time-efficient combination of multiple concepts with a single trained models.

Fast Test-Time Tuning. HyperNetworks[164] replaces the weight matrix in a large model by fine-tuning the structure of a small parameter model and has been applied to the cross-attention module of U-Net in stable diffusion models[64] for achieving personalization. LoRA[113] is a commonly used fine-tuning method. Both LoRA and HyperNetworks modify the cross-attention module of the U-Net to alter the style of generated images. However, LoRA adjusts the weights of the cross-attention module, while HyperNetworks[164] inserts additional modules. InstantBooth[225] and Taming[226] enable personalized output generation in different styles by introducing a new conditioning branch for the diffusion model. FasterComposer[227] addresses the problem of identity blending in multisubject generation by proposing to use an image encoder to predict subject-specific embeddings. SuTI[228] achieves personalized image generation without test-time finetuning by learning from a large dataset of paired images generated by subject-driven expert models. While SuTI mitigates the need for finetuning, the inference model does not fully maintain the original integrity of the text-to-image model and lacks high subject fidelity[229]. Fig.12 shows the results of customized generation for different styles and concepts.

Recently, encoder-based approaches such as ElITE[165], E4T[230], Blip[94], ProFusion[166], and Domain-Agnostic[231] have emerged. These approaches train neural networks to predict a latent representation that synthesizes new images of a given concept. They incorporate regularization techniques such as subject-specific segmentation masks[165], single-domain training, or contrastive-based regularization[231] to improve inference from a single image. Alternatively, the model proposed by [228] can synthesize new images from dual conditions, combining a textual prompt with a set of images depicting the target.

5 Future Direction

Multimodal controllable diffusion modeling enables the provision of high-quality, diverse, and innovative content tailored to meet users' specific needs and preferences. However, there are several areas where multimodal controllable diffusion models have room for improvement in both theory and practice. These include enhancing sampling efficiency and likelihood estimation, handling special data structures, integrating with other types of generative models, and customization for specific applications. Looking ahead, the future research direction of the diffusion model can be explored from the following perspectives: personalization, new architectural designs, advancing theoretical understanding, and expanding applications within the field of AI-driven content generation.

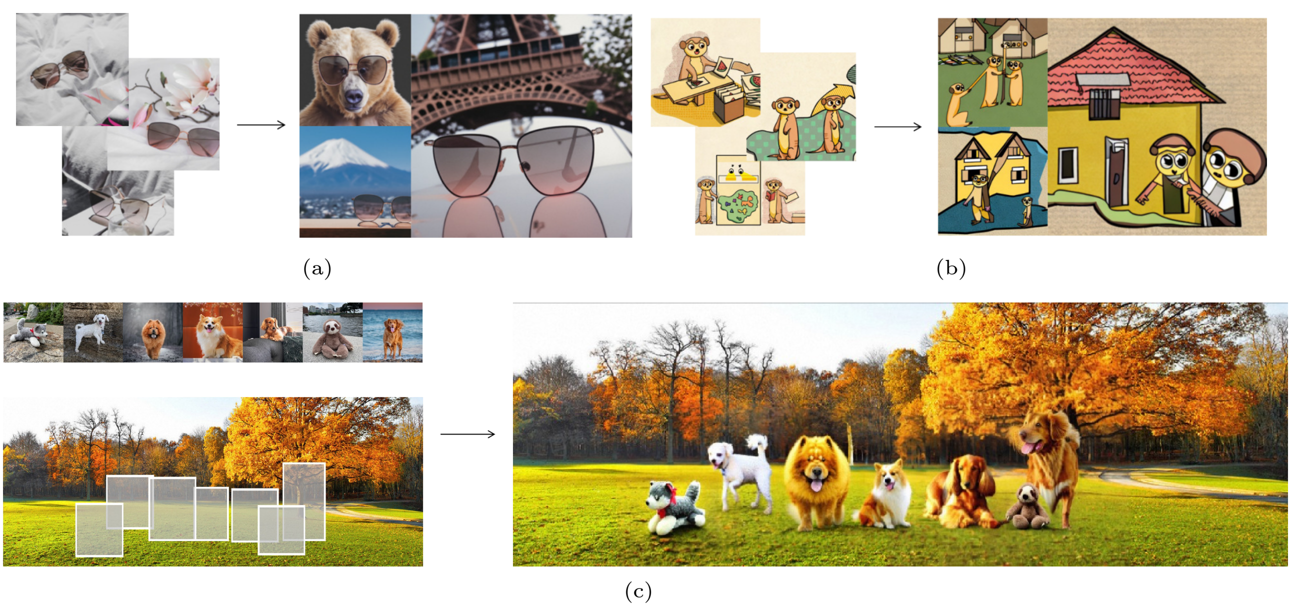


Fig.12. Personalization results of single concept object from (a) DreamBooth[112], (b) single concept style from Custom Diffusion[25], and (c) multi-concepts from Mix of show[167].

5.1 Architecture

The backbone and architecture of diffusion models hold significant potential for improvement. While U-Net and Transformer have demonstrated impressive results as denoising network backbones, they have inherent limitations in certain applications. Fortunately, the field of machine learning offers a diverse range of mature network architectures with attractive advantages. Leveraging and fine-tuning these architectures as denoising networks can bring additional benefits and unlock the full potential of diffusion models. Efforts are underway to compress architectures, reducing the number of parameters while maintaining performance.

5.2 Theory

Advancements in diffusion modeling can be achieved by developing new formulations for dimension destruction, establishing connections with well-established fields, and leveraging explainable techniques to enhance our understanding of diffusion models. Additionally, the success of diffusion models highlights the effectiveness of auto-regressive generation, which employs self-correction mechanisms to improve output quality. By delving into the information and structure embedded in random noise, diffusion modeling offers valuable insights and presents new possibilities and challenges for researchers in the field.

5.3 AIGC

The emergence of numerous fun-oriented mobile apps using AIGC is fascinating. While traditional tools like Photoshop are commonly used for image editing, they can be time-consuming and result in unnatural or unrealistic outputs. Similarly, video editing requires analyzing each clip and making editorial decisions based on both audio and visual content, a time-consuming process that requires careful consideration of every frame. Fortunately, some work has explored the utilization of diffusion, to the image[232, 233] or video editing[234], making the applications in AIGC such as face swapping and digital avatar possible.

6 Conclusions

In this comprehensive exploration, we delved into the realm of controllable diffusion models. We first provided a thorough understanding of diffusion model's formulations, sampling methods, and the key directions that drive their development. By highlighting the formulation of control, advancements in controllable technology, and the establishment of evaluation indicators, we have shed light on the intricacies of achieving controllability in diffusion models. Furthermore, our survey of applications across diverse domains has showcased the vast potential of diffusion models in addressing real-world challenges. Future research may witness more interdisciplinary collaborations to tackle complex problems specific to different domains. Establishing and refining evaluation metrics will be another key part of future research, aiding in the standardization of model performance comparisons and the selection of the most suitable models. By outlining future research avenues, we aim to inspire further advancements and provide readers with a valuable guide to the world of controllable diffusion models and their applications.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] Efros A A, Leung T K. Texture synthesis by non-parametric sampling. In *Proc. the 7th IEEE International Conference on Computer Vision*, Sept. 1999, pp.1033–1038. DOI: 10.1109/iccv.1999.790383.

[2] Heckbert P S. Survey of texture mapping. *IEEE Computer Graphics and Applications*, 1986, 6(11): 56–67. DOI: 10.1109/mcg.1986.276672.

[3] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. *Communications of the ACM*, 2020, 63(11): 139–144. DOI: 10.1145/3422622.

[4] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2019, pp.4396–4405. DOI: 10.1109/cvpr.2019.00453.

