State of the Art on Diffusion Models for Visual Computing

R. Po[1⋆] W. Yifan[1⋆] V. Golyanik[2⋆] K. Aberman[3] J. T. Barron[4] A. Bermano[5] E. Chan[1] T. Dekel[6] A. Holynski[4,7]
A. Kanazawa[7] C. K. Liu[1] L. Liu[8] B. Mildenhall[4] M. Nießner[9] B. Ommer[10] C. Theobalt[2] P. Wonka[11] G. Wetzstein[1]

[1]Stanford University [2]MPI for Informatics and VIA Center [3]Snap Inc. [4]Google Research [5]Tel Aviv University [6]Weizmann Institute of Science
[7]UC Berkley [8]University of Pennsylvania [9]TU Munich [10]LMU Munich [11]KAUST ⋆Equal contribution

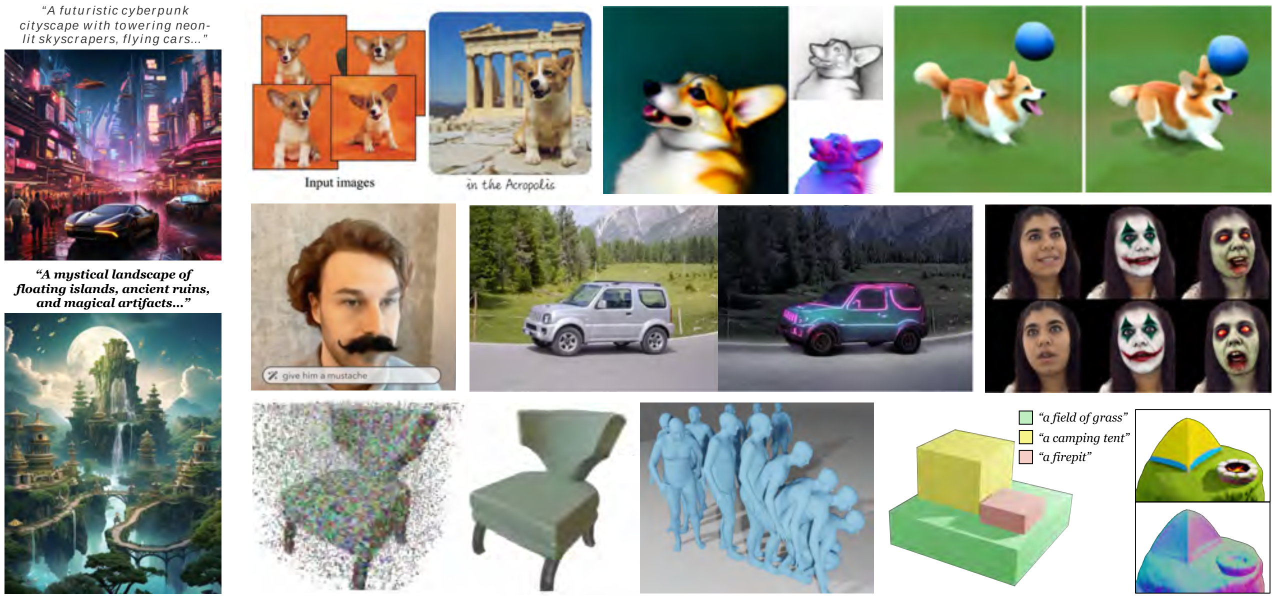

"A futuristic cyberpunk cityscape with towering neon-lit skyscrapers, flying cars..."
"A mystical landscape of floating islands, ancient ruins, and magical artifacts..."
Input images
in the Acropolis
give him a mustache
"a field of grass"
"a camping tent"
"a firepit"

Figure 1: *This state-of-the-art report discusses the theory and practice of diffusion models for visual computing. These models have recently become the de-facto standard for image, video, 3D, and 4D generation and editing. Images adapted from [PJBM22, DMGT23, SSP*23b, MSP*23, BTOAF*22, HTE*23, Lab23, PW23, RLJ*22, MPE*23, Arn23] ©2023 IEEE.*

Abstract

The field of visual computing is rapidly advancing due to the emergence of generative artificial intelligence (AI), which unlocks unprecedented capabilities for the generation, editing, and reconstruction of images, videos, and 3D scenes. In these domains, diffusion models are the generative AI architecture of choice. Within the last year alone, the literature on diffusion-based tools and applications has seen exponential growth and relevant papers are published across the computer graphics, computer vision, and AI communities with new works appearing daily on arXiv. This rapid growth of the field makes it difficult to keep up with all recent developments. The goal of this state-of-the-art report (STAR) is to introduce the basic mathematical concepts of diffusion models, implementation details and design choices of the popular Stable Diffusion model, as well as overview important aspects of these generative AI tools, including personalization, conditioning, inversion, among others. Moreover, we give a comprehensive overview of the rapidly growing literature on diffusion-based generation and editing, categorized by the type of generated medium, including 2D images, videos, 3D objects, locomotion, and 4D scenes. Finally, we discuss available datasets, metrics, open challenges, and social implications. This STAR provides an intuitive starting point to explore this exciting topic for researchers, artists, and practitioners alike.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics; Neural networks;***

1. Introduction

For decades, the computer graphics and 3D computer vision communities have been striving to develop physically accurate models to synthesize computer-generated imagery or infer physical properties of a scene from photographs. This methodology, which includes rendering, simulation, geometry processing, and photogrammetry, forms a cornerstone of several industries including visual effects, gaming, image and video processing, computer-aided design, virtual and augmented reality, data visualization, robotics, autonomous vehicles, remote sensing, among others.

The emergence of generative artificial intelligence (AI) marks a paradigm shift for visual computing. Generative AI tools enable the generation and editing of photorealistic and stylized images, videos, or 3D objects with little more than a text prompt or high-level user guidance as input. These tools automate many laborious processes in visual computing that had previously been reserved for experts with specialized domain knowledge, making them more broadly accessible.

The unprecedented capabilities of generative AI have been unlocked by foundation models for visual computing, such as Stable Diffusion [RBL*22], Imagen [SCS*22], Midjourney [Mid23], or DALL-E 2 [Ope23a] and DALL-E 3 [Ope23b]. Trained on hundreds of millions to billions of text–image pairs, these models have "seen it all" and, with an estimated few billion learnable parameters, are extremely large. After being trained on a massive cloud of high-end graphics processing units (GPUs), these models form the foundation of the aforementioned generative AI tools. The networks commonly used for image, video, and 3D object generation are typically variants of convolutional neural network (CNN)–based diffusion models that are combined in a multi-modal manner with text computed via transformer-based architectures, such as CLIP [RKH*21].

While much of the successful development and training of foundation models for 2D image generation has come from well-funded industry players using a massive amount of resources, there is still room for the academic community to contribute in major ways to the development of these tools for graphics and vision. For instance, it is not clear how to extend existing image foundation models to other, higher-dimensional domains, like video and 3D scene generation. This is largely due to the lack of certain types of training data. The web, for example, contains billions of 2D images but much fewer instances of high-quality and diverse 3D objects or scenes. Moreover, it is not obvious how to scale 2D image generation architectures to handle higher dimensions, as required for video, 3D scene, or 4D multi-view-consistent scene generation. Another example of an existing limitation is *computation*: diffusion models are rather slow at inference time due to the large size of their networks and their iterative nature, and even though massive amounts of (unlabeled) video data exists on the web, current network architectures are often too inefficient to be trained in a reasonable amount of time or on a reasonable amount of compute resources.

Despite the remaining open challenges, recent developments have spurred an explosion of diffusion models for visual computing over the last year (see representative examples in Fig. 1). The goals of this state-of-the-art report (STAR) are to introduce the fundamentals of diffusion models, to present a structured overview of the many recent works focusing on applications of diffusion models in visual computing, and to outline open challenges.

This STAR is structured as follows: Sec. 2 outlines the scope and refers interested readers to surveys on closely related topics that are not covered here; Sec. 3 gives an overview of the mathematical foundations of 2D diffusion; Sec. 4 discusses the challenge of moving beyond 2D images towards video, 3D, and higher-dimensional diffusion models; Sec. 5 outlines approaches to diffusion-based video synthesis and editing; Sec. 6 summarizes recent approaches to 3D object and scene generation; Sec. 7 includes a discussion on 4D spatio-temporal diffusion for multi-view consistent video, human motion and scene generation (e.g., using parametric human body models); Sec. 8 includes a brief discussion on available training data; Sec. 9 reviews metrics used for various generated content; Sec. 10 outlines open challenges; Sec. 11 discusses societal implications and ethical concerns; and Sec. 12 concludes the STAR.

2. Scope of this STAR

In this STAR, we focus on recent advances in applications of diffusion models in visual computing. Specifically, we discuss the role of diffusion models in the context of generating and editing images, videos, 3D objects or scenes, and multi-view consistent 4D dynamic scenes. We start by laying down the mathematical underpinnings of diffusion models. This includes a brief introduction to the general diffusion process as applied to 2D images. We then delve into how these techniques enable generative modeling of high-dimensional signals, as we provide a comprehensive overview of the leading methodologies of diffusion models for video, 3D and 4D data. This report aims to highlight techniques leveraging diffusion models to tackle problems on data beyond the image domain. With that in mind, we do not cover each and every method only applicable to 2D data. We also do not discuss works that leverage generative pipelines other than diffusion models.

Related Surveys. Other generative methods, such as GANs, are closely related to diffusion models. However, we consider them to be beyond the scope of this report. We refer readers to [GSW*21] for an in-depth discussion on GANs and [BTLLW21, LZW*23, SPX*22] for a broader review of other generative methods, or the use of different generative model architectures for multi-modal image synthesis and editing [ZYW*23]. Recently, the term *foundation model* has become analogous to a diffusion model trained on internet-scale data image [RBL*22]. While this report discusses methods that leverage such models, please refer to [BHA*21] for an introduction and overview of foundation models in the context of natural language processing, visual computing, and other domains. Last but not least, the explosive advances in text-to-image (T2I) generation has led to an intrinsic link between large language models (LLMs) and diffusion models; interested readers can refer to [ZZL*23] for a comprehensive survey on LLMs.

