Text-to-image Diffusion Models in Generative AI: A Survey

Chenshuang Zhang[a] Chaoning Zhang[b,1], Mengchun Zhang[a],
In So Kweon[a], Junmo Kim[a]

[a]*Korea Advanced Institute of Science and Technology (KAIST),*
[b]*Kyung Hee University,*

Abstract

This survey reviews the progress of diffusion models in generating images from text, *i.e.* text-to-image diffusion models. As a self-contained work, this survey starts with a brief introduction of how diffusion models work for image synthesis, followed by the background for text-conditioned image synthesis. Based on that, we present an organized review of pioneering methods and their improvements on text-to-image generation. We further summarize applications beyond image generation, such as text-guided generation for various modalities like videos, and text-guided image editing. Beyond the progress made so far, we discuss existing challenges and promising future directions.

Keywords: Generative models, Diffusion models, Text-to-image generation

1. Introduction

A picture is worth a thousand words. Images often convey stories more effectively than text alone. The ability to visualize from text enhances human understanding and enjoyment. Therefore, creating a system that generates realistic images from text descriptions, i.e., the text-to-image (T2I) task, is a significant step towards achieving human-like or general artificial intelligence. With the development of deep learning, text-to-image task has become one of the most impressive applications in computer
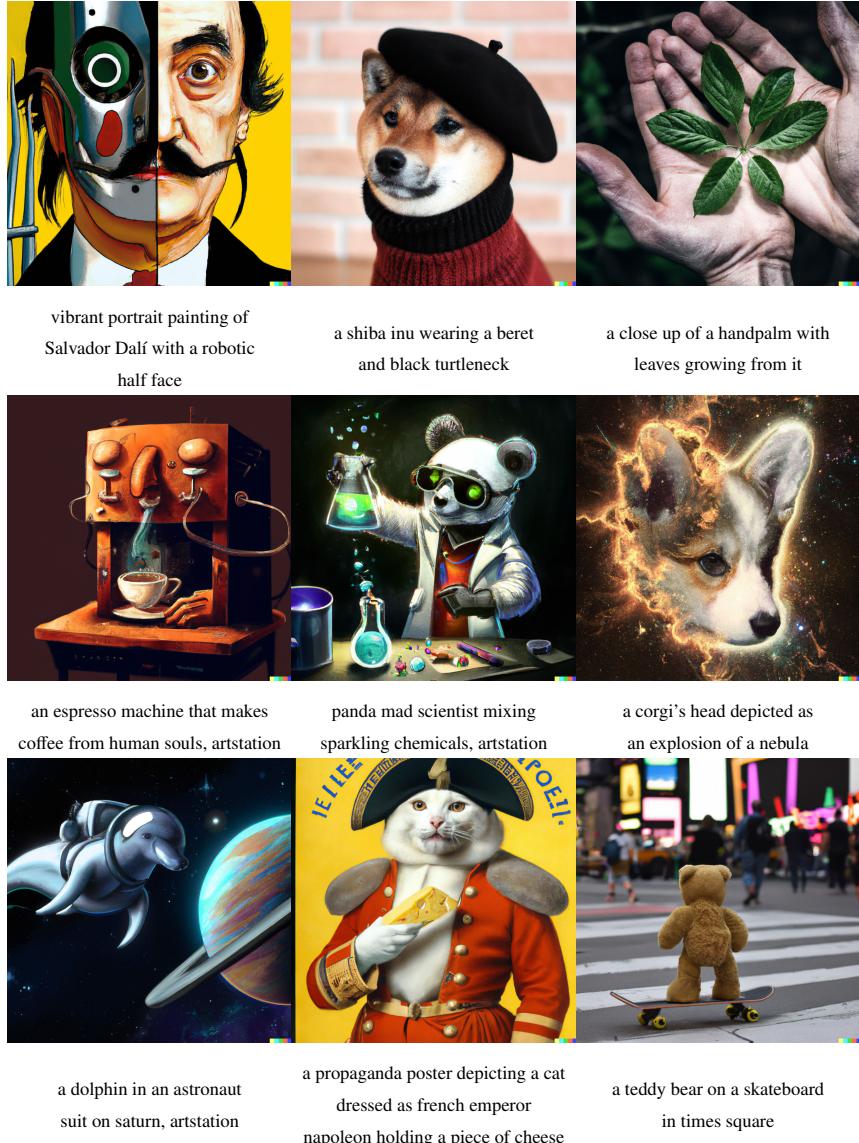
---

[1]Corresponding author.

vibrant portrait painting of
Salvador Dalí with a robotic
half face

a shiba inu wearing a beret
and black turtleneck

a close up of a handpalm with
leaves growing from it

an espresso machine that makes
coffee from human souls, artstation

panda mad scientist mixing
sparkling chemicals, artstation

a corgi's head depicted as
an explosion of a nebula

a dolphin in an astronaut
suit on saturn, artstation

a propaganda poster depicting a cat
dressed as french emperor
napoleon holding a piece of cheese

a teddy bear on a skateboard
in times square

Figure 1: Generated images by text-to-image diffusion models. These images are examples generated by the pioneering model DALL-E2 [1] from OpenAI. Based on user-input text prompts, the model can generate very imaginative images with high fidelity.

vision [2, 1].

We summarize the timeline of representative studies for text-to-image generation in Figure 2. AlignDRAW [3] marked a significant step by creating images from nat-
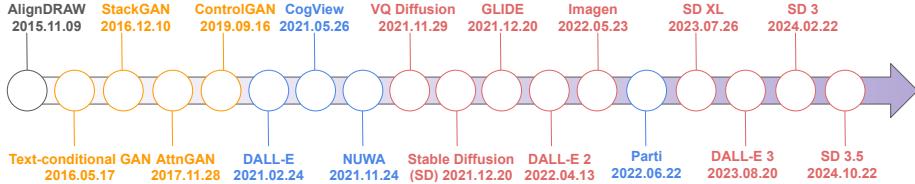
Figure 2: Representative works on text-to-image task over time. The GAN-based methods, autoregressive methods, and diffusion-based methods are masked in yellow, blue and red, respectively. We abbreviate Stable Diffusion as SD for brevity in this figure. As diffusion-based models have achieved unprecedented success in image generation, this work mainly discusses the pioneering studies for text-to-image generation using diffusion models.

ural language, albeit with limited realism. Text-conditional GAN [4] emerged as the first fully end-to-end differential architecture extending from character-level input to pixel-level output, but was always trained on small-scale data. Autoregressive methods further utilize large-scale training data for text-to-image generation, such as DALL-E [5] from OpenAI. However, autoregressive nature makes these methods [5, 6, 7, 8] suffer from high computation costs and sequential error accumulation.

More recently, diffusion models (DMs) have emerged as the leading method in text-to-image generation [9, 1]. Figure 1 shows example images generated by the pioneering text-to-image diffusion model DALL-E2 [1], demonstrating extraordinary fidelity and imagination. However, the vast amount of research in this field makes it difficult for readers to learn the key breakthroughs without a comprehensive survey. A branch of existing surveys [10, 11, 12] reviews the progress of the diffusion model in all fields, offering a limited introduction specifically on text-to-image synthesis. Other studies [13, 11, 14] focus on text-to-image tasks using GAN-based approaches, lacking the introduction of diffusion-based methods.

To our knowledge, this is the first survey to review the progress of diffusion-based text-to-image generation. The rest of the paper is organized as follows. We also summarize the paper outline in Figure 3. Section 2 introduces the background of diffusion models. Section 3 covers pioneering studies on text-to-image diffusion models, while Section 4 discusses the follow-up advancements. Section 5 discusses the evaluation of text-to–image diffusion models from the technical and ethical perspectives. Section 6
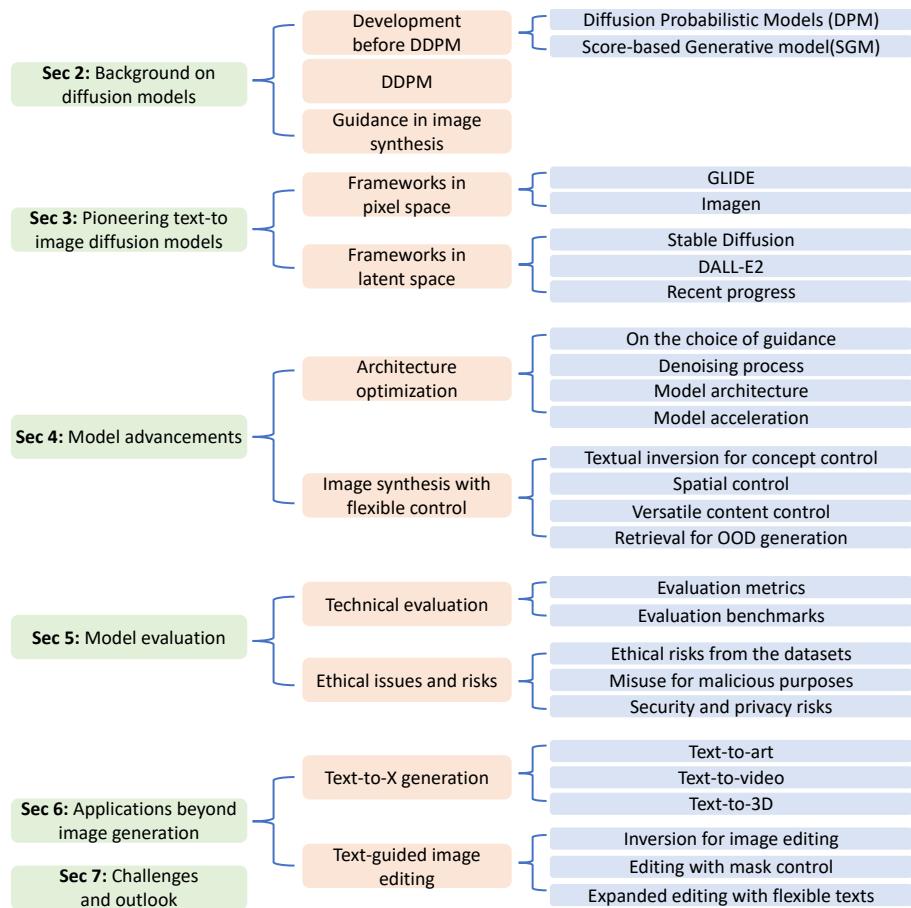
Sec 2: Background on diffusion models
- Development before DDPM
  - Diffusion Probabilistic Models (DPM)
  - Score-based Generative model(SGM)
- DDPM
- Guidance in image synthesis