[5] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. the 31st International Conference on Machine Learning*, Jun. 2014, pp.1278–1286.

[6] Rezende D J, Mohamed S. Variational inference with normalizing flows. In *Proc. the 32nd International Conference on Machine Learning*, Jul. 2015, pp.1530–1538.

[7] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. In *Proc. the 35th Conference on Neural Information Processing Systems*, Dec. 2021, pp.8780–8794.

[8] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium

thermodynamics. In *Proc. the 32nd International Conference on Machine Learning*, Jul. 2015, pp.2256–2265.

[9] Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. In *Proc. the 33rd International Conference on Neural Information Processing Systems*, Dec. 2019, Article No. 1067.

[10] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In *Proc. the 34th International Conference on Neural Information Processing Systems*, Dec. 2020, Article No. 574.

[11] Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. arXiv: 2011.13456, 2020. https://arxiv.org/abs/2011.13456, May 2024.

[12] Karras T, Aittala M, Aila T, Laine S. Elucidating the design space of diffusion-based generative models. arXiv: 2206.00364, 2022. https://arxiv.org/abs/2206.00364, May 2024.

[13] Gu S Y, Chen D, Bao J M, Wen F, Zhang B, Chen D D, Yuan L, Guo B N. Vector quantized diffusion model for text-to-image synthesis. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2022, pp.10686–10696. DOI: 10.1109/cvpr52688.2022.01043.

[14] Austin J, Johnson D D, Ho J, Tarlow D, van den Berg R. Structured denoising diffusion models in discrete state-spaces. In *Proc. the 35th Conference on Neural Information Processing Systems*, Dec. 2021, pp.17981–17993.

[15] Song J M, Meng C L, Ermon S. Denoising diffusion implicit models. arXiv: 2010.02502, 2020. https://arxiv.org/abs/2010.02502, May 2024.

[16] Bao F, Li C X, Zhu J, Zhang B. Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models. arXiv: 2201.06503, 2022. https://arxiv.org/abs/2201.06503, May 2024.

[17] Lu C, Zhou Y H, Bao F, Chen J F, Li C X, Zhu J. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv: 2211.01095, 2022. https://arxiv.org/abs/2211.01095, May 2024.

[18] Salimans T, Ho J. Progressive distillation for fast sampling of diffusion models. arXiv: 2202.00512, 2022. https://arxiv.org/abs/2202.00512, May 2024.

[19] Hu V T, Zhang D W, Asano Y M, Burghouts G J, Snoek C G M. Self-guided diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.18413–18422. DOI: 10.1109/cvpr52729.2023.01766.

[20] Cho W, Ravi H, Harikumar M, Khuc V, Singh K K, Lu J W, Inouye D I, Kale A. Towards enhanced controllability of diffusion models. arXiv: 2302.14368, 2023. https://arxiv.org/abs/2302.14368, May 2024.

[21] Deja K, Trzciński T, Tomczak J M. Learning data representations with joint diffusion models. In *Proc. the 2023 European Conference on Machine Learning and Knowledge Discovery in Databases*: *Research Track*, Sept. 2023, pp.543–559. DOI: 10.1007/978-3-031-43415-0_32.

[22] Zhang L M, Rao A Y, Agrawala M. Adding conditional control to text-to-image diffusion models. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.3813–3824. DOI: 10.1109/iccv51070.2023.00355.

[23] Ham C, Hays J, Lu J W, Singh K K, Zhang Z F, Hinz T. Modulating pretrained diffusion models for multimodal image synthesis. In *Proc. the 2023 Conference on Special Interest Group on Computer Graphics and Interactive Techniques*, Jul. 2023, Article No. 35. DOI: 10.1145/3588432.3591549.

[24] He Y F, Cai Z F, Gan X, Chang B B. DiffCap: Exploring continuous diffusion on image captioning. arXiv: 2305.12144, 2023. https://arxiv.org/abs/2305.12144, May 2024.

[25] Kumari N, Zhang B L, Zhang R, Shechtman E, Zhu J Y. Multi-concept customization of text-to-image diffusion. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.1931–1941. DOI: 10.1109/cvpr52729.2023.00192.

[26] Kumar Bhunia A, Khan S, Cholakkal H, Anwer R M, Laaksonen J, Shah M, Khan F S. Person image synthesis via denoising diffusion model. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.5968–5976. DOI: 10.1109/cvpr52729.2023.00578.

[27] Ju X, Zeng A L, Zhao C C, Wang J N, Zhang L, Xu Q. HumanSD: A native skeleton-guided diffusion model for human image generation. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.15942–15952. DOI: 10.1109/iccv51070.2023.01465.

[28] Cao H Q, Tan C, Gao Z Y, Xu Y L, Chen G Y, Heng P A, Li S Z. A survey on generative diffusion models. *IEEE Trans. Knowledge and Data Engineering*, 20241–20. DOI: 10.1109/tkde.2024.3361474.

[29] Yang L, Zhang Z L, Song Y, Hong S D, Xu R S, Zhao Y, Zhang W T, Cui B, Yang M H. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2024, 56(4): 105. DOI: 10.1145/3626235.

[30] Kazerouni A, Aghdam E K, Heidari M, Azad R, Fayyaz M, Hacihaliloglu I, Merhof D. Diffusion models for medical image analysis: A comprehensive survey. arXiv: 2211.07804, 2022. https://arxiv.org/abs/2211.07804, May 2024.

[31] Croitoru F A, Hondru V, Ionescu R T, Shah M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2023, 45(9): 10850–10869. DOI: 10.1109/tpami.2023.3261988.

[32] Zhang C S, Zhang C N, Zhang M C, Kweon I S. Text-to-image diffusion models in generative AI: A survey. arXiv: 2303.07909, 2023. https://arxiv.org/abs/2303.07909, May 2024.

[33] Zou H, Kim Z M, Kang D. A survey of diffusion models in natural language processing. arXiv: 2305.14671, 2023. https://arxiv.org/abs/2305.14671, May 2024.

[34] Anderson B D O. Reverse-time diffusion equation mod-

els. *Stochastic Processes and Their Applications*, 1982, 12(3): 313–326. DOI: 10.1016/0304-4149(82)90051-5.

[35] Lu C, Zhou Y H, Bao F, Chen J F, Li C X, Zhu J. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Proc. the 36th International Conference on Neural Information Processing Systems*, Nov. 28/Dec. 9, 2022, Article No. 418.

[36] Zhang Q S, Chen Y X. Fast sampling of diffusion models with exponential integrator. arXiv: 2204.13902, 2022. https://arxiv.org/abs/2204.13902, May 2024.

[37] Liu L P, Ren Y, Lin Z J, Zhao Z. Pseudo numerical methods for diffusion models on manifolds. arXiv: 2202.09778, 2022. https://arxiv.org/abs/2202.09778, May 2024.

[38] Zhang Q S, Tao M L, Chen Y X. gDDIM: Generalized denoising diffusion implicit models. arXiv: 2206.05564, 2022. https://arxiv.org/abs/2206.05564, May 2024.

[39] Ascher U M, Petzold L R. Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. Society for Industrial and Applied Mathematics, 1998.

[40] Bao F, Li C X, Sun J C, Zhu J, Zhang B. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. In *Proc. the 39th International Conference on Machine Learning*, Jul. 2022, pp.1555–1584.

[41] Lin Z H, Gong Y Y, Liu X, Zhang H, Lin C, Dong A L, Jiao J, Lu J W, Jiang D X, Majumder R, Duan N. PROD: Progressive distillation for dense retrieval. In *Proc. the 2023 ACM Web Conference*, Apr. 2023, pp.3299–3308. DOI: 10.1145/3543507.3583421.