Selection Scheme. This report covers papers published in the proceedings of major computer vision, machine learning, and computer graphics conferences, as well as preprints released on arXiv (2021–2023). The authors of this report have selected papers based on their relevance to the scope of this survey, as we aim to provide a comprehensive overview of the rapid advances in diffusion models in the context of visual computing. However, though this report serves as a list of state-of-the-art methods in a specific domain, we do not claim completeness and highly recommend that readers refer to cited works for in-depth discussions and details.

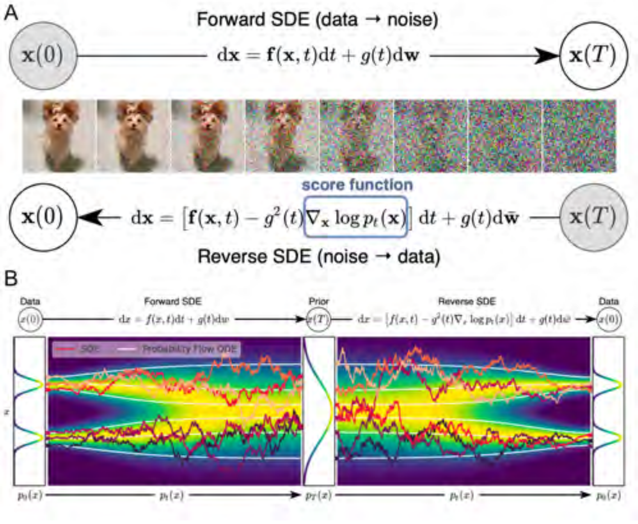

A
Forward SDE (data → noise)
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x},t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
$\mathbf{x}(T)$
score function
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x},t) - g^2(t)\nabla_\mathbf{x}\log p_t(\mathbf{x})]\,\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
$\mathbf{x}(T)$
Reverse SDE (noise → data)
B
Data
Forward SDE
Prior
Reverse SDE
Data
$x(0)$
$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w$
$x(T)$
$\mathrm{d}x = [f(x,t) - g^2(t)\nabla_x \log p_t(x)]\,\mathrm{d}t + g(t)\mathrm{d}\bar{w}$
$x(0)$
SDE
Probability Flow ODE
x
$p_0(x)$
$p_t(x)$
$p_T(x)$
$p_t(x)$
$p_0(x)$

Figure 2: ***Diffusion Process.*** *(A) The forward SDE transforms images to noise. The forward SDE can be reversed [And82] if we can predict the score function, enabling image synthesis. (B) The distributions of images and noise are linked with stochastic trajectories, modeled by SDEs, and deterministic trajectories, modeled by a probability flow ODE. Figures adapted from [SSDK*20].*

3. Fundamentals of Diffusion Models

In this section, we give a concise overview of the fundamentals of diffusion models. We introduce the mathematical preliminaries, discuss a practical implementation using the popular Stable Diffusion model as an example, and then overview important concepts for conditioning and guidance, before discussing concepts related to inversion, image editing, and customization. This section covers a large amount of references, so we focus on giving the reader a clear and high-level overview of the most important concepts of diffusion-based generation and editing in the context of 2D images.

3.1. Mathematical Preliminaries

Assume we are given a training dataset of examples where each example in the data is drawn independently from an underlying data distribution $p_{\text{data}}(\boldsymbol{x})$. We desire to fit a model to $p_{\text{data}}(\boldsymbol{x})$ so that we can synthesize novel examples by sampling from this distribution.

The general idea behind inference with denoising diffusion models is to sequentially denoise samples of random noise into samples from the data distribution. Consider a range of noise levels, $\sigma_{\max} > \ldots > \sigma_0 = 0$, and the corresponding noisy image distributions $p(\boldsymbol{x}, \sigma)$ defined as the distribution of adding Gaussian i.i.d. noise with variance σ^2 to the data. For sufficiently large $\sigma_{\max}$, the noise almost completely obscures the data and $p(\boldsymbol{x}, \sigma_{\max})$ is practically indistinguishable from Gaussian noise. Thus, we can sample an initial noise image $\boldsymbol{x}_T \sim \mathcal{N}(0, \sigma_{\max}^2)$ and sequentially denoise it such that at every step, $\boldsymbol{x}_i \sim p(\boldsymbol{x}, \sigma_i)$. The endpoint of this sampling chain, $\boldsymbol{x}_0$ is distributed according to the data.

However, instead of thinking about the denoising through a discrete collection of noise levels, it is useful to think of the noise level as a continuous, time-dependent function $\sigma(t)$ (a common choice is $\sigma(t) = t$). The noisy image sample $\boldsymbol{x}$ can move continuously through noise levels, following a trajectory—either forward in time, gradually adding noise, or backwards in time, gradually removing noise (see Fig. 2 (B)).

Song et al. [SSDK*20] introduce a stochastic differential equation (SDE) framework to model these trajectories. Ordinary differential equations give us tools to solve initial value problems – given an initial state and a differential equation that describes a function, we can solve for the function at a different time. As an example, given an object's initial position and known velocity, we can solve for the object's position at any time in the future. As Fig. 2 (A) depicts, in much the same way, noising an image can be thought of as picking an initial image $\boldsymbol{x}_0$ from the image domain and solving a differential equation forwards in time; denoising an image can be thought of as picking an initial noise image $\boldsymbol{x}_T \sim \mathcal{N}(0, \sigma_{\max}^2)$ and solving a differential equation backwards in time.

The gradual corruption of an image with noise over time is a diffusion process that can be modeled by an Itô stochastic differential equation (SDE; Eq. 1) [Itô50, Itô51], where $\mathbf{f}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector-valued function known as the drift coefficient, $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a scalar-valued function known as the diffusion coefficient, and $\boldsymbol{w}$ is the standard Wiener process:

$$\mathrm{d}\boldsymbol{x} = \mathbf{f}(\boldsymbol{x}, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}. \tag{1}$$

Implementing a diffusion model requires selecting $\mathbf{f}$ and g, and several specific choices have been explored by [SSDK*20]. The choices of $\mathbf{f}(\boldsymbol{x}, t) = 0$ and $g(t) = \sqrt{2\sigma(t)\frac{\mathrm{d}\sigma(t)}{\mathrm{d}t}}$ yield an SDE that describes noising an image by adding Gaussian noise of variance $\sigma^2(t)$. This SDE is known as the Variance Exploding SDE (Eq. 2), so-called because the variance continuously increases with increasing t. Noising an image can be thought of as selecting an initial clean image and solving Eq. 2 forward in time as

$$\mathrm{d}\boldsymbol{x} = \sqrt{2\sigma(t)\frac{\mathrm{d}\sigma(t)}{\mathrm{d}t}}\mathrm{d}\boldsymbol{w}. \tag{2}$$

The Variance Exploding SDE has the closed-form solution

$$p(\boldsymbol{x}_t | \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{x}_0, \left[\sigma^2(t) - \sigma^2(0)\right]\mathbf{I}\right). \tag{3}$$

In other words, to obtain a noisy image at timestep t, all we need to do is add Gaussian noise of variance $\sigma^2(t) - \sigma^2(0)$.

The work of Anderson [And82] enables the discovery of an SDE that reverses a diffusion process. Applied to Eq. 2, this produces a reverse-time SDE

$$\mathrm{d}\boldsymbol{x} = -2\sigma(t)\frac{\mathrm{d}\sigma(t)}{\mathrm{d}t}\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x}; \sigma(t))\mathrm{d}t + \sqrt{2\sigma(t)\frac{\mathrm{d}\sigma(t)}{\mathrm{d}t}}\mathrm{d}\boldsymbol{w}. \tag{4}$$

In Eq. 4, $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}; \sigma(t))$ is known as the *score function*, a vector field that points towards regions of higher data likelihood. To solve Eq. 4, we need to predict the score function with a neural network. Remarkably, for a denoiser function D that minimizes the L_2 denoising error $\mathbb{E}_{\boldsymbol{y} \sim p_{\text{data}}}\mathbb{E}_{\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})}\|D(\boldsymbol{y} + \boldsymbol{n}; \sigma) - \boldsymbol{y}\|_2^2$, the score function can be easily obtained from the model output as $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}; \sigma(t)) = (D(\boldsymbol{x}; \sigma) - \boldsymbol{x})/\sigma^2$. This means that by simply training a model to denoise images, we can extract a prediction of the score function. This is known as Denoising Score Matching and the above relationship was derived by [Vin11]. Note that it is common to parameterize the neural network with parameters ϕ or θ, as a noise-prediction network ϵ_ϕ, rather than as a denoiser D_θ; however, each is easily recoverable from the other as $\epsilon_\phi(\boldsymbol{x};\sigma) = \boldsymbol{x} - D_\theta(\boldsymbol{x};\sigma)$.

In order to sample an image, we need only start with some initial $\boldsymbol{x}_T \sim \mathcal{N}(0, \sigma_{\max}^2 \mathbf{I})$ and we can solve Eq. 4 backwards in time to arrive at a sample from the $p_{\text{data}}(\boldsymbol{x})$. However, as is the case with ordinary differential equations (ODE), only a small subset has closed-form solutions. Fortunately, as an alternative, we can approximate the solution to the SDE numerically.

Euler–Maruyama (Alg. 1) is an algorithm for approximating numerical solutions to SDEs. It is a simple extension of Euler's method, which is the most basic numerical ODE solver, and a technique with which many will be familiar. Like Euler's method, Euler–Maruyama approximates a trajectory by taking small steps tangent to the trajectory. Smaller steps enable approximation with greater precision. The sampling techniques of many common diffusion models [SSDK*20, SME20, ND21] can be viewed as modifications of Euler–Maruyama.