Sec 3: Pioneering text-to-image diffusion models
- Frameworks in pixel space
  - GLIDE
  - Imagen
- Frameworks in latent space
  - Stable Diffusion
  - DALL-E2
  - Recent progress

Sec 4: Model advancements
- Architecture optimization
  - On the choice of guidance
  - Denoising process
  - Model architecture
  - Model acceleration
- Image synthesis with flexible control
  - Textual inversion for concept control
  - Spatial control
  - Versatile content control
  - Retrieval for OOD generation

Sec 5: Model evaluation
- Technical evaluation
  - Evaluation metrics
  - Evaluation benchmarks
- Ethical issues and risks
  - Ethical risks from the datasets
  - Misuse for malicious purposes
  - Security and privacy risks

Sec 6: Applications beyond image generation
- Text-to-X generation
  - Text-to-art
  - Text-to-video
  - Text-to-3D
- Text-guided image editing
  - Inversion for image editing
  - Editing with mask control
  - Expanded editing with flexible texts

Sec 7: Challenges and outlook

Figure 3: Paper outline. We summarize each section in this figure. Our work not only offers a comprehensive overview of text-to-image diffusion models, but also provides readers a broader perspective by discussing related areas such as text-to-X generation.

4

explores tasks beyond text-to-image generation., such as video generation and 3D object generation. Finally, we discuss challenges and future opportunities in text-to-image generation tasks.

2. Background on diffusion models

Diffusion models (DMs), also widely known as diffusion probabilistic models [15], are a family of generated models that are Markov chains trained with variational inference [16]. The learning goal of DM is to reserve a process of perturbing the data with noise, *i.e.* diffusion, for sample generation [15, 16]. As a milestone work, denoising diffusion probabilistic model (DDPM) [16] was published in 2020 and sparked an exponentially increasing interest in the community of generative models afterwards. Here, we provide a self-contained introduction to DDPM by covering the most related progress before DDPM and how unconditional DDPM works with image synthesis as a concrete example. Moreover, we summarize how guidance helps in conditional DM, which is an important foundation for understanding text-conditional DM for text-to-image.

2.1. Development before DDPM

The advent of DDPM [16] can be mainly attributed to two early attempts: score-based generative models (SGM) [17] being investigated in 2019 and diffusion probabilistic models (DPM) [15] emerging as early as in 2015. Therefore, it is important to revisit the working mechanism of DPM and SGM before we introduce DDPM.

Diffusion Probabilistic Models (DPM). DPM [15] is the first work to model probability distribution by estimating the reversal of Markov diffusion chain which maps data to a simple distribution. Specifically, DPM defines a forward (inference) process which converts a complex data distribution to a much simpler one, and then learns the mapping by reversing this diffusion process. Experimental results on multiple datasets show the effectiveness of DPM when estimating complex data distribution. DPM can be viewed as the foundation of DDPM [16], while DDPM optimizes DPM with improved implementations.
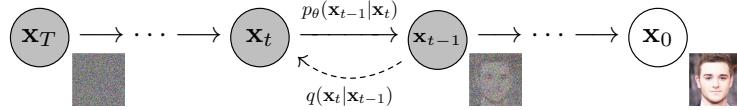
Figure 4: Diffusion process illustrated in [16]. Diffusion models include a forward pass that adds noises to a clean image, and a reverse pass that recovers the clean image from its noisy counterpart.

Score-based Generative model(SGM). Techniques for improving score-based generative models have also been investigated in [17]. SGM [17] proposes to perturb the data with random Gaussian noise of various magnitudes. With the gradient of log probability density as score function, SGM generates the samples towards decreasing noise levels and trains the model by estimating the score functions for noisy data distribution. Despite different motivations, SGM shares a similar optimization objective with DDPM during training, which is also discussed in [16] that the DDPM under a certain parameterization is equivalent to SGM during training.

$$E_{t \sim \mathcal{U}(1,T), \mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \lambda(t) \left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t) \right\|^2 \tag{1}$$

2.2. How does DDPM work for image synthesis?

Denoising diffusion probabilistic models (DDPMs) are defined as a parameterized Markov chain, which generates images from noise within finite transitions during inference. During training, the transition kernels are learned in a reversed direction of perturbing natural images with noise, where the noise is added to the data in each step and estimated as the optimization target. The diffusion processes are shown in Figure 4.

Forward pass. In the forward pass, DDPM is a Markov chain where Gaussian noise is added to data in each step until the images are destroyed. Given a data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, DDPM generates $\mathbf{x}_T$ successively with $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ [16]:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{2}$$

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I) \tag{3}$$

6

where T and β_t are the diffusion steps and hyper-parameters, respectively. We only discuss the case of Gaussian noise as transition kernels for simplicity, indicated as $\mathcal{N}$ in Eq. 3. With $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$, we can obtain noised image at arbitrary step t as follows [18]:

$$q(x_t|x_0) := \mathcal{N}(x_t;\ \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \tag{4}$$

Reverse pass. With the forward pass defined above, we can train the transition kernels with a reverse process. Starting from $p_\theta(T)$, we hope the generated $p_\theta(x_0)$ can follow the true data distribution $q(x_0)$. Therefore, the optimization objective of model is as follows(quoted from [18]):

$$E_{t\sim\mathcal{U}(1,T),\mathbf{x}_0\sim q(\mathbf{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\lambda(t)\left\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\right\|^2 \tag{5}$$

Considering the optimization objective similarities between DDPM and SGM, thy are unified in [19] from the perspective of stochastic differential equations, allowing more flexible sampling methods.

2.3. Guidance in diffusion-based image synthesis

Labels improve image synthesis. Early works on generative adversarial models (GAN) have shown that class labels can improve the quality of generated images by either providing a conditional input or guiding the image synthesis via an auxiliary classifier. These practices are also introduced to diffusion models:

Conditional diffusion model: A conditional diffusion model learns from additional information (e.g., class and text) by taking them as model input.

Guided diffusion model: During the training of a guided diffusion model, the class-induced gradients (e.g. through an auxiliary classfier) are involved in the sampling process.

Classifier-free guidance. Different from guided diffusion model, [20] found that the guidance can be obtained by generative model itself without a classifier, termed

as *classifier-free guidance*. Specifically, classifier-free guidance jointly trains a single model with unconditional score estimator $\epsilon_\theta(x)$ and conditional $\epsilon_\theta(x, c)$, where c denotes the class label. A null token $\varnothing$ is placed as the class label in the unconditional part, i.e., $\epsilon_\theta(x) = \epsilon_\theta(x, \varnothing)$. Experimental results in [20] show that classifier-free guidance achieves a trade-off between quality and diversity similar to that achieved by classifier guidance. Without resorting to a classifier, classifier-free diffusion facilitates more modalities, e.g., text in text-to-image, as guidance.

3. Pioneering text-to-image diffusion models

In this section, we introduce the pioneering text-to-image work based on diffusion models, which can be roughly categorized considering where the diffusion process is conducted, i.e., the pixel space or latent space. The first class of methods generates images directly from the high-dimensional pixel level, including GLIDE [9] and Imagen [21]. Another stream of works propose to first compress the image to a low-dimensional space, and then train the diffusion model on this latent space. Representative methods falling into the class of latent space include Stable Diffusion [2] and DALL-E 2 [1].

3.1. Frameworks in pixel space

GLIDE: the first T2I work on DM. In essence, text-to-image is text-conditioned image synthesis. Therefore, it is intuitive to replace the class label in class-conditioned DM with *text* for making the sampling generation conditioned on text. As discussed in Sec. 2.3, guided diffusion improves the photorealism of samples in conditional DM and its classifier-free variant [20] facilitates handling free-form prompts. Motivated by this, GLIDE [9] adopts classifier-free guidance in T2I by replacing original class label with text. GLIDE [9] also investigated CLIP guidance but is less preferred by human evaluators than classifier-free guidance for the sample photorealism and caption similarity. As an important component in their framework, the text encoder is set to a transformer with 24 residual blocks with a width of 2048 (roughly 1.2B parameters). Experimental results show that GLIDE [9] outperforms DALL-E [5] in both FID and human evaluation.
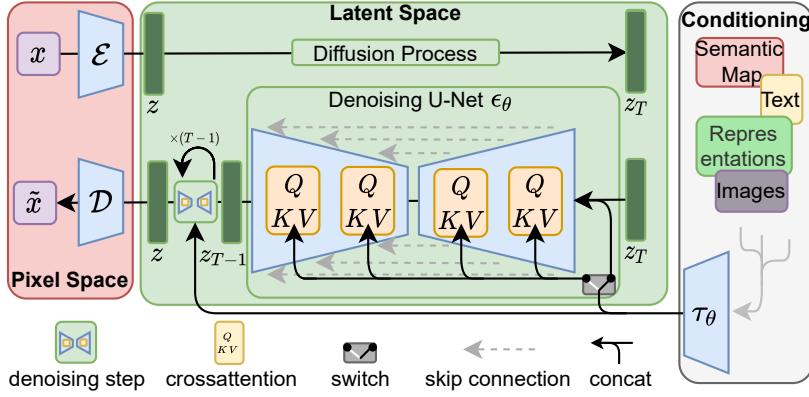
Figure 5: Model architecture of Stable Diffusion [2]. Stable Diffusion first converts the image to a latent space, where the diffusion process is performed. Stable Diffusion significantly improves the quality and efficiency of image generation compared to prior models.