[42] Huang R J, Zhao Z, Liu H D, Liu J L, Cui C Y, Ren Y. ProDiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proc. the 30th ACM International Conference on Multimedia*, Oct. 2022, pp.2595–2605. DOI: 10.1145/3503161.3547855.

[43] Luo W J. A comprehensive survey on knowledge distillation of diffusion models. arXiv: 2304.04262, 2023. https://arxiv.org/abs/2304.04262, May 2024.

[44] Luhman E, Luhman T. Knowledge distillation in iterative generative models for improved sampling speed. arXiv: 2101.02388,2021.https://arxiv.org/abs/2101.02388, May 2024.

[45] Zheng H K, Nie W L, Vahdat A, Azizzadenesheli K, Anandkumar A. Fast sampling of diffusion models via operator learning. In *Proc. the 40th International Conference on Machine Learning*, Jul. 2023, pp.42390–42402.

[46] Meng C L, Rombach R, Gao R Q, Kingma D, Ermon S, Ho J, Salimans T. On distillation of guided diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.14297–14306. DOI: 10.1109/cvpr52729.2023.01374.

[47] Berthelot D, Autef A, Lin J R, Yap D A, Zhai S F, Hu S Y, Zheng D, Talbott W, Gu E. TRACT: Denoising diffusion models with transitive closure time-distillation. arXiv:2303.04248,2023.https://arxiv.org/abs/2303.04248, May 2024.

[48] Daras G, Dagan Y, Dimakis A G, Daskalakis C. Score-guided intermediate layer optimization: Fast Langevin mixing for inverse problems. arXiv: 2206.09104, 2022. https://arxiv.org/abs/2206.09104, May 2024.

[49] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Oct. 2015, pp.234–241. DOI: 10.1007/978-3-319-24574-4_28.

[50] Salimans T, Kingma D P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Proc. the 30th International Conference on Neural Information Processing Systems*, Dec. 2016, pp.901–909.

[51] Wu Y X, He K M. Group normalization. *International Journal of Computer Vision*, 2020, 128(3): 742–755. DOI: 10.1007/s11263-019-01198-w.

[52] Chen C F R, Fan Q F, Panda R. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2021, pp.347–356. DOI: 10.1109/iccv48922.2021.00041.

[53] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models. In *Proc. the 38th International Conference on Machine Learning*, Jul. 2021, pp.8162–8171.

[54] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In *Proc. the 31st International Conference on Neural Information Processing Systems*, Dec. 2017, pp.6000–6010.

[55] Tamborrino A, Pellicanò N, Pannier B, Voitot P, Naudin L. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proc. the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp.3878–3887. DOI: 10.18653/v1/2020.acl-main.357.

[56] Wen Q S, Zhou T, Zhang C L, Chen W Q, Ma Z Q, Yan J C, Sun L. Transformers in time series: A survey. In *Proc. the 32nd International Joint Conference on Artificial Intelligence*, Aug. 2023, pp.6778–6786. DOI: 10.24963/ijcai.2023/759.

[57] Peebles W, Xie S N. Scalable diffusion models with transformers. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.4172–4182. DOI: 10.1109/iccv51070.2023.00387.

[58] Bao F, Nie S, Xue K W, Cao Y, Li C X, Su H, Zhu J. All are worth words: A ViT backbone for diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.22669–22679. DOI: 10.1109/cvpr52729.2023.02171.

[59] Gao S H, Zhou P, Cheng M M, Yan S C. Masked diffusion transformer is a strong image synthesizer. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.23107–23116. DOI:

10.1109/iccv51070.2023.02117.

[60] Hoogeboom E, Heek J, Salimans T. Simple diffusion: End-to-end diffusion for high resolution images. arXiv: 2301.11093, 2023. https://arxiv.org/abs/2301.11093, May 2024.

[61] Chen J W, Pan Y W, Yao T, Mei T. ControlStyle: Text-driven stylized image generation using diffusion priors. In *Proc. the 31st ACM International Conference on Multimedia*, Oct. 29/Nov. 3, 2023, pp.7540–7548. DOI: 10.1145/3581783.3612524.

[62] Blattmann A, Rombach R, Ling H, Dockhorn T, Kim S W, Fidler S, Kreis K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.22563–22575. DOI: 10.1109/cvpr52729.2023.02161.

[63] Avrahami O, Fried O, Lischinski D. Blended latent diffusion. *ACM Trans. Graphics*, 2023, 42(4): 149. DOI: 10.1145/3592450.

[64] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2022, pp.10674–10685. DOI: 10.1109/cvpr52688.2022.01042.

[65] Vlassis N N, Sun W, Alshibli K A, Regueiro R A. Synthesizing realistic sand assemblies with denoising diffusion in latent space. arXiv: 2306.04411, 2023. https://arxiv.org/abs/2306.04411, May 2024.

[66] Yu S, Sohn K, Kim S, Shin J. Video probabilistic diffusion models in projected latent space. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.18456–18466. DOI: 10.1109/cvpr52729.2023.01770.

[67] Braure T, Lazaro D, Hateau D, Brandon V, Ginsburger K. Conditioning generative latent optimization for sparse-view CT image reconstruction. arXiv: 2307.16670, 2023. https://arxiv.org/abs/2307.16670, May 2024.

[68] Koley S, Bhunia A K, Sain A, Chowdhury P N, Xiang T, Song Y Z. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2023, pp.6850–6861. DOI: 10.1109/cvpr52729.2023.00662.

[69] Do H, Yoo E, Kim T, Lee C, Choi J Y. Quantitative manipulation of custom attributes on 3D-aware image synthesis. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.8529–8538. DOI: 10.1109/cvpr52729.2023.00824.

[70] Hu V T, Zhang W, Tang M, Mettes P, Zhao D L, Snoek C. Latent space editing in transformer-based flow matching. In *Proc. the 38th AAAI Conference on Artificial Intelligence*, Feb. 2024, pp.2247–2255. DOI: 10.1609/aaai.v38i3.27998.

[71] Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents. arXiv: 2204.06125, 2022. https://arxiv.org/abs/2204.06125, May 2024.

[72] Liu H H, Chen Z H, Yuan Y, Mei X H, Liu X B, Mandic D, Wang W W, Plumbley M D. AudioLDM: Text-to-audio generation with latent diffusion models. arXiv: 2301.12503, 2023. https://arxiv.org/abs/2301.12503, May 2024.

[73] Schramowski P, Brack M, Deiseroth B, Kersting K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.22522–22531. DOI: 10.1109/cvpr52729.2023.02157.

[74] Ni H M, Shi C H, Li K, Huang S X, Min M R. Conditional image-to-video generation with latent flow diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.18444–18455. DOI: 10.1109/cvpr52729.2023.01769.

[75] Wang X, Yuan H J, Zhang S W, Chen D Y, Wang J N, Zhang Y Y, Shen Y J, Zhao D L, Zhou J R. VideoComposer: Compositional video synthesis with motion controllability. arXiv: 2306.02018, 2023. https://arxiv.org/abs/2306.02018, May 2024.

[76] Saharia C, Chan W, Saxena S, Li L L, Whang J, Denton E, Ghasemipour S K S, Ayan B K, Mahdavi S S, Gontijo-Lopes R, Salimans T, Ho J, Fleet D J, Norouzi M. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. the 36th International Conference on Neural Information Processing Systems*, Nov. 28/Dec. 9, 2022, Article No. 2643.

[77] Saharia C, Chan W, Chang H W, Lee C, Ho J, Salimans T, Fleet D, Norouzi M. Palette: Image-to-image diffusion models. In *Proc. the 2022 Conference on Special Interest Group on Computer Graphics and Interactive Techniques*, Aug. 2022, Article No. 15. DOI: 10.1145/3528233.3530757.