Algorithm 1 Euler–Maruyama Approximation

```
Input: An SDE of the form $\mathrm{d}\boldsymbol{x} = a(\boldsymbol{x},t)\mathrm{d}t + b(\boldsymbol{x},t)\mathrm{d}\boldsymbol{w}$; an initial condition $\boldsymbol{x}_0$; a time interval $[0,T]$. A finite number of subintervals $N$.
Output: A simulated trajectory $\{\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{N-1}\}$.
Initialize: Partition the time interval $[0,T]$ into $N$ equal subintervals $\tau_0 < \tau_1 < ... < \tau_N$, where $\tau_{i+1} - \tau_i = \Delta t = T/N$; $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0$; n = 0.
while n < N do
    $\hat{\boldsymbol{x}}_{n+1} = \hat{\boldsymbol{x}}_n + a(\hat{\boldsymbol{x}}_n, \tau_n)\Delta t + b(\hat{\boldsymbol{x}}_n, \tau_n)\Delta \boldsymbol{w}_n$,
        where $\Delta \boldsymbol{w}_n = \boldsymbol{w}_{\tau_{n+1}} - \boldsymbol{w}_{\tau_n}$
    $n = n + 1$.
end while
```

Viewing image synthesis with diffusion models through the lens of numerical SDE solvers can give us intuition about the behavior of different sampling schemes and some insight into how some works have improved the computational efficiency of generating images. Empirically, we observe that generating high-quality images with diffusion models requires many (often hundreds) of iterations; fewer iterations produce poor samples. In the language of numerical differential equation solvers, poor quality results with few iterations is a result of truncation error—the smaller we make our timesteps Δt, the more accurate our numerical approximation. For the same reason, higher-order differential equation solvers [KAAL22, DVK22] can reduce the error in our numerical approximation, allowing us to sample with greater accuracy or enabling equal quality with fewer network evaluations.

For any diffusion process, there exists a corresponding deterministic process, which can be described by an ODE, that recovers the same marginal probability densities $p(\boldsymbol{x},\sigma)$. Song et al. [SSDK*20] define an ODE that describes the deterministic process and name it the probability flow ODE (Eq. 5, Fig. 2b):

$$\mathrm{d}\boldsymbol{x} = \left[\mathbf{f}(\boldsymbol{x},t) - \frac{1}{2}g(t)^2 \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x};\sigma(t))\right] \mathrm{d}t. \tag{5}$$

This probability flow ODE enables deterministic image synthesis—instead of sampling random noise and simulating the reverse-time SDE to synthesize images, we can instead sample random noise and solve the reverse deterministic probability-flow ODE; doing so recovers the same distribution of images. Unlike stochastic sampling, where the final image is determined both by the initial noise image x_t and noise injected at every iteration (corresponding to the $d\boldsymbol{w}$ term in Eq. 4), an image created deterministically is defined only by the initial noise.

But we can also do the reverse: Instead of sampling noise and producing an image, we can draw an arbitrary image and follow the forward probability flow ODE to encode the image into noise. In fact, the probability flow ODE defines a bijective mapping between images and (noisy) latents. With sufficient accuracy in the ODE solver, one can encode an image into latent space by solving the probability flow ODE forward in time, arrive at a latent $\boldsymbol{x}_T$, and solve the probability flow ODE backwards in time to recover the original image. Moreover, one can edit an image by manipulating the corresponding latent. For example, interpolation in latent space may produce a compelling interpolation in image space, and scaling the latent can influence the temperature of the generated image [SSDK*20]. Several recent image editing works are built on the premise of a deterministic mapping between images and latents, and rely on the forward probability flow ODE to map real images into the latent space of diffusion models [HMT*22, MHA*23, SSME22].

Another advantage of deterministic sampling is that the probability flow ODE can also often be solved sufficiently accurately with fewer iterations [SSDK*20, SME20] of a numerical solver than the corresponding SDE; as a result, deterministic sampling may produce high-quality images with fewer network evaluations than stochastic sampling, accelerating inference.

Despite these advantages, stochastic sampling is still often preferred over deterministic sampling when many (generally hundreds to thousands) of denoising iterations are available, and image quality is paramount. Intuitively, stochasticity can be seen as a corrective force that repairs errors made earlier in sampling. Thus, stochastic samplers, when combined with many denoising iterations, often produce images that evaluate best according to metrics.

Karras et al. [KAAL22] illustrate this by decomposing the reverse-time SDE into the sum of a probability flow ODE, which deterministically moves a sample between noise levels, and a Langevin diffusion SDE, which stochastically "churns" a sample at a fixed noise level by adding and removing a small amount of noise. Here, the Langevin diffusion SDE, i.e., the stochasticity component, pushes a sample towards the marginal distribution $p(\boldsymbol{x},\sigma)$, correcting errors that may have been carried over in the purely deterministic setting. In practice, deterministic sampling and stochastic sampling have specific strengths and weaknesses, and the level of stochasticity that works best will depend on the task.

3.2. Latent Diffusion using the Stable Diffusion Model

Unlike generative models such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs), which generate images through a single forward pass, diffusion models necessitate recurrent forward passes. This characteristic imposes a higher computational burden during training, as the model must learn denoising across multiple noise scales. Additionally, the iterative nature of the multi-stage denoising process elongates inference time, making diffusion models computationally less efficient than their generative counterparts [KAAL22].

Generating high-resolution images with diffusion models is often infeasible on consumer-grade GPUs due to the excessive memory requirements. Even on high-end GPUs, the constraints on batch sizes prolong the training process, making it impractical for a large segment of the research community.

Perceptual Image Compression. To address these challenges, Rombach et al. [RBL*22] introduced latent diffusion models that operate in a compressed latent space, rather than directly on image pixels. This approach retains perceptually relevant details while significantly reducing computational cost. The compressed image space is obtained using an encoder–decoder architecture. Among the various techniques, VQ-GAN [ERO21] has emerged as the most common choice due to its impressive compression ability and preservation of perceptual quality. These latent diffusion models (LDMs) consist of a two-stage process: an initial autoencoder for image reconstruction and a subsequent denoising model operating on the latent codes, achieving superior performance with reduced computational demands (see Fig. 3).

Architecture. The architecture proposed by Rombach et al. [RBL*22] builds upon the U-Net framework [RFB15, HJA20, JMPTdCM20]. It incorporates attention mechanisms [VSP*17], specifically self-attention and cross-attention blocks, at various stages of the U-Net. In the self-attention block, features derived from intermediate U-Net outputs are projected into queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$. The output of the block is given by:

$$\mathbf{A}\cdot\mathbf{V} \quad \text{where} \quad A = \text{Attention}(\mathbf{Q},\mathbf{K}). \tag{6}$$

Here, the Attention mechanism captures contextual information between the d-dimensional $\mathbf{Q}$ and $\mathbf{V}$ projection matrices via

$$\text{Attention}(\mathbf{Q},\mathbf{K}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d}). \tag{7}$$

Cross-attention blocks operate similarly, enabling controlled generation by injecting conditioning signals such as text prompts (see Sec. 3.3).

Retrieval-augmentation Mechanism (RDM). To further optimize computational efficiency, Blattmann et al. [BRO*22] introduced a Retrieval-augmentation Mechanism (RDM) that fetches relevant image patches from an external database during the generative process. These patches are selected based on latent codes from a pre-trained autoencoder and undergo simple augmentations. Since details from the retrieved image patches need not be saved in model parameters anymore, this mechanism leads to a more streamlined denoising model and hence accelerates both training and inference, albeit at the cost of added computational complexity and dependency on a well-trained autoencoder.

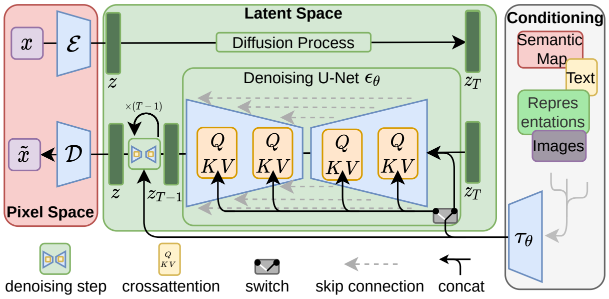


Figure 3: ***Stable Diffusion.*** *This schematic shows an overview of the latent diffusion approach, including encoder $\mathcal{E}$, decoder $\mathcal{D}$, and conditioning using a cross-attention mechanism. Figure adapted from [RBL*22].*

3.3. Conditioning and Guidance

Conditioning. Perhaps the most important property of a generative model is the ability to control generation through user-defined conditions. Such conditions include text [RAY*16], semantic maps [PLWZ19], sketches [VACO23], multi-modal combinations of conditions [ZYW*23], and other image-to-image translation tasks [IZZE18, SCC*22]. Formally, instead of sampling data from an unconditional distribution $p(\boldsymbol{x})$, we would like to sample from a conditional distribution $p(\boldsymbol{x}|\boldsymbol{c})$ given some conditioning signal $\boldsymbol{c}$.

To accommodate the variety of conditioning modalities, a flexible set of conditioning mechanisms has been developed for diffusion models. The most simple of these methods is concatenation [SCC*22], where the condition is directly concatenated with intermediate denoising targets and passed through the score estimator as input. Concatenation can be performed along with the diffusion model input during different stages of the model architecture. It is also applicable to nearly all conditioning modalities. Most notably, Palette [SCC*22] tackles various image-to-image translation tasks such as in-painting, colorization, uncropping and image restoration using conditioning by concatenation.

Another effective method is to inject conditioning signals through cross-attention. Rombach et al. [RBL*22] modifies the U-Net architecture [RFB15] for conditioning control with cross-attention mechanisms. To control the image synthesis, a conditioning signal $\boldsymbol{c}$, for example a guiding text prompt, is first preprocessed by a domain-specific encoder τ to an intermediate projection $\tau(\boldsymbol{c})$. The projected conditioning signal is then injected into the intermediate layers of the denoising U-Net by means of cross attention [VSP*17], via Eq. 7, with

$$\mathbf{Q} = \mathbf{W}_Q\cdot\varphi(z_t),\ \mathbf{K} = \mathbf{W}_K\cdot\tau(\boldsymbol{c}),\ \mathbf{V} = \mathbf{W}_V\cdot\tau(\boldsymbol{c}), \tag{8}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices, and $\varphi(z_t)$ represents an intermediate result from the denoising U-Net; see Fig. 3 for a detailed visualization. Intuitively, $\mathbf{Q}$ is the projection of activations of intermediate U-Net layers, while $\mathbf{K}$ and $\mathbf{V}$ are obtained via projection of the given condition.