Imagen: encoding text with pretrained language model. Following GLIDE [9], Imagen [21] adopts classifier-free guidance for image generation. A core difference between GLIDE and Imagen lies in their choice of text encoder. Specifically, GLIDE trains the text encoder together with the diffusion prior with paired image-text data, while Imagen [21] adopts a pretrained and frozen large language model as the text encoder. Since the text-only corpus is significantly larger than paired image-text data, such as 800GB used in T5 [22], the pretrained large language models are exposed to text with a rich and wide distribution. With different T5 [22] variants as the text encoder, [21] reveals that increasing the size of language model improves the image fidelity and image-text alignment more than enlarging the diffusion model size in Imagen. Moreover, freezing the weights of pretrained encoder facilitates offline text embedding, which reduces negligible computation burden to the online training of the text-to-image diffusion prior.

3.2. Frameworks in latent space

Stable diffusion: a milestone work on latent space. A representative framework that trains the diffusion models on latent space is Stable Diffusion, which is a scaled-up version of Latent Diffusion Model (LDM) [2]. Following Dall-E [5] that adopts a
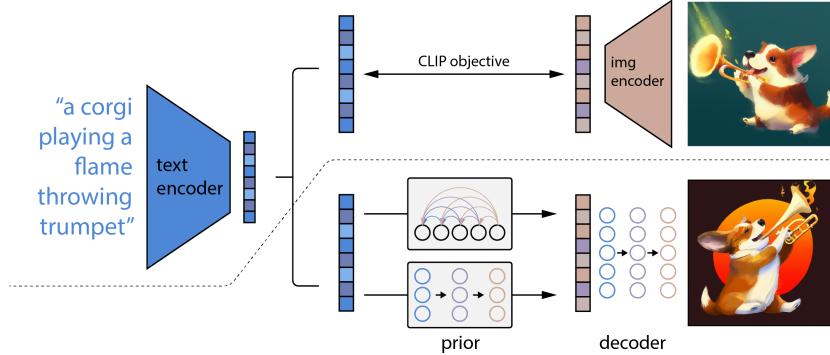
Figure 6: Model architecture of DALLE-2 [1]. DALLE-2 uses CLIP [23] model to project the image and text to latent space.

VQ-VAE to learn a visual codebook, Stable diffusion applies VQ-GAN for the latent representation in the first stage. Notably, VQ-GAN improves VQ-VAE by adding an adversarial objective to increase the naturalness of synthesized images. With the pretrained VAE, stable diffusion reverses a forward diffusion process that perturbs latent space with noise. Stable diffusion also introduces cross-attention as general-purpose conditioning for various condition signals like text. The model architecture of Stable Diffusion is shown in Figure 5. Experimental results in [2] highlight that performing diffusion modeling on the latent space significantly outperforms that on the pixel space in terms of complexity reduction and detail preservation. A similar approach has also been investigated in VQ-diffusion with a mask-then-replace diffusion strategy. Resembling the finding in pixel-space method, classifier-free guidance also significantly improves the text-to-image diffusion models in latent space [2].

DALL-E2: with multimodal latent space. Another stream of text-to-image diffusion models in latent space relies on multimodal contrasitve models [23], where image embedding and text encoding are matched in the same representation space. For example, CLIP [23] is a pioneering work learning the multimodal representations and has been widely used in numerous text-to-image models [1]. A representative work applying CLIP is DALL-E 2, also known as unCLIP [1], which adopts the CLIP text encoder but inverts the CLIP image encoder with a diffusion model that generates images from CLIP latent space. Such a combination of encoder and decoder resembles the structure

10

of VAE adopted in LDM, even though the inverting decoder is non-deterministic [1]. Therefore, the remaining task is to train a prior to bridge the gap between CLIP text and image latent space, and we term it as *text-image latent prior* for brevity. DALL-E2 [1] finds that this prior can be learned by either autoregressive method or diffusion model, but diffusion prior achieves superior performance. Moreover, experimental results show that removing this *text-image latent prior* leads to a performance drop by a large margin [1], which highlights the importance of learning the *text-image latent prior*. We show image examples generated by DALLE-2 in Figure 1.

Recent progress of Stable Diffusion and DALL-E family. Since the publication of Stable Diffusion [2], multiple versions of models have been released, including Stable Diffusion 1.4, 1.5, 2.0, 2.1, XL, and 3. Starting from Stable Diffusion 2.0 [24], a notable feature is negative prompts, which allow users to specify what they do not wish to generate in the output image. Stable Diffusion XL [25] enhances capabilities beyond previous versions by incorporating a larger Unet architecture, leading to improved abilities such as face generation, richer visuals, and more impressive aesthetics. Stable Diffusion 3 is built on diffusion transformer architecture [26] and use two separate sets of weights to model text and image modality. Stable Diffusion 3 improve overall comprehension and typography of generated images. On the other hand, the evolution of the DALL-E model has progressed from the autoregressive DALL-E [5], to the diffusion-based DALL-E2 [1], and most recently, DALLE-3 [27]. Integrated into the GPT-4 API, DALLE-3 showcases superior performance in capturing intricate nuances and details.

4. Model advancements

Numerous works attempt to improve the text-to-image diffusion models, which we roughly categorize into architecture optimization and versatile use.

4.1. Architecture optimization

On the choice of guidance. Beyond the classifier-free guidance, some works [9] have also explored cross-modal guidance with CLIP [23]. Specifically, GLIDE [9]

finds that CLIP-guidance underperforms the classifier-free variant of guidance. By contrast, another work UPainting [28] points out that lacking of a large-scale transformer language model makes these models with CLIP guidance difficult to encode text prompts and generate complex scenes with details. By combing large language model and cross-modal matching models, UPainting [28] significantly improves the sample fidelity and image-text alignment of generated images. The general image synthesis capability enables UPainting [28] to generate images in both simple and complex scenes.

Denoising process. By default, DM during inference repeats the denoising process on the same denoiser model, which makes sense for an unconditional image synthesis since the goal is only to get a high-fidelity image. In the task of text-to-image synthesis, the generated image is also required to align with the text, which implies that the denoiser model has to make a trade-off between these two goals. Specifically, two recent works [29, 30] point out a phenomenon: the early sampling stage strongly relies on the text prompt for the goal of aligning with the caption, but the later stage focuses on improving image quality while almost ignoring the text guidance. Therefore, they abort the practice of sharing model parameters during the denoising process and propose to adopt multiple denoiser models which are specialized for different generation stages. Specifically, ERNIE-ViLG 2.0 [29] also mitigates the problem of object-attribute by the guidance of a text parser and object detector, improving the fine-grained semantic control.

Model architecture. A branch of studies enhances text-to-image generation by improving the denoising model. For instance, Free-U [31] strategically re-weights the contributions sourced from the U-Net's skip connections and backbone feature maps, which improves image generation quality without additional training or fine-tuning. The pioneering work DiT [26] proposes a diffusion transformer architecture as the denoising model of diffusion models, which replaces the commonly-used U-Net backbone (see Figure 7). Pixart-α [32] is a pioneering work that adopts a transformer-based backbone and supports high-resolution image synthesis up to 1024×1024 resolution with low training cost.

Model acceleration. Diffusion models have achieved great success in image gen-
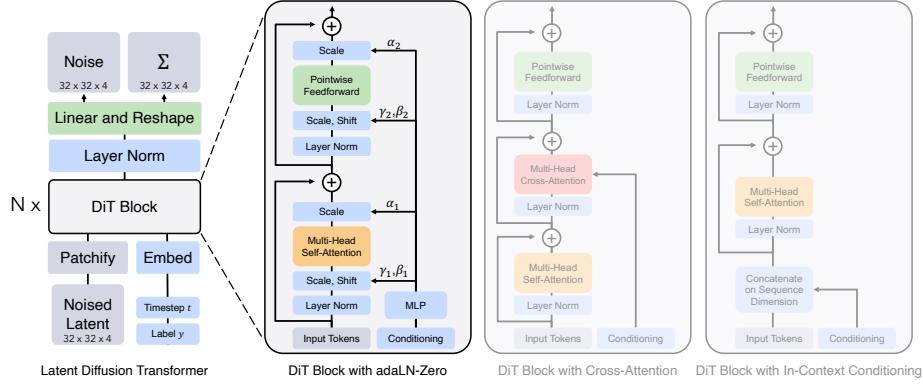
12

Figure 7: Transformer architecture of DiT [26]. DiT trains conditional latent diffusion models with transformer blocks. Adaptive layer norm works best among all block types.

eration, outperforming GAN. However, one drawback of diffusion models is their slow sampling process, which requires hundreds or thousands of iterations to generate an image. V-prediction [33] improves the sampling speed by distilling a pre-trained diffusion model with N-step DDIM sampler to a new model of N/2 sampling steps, without hurting generation quality. Flow Matching (FM) [34] finds that employing FM with diffusion paths results in a more robust and stable alternative for training diffusion models. A recent work REPresentation Alignment (REPA) [35] emphasizes the key role of representations in training large-scale diffusion models, and introduces the representations from self-supervised models (DINO v2 [36]) to the training of diffusion models like DiT [26] and SiT [37]. REPA [35] achieves significant acceleration results by speeding up SiT [37] training by over $17.5\times$.