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

[79] Chang H W, Zhang H, Barber J, Maschinot A J, Lezama J, Jiang L, Yang M H, Murphy K, Freeman W T, Rubinstein M, Li Y Z, Krishnan D. Muse: Text-to-image generation via masked generative transformers. arXiv: 2301.00704, 2023. https://arxiv.org/abs/2301.00704, May 2024.

[80] Saharia C, Ho J, Chan W, Salimans T, Fleet D J, Norouzi M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2023, 45(4): 4713–4726. DOI: 10.1109/tpami.2022.3204461.

[81] Balaji Y, Nah S, Huang X, Vahdat A, Song J M, Zhang Q S, Kreis K, Aittala M, Aila T, Laine S, Catanzaro B, Karras T, Liu M Y. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv: 2211.01324, 2022. https://arxiv.org/abs/2211.01324, May 2024.

[82] Kim S, Jung S, Kim B, Choi M, Shin J, Lee J. Towards safe self-distillation of Internet-scale text-to-image diffusion models. arXiv: 2307.05977, 2023. https://arxiv.org/abs/2307.05977, May 2024.

[83] Li Y H, Liu H T, Wu Q Y, Mu F Z, Yang J W, Gao J F, Li C Y, Lee Y J. GLIGEN: Open-set grounded text-to-image generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2023, pp.22511–22521. DOI: 10.1109/cvpr52729.2023.02156.

[84] Mou C, Wang X T, Xie L B, Wu Y Z, Zhang J, Qi Z A, Shan Y. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proc. the 38th AAAI Conference on Artificial Intelligence*, Feb. 2024, pp.4296–4304. DOI: 10.1609/aaai.v38i5.28226.

[85] Chen D, Qi X D, Zheng Y, Lu Y Z, Huang Y B, Li Z J. Deep data augmentation for weed recognition enhancement: A diffusion probabilistic model and transfer learning based approach. In *Proc. the 2023 ASABE Annual International Meeting*, Jul. 2023. DOI: 10.13031/aim.202300108.

[86] Ding K Z, Xu Z, Tong H H, Liu H. Data augmentation for deep graph learning: A survey. *ACM SIGKDD Explorations Newsletter*, 2022, 24(2): 61–77. DOI: 10.1145/3575637.3575646.

[87] Zheng G C, Zhou X P, Li X W, Qi Z A, Shan Y, Li X. LayoutDiffusion: Controllable diffusion model for layout-to-image generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.22490–22499. DOI: 10.1109/cvpr52729.2023.02154.

[88] Inoue N, Kikuchi K, Simo-Serra E, Otani M, Yamaguchi K. LayoutDM: Discrete diffusion model for controllable layout generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.10167–10176. DOI: 10.1109/cvpr52729.2023.00980.

[89] Avrahami O, Hayes T, Gafni O, Gupta S, Taigman Y, Parikh D, Lischinski D, Fried O, Yin X. SpaText: Spatio-textual representation for controllable image generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.18370–18380. DOI: 10.1109/cvpr52729.2023.01762.

[90] Yang Z Y, Wang J F, Gan Z, Li L J, Lin K, Wu C F, Duan N, Liu Z C, Liu C, Zeng M, Wang L J. ReCo: Region-controlled text-to-image generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.14246–14255. DOI: 10.1109/cvpr52729.2023.01369.

[91] Xie J H, Li Y X, Huang Y W, Liu H Z, Zhang W T, Zheng Y F, Shou M Z. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.7418–7427. DOI: 10.1109/iccv51070.2023.00685.

[92] Voynov A, Aberman K, Cohen-Or D. Sketch-guided text-to-image diffusion models. In *Proc. the 2023 Conference on Special Interest Group on Computer Graphics and Interactive Techniques*, Jul. 2023, Article No. 55. DOI: 10.1145/3588432.3591560.

[93] Yu J W, Wang Y H, Zhao C, Ghanem B, Zhang J. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.23117–23127. DOI: 10.1109/iccv51070.2023.02118.

[94] Li D X, Li J N, Hoi S C H. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv: 2305.14720, 2023. https://arxiv.org/abs/2305.14720, May 2024.

[95] Zhao S H, Chen D D, Chen Y C, Bao J M, Hao S Z, Yuan L, Wong K Y K. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *Proc. the 37th Conference on Neural Information Processing Systems*, Dec. 2023.

[96] Qin C, Zhang S, Yu N, Feng Y H, Yang X Y, Zhou Y B, Wang H, Niebles J C, Xiong C M, Savarese S, Ermon S, Fu Y, Xu R. UniControl: A unified diffusion model for controllable visual generation in the wild. arXiv: 2305.11147, 2023. https://arxiv.org/abs/2305.11147, May 2024.

[97] Huang L H, Chen D, Liu Y, Shen Y J, Zhao D L, Zhou J R. Composer: Creative and controllable image synthesis with composable conditions. arXiv: 2302.09778, 2023. https://arxiv.org/abs/2302.09778, May 2024.

[98] Cao Z, Simon T, Wei S E, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jul. 2017, pp.1302–1310. DOI: 10.1109/cvpr.2017.143.

[99] Yang R H, Srivastava P, Mandt S. Diffusion probabilistic modeling for video generation. *Entropy*, 2023, 25(10): 1469. DOI: 10.3390/e25101469.

[100] Mo S C, Mu F Z, Lin K H, Liu Y L, Guan B C, Li Y, Zhou B L. FreeControl: Training-free spatial control of any text-to-image diffusion model with any condition. arXiv: 2312.07536, 2023. https://arxiv.org/abs/2312.07536, May 2024.

[101] Patashnik O, Wu Z Z, Shechtman E, Cohen-Or D, Lischinski D. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2021, pp.2065–2074. DOI: 10.1109/iccv48922.2021.00209.

[102] Wu Z Z, Lischinski D, Shechtman E. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2021, pp.12858–12867. DOI: 10.1109/cvpr46437.2021.01267.

[103] Liu Z H, Feng R L, Zhu K, Zhang Y F, Zheng K C, Liu Y, Zhao D L, Zhou J R, Cao Y. Cones: Concept neurons in diffusion models for customized generation. arXiv: 2303.05125, 2023. https://arxiv.org/abs/2303.05125, May 2024.

[104] Yang B X, Gu S Y, Zhang B, Zhang T, Chen X J, Sun X Y, Chen D, Wen F. Paint by example: Exemplar-based image editing with diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.18381–18391. DOI: 10.1109/cvpr52729.2023.01763.

[105] Song Y Z, Zhang Z F, Lin Z, Cohen S, Price B, Zhang J M, Kim S Y, Aliaga D. ObjectStitch: Object compositing with diffusion model. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.18310–18319. DOI: 10.1109/cvpr52729.2023.01756.

[106] Pan Z H, Zhou X, Tian H. Arbitrary style guidance for enhanced diffusion-based text-to-image generation. In *Proc. the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Jan. 2023, pp.4450–4460. DOI: 10.1109/wacv56688.2023.00444.

[107] Kang M, Han W, Hwang S J, Yang E. ZET-Speech: Zero-shot adaptive emotion-controllable text-to-speech synthesis with diffusion and style-based models. In *Proc. the 2023 INTERSPEECH*, Aug. 2023, pp.4339–4343. DOI: 10.21437/interspeech.2023-754.

[108] Huang N S, Zhang Y X, Tang F, Ma C Y, Huang H B, Dong W M, Xu C S. DiffStyler: Controllable dual diffusion for text-driven image stylization. *IEEE Trans. Neural Networks and Learning Systems*, 2024. DOI: 10.1109/tnnls.2023.3342645. (early access)

[109] Tarrés G C, Ruta D, Bui T, Collomosse J. PARASOL: Parametric style control for diffusion image synthesis. arXiv: 2303.06464, 2023. https://arxiv.org/abs/2303.06464, May 2024.