During inference, additional techniques can be applied to condition the network on different modalities, including

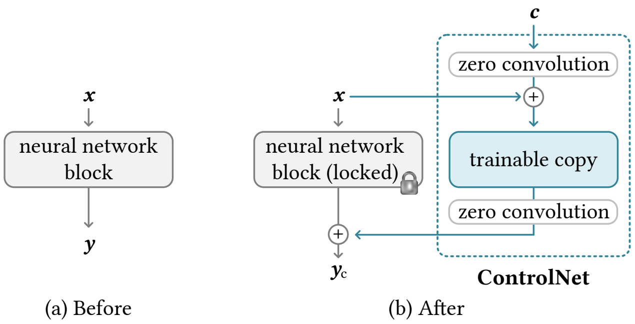


Figure 4: ***Overview of ControlNet.*** *ControlNet [ZA23] modifies existing network architectures by duplicating network blocks and connecting them through zero convolutions. The auxiliary module takes some conditioning* ***c****, allowing the model to learn additional control handles.*

sketches [VACO23] and spatial layout [ZA23, MWX*23]. Among these, adapter methods, exemplified by the popular ControlNet [ZA23] (see Fig. 4), provide an effective and flexible route to add new condition modality without altering the pre-trained diffusion model by embedding new module layers into the existing network architecture. Specifically, it proposes a backbone for learning diverse control handles for large pre-trained diffusion models through the addition of auxiliary network modules. New network modules are initialized by duplicating encoding layers from the pre-trained network and connected to the original model through "zero-convolutions", a mechanism that initializes layer parameters to zero, ensuring that no harmful noise is learned during the fine-tuning process. The auxiliary network is fine-tuned on a given set of condition-output pairs, while the original network layers remain unchanged as illustrated in Fig. 4.

Guidance. While conditioning affords a level of control over the sampled distribution, it falls short in fine-tuning the strength of the conditioning signal within the model. Guidance emerges as an alternative, generally applied post-training, to more precisely steer the diffusion trajectory.

Dhariwal et al. [DN21] observed that an auxiliary classifier can steer an unconditional generative model. This technique, termed classifier guidance, alters the original diffusion score by incorporating the gradient of the log-likelihood from a pre-trained classifier model $p_\phi(\boldsymbol{c}|\boldsymbol{x})$ that estimates $\boldsymbol{c}$ from a given image $\boldsymbol{x}$. Using Bayes' theorem, the score estimator for $p(\boldsymbol{x}|\boldsymbol{c})$ is given by:

$$\nabla_{\boldsymbol{x}_t} \log(p_\theta(\boldsymbol{x}_t) p_\phi(\boldsymbol{c}|\boldsymbol{x}_t)) = \nabla_{\boldsymbol{x}_t} \log p_\theta(\boldsymbol{x}_t) + \nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{c}|\boldsymbol{x}_t). \quad (9)$$

We can then define a new score estimator $\tilde{\varepsilon}(\boldsymbol{x}_t)$ corresponding to the joint distribution, giving:

$$\tilde{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) = \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - w\sigma_t \nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{c}|\boldsymbol{x}_t), \quad (10)$$

where w is a tunable parameter controlling guidance strength. Despite its versatility, classifier guidance has limitations, such as the need for a noise-robust auxiliary classifier and the risk of poorly defined gradients due to irrelevant information in $\boldsymbol{x}_t$.

To circumvent the limitations of classifier guidance, Ho and Salimans introduced classifier-free guidance [HS22]. This method directly alters the training regimen, utilizing a single neural network to represent both an unconditional and a conditional model. These models are jointly trained, with the unconditional model parameterized by a null token $\boldsymbol{c} = \varnothing$. The model can then be sampled as follows, using the previously introduced guidance scale w:

$$\hat{\epsilon}_\theta(\boldsymbol{x}_t) = (1+w)\epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - w\epsilon_\theta(\boldsymbol{x}_t, \varnothing). \quad (11)$$

Controlling the strength of the conditional signal leads to a trade-off between diversity and sample quality [HS22]. As guidance scale w increases, the diversity of the resulting sample decreases in return for higher sample quality. However, it is often observed that models trained using classifier-free guidance tend to generate low-quality samples for very low or high guidance scales. For example, Stable Diffusion [RBL*22] generates empty gray images when sampled using only the unconditional score estimator ($w = -1$), and outputs images with saturation artifacts at higher guidance values ($w > 10$).

3.4. Editing, Inversion and Customization

A pretrained diffusion model essentially provides an expressive generative prior, which can be leveraged to allow average users to perform various image manipulation tasks without any experience on pixel-level crafting skills. Many recent work have investigated the use of text-to-image diffusion models for *editing*, and generation of personalized images, also known as *customization*. This section surveys the main works in these two categories, as well as a crucial technical component, *inversion*, which is often used as a building block for editing.

Editing. Owing to its inherently progressive and attention-based architecture, diffusion models offer a unique platform for fine-grained image editing by facilitating adjustments to various network phases and components to manipulate both spatial layout and visual aesthetics. The research trajectory in this domain is geared towards enhancing editing controllability and flexibility, while simultaneously ensuring an intuitive user interface. A prevalent use-case involves altering the visual attributes of an image while retaining its spatial configuration. In this context, SDEdit [MHS*21] presents a straightforward approach that introduces a calibrated level of noise to an image, resulting in a partially-noised image, followed by a reverse diffusion process with a new conditioning signal as guidance. This foundational approach has been further extended by [KKY22] to manipulate global characteristics by direct text prompt modification, while localized editing is accomplished through the incorporation of auxiliary masks in the diffusion process [ALF22, NDR*21]. Similarly, an additional guidance signal (i.e., any gradient function, as a classifier is in classifier guidance) can be used to modify a sampling trajectory to perform manipulative edits like changing object appearances or rearranging content in the scene [EJP*23]. Another editing strategy involves constraining specific feature maps derived from a different generative processes. For example, Prompt-to-Prompt [HMT*22] employs fixed cross-attention layers to selectively modify image regions corresponding to specific textual cues. Plug-and-Play [TGBD23] explores the injection of spatial features and self-attention maps to maintain the overall structural integrity of the image. [CWQ*23] advocates for leveraging the self-attention mechanism to enable consistent, non-rigid image editing without the necessity for manual tuning. Moreover, the text prompt itself serves as a critical determinant of editing quality. Imagic [KZL*23] refines the text prompts through an optimization of textual embeddings coupled with model fine-tuning, thereby enabling diverse, spatially non-rigid image editing. Delta Denoising Score [HACO23] ingeniously utilizes the generative prior of Text-to-Image (T2I) diffusion models as a loss term in an optimization framework to guide image transformations based on textual directives. InstructPix2Pix [BHE23] simplifies the user's text input from descriptive target image annotations to more intuitive editing directives by fine-tuning a T2I model on a generated dataset of image pairs aligned with editing instructions, created using a combination of Prompt-to-Prompt and a large language model. [PKSZ*23] obviates the need for text prompts altogether by introducing an automated mechanism for discovering editing directions from exemplar image pairs.

Following similar work in the GAN literature [PTL*23], a number of diffusion-based methods aim to perform edits driven by sparse user-annotated correspondences, where the appearance and identity of objects are preserved, and only their layout or orientation are changed [MWS*23, SXP*23].

Inversion. Many methods for editing an existing (i.e., real) image through generative models often involve an "inversion" task, which identifies a specific input latent code or sequence of latents that, when fed into the model, reproduces a given image. Inversion allows manipulation in the latent space, which enables the use of generative priors already learned by the model. In the context of diffusion, DDIM inversion [SME20] is a fundamental technique that adds small noise increments to a given image to approximate the corresponding input noise. When running a reverse diffusion using DDIM with this noise, the original image is reproduced. In the case of text-to-image diffusion, when provided with a specific text–image pairing, the DDIM inversion method tends to accumulate small errors, especially with classifier-free guidance [HS22]. Null-Text Inversion [MHA*23] compensates for the timestep drift by optimizing the input null-text embedding for each timestep. EDICT [WGN23] achieves precise DDIM inversion using two coupled noise vectors. [WDlT22] showcased a DDPM-inversion method, recovering noise vectors for an accurate image reconstruction within the DDPM sampling framework.

Beyond image-to-noise inversion, in the context of text-to-image models, "textual inversion" [GAA*22] offers a framework to convert image(s) into token embeddings. The original work proposed converting a concept that appears in a few images into a single token using optimization. Follow-up works also demonstrated the inversion of a single image into a token using an encoder [GAA*23b], or converting the concept into a sequence of per-layer tokens to improve the concept's reconstruction [VCCOA23].

Customization. Recent works have extensively explored the customization of T2I diffusion models, i.e., adapt a pretrained diffusion model to generate better outputs for a specific person or object. The pioneering work DreamBooth [RLJ*22] achieves this by optimizing the network weights to represent a subject shown in a set of images by a unique token. One line of follow-up works has focused on fine-tuning only specific parts of the network. CustomDiffusion [KZZ*23] modifies only the cross-attention layers, SVDiff [HLZ*23] refines the singular values of weights, LoRA [HSW*21] targets optimizing low-rank approximations of weight residuals [HSW*21] and StyleDrop [SRL*23] adopts adapter tuning [HGJ*19] to fine-tune a selected set of adapter weights for style customization. Similar techniques have been applied to other problem statements besides text-to-image generation, such as image inpainting or outpainting [TRC*23].

Another line of research is dedicated to accelerating the customization process. [GAA*23a] and [WZJ*23] employ encoders to determine initial text embeddings and subsequently fine-tune them to enhance subject fidelity. [RLJ*23] predicts low-rank network residuals tailored to specific subjects directly. SuTI [CHL*23] initially constructs a comprehensive paired dataset using images and their recontextualized counterparts produced by DreamBooth, and uses it to train a network that can execute personalized image generation in a feed-forward manner. InstantBooth [SXLJ23] and Taming Encoder [JZC*23] introduce a conditional branch to the diffusion model which allows conditioning using a minimal set of images or even just one, facilitating the generation of personalized outputs in various styles. Break-A-Scene [AAF*23] customizes a model to support a few subjects depicted in a single image, while FastComposer [XYF*23] leverages an image encoder to project subject-specific embeddings for multi-subject generation.

4. The Challenge of Moving Beyond Images

Diffusion models have garnered significant attention and success in the realm of image processing, owing to a confluence of factors that have made them particularly well-suited for this domain. One of the primary reasons for their efficacy lies in the maturity of network architectures tailored for image processing, particularly in the area of denoising. Diffusion models in the 2D image domain have capitalized on these advancements, incorporating well-defined building blocks such as convolutional layers, attention mechanism and U-Net structures as their backbone. Advances in Transformers [VSP*17] and large language models [DCLT18] have further enhanced these models by enabling approximate controllability through natural language prompts, facilitated by the pairing of images with text descriptions [RKH*21].