4.2. Image synthesis with flexible control

Textual inversion for concept control. Pioneering works on text-to-image generation [9, 21, 2, 1] rely on natural language to describe the content and styles of generated images. However, there are cases when the text cannot exactly describe the desired semantics by users, e.g., generating a new subject. In order to synthesize novel scenes with certain concepts or subjects, [39, 38] introduces several reference images with the desired concepts, then inverts the reference images to the textual descriptions. Specifically, [39] inverts the shared concept in a couple of reference images into the

Figure 8: Textual inversion for concept control in Dreambooth [38]. Based on the user input images, Dreambooth [38] finetunes a pretrained model to learn the key concept of subject in input images. The users can further control the status of the subject by prompts such as "getting a haircut".

text (embedding) space, i.e. "pseudo-words". The generated "pseudo-words" can be used for personalized generation. DreamBooth [38] adopts a similar technique and mainly differs by fine-tuning (instead of freezing) the pretrained DM model for preserving key visual features from the subject identity. With the learned subject, the DreamBooth [38] allows users to control the status of subject in the generated images by specifying in the input prompt. Figure 8 shows generated images by DreamBooth with dog images as well as the prompt as model inputs. Textual inversion has also been applied in other applications, such as to control the spatial relationship in [40].

Spatial control. Despite their unprecedented high image fidelity and caption similarity, most text-to-image DMs like Imagen [21] and DALL-E2 [1] do not provide fine-grained control of spatial layout. To this end, SpaText [41] introduces spatio-textual (ST) representation which can be included to finetune a SOTA DM by adapting its decoder. Specifically, the new encoder conditions both local ST and existing global text. Therefore, the core of SpaText [41] lies in ST where the diffusion prior in trained separately to convert the image embeddings in CLIP to its text embeddings. During training, the ST is generated directly by using the CLIP image encoder taking the segmented image object as input. A concurrent work [42] proposes to realize fine-grained local control through a simple sketch image. Core to their approach is a Latent Guidance Predictor (LGP)that is a pixel-wise MLP mapping the latent feature of a noisy image to that of its corresponding sketch input. After being trained (see [42] for more training details), the LGP can be deployed to the pretrained text-to-image DM with-

14

"As the _aurora_ lights up the sky, a herd of _reindeer_ leisurely wanders on the grassy _meadow_, admiring the breathtaking view, a serene _lake_ quietly reflects the magnificent display, and in the distance, a snow-capped _mountain_ stands majestically, fantasy, 8k, highly detailed"
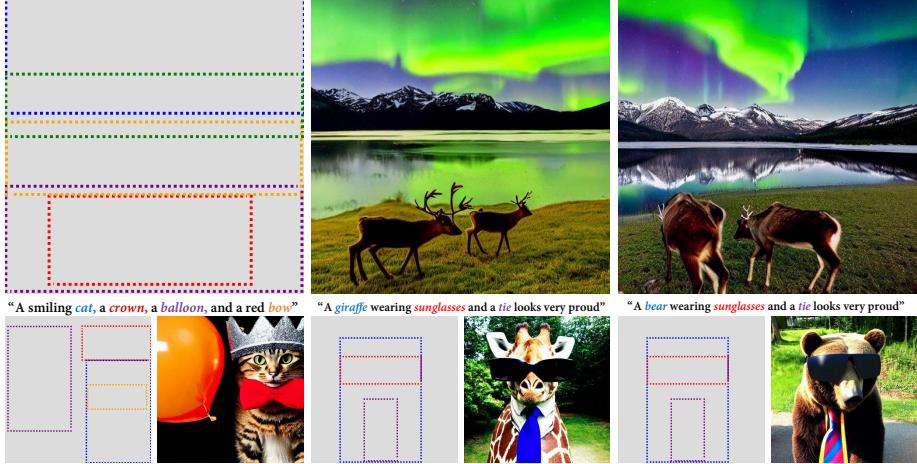
"A smiling _cat_, a _crown_, a _balloon_, and a red _bow_" "A _giraffe_ wearing _sunglasses_ and a _tie_ looks very proud" "A _bear_ wearing _sunglasses_ and a _tie_ looks very proud"

Figure 9: Spatial control in BoxDiff [44]. BoxDiff enables to control the layout of generated images with provided boxes or scribbles.

out the need for fine-tuning. Other representative studies for spatial control includes BoxDiff [43] that uses the provided box or scribble to control the layout of generated images, as shown in Figure 9.

Versatile content control. ControlNet [44] has achieved great attention due to its powerful ability to add various conditioning controls to large pretrained models. ControlNet [44] reuses the pretrained encoding layers as a strong backbone, and a zero convolution architecture is proposed to ensure no harm noise could affect the finetuning. ControlNet [44] achieves outstanding results with various conditioning signals, such as edges, depth, and segmentation. Figure 10 shows an example from [44] that use canny edge and human as condition to control the image generation of Stable Diffusion model. There are also other widely used methods unifies various signals in one model for content control, such as T2i-adapter [45], Uni-ControlNet [46], GLIGEN [47], and Composer [48]. HumanSD [49] and HyperHuman [50] focuses on the generation of human images by taking human skeleton as model inputs.

Retrieval for out-of-distribution generation. State-of-the-art text-to-image models assume sufficient exposure to descriptions of common entities and styles from
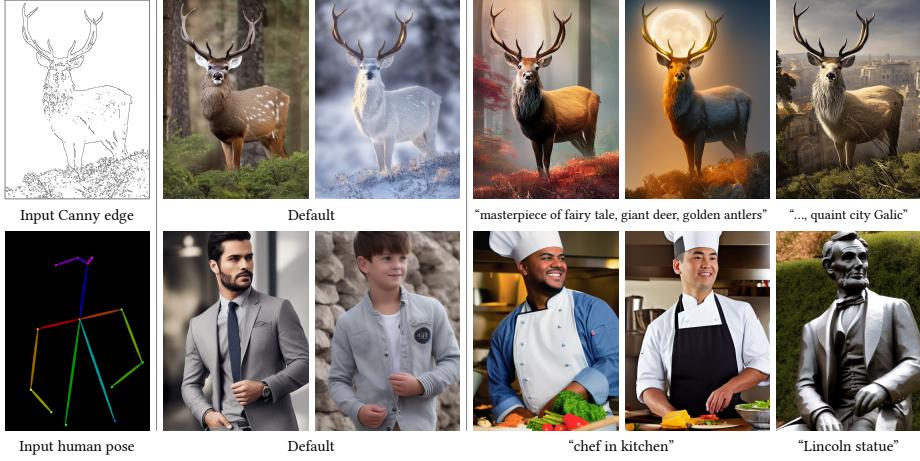
15

Figure 10: Control Stable Diffusion with conditions [44]. ControlNet [44] allows users to specify conditions, such as canny edges, in image generation of large-scale pretrained diffusion models. For example, the default prompt is "a high-quality, detailed, and professional image", while the users can add conditions such as "quaint city Galic".

training. This assumption breaks down with rare entities or vastly different styles, leading to performance drops. To counter this, several studies [51, 52, 53, 54] use external databases for retrieval, a semi-parametric approach adapted from NLP [55, 56] and GAN-based synthesis [57]. Retrieval-Augmented Diffusion Models (RDMs) [51] use k-nearest neighbors (KNN) based on CLIP distance for enhanced diffusion guidance, while KNN-diffusion [52] improves quality by adding text embeddings. Re-Imagen [54] refines this with a single-stage framework, retrieving both images and text in latent space, outperforming KNN-diffusion on the COCO benchmark.

5. Model evaluation

5.1. Technical evaluation

Evaluation metrics. A common metric to evaluate image quality quantitatively is Fréchet Inception Distance (FID), which measures the Fréchet distance (also known as Wasserstein-2 distance) between synthetic and real-world images. We summarize the evaluation results of representative methods on MS-COCO dataset in Table 1 for

reference. The smaller the FID is, the higher the image fidelity. To measure the text-image alignment, CLIP scores are widely applied, which trades off against FID. There are also other metrics for text-to-image evaluation, including Inception score (IS) [58] for image quality and R-precision for text-to-image generation.

Evaluation benchmarks. Apart from the automatic metrics discussed above, multiple works involve human evaluation and propose their new evaluation benchmarks [60, 21, 8, 28, 61, 54, 62]. We summarize representative benchmarks in Table 2. For a better evaluation of fidelity and text-image alignment, DrawBench[21], PartiPropts [8] and UniBench [28] ask the human raters to compare generated images from different models. Specifically,

Table 1: Image quality comparison of autoregressive and diffusion models. Diffusion models outperforms autoregressive models in image quality, with lower FID on MS-COCO dataset.

	model	FID ($\downarrow$)
Autoregressive	CogView [6]	27.10
	LAFITE [59]	26.94
	DALLE [5]	17.89
Diffusion models	GLIDE [9]	12.24
	Imagen [21]	7.27
	Stable Diffusion [2]	12.63
	DALL-E 2 [1]	10.39
	Upainting [28]	8.34
	ERNIE-ViLG 2.0 [29]	6.75
	eDiff-I [30]	6.95

UniBench [28] proposes to evaluate the model on both simple and complex scenes and includes both Chinese and English prompts. PartiPropts [8] introduces a diverse set of over 1600 (English) prompts and also proposes a challenge dimension that highlights why this prompt is difficult. To evaluate the model from various aspects, PaintSKills [60] evaluates the *visual reasoning skills* and *social biases* apart from image quality and text-image alignment. However, PaintSKills [60] only focuses on unseen object-color and object-shape scenario [28]. EntityDrawBench [54] further evaluates the model with various infrequent entities in different scenes. Compared to PartiPropts [8] with prompts at different difficulty levels, Multi-Task Benchmark [61] proposes thirty-two tasks that evaluate different capabilities and divides each task into three difficulty levels.