[110] Nair N G, Cherian A, Lohit S, Wang Y, Koike-Akino T, Patel V M, Marks T K. Steered diffusion: A generalized framework for plug-and-play conditional image synthesis. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.20793–20803. DOI: 10.1109/iccv51070.2023.01906.

[111] Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano A H, Chechik G, Cohen-Or D. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv: 2208.01618, 2022. https://arxiv.org/abs/2208.01618, May 2024.

[112] Ruiz N, Li Y z, Jampani V, Pritch Y, Rubinstein M, Aberman K. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.22500–22510. DOI: 10.1109/cvpr52729.2023.02155.

[113] Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, Wang L, Chen W Z. LoRA: Low-rank adaptation of large language models. arXiv: 2106.09685, 2021. https://arxiv.org/abs/2106.09685, May 2024.

[114] Lu H M, Tunanyan H, Wang K, Navasardyan S, Wang Z Y, Shi H. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.14267–14276. DOI: 10.1109/cvpr52729.2023.01371.

[115] Yu J H, Xu Y Z, Koh J Y, Luong T, Baid G, Wang Z R, Vasudevan V, Ku A, Yang Y F, Ayan B K, Hutchinson B, Han W, Parekh Z, Li X, Zhang H, Baldridge J, Wu Y H. Scaling autoregressive models for content-rich text-to-image generation. arXiv: 2206.10789, 2022. https://arxiv.org/abs/2206.10789, May 2024.

[116] Meng C L, He Y T, Song Y, Song J M, Wu J J, Zhu J Y, Ermon S. SDEdit: Guided image synthesis and editing with stochastic differential equations. arXiv: 2108.01073, 2021. https://arxiv.org/abs/2108.01073, May 2024.

[117] Zhu Y Z, Li Z H, Wang T W, He M C, Yao C. Conditional text image generation with diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.14235–14244. DOI: 10.1109/cvpr52729.2023.01368.

[118] Huang Z Q, Chan K C K, Jiang Y M, Liu Z W. Collaborative diffusion for multi-modal face generation and editing. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.6080–6090. DOI: 10.1109/cvpr52729.2023.00589.

[119] Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv: 2112.10741, 2021. https://arxiv.org/abs/2112.10741, May 2024.

[120] Liu X H, Park D H, Azadi S, Zhang G, Chopikyan A, Hu Y X, Shi H, Rohrbach A, Darrell T. More control for free! Image synthesis with semantic diffusion guidance. In *Proc. the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Jan. 2023, pp.289–299. DOI: 10.1109/wacv56688.2023.00037.

[121] Xifara T, Sherlock C, Livingstone S, Byrne S, Girolami M. Langevin diffusions and the metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 2014, 91: 14–19. DOI: 10.1016/j.spl.2014.04.002.

[122] Luo C. Understanding diffusion models: A unified perspective. arXiv: 2208.11970, 2022. https://arxiv.org/abs/2208.11970, May 2024.

[123] Ho J, Salimans T. Classifier-free diffusion guidance. arXiv: 2207.12598, 2022. https://arxiv.org/abs/2207.12598, May 2024.

[124] Hosseini H, Xiao B C, Poovendran R. Google's cloud vision API is not robust to noise. In *Proc. the 16th IEEE International Conference on Machine Learning and Applications* (*ICMLA*), Dec. 2017, pp.101–105. DOI: 10.1109/icmla.2017.0-172.

[125] Wallace B, Gokul A, Ermon S, Naik N. End-to-end diffusion latent optimization improves classifier guidance. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.7246–7256. DOI: 10.1109/iccv51070.2023.00669.

[126] Bansal A, Borgnia E, Chu H M, Li J S, Kazemi H, Huang F R, Goldblum M, Geiping J, Goldstein T. Cold diffusion: Inverting arbitrary image transforms without noise. arXiv: 2208.09392, 2022. https://arxiv.org/abs/

2208.09392, May 2024.

[127] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In *Proc. the 38th International Conference on Machine Learning*, Jul. 2021, pp.8748–8763.

[128] Hertz A, Mokady R, Tenenbaum J, Aberman K, Pritch Y, Cohen-Or D. Prompt-to-prompt image editing with cross attention control. arXiv: 2208.01626, 2022. https://arxiv.org/abs/2208.01626, May 2024.

[129] Mokady R, Hertz A, Aberman K, Pritch Y, Cohen-Or D. Null-text inversion for editing real images using guided diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.6038–6047. DOI: 10.1109/cvpr52729.2023.00585.

[130] Feng W X, He X H, Fu T J, Jampani V, Akula A, Narayana P, Basu S, Wang X E, Wang W Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv: 2212.05032, 2022. https://arxiv.org/abs/2212.05032, May 2024.

[131] Chen M H, Laina I, Vedaldi A. Training-free layout control with cross-attention guidance. In *Proc. the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Jan. 2024, pp.5331–5341. DOI: 10.1109/wacv57701.2024.00526.

[132] He Y T, Salakhutdinov R, Kolter J Z. Localized text-to-image generation for free via cross attention control. arXiv: 2306.14636, 2023. https://arxiv.org/abs/2306.14636, May 2024.

[133] Parmar G, Singh K K, Zhang R, Li Y J, Lu J W, Zhu J Y. Zero-shot image-to-image translation. In *Proc. the 2023 Conference on Special Interest Group on Computer Graphics and Interactive Techniques*, Jul. 2023, Article No. 11. DOI: 10.1145/3588432.3591513.

[134] Mou C, Wang X T, Song J C, Shan Y, Zhang J. DragonDiffusion: Enabling drag-style manipulation on diffusion models. arXiv: 2307.02421, 2023. https://arxiv.org/abs/2307.02421, May 2024.

[135] Choi J, Kim S, Jeong Y, Gwon Y, Yoon S. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2021, pp.14347–14356. DOI: 10.1109/ICCV48922.2021.01410.

[136] Kawar B, Elad M, Ermon S, Song J M. Denoising diffusion restoration models. In *Proc. the 36th International Conference on Neural Information Processing Systems*, Nov. 28/Dec. 9, 2022, Article No. 1714.

[137] Lugmayr A, Danelljan M, Romero A, Yu F, Timofte R, Van Gool L. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2022, pp.11451–11461. DOI: 10.1109/cvpr52688.2022.01117.

[138] Wang Y H, Yu J W, Zhang J. Zero-shot image restoration using denoising diffusion null-space model. arXiv: 2212.00490, 2022. https://arxiv.org/abs/2212.00490, May 2024.

[139] Wang Y H, Hu Y J, Yu J W, Zhang J. GAN prior based null-space learning for consistent super-resolution. In *Proc. the 37th AAAI Conference on Artificial Intelligence*, Feb. 2023, pp.2724–2732. DOI: 10.1609/aaai.v37i3.25372.

[140] Chen D D, Davies M E. Deep decomposition learning for inverse imaging problems. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.510–526. DOI: 10.1007/978-3-030-58604-1_31.

[141] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. arXiv: 1809.11096, 2018. https://arxiv.org/abs/1809.11096, May 2024.

[142] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. the 31st International Conference on Neural Information Processing Systems*, Dec. 2017, pp.6629–6640.

[143] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In *Proc. the 30th International Conference on Neural Information Processing Systems*, Dec. 2016, pp.2234–2242.

[144] Cho J, Li L J, Yang Z Y, Gan Z, Wang L J, Bansal M. Diagnostic benchmark and iterative inpainting for layout-guided image generation. arXiv: 2304.06671, 2023. https://arxiv.org/abs/2304.06671, May 2024.