Moreover, the ubiquity of mobile phones and social norms have democratized the capture, storage, and sharing of 2D images, resulting in a near-infinite supply of freely available images. In summary, by the end of the 2010s, all the essential ingredients for the success of 2D diffusion models were in place: a well-defined mathematical framework, flexible function approximators capable of learning expressive image transformations, and an abundant supply of training data.

Though 2D image synthesis has been blessed by a happy coincidence of technological progress and available data, this has not been the case for higher-dimensional signals. The task of synthesizing higher-dimensional content, such as video and 3D content, is substantially more difficult than 2D image synthesis, limited by a number of additional issues described further in this section.

Models. The network architecture for processing high-dimensional data is still an open question. Unlike images, which can be efficiently represented and processed using discrete pixel values, higher-dimensional data often require more complex representations. This complexity is exacerbated by the need to handle long-range information flow, which is critical for understanding the temporal dynamics in videos or the spatial relationships in 3D structures. As of now, there is no consensus on a network architecture that can serve as a reliable and scalable backbone for diffusion models in these domains. The computational and memory costs associated with processing high-dimensional data further complicate the issue, making it challenging to find an efficient yet expressive solution.

Data. The availability and quality of data pose another significant challenge. For 3D structures, the process of creating a single 3D model involves multiple steps, including scanning, processing, and reconstruction, each of which requires specialized expertise and resources. This makes the data acquisition process both time consuming and expensive. In the case of videos—although raw data may be abundant—annotating these data, especially for capturing motion and temporal dependencies, is far from trivial. This scarcity of high-quality annotated data hampers the training of robust and generalizable diffusion models for higher-dimensional domains.

5. Video Generation and Editing

Despite the tremendous progress in *image* diffusion models, and the remarkable breakthroughs in T2I generation, expanding this progress to the domain of *video* is still in nascent stages, due to two main challenges.

First, learning from videos requires orders of magnitude more data than images, due to the complexity and diversity of our dynamic world. While online videos are abundant, curating high-quality video datasets remains a difficult task that typically involves significant engineering efforts and requires dedicated automatic curation tools.

Another substantial challenge arises from the high dimensionality of raw video data (e.g., a two-minute video with 30 fps consists of 3600× more pixels than a single frame). This makes the extension of 2D architectures to the temporal domain very challenging and computationally expensive. We next discuss how these challenges have been tackled in the context of diffusion models for video generation.

5.1. Unconditional and Text-Conditioned Video Diffusion

There have been significant research efforts in extending diffusion models to the temporal domain, aiming to capture the vast distribution of natural motion from a large-scale video dataset. Ho et al. [HSG*22] introduced the first Video Diffusion Model (VDM), extending the 2D U-Net backbone to the temporal domain. This is achieved through factorized space and time modules, enabling more efficient computation as well as joint training on both individual images, videos, and text. This approach has been scaled up by Imagen Video [HCS*22], a cascaded text-to-video (T2V) model with 11 billion parameters, which comprises a low spatio-temporal resolution base model, followed by multiple cascaded super-resolution models that increase both the spatial resolution and the effective framerate. Imagen Video is *trained from scratch*, using a large corpus of high-quality video and corresponding captions as well as a number of text-image datasets.

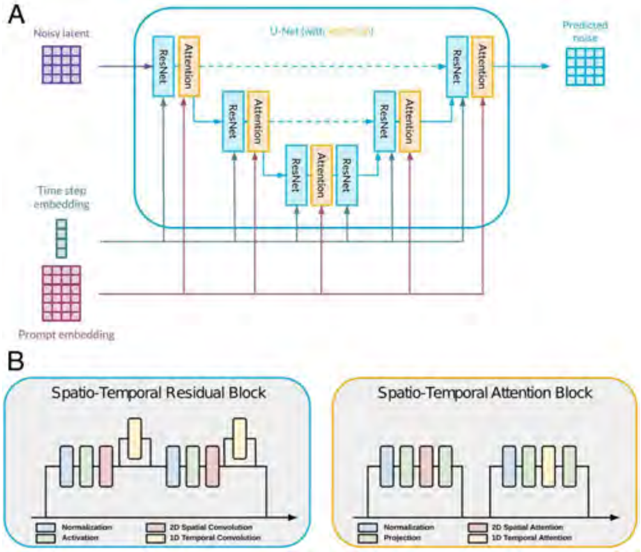


Figure 5: ***Denoising U-Net with Spatio-temporal Attention.*** *(A) Common attention U-Net used for the denoising step in image and video diffusion models comprising residual convolution blocks and attention blocks as well as concatenated time step and text prompt–embeddings. Figure adapted from [KDSD23]. (B) Temporal structure in video diffusion and editing is commonly modeled by adding 1D temporal convolutions in the residual blocks (left) as well as 1D temporal attention blocks after each of the 2D spatial attention blocks (right). Figure adapted from [ECA*23].*

Aiming to re-use learned image priors for video generation, Make-A-Video [SPH*22] builds their framework on a pre-trained T2I model, extending it to videos by adding spatio-temporal convolution and attention layers to the existing T2I model, followed by spatial and temporal super-resolution models. A key property of this approach is that each component can be trained separately: The T2I model is pre-trained over image-text pairs, while the entire T2V is fine-tuned on a large-scale corpus of *unlabeled* videos, thus bypassing the need to have video–caption-paired training data. This effectiveness of inflating and finetuning a T2I model for video generation has been also demonstrated in autoregressive Transformer-based models (e.g., [HDZ*22]).

These two components—using factorized/separable spatio-temporal modules (Fig. 5) and building upon a pre-trained text-to-image model—have been also used to expand an image Latent Diffusion Model (LDM), e.g., Stable Diffusion [RBL*22], to videos, i.e., learning the video distribution in a low-dimensional space (e.g., [BRL*23, LCW*23, ZWY*22, WYC*23]). Here too, compared to vanilla Video Latent Models (e.g., [HYZ*22, YSKS23]), leveraging a pre-trained image model allows efficient training (with less data), while harnessing the rich 2D priors learned by T2I LDMs. Another advantage of this approach is the ability to transfer the learned motion modules to other image models derived from the same base T2I. For example, plugging in a personalized tuned version of the T2I model thus synthesizing videos in a specific style or depicting specific objects [BRL*23, GYR*23].

As illustrated in Fig. 5, the factorized spatio-temporal modules expand both the convolutional block by a 1D convolution in time and also the self-attention block to model dynamics. A common approach to "inflating" self-attention (see Sec. 3) to multiple frames, aka extended or cross-frame attention, expands the self-attention across all or a subset of the frames of a video as

$$\text{Softmax}\left(\frac{\mathbf{Q}^i\left[\mathbf{K}^1,\ldots,\mathbf{K}^I\right]^T}{\sqrt{d}}\right)\cdot\left[\mathbf{V}^1,\ldots,\mathbf{V}^I\right], \qquad (12)$$

where $\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i$ are the queries, keys, and values of frame $i = \{1 \ldots I\}$. This inflation mechanism is prevalent among many methods discussed in this section. While it promotes temporal consistency, it does not guarantee it.

Other approaches for video generation include MCVD [VJMP22], which autoregressively generates videos in a 3D latent space by conditioning new frames on previously generated ones; VideoFusion [LCZ*23], which decomposes the noisy video latents into the sum of a (static) base noise shared across all frames, and a (dynamic) residual per-frame noise; and Generative Image Dynamics [LTSH23], which instead of directly predicting video content, learns to generate motion trajectories for pixels in an image, such that images can be animated to arbitrary length videos with oscillating motion.

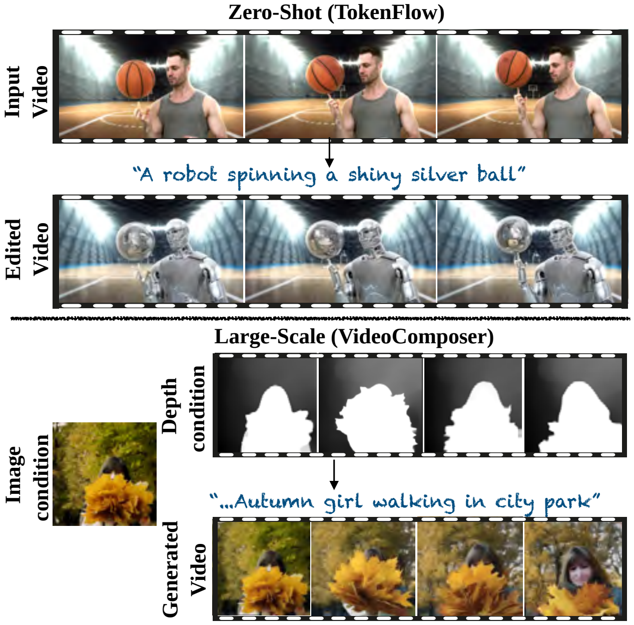

Zero-Shot (TokenFlow)
Input Video
"A robot spinning a shiny silver ball"
Edited Video
Large-Scale (VideoComposer)
Depth condition
Image condition
"...Autumn girl walking in city park"
Generated Video

Figure 6: ***Text-driven video editing.*** *Top: Given an input video, TokenFlow [GBTBD23]—a zero-shot method that leverages a pre-trained T2I model—enables consistent editing according to a given text prompt. Bottom: VideoComposer [WYZ*23]—a conditional video diffusion model trained on a large-scale video dataset—enables various controls, including conditioning video generation on a given image and per-frame depth maps.*

5.2. Controlled Video Generation and Editing

Similar to images, an important aspect in harnessing diffusion models for real-world content creation tasks, is the ability to provide users with controls over various attributes of the generated content, ranging from texture/appearance to editing motion and actions in video. While powerful methods have been developed to control various image attributes in T2I models, video editing poses additional challenges. First, any edit has to be applied in a consistent manner to all video frames. Second, while the community has developed rich and powerful representations for "static image attributes", the question of how to represent motion and time-varying signals in videos still remains open.