5.2. Ethical issues and risks

Ethical risks from the datasets. Text-to-image generation is a highly data-driven task, and thus models trained on large-scale unfiltered data may suffer from even rein-

Table 2: Benchmarks for text-to-image generation task.

Benchmark	Measurement	Metric	Auto-eval	Human-eval	Language
DrawBench[21]	Fidelity, alignment	User preference rates	N	Y	English
UniBench [28]	Fidelity, alignment	User preference rates	N	Y	English, Chinese
PartiPrompts [8]	Fidelity, alignment	Qualitative	N	Y	English
PaintSKills [60]	Visual reasoning skills, social biases	Statistics	Y	Y	English
EntityDrawBench [54]	Entity-centric faithfulness	Human rating	N	Y	English
Multi-Task Benchmark [61]	Various capabilities	Human rating	N	Y	English

force the biases from the dataset, leading to ethical risks. [63] finds a large amount of inappropriate content in the generated images by Stable diffusion [2] (e.g., offensive, insulting, or threatening information), and first establishes a new test bed to evaluate them. Moreover, it proposes Safe Latent Diffusion, which successfully removes and suppresses inappropriate content with additional guidance. Another ethical issue, the fairness of social group, is studied in [64, 65]. Specifically, [64] finds that simple homoglyph replacements in the text descriptions can induce culture bias in models, i.e., generating images from different cultures. [65] introduce an Ethical NaTural Language Interventions in Text-to-Image GENeration (ENTIGEN) benchmark dataset, which can evaluate the change of generated images with ethical interventions by three axes: gender, skin color, and culture. With intervened text prompts, [65] improves diffusion models (e.g., Stable diffusion [2]) from the social diversity perspective. Fair Diffusion [66] evaluates the fairness problem of diffusion models and mitigates this problem at the deployment stage of diffusion models. Specifically, Fair Diffusion [66] instructs the diffusion models on fairness with textual guidance. Another work [67] finds that a broad range of prompts to text-to-image diffusion models could produce stereotypes, such as simply mentioning traits, descriptors, occupations, or objects.

Misuse for malicious purposes. Text-to-image diffusion models have shown their power in generating high-quality images. However, this also raises great concern that the generated images may be used for malicious purposes, e.g., falsifying electronic evidence [68]. DE-FAKE [68] is the first to conduct a systematical study on visual forgeries of the text-to-image diffusion models, which aims to distinguish generated images from the real ones, and also further track the source model of each fake image. To achieve these two goals, DE-FAKE [68] analyzes from visual modality perspective,

and finds that images generated by different diffusion models share common features and also present unique model-wise fingerprints. Two concurrent works [69, 70] approach the detection of faked images both by evaluating the existing detection methods on images generated by the diffusion model, and also analyzing the frequency discrepancy of images by GAN and diffusion models. [69, 70] find that the performance of existing detection methods drops significantly on generated images by diffusion models compared to GAN. Moreover, [69] attributes the failure of existing methods to the mismatch of high frequencies between images generated by diffusion models and GAN. Another work [71] discusses the concern of artistic image generation from the perspective of artists. Although agreeing that the artistic image generation may be a promising modality for the development of art, [71] points out that the artistic image generation may cause plagiarism and profit shifting (profits in the art market shift from artists to model owners) problems if not properly used.

Security and privacy risks. While text-to-image diffusion models have attracted great attention, the security and privacy risks have been neglected so far. Two pioneering works [72, 73] discuss the backdoor attack and privacy issues, respectively. Inspired by the findings in [64] that a simple word replacement can invert culture bias to models, Rickrolling the Artist [72] proposes to inject the backdoors into the pre-trained text encoders, which will force the generated image to follow a specific description or include certain attributes if the trigger exists in the text prompt. [73] is the first to analyze the membership leakage problem in text-to-image generation models, where whether a certain image is used to train the target text-to-image model is inferred. Specifically, [73] proposes three intuitions on the membership information and four attack methods accordingly. Experiments show that all the proposed attack methods achieve impressive results, highlighting the threat of membership leakage.

6. Applications beyond image generation

The advancement of diffusion models has inspired various applications beyond image generation, such as text-to-X where X refers to a modality such as video, and text-guided image editing. We introduce pioneering work as follows.
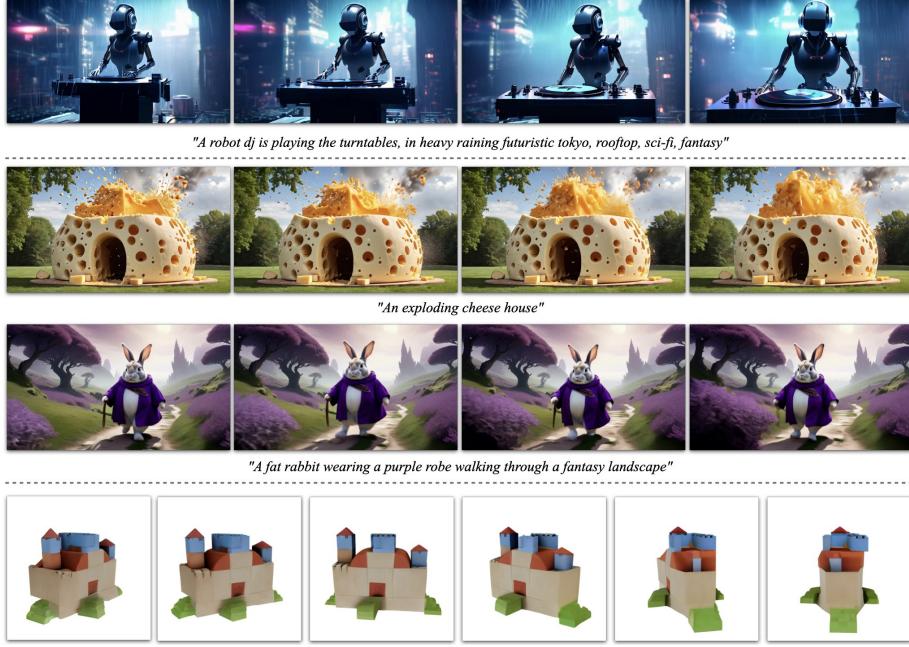
19

"A robot dj is playing the turntables, in heavy raining futuristic tokyo, rooftop, sci-fi, fantasy"

"An exploding cheese house"

"A fat rabbit wearing a purple robe walking through a fantasy landscape"

Figure 11: Text-to-video generation by Stable Video Diffusion [74] from Stable AI.

6.1. Text-to-X generation

6.1.1. Text-to-art

Artistic painting is an interesting and imaginative area that benefits from the success of generative models. Despite the progress of GAN-based painting[75], they suffer from the unstable training and model collapse problem brought by GAN. Recently, multiple works have presented impressive images of paintings based on diffusion models, investigating improved prompts and different scenes. Multimodal guided artwork diffusion (MGAD) [76] refines the generative process of diffusion model with multimodal guidance (text and image) and achieves excellent results regarding both the diversity and quality of generated digital artworks. In order to maintain the global content of the input image, DiffStyler [77] proposes a controllable dual diffusion model with learnable noise in the diffusion process of the content image. During inference, explicit content and abstract aesthetics can both be learned with two diffusion models. Experimental results show that DiffStyler [77] achieve excellent results on both quan-

titative metrics and manual evaluation. To improve the creativity of Stable Diffusion model, [78] proposes two directions of textual condition extension and model retraining with the Wikiart dataset, enabling the users to ask the famous artists to draw novel images. [79] personalizes text-to-image generation by customizing the aesthetic styles with a set of images, while [80] extends generated images to Scalable Vector Graphics (SVGs) for digital icons or arts. In order to improve computation efficiency, [53] proposes to generate artistic images based on retrieval-augmented diffusion models. By retrieving neighbors from specialized datasets (e.g., Wikiart), [53] obtains fine-grained control of the image style. In order to specify more fine-grained style features (e.g., color distribution and brush strokes), [81] proposes supervised style guidance and self-style guidance method, which can generate images of more diverse styles.

6.1.2. Text-to-video

Early studies. Since video is just a sequence of images, a natural application of text-to-image is to make a video conditioned on the text input. Conceptually, text-to-video DM lies in the intersection between text-to-image DM and video DM. Make-A-Video [82] adopts a pretrained text-to-image DM to text-to-video and Video Imagen [83] extends an existing video DM method to text-to-video. Other representative text-to-video diffusion models include ModelScope [84], Tune-A-Video [85], and VideoCrafter [86, 87, 88]. The success of text-to-video naturally inspires a future direction of movie generation based on text inputs. Different from general text-to-video tasks, story visualization requires the model to *reason* at each frame about whether to maintain the consistency of actors and backgrounds between frames or scenes, based on the story progress [89]. Make-A-Story [89] uses an autoregressive diffusion-based framework and visual memory module to maintain consistency of actors and backgrounds across frames, while AR-LDM [90] leverages image-caption history for coherent frame generation. Moreover, AR-LDM [90] shows the consistency for unseen characters, and also the ability for real-world story synthesis on a newly introduced dataset VIST [91].