[145] Li H Y, Yang Y F, Chang M, Chen S Q, Feng H J, Xu Z H, Li Q, Chen Y T. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022, 479: 47–59. DOI: 10.1016/j.neucom.2022.01.029.

[146] Fei B, Lyu Z Y, Pan L, Zhang J Z, Yang W D, Luo T Y, Zhang B, Dai B. Generative diffusion prior for unified image restoration and enhancement. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.9935–9946. DOI: 10.1109/cvpr52729.2023.00958.

[147] Zheng G C, Li S M, Wang H, Yao T P, Chen Y, Ding S H, Li X. Entropy-driven sampling and training scheme for conditional diffusion generation. In *Proc. the 17th European Conference on Computer Vision*, Oct. 2022, pp.754–769. DOI: 10.1007/978-3-031-20047-2_43.

[148] Harvey W, Naderiparizi S, Masrani V, Weilbach C, Wood F. Flexible diffusion modeling of long videos. In *Proc. the 36th International Conference on Neural Information Processing Systems*, Nov. 28/Dec. 9, 2022, Article No. 2027.

[149] Voleti V, Jolicoeur-Martineau A, Pal C. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *Proc. the 36th International Conference on Neural Information Processing Systems*, Nov. 28/Dec. 9, 2022, Article No. 1698.

[150] Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S Y, Hu Q Y, Yang H, Ashual O, Gafni O, Parikh D, Gupta S, Taigman Y. Make-A-Video: Text-to-video generation without text-video data. arXiv: 2209.14792, 2022. https://

arxiv.org/abs/2209.14792, May 2024.

[151] Xing J B, Xia M H, Liu Y X, Zhang Y C, Zhang Y, He Y Q, Liu H Y, Chen H X, Cun X D, Wang X T, Shan Y, Wong T T. Make-Your-Video: Customized video generation using textual and structural guidance. *IEEE Trans. Visualization and Computer Graphics*, 20241–15. DOI: 10.1109/tvcg.2024.3365804.

[152] Ma W D K, Lahiri A, Lewis J P, Leung T, Kleijn W B. Directed diffusion: Direct control of object placement through attention guidance. In *Proc. the 38th AAAI Conference on Artificial Intelligence*, Feb. 2024, pp.4098–4106. DOI: 10.1609/aaai.v38i5.28204.

[153] Zhang Y B, Wei Y X, Jiang D S, Zhang X P, Zuo W M, Tian Q. ControlVideo: Training-free controllable text-to-video generation. arXiv: 2305.13077, 2023. https://arxiv.org/abs/2305.13077, May 2024.

[154] Luo Z X, Chen D Y, Zhang Y Y, Huang Y, Wang L, Shen Y J, Zhao D L, Zhou J R, Tan T N. Notice of removal: VideoFusion: Decomposed diffusion models for high-quality video generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2023, pp.10209–10218. DOI: 10.1109/CVPR52729.2023.00984.

[155] Poole B, Jain A, Barron J T, Mildenhall B. DreamFusion: Text-to-3D using 2D diffusion. arXiv: 2209.14988, 2022. https://arxiv.org/abs/2209.14988, May 2024.

[156] Lin C H, Gao J, Tang L M, Takikawa T, Zeng X H, Huang X, Kreis K, Fidler S, Liu M Y, Lin T Y. Magic3D: High-resolution text-to-3D content creation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.300–309. DOI: 10.1109/cvpr52729.2023.00037.

[157] Chen R, Chen Y W, Jiao N X, Jia K. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.22189–22199. DOI: 10.1109/iccv51070.2023.02033.

[158] Liu R S, Wu R D, Van Hoorick B, Tokmakov P, Zakharov S, Vondrick C. Zero-1-to-3: Zero-shot one image to 3D object. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.9264–9275. DOI: 10.1109/iccv51070.2023.00853.

[159] Qian G C, Mai J J, Hamdi A, Ren J, Siarohin A, Li B, Lee H Y, Skorokhodov I, Wonka P, Tulyakov S, Ghanem B. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. arXiv: 2306.17843, 2023. https://arxiv.org/abs/2306.17843, May 2024.

[160] Liu Y, Lin C, Zeng Z J, Long X X, Liu L J, Komura T, Wang W P. SyncDreamer: Generating multiview-consistent images from a single-view image. arXiv: 2309.03453, 2023. https://arxiv.org/abs/2309.03453, May 2024.

[161] Zheng X Y, Pan H, Wang P S, Tong X, Liu Y, Shum H Y. Locally attentional SDF diffusion for controllable 3D shape generation. *ACM Trans. Graphics*, 2023, 42(4): 91. DOI: 10.1145/3592103.

[162] Han L G, Li Y X, Zhang H, Milanfar P, Metaxas D, Yang F. SVDiff: Compact parameter space for diffusion fine-tuning. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.7289–7300. DOI: 10.1109/iccv51070.2023.00673.

[163] Tewel Y, Gal R, Chechik G, Atzmon Y. Key-locked rank one editing for text-to-image personalization. In *Proc. the 2023 Conference on Special Interest Group on Computer Graphics and Interactive Techniques*, Jul. 2023, Article No. 12. DOI: 10.1145/3588432.3591506.

[164] Shamsian A, Navon A, Fetaya E, Chechik G. Personalized federated learning using hypernetworks. In *Proc. the 38th International Conference on Machine Learning*, Jul. 2021, pp.9489–9502.

[165] Wei Y X, Zhang Y B, Ji Z L, Bai J F, Zhang L, Zuo W M. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.15897–15907. DOI: 10.1109/iccv51070.2023.01461.

[166] Zhou Y F, Zhang R Y, Sun T, Xu J H. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. arXiv: 2305.13579, 2023. https://arxiv.org/abs/2305.13579, May 2024.

[167] Gu Y C, Wang X T, Wu J Z, Shi Y J, Chen Y P, Fan Z H, Xiao W Y, Zhao R, Chang S N, Wu W J, Ge Y X, Shan Y, Shou M Z. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. arXiv: 2305.18292, 2023. https://arxiv.org/abs/2305.18292, May 2024.

[168] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 2004, 13(4): 600–612. DOI: 10.1109/TIP.2003.819861.

[169] Horé A, Ziou D. Image quality metrics: PSNR vs. SSIM. In *Proc. the 20th International Conference on Pattern Recognition*, Aug. 2010, pp.2366–2369. DOI: 10.1109/icpr.2010.579.

[170] Zhang R, Isola P, Efros A A, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.586–595. DOI: 10.1109/cvpr.2018.00068.

[171] Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S. FVD: A new metric for video generation. In *Proc. the 2019 International Conference on Learning Representations*, May 2019.

[172] Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp.7514–7528. DOI: 10.18653/v1/2021.emnlp-main.595.

[173] Sajjadi M S M, Bachem O, Lucic M, Bousquet O, Gelly S. Assessing generative models via precision and recall. In *Proc. the 32nd International Conference on Neural Information Processing Systems*, Dec. 2018, pp.5234–5243.

[174] Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T.

Improved precision and recall metric for assessing generative models. In *Proc. the 33rd International Conference on Neural Information Processing Systems*, Dec. 2019, Article No. 353.

[175] Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R, Jitsev J. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proc. the 36th International Conference on Neural Information Processing Systems*, Nov. 28/Dec. 9, 2022, Article No. 1833.

[176] Zhou Y F, Liu B C, Zhu Y Z, Yang X, Chen C Y, Xu J H. Shifted diffusion for text-to-image generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.10157–10166. DOI: 10.1109/cvpr52729.2023.00979.

[177] Feng Z D, Zhang Z Y, Yu X T, Fang Y W, Li L X, Chen X Y, Lu Y X, Liu J X, Yin W C, Feng S K, Sun Y, Chen L, Tian H, Wu H, Wang H F. ERNIE-ViLG 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.10135–10145. DOI: 10.1109/cvpr52729.2023.00977.