Conditional Video Models. One approach for controlled video generation is to design and train a conditional video model that directly takes control signals as input. Most existing works have been focused on video-to-video translation, where the overall motion and layout are extracted from a driving video. For instance, Runway's Gen-1 model [ECA*23] takes as input per-frame depth maps, which control the spatial layout of each frame, and a CLIP embedding, which controls the global appearance and semantic content of the video. Control-A-Video [CWX*23] and VideoComposer [WYZ*23] enable video generation conditioned on various other image-space control signals, such as edge maps, sketches or pose, by extending ideas from controlled image generation to the realm of videos [ZA23, HCL*23]. All the above models are based on inflated T2I models, which typically incorporate additional temporal layers, and are trained on large-scale video datasets. Similar concepts have been explored for more limited domain videos, for example models aimed at animating videos of humans, conditioned on an input image and a sequence of target poses [KHWKS23]. MagicEdit [LYZ*23] proposes a modular framework, which only trains the added temporal layers and not the pre-trained T2I model. Similar to AnimateDiff [GYR*23], this enables plugging the trained temporal layers into a given conditional T2I model (e.g., ControlNet [ZA23]) for controllable video generation.

Few-shot Methods for Video Editing. On the other side of the spectrum, a surge of methods suggest to leverage a pre-trained T2I model in a one-shot setting, i.e., by fine-tuning on a single test video (e.g, [WGW*22, LZL*23c]), or in a zero-shot setting, i.e., given no additional training data (e.g., [QCZ*23, CHM23, KMT*23, GBTBD23, FAKD23]).

For instance, Tune-A-Video [WGW*22] observes that simply extending the spatial self-attention in the T2I model from one image to multiple images produces consistent content and appearance across the generated images. Based on this finding, they design an inflated version of the T2I model, and finetune it on the test video.

At inference, given a text prompt, the model can be used to change the object category or stylize the video.

While tuning a T2I model on a single video yields surprisingly promising results, it is prone to overfitting (i.e., forgetting the T2I prior) and demands costly computation, limiting its use to generating text-driven variations of short, sub-sampled clips. Alternatively, a surge of video editing methods leverage a T2I model in a zero-shot fashion, by *directly manipulating its internal features*. Following feature-based image editing methods [TGBD23, HMT*22], Fate-Zero [QCZ*23] extracts the attention maps from an input video, and injects them into the T2I model during the generation of the edited video; this operation allows them to preserve the spatial layout of each frame, as discussed in Sec. 3; appearance consistency is achieved by extending the self-attention to operate on multiple frames, as in [WGW*22]. Concurrent works [CHM23, KMT*23] combine the self-attention extension with a conditional T2I (e.g, ControlNet [ZA23]), thus allowing to condition the generation on various spatial controls. However, achieving highly consistent edits remains challenging, as it is only implicitly encouraged via the inter-frame attention module.

Re-Render A Video [YZLL23] aims to improve consistency via a two-step approach by editing keyframes, and applying off-the-shelf video propagation methods to apply the edit on the rest of the frames. Their method heavily relies on accurate optical flow in both keyframe editing and propagation, which limits their use to rather simple motion and short clips. Concurrently, TokenFlow [GBTBD23] observes that consistency in RGB space is directly affected by the consistency of the diffusion features across frames. Thus, their method ensures temporal consistency by explicitly preserving the inter-frame feature correspondences of the original video. Notably, TokenFlow operates entirely in the diffusion feature space, thus does not exhibit optical-flow restrictions.

Neural Video Representations for Consistent Synthesis. Videos inherently contain highly redundant information across time, i.e., often share the same objects, scene background and overall appearance across frames. Thus, synthesizing or editing video content in a frame-by-frame manner is a tedious process that is susceptible to introducing temporal inconsistencies. This has motivated the development of *neural video representations* that facilitate effective and consistent synthesis and editing. Such representations range from video-to-layered-image decomposition to full 4D dynamic scene representations.

For example, representations like Layered Neural Atlases (LNA) [KOWD21] and Deformable Sprits [YLT*22] decompose a video into a set of layered canonical images (atlases), with corresponding per-frame deformation fields. This approach allows to significantly simplify video editing: 2D edits applied to a single canonical image are automatically propagated to all video frames. Text2LIVE [BTOAF*22] leverages LNA representation in conjunction with CLIP to perform text-driven video editing. This approach has been extended in [LJC*23] to leverage a pre-trained T2I model and to enable shape deformations. While temporal consistency is guaranteed by design, these methods typically require heavy optimization (hours of training per video). To tackle this issue, CoDeF [OWX*23] proposes an efficient hash-based representation to both canonical images and deformation fields.

To ensure both temporal and geometric consistency, Make-a-Video3D [SSP*23b] synthesize a video by leveraging a full 4D video representation, i.e., a 4D dynamic NeRF [CJ23], representing the 3D position of each scene point, and its 3D motion throughout the video. Given a target text prompt, this representation is optimized using a score-distillation-based objective that combines 2D image priors learned by a pre-trained T2I, and motion priors learned by a pre-trained T2V (see more details in Sec. 7).

6. 3D Diffusion

Beyond videos, the advent of diffusion models has also ushered in a transformative era for visual computing in the domain of 3D content generation. Although videos can be seen as 3D data (2D frames stacked sequentially in time), here we use the term “3D” to refer to spatial structure, i.e., 3D geometry. While 3D scenes can be visually presented as 2D video content by rendering a scene from multiple viewpoints, the rendered videos differ from general video content in that they only contain camera motion (as long as there exists a single 3D-consistent scene geometry). Generating this scene geometry (and the accompanying appearance or texture) is the primary focus of 3D generative tasks.

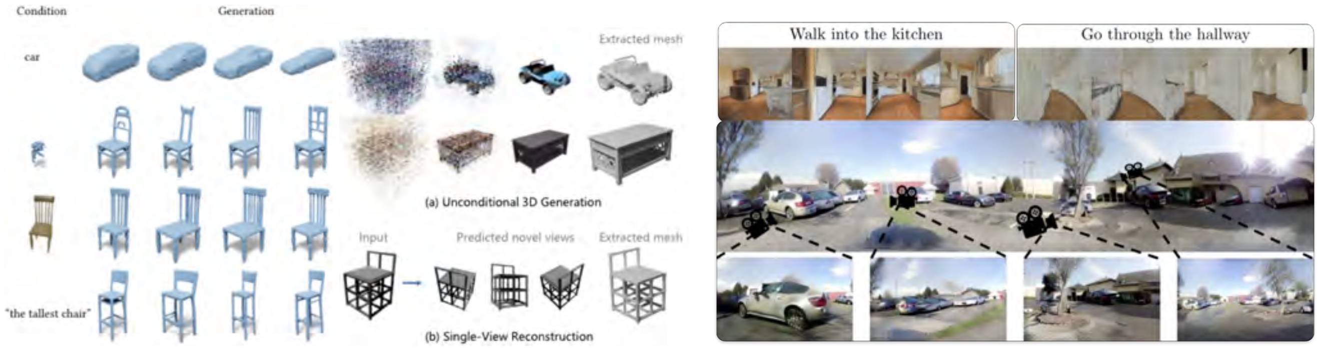
As elaborated in Sec. 4, the application of diffusion models to higher-dimensional data faces inherent difficulties, with data scarcity being particularly acute in the 3D domain. Existing 3D datasets are not only orders of magnitude smaller than their 2D counterparts, but they also exhibit wide variations in quality and complexity (see Sec. 8). Many of the available datasets focus on individual objects with limited detail and texture, limiting their generalization and usefulness for practical applications. Fundamentally, this issue stems from the fact that 3D content cannot be captured (or created) with the same ease as an image or video, resulting in orders of magnitude less total data. Furthermore, the culture for sharing 3D content is nowhere near that for 2D, exacerbating the scarcity of annotated data.

A unique challenge specific to 3D diffusion is the lack of a standardized 3D data representation. Unlike in image and video processing, where consensus has emerged around certain data formats, the 3D field is still wrestling with multiple competing representations, i.e., meshes, points, voxels, and implicit functions, each with its own merits and limitations, and each necessitating different network designs. Yet, none has emerged as a definitive solution that balances scalability with granularity.

To navigate these complexities, this section considers two distinct approaches: 1) *Direct 3D Generation via Diffusion Models*, which aims to model the distribution of 3D shapes, such that 3D content can be generated directly, putting geometry at the forefront and potentially serving as a foundation for tasks like 3D reconstruction and shape retrieval; and 2) *Multiview 2D-to-3D Generation via Diffusion Models*, which offers a more practical route by leveraging high-quality 2D generative models to create textured, consumer-ready 3D content. Both approaches offer valuable insights and capabilities, each addressing different facets of the challenges and opportunities in 3D content generation via diffusion models.

Output	Method	3D Repr.	Diffusion Repr.	Latent Struct.	Diffusion Arch.	Supervision	Hierarchical	Optional Conditioning	Data
object geom.	DPM [LH21]	points	points	-	PointNet	3D	✗	NA	ShapeNet
	PVD [ZDW21]	points	points	-	PVCNN	3D	✗	d	ShapeNet
	NeuralWavelet [HLHF22]	TSDF grid	wavelet coefficients	-	3D U-Net	3D	✓	NA	ShapeNet
	LAS-Diffusion [ZPW*23]	Occ. & SDF grid	Occ. grid & SDF octree	-	3D U-Net	3D	✓	k, c	ShapeNet
	LION [ZVW*22]	points	latents	points	PVCNN	3D	✓	s	ShapeNet
	SDFusion [CLT*23]	TSDF grid	latents	voxel	3D U-Net	3D	✗	s, i, t	ShapeNet, BuildingNet, Pix3D, Text2shape
	HyperDiffusion [EMS*23]	SDF	net weights	1D	Transformer	3D	✗	NA	ShapeNet
	Diffusion-SDF [LDZL23]	TSDF grid	latents	voxel	3D U-Net	3D	✗	s, i, t	ShapeNet, Text2shape
	NFD [SCP*23]	Occ.	latents	triplane	2D U-Net	3D	✗	NA	ShapeNet
	3DShape2VecSet [ZTNW23]	SDF	latents	set	1D U-Net	3D	✗	s, c, i, t	ShapeNet, ShapeGlot
	Michelangelo [ZLC*23]	occupancy	latents	set	1D U-Net	3D	✗	i, t	ShapeNet 3D Cartoon Monster (not public)
object geom.+ appear.	Point-E [NJD*22]	colored points	latents	points	Transformer	3D	✓	t	proprietary
	Shap-E [JN23]	radiance field	net weights	1D	Transformer	3D	✗	t	proprietary
	3DGen [GXN*23]	textured mesh	latents	triplane	U-Net	3D	✗	i, t	ShapeNet, Objaverse
	DiffRF [MSP*23]	radiance field	latents	voxel	U-Net	3D U-Net	✗	i	Photoshape Chairs, ABO
	RenderDiffusion [AXF*23]	radiance field	latents	triplane	U-Net	2D	✗	i	FFHQ, AFHQv2, CLEVR, ShapeNet
	HoloDiffusion [KMVN23]	radiance field	latents	voxel	3D U-Net	2D	✗	i	CO3Dv2
	HoloFusion [KVNM23]	radiance field	latents	voxel	3D U-Net	2D	✓	i	CO3Dv2
	SSDNeRF [CGC*23]	radiance field	latents	triplane	U-Net	2D	✗	i	ShapeNet, ABO
scene geom. + appear.	GAUDI [BGA*22]	radiance field	latents	1D	1D U-Net	2D	✗	i, t, c	VizDoom, Replica, VLN-CE, ARKitScenes
	NF-LDM [KBY*23]	radiance field	latents	hybrid	1-2D U-Net	2+3D	✓	m	VizDoom, Replica, Carla, AVD (not public)