Recent process. More recently, Stable Video Diffusion [74] from Stable AI achieves significant performance improvements for text-to-video and image-to-video genera-
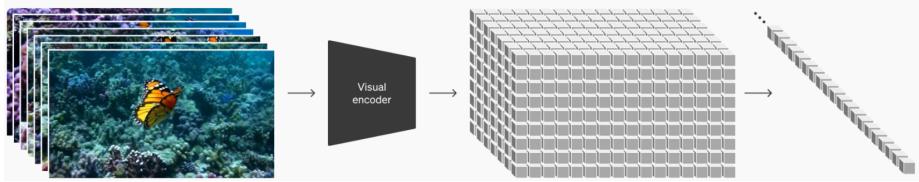
Figure 12: Sora [92] from Open AI. Sora represents video frames by compressing them to image patches with a transformer-base backbone.

tion. It identifies and evaluates different training stages of applying diffusion models to video synthesis, and introduces a systematic training process including the captioning and filtering strategies. OpenAI launches the state-of-the-art video generation model Sora [92], which is capable of generating a minute of high-fidelity video. Inspired by large language models that turn different data (like text and code) into tokens, Sora [92] first unifies diverse types of videos and images as patches and compresses them to a lower-dimensional latent space, as shown in Figure 12. Sora then decoposes the representations into spacetime patches and performs the diffusion process based on the transformer backbone. As Sora [92] is not open-sourced yet, some studies aim to provide open access to advanced video generation models, such as Open-Sora [93] and Open-Sora-Plan [94].

6.1.3. Text-to-3D

3D object generation is evidently much more sophisticated than 2D image synthesis task. DeepFusion [95] is the first work that successfully applies diffusText-guided creative generation models to 3D object synthesis. Inspired by Dream Fields [96] which applies 2D image-text models (i.e., CLIP) for 3D synthesis, DeepFusion [95] trains a randomly initialized NeRF [97] with the distillation of a pretrained 2D diffusion model (i.e., Imagen). However, according to Magic3D [98], the low-resolution image supervision and extremely slow optimization of NeRF result in low-quality generation and long processing time of DeepFusion [95]. For higher-resolution results, Magic3D [98] proposes a coarse-to-fine optimization approach with coarse representation as initialization as the first step, and optimizing mesh representations with high-resolution diffusion priors. Magic3D [98] also accelerates the generation process with a sparse

3D hash grid structure. 3DDesigner [99] focuses on another topic of 3D object genera-
tion, *consistency*, which indicates the cross-view correspondence. With low-resolution
results from NeRF-based condition module as the prior, a two-stream asynchronous
diffusion module further enhances the consistency and achieves 360-degree consistent
results. Apart from 3D object generation from text, recent work Zero-1-to-3 [100]
has achieved great attention by enabling zero-shot novel view synthesis and 3D recon-
struction from a single image, inspiring various follow-up studies such as Vivid-1-to-
3 [101].

6.2. Text-guided image editing

Diffusion models not only significantly improve the quality of text-to-image syn-
thesis, but also enhance text-guided image editing. Before DM gained popularity, zero-
shot image editing had been dominated by GAN inversion methods [102, 103, 104, 105,
106, 107] combined with CLIP. However, GAN is often constrained to have limited in-
version capability, causing unintended changes to the image content. In this section,
we discuss pioneering studies for image editing based on diffusion models.

Inversion for image editing. A branch of studies edits the images by modifying
the noisy signals in the diffusion process. SDEdit[108] is a pioneering work that ed-
its images by iteratively denoising through a stochastic differential equation (SDE).
Without any task-specific training, SDEdit[108] first add noises to the input (such as
stroke painting), then subsequently denoises the noisy image through the SDE prior
to increase image realism. DiffusionCLIP [109] further adds text control in the edit-
ing process by fine-tuning the diffusion model at the reverse DDIM [110] process with
CLIP-based loss. Due to the local linearization assumptions, DDIM may lead to incor-
rect image reconstruction with the error propagation [111]. To mitigate this problem,
Exact Diffusion Inversion via Coupled Transformations (EDICT) [111] proposes to
maintain two coupled noise vectors in the diffusion process and achieves higher recon-
struction quality than DDIM [110]. Another work [112] introduces an accurate inver-
sion technique for text-guided editing by pivotal inversion and null-text optimization,
showing high-fidelity editing of various real images. To improve the editing efficiency,
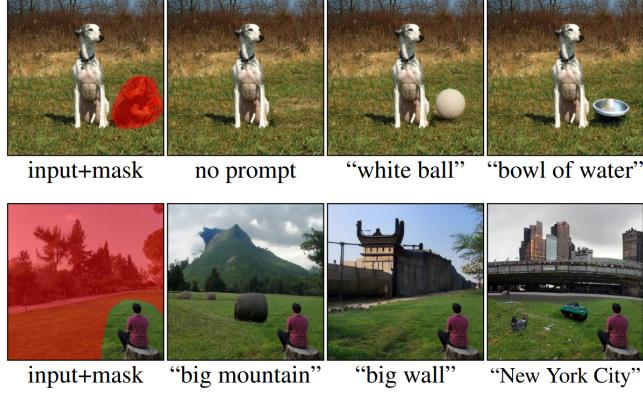LEDITS++ [113] proposes a novel inversion approach without tuning and optimiza-

Figure 13: Edit images by mask control in Blended Diffusion [114].

tion, which could produce high-fidelity results with a few diffusion steps.

Editing with mask control. A branch of work manipulates the image mainly on a local (masked) region [114], as shown in Figure 13. The difficulty lies in guaranteeing a seamless coherence between the masked region and the background. To guarantee the seamless coherence between the edited region and the remaining part, Blended diffusion [114] spatially blends noisy image with the local text-guided diffusion latent in a progressive manner. This approached is further improved for a blended latent diffusion model in [115] and a multi-stage variant in [116]. Different from [114, 115, 116] that requires a manually designed mask, DiffEdit [117] proposes to automatically generate the mask to indicate which part to be edited.

Expanded editing with flexible texts. Some studies enable more types of image editing with flexible text inputs. Imagic [118] is the first to perform text-based semantic edits to a single image, such as postures or composition of multiple objects. Specifically, Imagic [118] first obtains an optimized embedding for the target text, then linearly interpolates between the target text embedding and the optimized one. This generated representation is then sent to the fine-tuned model and generates the edited images. To solve the problem that a simple modification of text prompt may leads to a different output, Prompt-to-Prompt [119] proposes to use a cross-attention map during the diffusion progress, which represents the relation between each image pixel and word in the text prompt. InstructPix2Pix [120] works on the task of editing the image
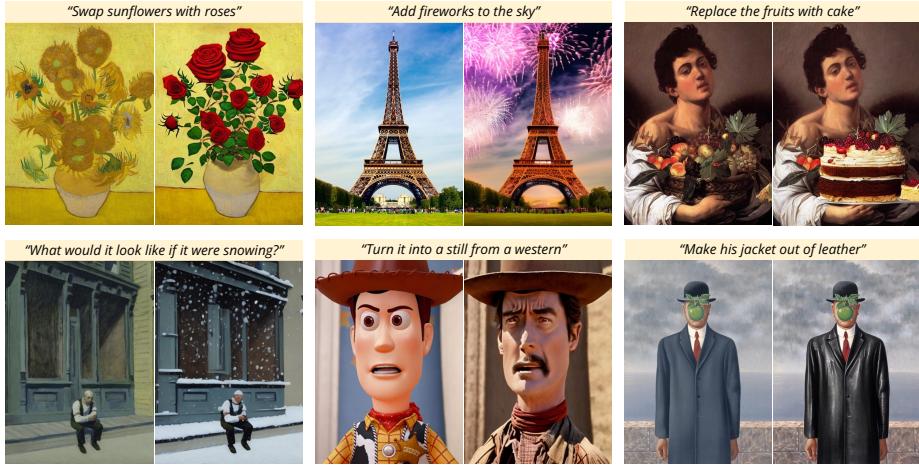
Figure 14: Image editing by InstructPix2Pix [120]. InstructPix2Pix [120] allows the users to edit an existing image by simply giving textural instructions, such as "Swap sunflowers with roses".

with human-written instructions. Based on a large model (GPT-3) and a text-to-image model (Stable diffusion), [120] first generates a dataset for this new task, and trains a conditional diffusion model InstructPix2Pix which generalizes to real images well. However, it is admitted in [120] there still remain some limitations, e.g., the visual quality of generated datasets limits the model.

There are also other interesting tasks regarding image editing. Paint by example [121] proposes a semantic image composition problem reference image is semantically transformed and harmonized before blending into another image [121]. MagicMix [122] proposes a new task called semantic mixing, which blends two different semantics (e.g.,corgi and coffee machine) to create a new concept (corgi-like coffee machine). SEGA [123] allows for subtle and extensive image edits by instructing the diffusion models with semantic control. Specifically, it interacts with the diffusion process to flexibly steer it along semantic directions.

7. Challenges and outlook

7.1. Challenges

Challenges on ethical issues and dataset bias. Text-to-image models trained on large-scale unfiltered data may suffer from even reinforce the biases from the training dataset, leading to the generation of inappropriate (e.g., offensive, insulting, or threatening information) [124] or unfair content regarding social groups [64, 65]. Moreover, the current models predominantly adopt English as the default language for input text. This might further put those people who do not understand English in an unfavored situation [125, 126].

Challenges on security risks. As the diffusion models improve, it is becoming more difficult to detect generated images from the real ones. This brings security risks since the generated images may be used for malicious purposes like falsifying electronic evidence [68]. Moreover, the diffusion models suffer from security risks, such as backdoor attack [72] and privacy issues [73].

Challenges on data and computation. As widely recognized, the success of deep learning heavily depends on the scaled training data and huge computation resources. In the context of text-to-image DM, this is especially true. For example, the major frameworks, such as Stable Diffusion [2] and DALL-E 2 [1], are all trained with hundreds of millions of image-text pairs. Moreover, the computation overhead is so large that it renders the opportunities to train such a model from scratch to large companies, such as OpenAI [1] and Google [21]. Despite the advancements to accelerate the training and inference of diffusion models, it is still challenging to effectively adopt the diffusion models in efficiency-oriented environments, such as edge devices.