[178] Wei C, Mangalam K, Huang P Y, Li Y H, Fan H Q, Xu H, Wang H Y, Xie C H, Yuille A, Feichtenhofer C. Diffusion models as masked autoencoders. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.16238–16248. DOI: 10.1109/iccv51070.2023.01492.

[179] Deng J, Dong W, Socher R, Li L J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In *Proc. the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp.248–255. DOI: 10.1109/cvpr.2009.5206848.

[180] Pan X G, Zhan X H, Dai B, Lin D H, Loy C C, Luo P. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2022, 44(11): 7474–7489. DOI: 10.1109/tpami.2021.3115428.

[181] Kawar B, Vaksman G, Elad M. SNIPS: Solving noisy inverse problems stochastically. In *Proc. the 35th Conference on Neural Information Processing Systems*, Dec. 2021, pp.21757–21769.

[182] Romano Y, Elad M, Milanfar P. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 2017, 10(4): 1804–1844. DOI: 10.1137/16m1102884.

[183] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. arXiv: 1710.10196, 2017. https://arxiv.org/abs/1710.10196, May 2024.

[184] Cun X D, Pun C M, Shi C. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *Proc. the 34th AAAI Conference on Artificial Intelligence*, Feb. 2020, pp.10680–10687. DOI: 10.1609/aaai.v34i07.6695.

[185] Luo Z W, Gustafsson F K, Zhao Z, Sjölund J, Schön T B. Image restoration with mean-reverting stochastic differential equations. arXiv: 2301.11699, 2023. https://arxiv.org/abs/2301.11699, May 2024.

[186] Luo Z W, Gustafsson F K, Zhao Z, Sjölund J, Schön T B. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Jun. 2023, pp.1680–1691. DOI: 10.1109/cvprw59228.2023.00169.

[187] Wei C, Wang W J, Yang W H, Liu J Y. Deep retinex decomposition for low-light enhancement. arXiv: 1808.04560, 2018. https://arxiv.org/abs/1808.04560, May 2024.

[188] Li C Y, Guo J C, Porikli F, Pang Y W. LightenNet: A convolutional neural network for weakly illuminated image enhancement. *Pattern Recognition Letters*, 2018, 104: 15–22. DOI: 10.1016/j.patrec.2018.01.010.

[189] Jiang Y F, Gong X Y, Liu D, Cheng Y, Fang C, Shen X H, Yang J C, Zhou P, Wang Z Y. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Trans. Image Processing*, 2021, 30: 2340–2349. DOI: 10.1109/tip.2021.3051462.

[190] Zhang Y H, Zhang J W, Guo X J. Kindling the darkness: A practical low-light image enhancer. In *Proc. the 27th ACM International Conference on Multimedia*, Oct. 2019, pp.1632–1640. DOI: 10.1145/3343031.3350926.

[191] Liu J Y, Xu D J, Yang W H, Fan M H, Huang H F. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 2021, 129(4): 1153–1184. DOI: 10.1007/s11263-020-01418-8.

[192] Sauer A, Schwarz K, Geiger A. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *Proc. the 2022 Conference on Special Interest Group on Computer Graphics and Interactive Techniques*, Aug. 2022, Article No. 49. DOI: 10.1145/3528233.3530738.

[193] Hang T K, Gu S Y, Li C, Bao J M, Chen D, Hu H, Geng X, Guo B N. Efficient diffusion training via min-SNR weighting strategy. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision* (*ICCV*), Oct. 2023, pp.7407–7417. DOI: 10.1109/iccv51070.2023.00684.

[194] Choi J, Lee J, Shin C, Kim S, Kim H, Yoon S. Perception prioritized training of diffusion models. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2022, pp.11462–11471. DOI: 10.1109/cvpr52688.2022.01118.

[195] Yang X Y, Zhou D Q, Feng J S, Wang X C. Diffusion probabilistic model made slim. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.22552–22562. DOI: 10.1109/cvpr52729.2023.02160.

[196] Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf, May 2024.

[197] Vahdat A, Kreis K, Kautz J. Score-based generative modeling in latent space. In *Proc. the 35th Conference on Neural Information Processing Systems*, Dec. 2021, pp.11287–11302.

[198] Tan F W, Feng S, Ordonez V. Text2Scene: Generating compositional scenes from textual descriptions. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2019, pp.6703–6712. DOI: 10.1109/cvpr.2019.00687.

[199] Hinz T, Heinrich S, Wermter S. Semantic object accuracy for generative text-to-image synthesis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2022, 44(3): 1552–1565. DOI: 10.1109/tpami.2020.3021209.

[200] Yu J H, Li X, Koh J Y, Zhang H, Pang R M, Qin J, Ku A, Xu Y Z, Baldridge J, Wu Y H. Vector-quantized image modeling with improved VQGAN. arXiv: 2110.04627, 2021. https://arxiv.org/abs/2110.04627, May 2024.

[201] Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft COCO: Common objects in context. In *Proc. the 13th European Conference on Computer Vision*, Sept. 2014, pp.740–755. DOI: 10.1007/978-3-319-10602-1_48.

[202] Zhou Y F, Zhang R Y, Chen C Y, Li C Y, Tensmeyer C, Yu T, Gu J X, Xu J H, Sun T. Towards language-free training for text-to-image generation. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2022, pp.17886–17896. DOI: 10.1109/cvpr52688.2022.01738.

[203] Ding M, Yang Z Y, Hong W Y, Zheng W D, Zhou C, Yin D, Lin J Y, Zou X, Shao Z, Yang H X, Tang J. CogView: Mastering text-to-image generation via transformers. In *Proc. the 35th Conference on Neural Information Processing Systems*, Dec. 2021, pp.19822–19835.

[204] Ho J, Chan W, Saharia C, Whang J, Gao R Q, Gritsenko A, Kingma D P, Poole B, Norouzi M, Fleet D J, Salimans T. Imagen video: High definition video generation with diffusion models. arXiv: 2210.02303, 2022. https://arxiv.org/abs/2210.02303, May 2024.

[205] Molad E, Horwitz E, Valevski D, Acha A R, Matias Y, Pritch Y, Leviathan Y, Hoshen Y. Dreamix: Video diffusion models are general video editors. arXiv: 2302.01329, 2023. https://arxiv.org/abs/2302.01329, May 2024.

[206] Mei K F, Patel V. VIDM: Video implicit diffusion models. In *Proc. the 37th AAAI Conference on Artificial Intelligence*, Feb. 2023, pp.9117–9125. DOI: 10.1609/aaai.v37i8.26094.

[207] Zhou D Q, Wang W M, Yan H S, Lv W W, Zhu Y Z, Feng J S. MagicVideo: Efficient video generation with latent diffusion models. arXiv: 2211.11018, 2022. https://arxiv.org/abs/2211.11018, May 2024.

[208] Deng Z J, He X T, Peng Y X, Zhu X W, Cheng L L. MV-Diffusion: Motion-aware video diffusion model. In *Proc. the 31st ACM International Conference on Multimedia*, Oct. 29/Nov. 3, 2023, pp.7255–7263. DOI: 10.1145/3581783.3612405.

[209] Deng Z J, He X T, Peng Y X. Efficiency-optimized video diffusion models. In *Proc. the 31st ACM International Conference on Multimedia*, Oct. 29/Nov. 3, 2023, pp.7295–7303. DOI: 10.1145/3581783.3612406.

[210] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012. https://arxiv.org/abs/1212.0402, May 2024.

[211] Hong W Y, Ding M, Zheng W D, Liu X H, Tang J. CogVideo: Large-scale pretraining for text-to-video generation via transformers. arXiv: 2205.15868, 2022. https://arxiv.org/abs/2205.15868, May 2024.