Table 1: ***Geometry Generation with Diffusion Models.*** *We divide the table into three sections corresponding to the generation of object-level geometry, object-level geometry and appearance, and scene-level geometry and appearance. The conditioning column uses t (text), i (image), d (depth map), k (sketch), m (segmentation map), c (category), s (partial or coarse shape) and NA (not applicable).*



(a) *Object-level geometry* **(b)** *Object-level geometry and appearance* **(c)** *Scene-level geometry and appearance*

Figure 7: ***Direct 3D Generation.*** *Representative examples of direct 3D generation via diffusion models. The examples from left to right depict the state-of-the-art results of the generation of object-level geometry [ZTNW23], object-level geometry and appearance [CGC*23], and scene-level geometry and appearance generation [BGA*22, KBY*23].*

6.1. Direct 3D Generation via Diffusion Models

Due to the aforementioned challenges inherent to 3D data representation and spatial reasoning, in the realm of "Direct 3D Generation via Diffusion Models", the design space is notably intricate, necessitating a nuanced exploration of various design factors that significantly distinguish existing methodologies. This has led to a diverse array of design choices, each with its own merits and limitations. The ensuing discussion will systematically explore these pivotal design vectors, elucidating their impact on the quality and applicability of generated 3D content. Table 1 provides a comprehensive summary of the methods reviewed herein.

Type of Output. The first way to distinguish different methods is to look at the type of output they generate. Current methods either generate object-level geometry, object-level geometry and appearance, or scene-level geometry and appearance (see Fig. 7 for an example of the output). The choice of output type is often dictated by the type of data that is available for training. Many works [LH21, ZDW21, HLHF22, ZPW*23, ZVW*22, CLT*23, EMS*23, LDZL23, SCP*23, ZTNW23] make use of the few existing large-scale 3D datasets, such as ShapeNet [CFG*15], which include object-scale synthetic 3D models with geometry and material provided. These works use these datasets to investigate various design choices to adapt 2D diffusion models to 3D content; they have demonstrated the potential of diffusion models for 3D generation with satisfying geometric details—as the training data permits. However, due to the limited size, diversity, and complexity of the data, the applicability of these methods in practice is relatively low. Especially the lack of diversity and appearance information is a major limitation. To address this, several works, including [NJD*22, JN23, GXN*23, ZLC*23], have started to use more complex datasets, most notably [RSH*21, DSS*22, DLW*23] with in-the-wild objects and their appearance.

More recently, several authors have started to explore the generation of large indoor and outdoor 3D scenes. The scale and complexity of these scenes are much higher than the object-level datasets, though, the diversity is still limited. Interesting, likely due to the lack of accurate 3D ground truth and/or suitable architectures to process large-scale 3D structures, existing scene-level generation methods [BGA*22, KBY*23] opt to generate video renderings of the scene, which includes both the static 3D scene and the camera trajectory. While this is seemingly harder than generating the 3D scene only, due to the additional output mode, these approaches can conveniently hide low-quality geometry through visually plausible textures and appearance. Normally, the generated 3D scenes are coarse and incomplete, which leads to uncanny warping effects in the generated video, creating an illusion that the shape is deforming over time. In addition, the generated scenes are still relatively simple and lack the diversity and complexity of real-world scenes.

3D Shape Representation. Given the type of output, there can be multiple suitable representations, for example, point clouds, voxels, meshes, continuous or discrete samples of signed distance functions (SDFs), occupancy, and radiance fields. For object-level geometry, points, meshes, occupancy and SDFs—both as continuous functions or discrete samples—are common choices. When appearance is also generated, points and meshes can be augmented with per-point colors [NJD*22] and textures [GXN*23], and the occupancy and signed distance function can also be augmented with texture field to form some variations of radiance field [JN23, ZLC*23]. For scene-scale outputs, due to the lack of high-quality data in points and meshes representations and relatively easier acquisition in RGBD format, radiance fields become a more viable choice.

Diffusion Space. One very important way to differentiate the existing frameworks is based on the representation they employ inside the diffusion process. A natural choice is to apply the diffusion process directly to the 3D shape representation. The most popular representation in this category is points [LH21, ZDW21] and voxelized occupancy and SDF [ZPW*23], as the former can be conveniently linked to the physics interpretation of diffusion models with Langevin dynamics and the latter, deploying voxel structures, can take advantage of tested 2D diffusion architectures by simply adding a spatial dimension to all network operations. However, this representation simply captures raw and uncompressed information, it is very inefficient in terms of memory and computation and thus fails to capture high-frequency details with a fixed spatial resolution. A popular alternative which addresses this issue is applying the diffusion process on a more expressive and compact (latent) representation, extracted from the 3D data through some transformation. The transformation can be deterministic such as a wavelet transform used in Neural Wavelet [HLHF22], or learned as a high-dimensional latent code [ZVW*22, ZDW21, LDZL23, SCP*23, ZTNW23, NJD*22, JN23, GXN*23, ZLC*23, BGA*22, KBY*23] or the weights of the trainable neural implicit representation themselves, for example neural occupancy fields or SDFs, such as in [EMS*23, JN23]. These methods typically consist of two steps: First, a piecewise constant or smooth latent code is learned from the 3D data; Subsequently, a diffusion model is trained to denoise these latent codes, while the autoencoder model remains fixed. The smoothness of the latent space is crucial for the success of this approach. It is often achieved by imposing a regularization term on the latent codes, such as the total variation loss [FKYT*22, SCP*23], or deploying a variational framework such as (vector quantized) variational autoencoder [MCST22]. While the two-step approach has been predominant, a series of concurrently developed works [AXF*23, KVNM23, KMVN23, CGC*23] demonstrate how these two steps can be unified into end-to-end training, which is not only more efficient in a practical sense but also mutually beneficial for the two steps. We will discuss this in greater detail in the **Types of Supervision** paragraph.

Structure of the Latent Codes. While the simplest form of latent code takes the form of a 1D vector [EMS*23, JN23, BGA*22], most approaches opt for a more spatially aligned latent code to improve the locality and capability of the latent representation. This structure governs computational efficiency and memory requirements and, more importantly, it is coupled with the choice of diffusion architecture. In fact, the popularity of well-established convolution-based diffusion architectures has led to the widespread use of voxel grids [CLT*23, LDZL23, KVNM23, KMVN23]. Still using convolution, the triplane structure [CLC*22] is also a popular choice [SCP*23, GXN*23, AXF*23, WZZ*23], which factorizes a 3D volume to three axis-aligned orthogonal planes [PNM*20, CLC*22] and thereby significantly reduces the memory and computation cost.

In an alternative approach, some methods have adopted point clouds as the data structure for latent codes [ZVW*22, NJD*22]. In this paradigm, each latent code is associated with a specific point in the 3D space, thereby enabling localized information storage. The spatial locations of these points are either learned or inferred from the input shape during the initial latent learning phase. In a more abstract vein, 3DShape2VecSet [ZTNW23] eschews spatial information altogether by eliminating the coordinates from the latent codes. Consequently, the latent codes become a set of unbounded codes, offering greater flexibility to model long-range dependencies and self-similarities at the expense of spatial locality. These models often employ architectures like common point-processing networks such as transformers and point-voxel CNN [LTLH19], where the latter synergizes PointNet with voxel partitions to impose spatial locality. Despite its premise of sparsity and scalability, this type of structure has not been adopted for large-scale scene-level 3D generation.

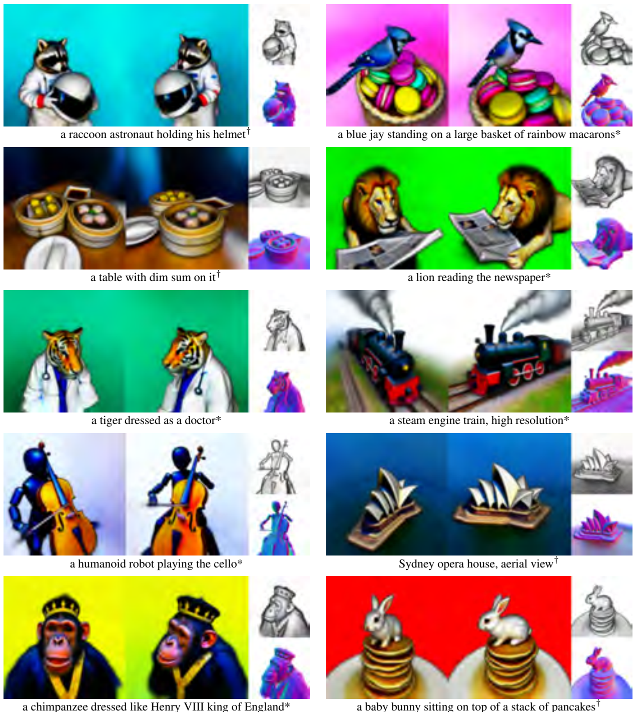
Hierarchy. One design choice inherited from 2D diffusion models is if a method employs a multi-stage (typically two-stage) diffusion process to achieve high-resolution generation while under memory and computation constraints. As exemplified in [HLHF22, ZPW*23, ZVW*22, NJD*22, KMVN23, KBY*23], the first stage generates a coarse representation and the second stage refines the output to a higher resolution. It is noteworthy that such a two-stage approach can use different representations for the lower and the higher resolution, for example [ZPW*23] uses an occupancy grid in the first stage and an SDF octree in the second stage.