7.2. Outlook

Safe and fair applications. With the wide application of text-to-image models, how to mitigate the ethical issues and security risks of current text-to-image models is demanding and challenging. Possible directions include a more diverse and balanced dataset to mitigate issues like race and gender, advanced methods for detecting generated images, and robust diffusion models against various attacks.

Unified multi-modality framework. Text-to-image generation can be seen as part of the multi-modality learning. Most works focus on the single task of text-to-image generation, but unifying multiple task into a single model can be a promising trend. For example, UniD3 [127] unify text-to-image generation and image captioning with a single diffusion model. The unified multi-modality model can boost each task by learning representations from each modality better, and may bring more inspirations of how model understands the multi-modality data.

Collaboration with other fields. In the past few years, deep learning has made great progress in multiple areas, such as multi-modal GPT-4 [128]. Prior studies have investigated how to collaborate diffusion models with models from other fields, such as the recent work that deconstructs a diffusion model to an autoencoder [129] and the adoption of GPT-3 [130] in InstructPix2Pix [120]. There are also studies applying diffusion models in vision applications, such as image restoration [131], depth estimation [132, 133], image enhancement [134] and classification [135]. Further collaborations between text-to-image diffusion model and recent findings in active research fields are an exciting topic to be explored.

References

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

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[3] E. Mansimov, E. Parisotto *et al.*, Generating images from captions with attention, ICLR (2016).

[4] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: ICML, PMLR, 2016, pp. 1060–1069.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: ICML, 2021.

[6] M. Ding, Z. Yang *et al.*, Cogview: Mastering text-to-image generation via trans-formers, NeurIPS 34 (2021) 19822–19835.

[7] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, N. Duan, Nüwa: Visual synthesis pre-training for neural visual world creation, in: European Conference on Computer Vision, Springer, 2022, pp. 720–736.

[8] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al., Scaling autoregressive models for content-rich text-to-image generation, arXiv:2206.10789 (2022).

[9] A. Nichol, P. Dhariwal *et al.*, Glide: Towards photorealistic image generation and editing with text-guided diffusion models, ICML (2022).

[10] F.-A. Croitoru, V. Hondru *et al.*, Diffusion models in vision: A survey, IEEE TPAMI (2023).

[11] A. Ulhaq, N. Akhtar, G. Pogrebna, Efficient diffusion models for vision: A survey, arXiv:2210.09292 (2022).

[12] H. Cao, C. Tan *et al.*, A survey on generative diffusion model, arXiv:2209.02646 (2022).

[13] S. Frolov, T. Hinz *et al.*, Adversarial text-to-image synthesis: A review, Neural Networks 144 (2021) 187–209.

[14] R. Zhou, C. Jiang, Q. Xu, A survey on generative adversarial network-based text-to-image synthesis, Neurocomputing 451 (2021) 316–336.

[15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsuper-vised learning using nonequilibrium thermodynamics, in: ICML, PMLR, 2015, pp. 2256–2265.

[16] J. Ho, A. Jain *et al.*, Denoising diffusion probabilistic models, NeurIPS 33 (2020) 6840–6851.

[17] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution, in: NeurIPS, volume 32, 2019.

[18] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, arXiv:2209.00796 (2022).

[19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, in: ICLR, 2020.

[20] J. Ho, T. Salimans, Classifier-free diffusion guidance, arXiv:2207.12598 (2022).

[21] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, NeurIPS 35 (2022) 36479–36494.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer., J. Mach. Learn. Res. 21 (2020) 1–67.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021.

[24] StableAI, Stable diffusion 2.0 release, 2022. URL: `https://stability.ai/news/stable-diffusion-v2-release`.

[25] StableAI, Stable diffusion v2.1, 2023. URL: `https://stablediffusionxl.com/`.

[26] W. Peebles, S. Xie, Scalable diffusion models with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4195–4205.

[27] OpenAI, Dall·e 3, 2023. URL: `https://openai.com/dall-e-3`.

[28] W. Li, X. Xu *et al.*, Upainting: Unified text-to-image diffusion generation with cross-modal guidance, arXiv:2210.16031 (2022).

[29] Z. Feng, Z. Zhang *et al.*, Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 10135–10145.

[30] Y. Balaji, S. Nah *et al.*, ediffi: Text-to-image diffusion models with an ensemble of expert denoisers, arXiv:2211.01324 (2022).

[31] C. Si, Z. Huang, Y. Jiang, Z. Liu, Freeu: Free lunch in diffusion u-net, in: CVPR, 2024, pp. 4733–4743.

[32] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, et al., Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, arXiv preprint arXiv:2310.00426 (2023).

[33] T. Salimans, J. Ho, Progressive distillation for fast sampling of diffusion models, in: ICLR, 2021.

[34] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, M. Le, Flow matching for generative modeling, in: The Eleventh International Conference on Learning Representations, ????

[35] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, S. Xie, Representation alignment for generation: Training diffusion transformers is easier than you think, arXiv preprint arXiv:2410.06940 (2024).

[36] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, Transactions on Machine Learning Research Journal (2024) 1–31.

[37] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, S. Xie, Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, arXiv preprint arXiv:2401.08740 (2024).

[38] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 22500–22510.

[39] R. Gal, Y. Alaluf *et al.*, An image is worth one word: Personalizing text-to-image generation using textual inversion, arXiv:2208.01618 (2022).

[40] Z. Huang, T. Wu *et al.*, Reversion: Diffusion-based relation inversion from images, arXiv:2303.13495 (2023).

[41] O. Avrahami, T. Hayes *et al.*, Spatext: Spatio-textual representation for controllable image generation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 18370–18380.

[42] A. Voynov, K. Aberman, D. Cohen-Or, Sketch-guided text-to-image diffusion models, arXiv:2211.13752 (2022).

[43] J. Xie, Y. Li, Y. Huang, H. Liu, W. Zhang, Y. Zheng, M. Z. Shou, Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7452–7461.

[44] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.

[45] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 4296–4304.

[46] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, K.-Y. K. Wong, Unicontrolnet: All-in-one control to text-to-image diffusion models, Advances in Neural Information Processing Systems 36 (2024).

[47] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, Y. J. Lee, Gligen: Open-set grounded text-to-image generation, in: CVPR, 2023, pp. 22511–22521.

[48] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, J. Zhou, Composer: Creative and controllable image synthesis with composable conditions, arXiv preprint arXiv:2302.09778 (2023).

[49] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, Q. Xu, Humansd: A native skeleton-guided diffusion model for human image generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15988–15998.

[50] X. Liu, J. Ren, A. Siarohin, I. Skorokhodov, Y. Li, D. Lin, X. Liu, Z. Liu, S. Tulyakov, Hyperhuman: Hyper-realistic human generation with latent structural diffusion, in: The Twelfth International Conference on Learning Representations, 2024. URL: `https://openreview.net/forum?id=duyA42HlCK`.

[51] A. Blattmann, R. Rombach, K. Oktay, B. Ommer, Retrieval-augmented diffusion models, arXiv:2204.11824 (2022).

[52] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, Y. Taigman, Knn-diffusion: Image generation via large-scale retrieval, arXiv:2204.02849 (2022).

[53] R. Rombach, A. Blattmann, B. Ommer, Text-guided synthesis of artistic images with retrieval-augmented diffusion models, arXiv:2207.13038 (2022).

[54] W. Chen, H. Hu, C. Saharia, W. W. Cohen, Re-imagen: Retrieval-augmented text-to-image generator, arXiv:2209.14491 (2022).

[55] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, M. Lewis, Generalization through memorization: Nearest neighbor language models, arXiv:1911.00172 (2019).

[56] K. Guu, K. Lee, Z. Tung, P. Pasupat, M. Chang, Retrieval augmented language model pre-training, in: International Conference on Machine Learning, PMLR, 2020, pp. 3929–3938.

[57] B. Li, P. H. Torr, T. Lukasiewicz, Memory-driven text-to-image generation, arXiv:2208.07022 (2022).

[58] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, NeurIPS (2016).

[59] Y. Zhou, R. e. a. Zhang, Towards language-free training for text-to-image generation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2022, pp. 17907–17917.

[60] J. Cho, A. Zala, M. Bansal, Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers, arXiv:2202.04053 (2022).

[61] V. Petsiuk, A. E. Siemenn, S. Surbehera, Z. Chin, K. Tyser, G. Hunter, A. Raghavan, Y. Hicke, B. A. Plummer, O. Kerret, et al., Human evaluation of text-to-image models on a multi-task benchmark, arXiv:2211.12112 (2022).

[62] P. Liao, X. Li, X. Liu, K. Keutzer, The artbench dataset: Benchmarking generative models with artworks, arXiv:2206.11404 (2022).

[63] P. Schramowski, M. Brack, B. Deiseroth, K. Kersting, Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models, arXiv preprint arXiv:2211.05105 (2022).

[64] L. Struppek, D. Hintersdorf, K. Kersting, The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models, arXiv:2209.08891 (2022).

[65] H. Bansal, D. Yin *et al.*, How well can text-to-image generative models understand ethical natural language interventions?, arXiv:2210.15230 (2022).

[66] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, K. Kersting, Fair diffusion: Instructing text-to-image generation models on fairness, arXiv preprint arXiv:2302.10893 (2023).

[67] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, A. Caliskan, Easily accessible text-to-image

generation amplifies demographic stereotypes at large scale, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1493–1504.

[68] Z. Sha, Z. Li, N. Yu, Y. Zhang, De-fake: Detection and attribution of fake images generated by text-to-image generation models, in: ACM SIGSAC CCS, 2023, pp. 3418–3432.

[69] J. Ricker, S. Damm, T. Holz, A. Fischer, Towards the detection of diffusion model deepfakes, arXiv:2210.14571 (2022).