[212] Xu J, Mei T, Yao T, Rui Y. MSR-VTT: A large video description dataset for bridging video and language. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2016, pp.5288–5296. DOI: 10.1109/cvpr.2016.571.

[213] Wu C F, Huang L, Zhang Q X, Li B Y, Ji L, Yang F, Sapiro G, Duan N. GODIVA: Generating open-domain videos from natural descriptions. arXiv: 2104.14806, 2021. https://arxiv.org/abs/2104.14806, May 2024.

[214] Wu C F, Liang J, Ji L, Yang F, Fang Y J, Jiang D X, Duan N. NÜWA: Visual synthesis pre-training for neural visual world creation. In *Proc. the 17th European Conference on Computer Vision*, Oct. 2022, pp.720–736. DOI: 10.1007/978-3-031-19787-1_41.

[215] Xu J L, Wang X T, Cheng W H, Cao Y P, Shan Y, Qie X H, Gao S H. Dream3D: Zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.20908–20918. DOI: 10.1109/cvpr52729.2023.02003.

[216] Wang H C, Du X D, Li J H, Yeh R A, Shakhnarovich G. Score jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.12619–12629. DOI: 10.1109/cvpr52729.2023.01214.

[217] Long X X, Guo Y C, Lin C, Liu Y, Dou Z Y, Liu L J, Ma Y X, Zhang S H, Habermann M, Theobalt C, Wang W P. Wonder3D: Single image to 3D using cross-domain diffusion. arXiv: 2310.15008, 2023. https://arxiv.org/abs/2310.15008, May 2024.

[218] Shi Y C, Wang P, Ye J L, Long M, Li K J, Yang X. MVDream: Multi-view diffusion for 3D generation. arXiv: 2308.16512, 2023. https://arxiv.org/abs/2308.16512, May 2024.

[219] Wang T F, Zhang B, Zhang T, Gu S Y, Bao J M, Baltrusaitis T, Shen J J, Chen D, Wen F, Chen Q F, Guo B N. RODIN: A generative model for sculpting 3D digital avatars using diffusion. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.4563–4573. DOI: 10.1109/cvpr52729.2023.00443.

[220] Downs L, Francis A, Koenig N, Kinman B, Hickman R, Reymann K, McHugh T B, Vanhoucke V. Google

scanned objects: A high-quality dataset of 3D scanned household items. In *Proc. the 2022 International Conference on Robotics and Automation* (*ICRA*), May 2022, pp.2553–2560. DOI: 10.1109/icra46639.2022.9811809.

[221] Melas-Kyriazi L, Laina I, Rupprecht C, Vedaldi A. RealFusion 360°; reconstruction of any object from a single image. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.8446–8455. DOI: 10.1109/cvpr52729.2023.00816.

[222] Liu M H, Xu C, Jin H A, Chen L H, Varma T M, Xu Z X, Su H. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. arXiv: 2306.16928, 2023. https://arxiv.org/abs/2306.16928, May 2024.

[223] Jun H, Nichol A. Shap-E: Generating conditional 3D implicit functions. arXiv: 2305.02463, 2023. https://arxiv.org/abs/2305.02463, May 2024.

[224] Voynov A, Chu Q H, Cohen-Or D, Aberman K. P+: Extended textual conditioning in text-to-image generation. arXiv: 2303.09522, 2023. https://arxiv.org/abs/2303.09522, May 2024.

[225] Shi J, Xiong W, Lin Z, Jung H J. InstantBooth: Personalized text-to-image generation without test-time finetuning. arXiv: 2304.03411, 2023. https://arxiv.org/abs/2304.03411, May 2024.

[226] Jia X H, Zhao Y, Chan K C K, Li Y D, Zhang H, Gong B Q, Hou T B, Wang H S, Su Y C. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv: 2304.02642, 2023. https://arxiv.org/abs/2304.02642, May 2024.

[227] Xiao G X, Yin T W, Freeman W T, Durand F, Han S. FastComposer: Tuning-free multi-subject image generation with localized attention. arXiv: 2305.10431, 2023. https://arxiv.org/abs/2305.10431, May 2024.

[228] Chen W H, Hu H X, Li Y D, Ruiz N, Jia X H, Chang M W, Cohen W W. Subject-driven text-to-image generation via apprenticeship learning. arXiv: 2304.00186, 2023. https://arxiv.org/abs/2304.00186, May 2024.

[229] Ruiz N, Li Y Z, Jampani V, Wei W, Hou T B, Pritch Y, Wadhwa N, Rubinstein M, Aberman K. HyperDreamBooth: Hypernetworks for fast personalization of text-to-image models. arXiv: 2307.06949, 2023. https://arxiv.org/abs/2307.06949, May 2024.

[230] Gal R, Arar M, Atzmon Y, Bermano A H, Chechik G, Cohen-Or D. Designing an encoder for fast personalization of text-to-image models. arXiv: 2302.12228, 2023. https://arxiv.org/abs/2302.12228, May 2024.

[231] Arar M, Gal R, Atzmon Y, Chechik G, Cohen-Or D, Shamir A, Bermano A H. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *Proc. the 2023 Conference on SIGGRAPH Asia*, Dec. 2023, Article No. 72. DOI: 10.1145/3610548.3618173.

[232] Brooks T, Holynski A, Efros A A. InstructPix2Pix: Learning to follow image editing instructions. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.18392–18402. DOI: 10.1109/cvpr52729.2023.01764.

[233] Kawar B, Zada S, Lang O, Tov O, Chang H W, Dekel T, Mosseri I, Irani M. Imagic: Text-based real image editing with diffusion models. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Jun. 2023, pp.6007–6017. DOI: 10.1109/cvpr52729.2023.00582.

[234] Liu S T, Zhang Y C, Li W B, Lin Z, Jia J Y. Video-P2P: Video editing with cross-attention control. arXiv: 2303.04761, 2023. https://arxiv.org/abs/2303.04761, May 2024.

Rui Jiang is a Ph.D. student at Zhejiang University, Hangzhou. He obtained his Master's degree in computer science and technology from Wuhan Institute of Technology, Wuhan, in 2020. His major research interest is in computer vision, and now he is working on research topics about image generation.

Guang-Cong Zheng received his Bachelor's degree in computer science and technology from Northeastern University, Shenyang, in 2020. He is currently pursuing his Ph.D. degree in computer science and technology, Zhejiang University, Hangzhou. His current research interests include scene graph generation, image/video/3D generation, and diffusion model.

Teng Li is a master student at Zhejiang University, Hangzhou. Li received his B.S. degree in computer science from Zhejiang University, Hangzhou, in 2023. His research interests include computer vision and generative models.

Tian-Rui Yang is an undergraduate student at Nanjing University, Nanjing. She visited the Computer Vision Laboratory mentored by Professor Xi Li at Zhejiang University in 2023. Her research interests include image generation, speech synthesis, and medical image segmentation.

Jing-Dong Wang received his Bachelor's degree in automation and Master's degree in control science and engineering from Tsinghua University, Beijing, in 2001 and 2004, respectively, and his Ph.D. degree in computer science from Hong Kong University of Science and Technology, Hong Kong, in 2007. He joined Microsoft Research Asia, Beijing, in 2007, and is currently the chief architect of Baidu Computer Vision.

Xi Li received his Ph.D. degree in computer science and technology from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, in 2009. From 2009 to 2010, he was a postdoctoral researcher with CNRS, Telecom ParisTech, Paris. He was a senior researcher with the University of Adelaide, Adelaide. He is currently a full professor with Zhejiang University, Hangzhou. His research interests include visual tracking, compact learning, motion analysis, face recognition, data mining, and image retrieval.