[70] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[71] A. Ghosh, G. Fossas, Can there be art without an artist?, arXiv:2209.07667 (2022).

[72] L. Struppek, D. Hintersdorf, K. Kersting, Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models, arXiv:2211.02408 (2022).

[73] Y. Wu, N. Yu, Z. Li, M. Backes, Y. Zhang, Membership inference attacks against text-to-image generation models, arXiv:2210.00968 (2022).

[74] A. Blattmann, T. Dockhorn *et al.*, Stable video diffusion: Scaling latent video diffusion models to large datasets, arXiv preprint arXiv:2311.15127 (2023).

[75] A. Jabbar, X. Li, B. Omar, A survey on generative adversarial networks: Variants, applications, and training, ACM computing surveys 54 (2021) 1–49.

[76] N. Huang, F. Tang, W. Dong, C. Xu, Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion, in: ACM Multimedia, 2022, pp. 1085–1094.

[77] N. Huang, Y. Zhang, F. Tang, C. Ma, H. Huang, W. Dong, C. Xu, Diffstyler: Controllable dual diffusion for text-driven image stylization, IEEE Transactions on Neural Networks and Learning Systems (2024).

[78] X. Wu, Creative painting with latent diffusion models, arXiv:2209.14697 (2022).

[79] V. Gallego, Personalizing text-to-image generation via aesthetic gradients, arXiv:2209.12330 (2022).

[80] A. Jain, A. Xie, P. Abbeel, Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 1911–1920.

[81] Z. Pan, X. Zhou, H. Tian, Arbitrary style guidance for enhanced diffusion-based text-to-image generation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4461–4471.

[82] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al., Make-a-video: Text-to-video generation without text-video data, arXiv:2209.14792 (2022).

[83] J. Ho, W. Chan *et al.*, Imagen video: High definition video generation with diffusion models, arXiv:2210.02303 (2022).

[84] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, S. Zhang, Modelscope text-to-video technical report, arXiv preprint arXiv:2308.06571 (2023).

[85] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, M. Z. Shou, Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7623–7633.

[86] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C. Weng, Y. Shan, Videocrafter1: Open diffusion models for high-quality video generation, 2023. `arXiv:2310.19512`.

[87] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, Y. Shan, Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. `arXiv:2401.09047`.

[88] J. Xing, M. Xia, Y. Zhang, H. Chen, X. Wang, T.-T. Wong, Y. Shan, Dynamicrafter: Animating open-domain images with video diffusion priors (2023). `arXiv:2310.12190`.

[89] T. Rahman, H.-Y. Lee, J. Ren, S. Tulyakov, S. Mahajan, L. Sigal, Make-a-story: Visual memory conditioned consistent story generation, arXiv:2211.13319 (2022).

[90] X. Pan, P. Qin, Y. Li, H. Xue, W. Chen, Synthesizing coherent story with autoregressive latent diffusion models, arXiv:2211.10950 (2022).

[91] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, et al., Visual storytelling, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2016, pp. 1233–1239.

[92] OpenAI, Video generation models as world simulators, 2024. URL: `https://openai.com/index/video-generation-models-as-world\ -simulators/`.

[93] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, Y. You, Opensora: Democratizing efficient video production for all, 2024. URL: `https:// github.com/hpcaitech/Open-Sora`.

[94] P.-Y. Lab, T. A. etc., Open-sora-plan, 2024. URL: `https://doi.org/10. 5281/zenodo.10948109`. doi:`10.5281/zenodo.10948109`.

[95] B. Poole, A. Jain, J. T. Barron, B. Mildenhall, Dreamfusion: Text-to-3d using 2d diffusion, arXiv:2209.14988 (2022).

[96] A. Jain, B. Mildenhall *et al.*, Zero-shot text-guided object generation with dream fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2022, pp. 867–876.

[97] B. Mildenhall, P. P. e. a. Srinivasan, Nerf: Representing scenes as neural radiance fields for view synthesis 65 (2021) 99–106.

[98] C.-H. Lin, J. Gao *et al.*, Magic3d: High-resolution text-to-3d content creation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 300–309.

[99] G. Li, H. Zheng *et al.*, 3ddesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models, arXiv:2211.14108 (2022).

[100] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, C. Vondrick, Zero-1-to-3: Zero-shot one image to 3d object, 2023. `arXiv:2303.11328`.

[101] J.-g. Kwak, E. Dong, Y. Jin, H. Ko, S. Mahajan, K. M. Yi, Vivid-1-to-3: Novel view synthesis with video diffusion models, in: CVPR, 2024, pp. 6775–6785.

[102] R. Abdal, Y. Qin, P. Wonka, Image2stylegan++: How to edit the embedded images?, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2020, pp. 8296–8305.

[103] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, A. Torralba, Semantic photo manipulation with a generative image prior, arXiv:2005.07727 (2020).

[104] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, D. Cohen-Or, Encoding in style: a stylegan encoder for image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2021, pp. 2287–2296.

[105] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, D. Cohen-Or, Designing an encoder for stylegan image manipulation, ACM Transactions on Graphics (TOG) 40 (2021) 1–14.

[106] M. Pernuš, C. Fookes, V. Štruc, S. Dobrišek, Fice: Text-conditioned fashion-image editing with guided gan inversion, Pattern Recognition 158 (2025) 111022.

[107] V. V. Dere, A. Shinde, P. Vast, Conditional reiterative high-fidelity gan inversion for image editing, Pattern Recognition 147 (2024) 110068.

[108] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, S. Ermon, SDEdit: Guided image synthesis and editing with stochastic differential equations, in: International Conference on Learning Representations, 2022.

[109] G. Kim, T. Kwon, J. C. Ye, Diffusionclip: Text-guided diffusion models for robust image manipulation, in: CVPR, 2022, pp. 2426–2435.

[110] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv:2010.02502 (2020).

[111] B. Wallace, A. Gokul, N. Naik, Edict: Exact diffusion inversion via coupled transformations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 22532–22541.

[112] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, D. Cohen-Or, Null-text inversion for editing real images using guided diffusion models, in: CVPR, 2023, pp. 6038–6047.

[113] M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting, A. Passos, Ledits++: Limitless image editing using text-to-image models, in: CVPR, 2024, pp. 8861–8870.

[114] O. Avrahami, D. Lischinski *et al.*, Blended diffusion for text-driven editing of natural images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2022, pp. 18208–18218.

[115] O. Avrahami, O. Fried *et al.*, Blended latent diffusion, ACM Transactions on Graphics (TOG) 42 (2023) 1–11.

[116] J. Ackermann, M. Li, High-resolution image editing via multi-stage blended diffusion, arXiv:2210.12965 (2022).

[117] G. Couairon, J. Verbeek *et al.*, Diffedit: Diffusion-based semantic image editing with mask guidance, arXiv:2210.11427 (2022).

[118] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, M. Irani, Imagic: Text-based real image editing with diffusion models, arXiv:2210.09276 (2022).

[119] A. Hertz, R. Mokady *et al.*, Prompt-to-prompt image editing with cross attention control, arXiv:2208.01626 (2022).

[120] T. Brooks, A. Holynski *et al.*, Instructpix2pix: Learning to follow image editing instructions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 18392–18402.

[121] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, F. Wen, Paint by example: Exemplar-based image editing with diffusion models, arXiv:2211.13227 (2022).

[122] J. H. Liew, H. Yan, D. Zhou, J. Feng, Magicmix: Semantic mixing with diffusion models, arXiv:2210.16056 (2022).

[123] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, K. Kersting, Sega: Instructing text-to-image models using semantic guidance, Advances in Neural Information Processing Systems 36 (2023) 25365–25389.

[124] P. Schramowski, M. Brack, B. Deiseroth, K. Kersting, Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 22522–22531.

[125] F. Friedrich, K. Hämmerl, P. Schramowski, J. Libovicky, K. Kersting, A. Fraser, Multilingual text-to-image generation magnifies gender stereotypes and prompt engineering may not help you, arXiv preprint arXiv:2401.16092 (2024).

[126] L. Struppek, D. Hintersdorf, F. Friedrich, P. Schramowski, K. Kersting, et al., Exploiting cultural biases via homoglyphs in text-to-image synthesis, Journal of Artificial Intelligence Research 78 (2023) 1017–1068.

[127] M. Hu, C. Zheng *et al.*, Unified discrete diffusion for simultaneous vision-language generation, arXiv:2211.14842 (2022).

[128] OpenAI, Gpt-4, 2024. URL: `https://openai.com/index/gpt-4/`.

[129] X. Chen, Z. Liu, S. Xie, K. He, Deconstructing denoising diffusion models for self-supervised learning, arXiv:2401.14404 (2024).

[130] OpenAI, Gpt-3 applications, 2024. URL: `https://openai.com/index/gpt-3-apps/`.

[131] Y. Liu, J. He, Y. Liu, X. Lin, F. Yu, J. Hu, Y. Qiao, C. Dong, Adaptbir: Adaptive blind image restoration with latent diffusion prior for higher fidelity, Pattern Recognition (2024) 110659.

[132] G. Kim, W. Jang, G. Lee, S. Hong, J. Seo, S. Kim, Depth-aware guidance with self-estimated depth representations of diffusion models, Pattern Recognition 153 (2024) 110474.

[133] Y. Xu, S. Wu, B. Wang, M. Yang, Z. Wu, Y. Yao, Z. Wei, Two-stage fine-grained image classification model based on multi-granularity feature fusion, Pattern Recognition 146 (2024) 110042.

[134] S. Panagiotou, A. S. Bosman, Denoising diffusion post-processing for low-light image enhancement, Pattern Recognition 156 (2024) 110799.

[135] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, D. Pathak, Your diffusion model is secretly a zero-shot classifier, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2206–2217.