Types of Supervision. While most of the discussed techniques try to learn 3D shapes or scenes from datasets of 3D shapes or scenes, a different approach is to follow the success of 3D GANs to train a diffusion model directly from datasets of 2D images. These methods can be considered as a special form of latent diffusion, in which the latent codes capture 3D information, yet the decoder converts the latent codes to 2D observations. Examples are RenderDiffusion [AXF*23], Holodiffusion [KVNM23], SSDNeRF [CGC*23], GAUDI [BGA*22], Viewset Diffusion [SRV23], and Diffusion with Forward Models [TYC*23]. In particular, the first four methods adopt an end-to-end training strategy, integrating the learning of the latent codes and the diffusion process into a single training process. As such, these methods are able to learn the latent codes that are most suitable for the diffusion process, while using the diffusion prior in-the-loop to regularize and improve the convergence of the latent learning. Notably, such an end-to-end approach has a significant meaning, as it eliminates the need to first train the ground truth "clean" 3D representation, which often requires 3D or dense 2D image supervision. Control3Diff [GGZ*23] merges diffusion models and GANs to enable training on single-view datasets, such as FFHQ and AFHQ. It leverages EG3D [CLC*22] to generate numerous pairs of control signals and tri-planes, employing a diffusion model with optional image guidance to learn the prior distributions of tri-planes and camera poses from input images. In future work, one could leverage more diverse large-scale 2D image datasets, such as LAION, which can be orders of magnitude larger than the largest 3D datasets.

Conditioning. Multiple types of conditioning have been used in the context of 3D diffusion models: text, images, depth maps, sketches, segmentation maps, shape category, or partial or coarse shapes. Most importantly, the task of text conditioning in these models presents unique challenges, primarily due to the scarcity of large-scale 3D datasets with corresponding text descriptions. Existing datasets, such as Text2Shape [CCS*19], ShapeGlot [AFH*19], PartIt [HLZH21], SNARE [TSB*22], and ShapeTalk [AHS*23], are predominantly confined to a limited range of object-level shape categories. This constraint significantly curtails the complexity and diversity of the generative models, thereby limiting their applicability. To circumvent the dearth of text-shape paired datasets, some studies have resorted to using images as surrogate data. These methods exploit the shared embedding space between text and images to condition the diffusion process. During the training phase, an image representation of the 3D shape is initially obtained through a rendering process. Subsequently, models such as CLIP are employed to derive an image embedding that serves as a conditioning variable for the diffusion model. At test time, a text prompt is processed through CLIP to obtain a text embedding, which can then be used interchangeably with the image embedding. This approach is agnostic to the choice of the underlying generative model. For instance, when initially proposed by Sanghi et al. [SCL*22], the generative model was based on normalizing flows. Diffusion-based 3D generative methods, such as those presented in [SFL*23, NJD*22, ZLC*23, JN23], have successfully adopted this strategy.

Datasets. Table 1 also includes the specific datasets that were used by the different papers. We do not provide an in-depth discussion of these datasets but refer the reader to Sec. 8 that reviews the most important datasets used by diffusion techniques overall.

Summary. The complexity of the design space in "Direct 3D Generation via Diffusion Models" stems from the lack of standardized data representations and architectures for 3D content. The 3D data landscape is characterized by its problematic sparsity coupled with a large spatial span, making it a challenging candidate for any single, standardized architecture capable of effective spatial reasoning. Consequently, researchers have ventured into a multitude of design choices, ranging from the type of output and 3D shape representation to the architecture of the diffusion model and supervision strategy. In this section, we have attempted to identify the most fundamental design choices explored by recent publications, shedding a light on their impact and the tradeoffs involved. Despite these advancements, it is imperative to acknowledge that 3D assets remain substantially less abundant compared to their 2D counterparts. While future work will be able to build on larger datasets, such as Objaverse [DSS*22] and Objaverse-XL [DLW*23], there is still a large discrepancy in dataset size between 2D and 3D diffusion models. This discrepancy underscores the need for innovative methodologies that can bridge this gap. In the ensuing section, we will explore how 2D diffusion models can be leveraged to facilitate the generation of 3D content.



a raccoon astronaut holding his helmet†

a blue jay standing on a large basket of rainbow macarons*

a table with dim sum on it†

a lion reading the newspaper*

a tiger dressed as a doctor*

a steam engine train, high resolution*

a humanoid robot playing the cello*

Sydney opera house, aerial view†

a chimpanzee dressed like Henry VIII king of England*

a baby bunny sitting on top of a stack of pancakes†

Figure 8: ***Text-to-3D Generation.*** *Example text-to-3D generations from DreamFusion [PJBM22] with corresponding input text-prompts. Each synthesized 3D model is rendered in two views with untextured renders and normals shown to the right. Image adapted from [PJBM22].*

6.2. Leveraging 2D Diffusion Models for 3D Generation

The exploration of 2D diffusion models for 3D generation stems from the notable advancements in image synthesis conditioned on text. These strides owe their success to abundant text-aligned image datasets and scalable model architectures. While similar techniques have been attempted for 3D generation (as described in the previous section), the scarcity of 3D data and the lack of well-explored denoising architectures present significant challenges. Notably, large-scale datasets such as LAION-5B [SBV*22], containing 5 billion text-image pairs, dwarf the largest available text-3D dataset, Objaverse-XL [DLW*23], with its mere 10 million paired samples.

To surmount the limitations imposed by limited 3D data and architectures, innovative strategies leveraging 2D image priors for 3D generation have emerged. These techniques distill multi-view geometry understanding, either implicitly learned from large-scale image diffusion models or explicitly via image diffusion models additionally conditioned on input camera poses and trained or fine-tuned on multi-view datasets.

6.2.1. Methods

Text-to-3D using Pre-trained Image Diffusion Models. Utilizing pre-trained image diffusion priors, DreamFusion [PJBM22] stands as a prime example, achieving groundbreaking text-to-3D generation. This innovative approach harnesses established image priors, allowing for zero-shot synthesis of intricate 3D objects, as vividly demonstrated in Fig. 8. The method revolves around optimizing a 3D model represented as a Neural Radiance Field (NeRF) through a specialized image space loss. This loss function, tailored to leverage pre-trained image diffusion models as priors, assigns lower values to plausible images, ensuring the coherence of generated 3D objects. This process of sampling through gradient descent optimization is known as Score Distillation Sampling (SDS). Formally, a 3D scene representation parameterized by ϕ is rendered by a differentiable generator g at a sampled camera pose, generating an image $g(\phi)$. Gaussian noise η is then injected into the image and passed through an image diffusion prior parameterized by θ, with text condition $\boldsymbol{c}$. 3D scene parameters are then updated via the gradient:

$$\nabla_{\phi}\mathcal{L}_{\mathrm{SDS}}(\theta, g(\phi)) = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_{\theta}(\boldsymbol{x}_t;\boldsymbol{c},t)-\eta)\frac{\partial g(\phi)}{\partial \phi}\right]. \quad (13)$$

Importantly, this technique has broader applicability, extending beyond 3D synthesis as explored by Hertz et al. [HACO23] for image editing. Crucially, this optimization through 3D model and differentiable rendering guarantees that each optimized image adheres to multi-view constraints, ensuring a coherent 3D model. This methodology, embraced by subsequent works such as Score Jacobian Chaining [WDL*22], Magic3D [LGT*23], and Latent-NeRF [MRP*23], maintains a similar philosophy while exploring diverse underlying representations for the optimized 3D model.

However, direct SDS application for 3D generation poses challenges, notably the *Janus problem*, where radially asymmetric objects exhibit unintended symmetries, like multiple faces on a human head. To mitigate this artifact, researchers in works like DreamFusion [PJBM22], Score Jacobian Chaining [WDL*22], Latent-NeRF [MRP*23], and Magic3D [LGT*23] employ strategies like view-dependent prompting, e.g., *"top-view of ..."*, and additional regularization terms. Alternatively, learning 3D representations from multi-view diffusion models ensures consistency across diverse views.

Additionally, SDS-based 3D generation demands operation at unusually high guidance scales, influenced by the mode-seeking behavior of the loss function. Operating beyond the typical range of the pre-trained image prior leads to saturation artifacts and limited diversity in outputs. ProlificDreamer [WLW*23] tackles these challenges through Variational Score Distillation, a generalized SDS version. This method incorporates finetuning of the image model using LoRA [HSW*21] during per-scene optimization, with the intention of mitigating both the Janus problem and the need for high guidance scales.

Adapting Image Models for Multi-view Synthesis. The efficacy of image diffusion models for 3D generation, as exemplified by DreamFusion [PJBM22], has prompted subsequent investigations. Follow-up studies have underscored the enhanced generation quality achievable through the finetuning of pre-trained image priors with 3D data. Zero-1-to-3 [LWH*23] finetunes a pre-trained text-to-image model to add camera pose conditioning. This finetuning is done on pairs of multi-view images rendered from the Objaverse dataset [DSS*22]. The resulting model takes as input an image and the camera parameters of a new viewpoint, and returns an image rendered at the novel view. This method has also been proven to work at scale, as Deitke et al. [DLW*23] train Zero-1-to-3-XL in a similar fashion using the larger Objaverse-XL dataset. This flavor of pose-conditioned image diffusion model can also be trained from scratch on multi-view image data, as shown by 3DiM [WCMB*22]. Performing DreamFusion-like 3D aggregation using a diffusion model with these added pose conditioning signals helps reduce the need for certain tricks like view-dependent text prompting. Unfortunately, these single-view models still suffer from many of the same artifacts as standard DreamFusion due to the fact that only a single image is queried a time (and, therefore, multiple queried viewpoints may propagate contradictory signals into the 3D model). Newer methods aim to resolve this: MVDream [SWY*23] finetunes a pre-trained image diffusion model to create a multi-view diffusion model, capable of generating a *set* of geometrically consistent images of the same object at four fixed camera poses, from an input text prompt. This is achieved through the addition of a 3D self-attention module trained on multi-view images rendered from the Objaverse [DSS*22] 3D dataset. The resulting multi-view diffusion model can be directly used for 3D generation through SDS. Since the multi-view diffusion model outputs images from four orthogonal azimuth angles, rather than from a single view at a time, this method provides a principled remedy to the aforementioned Janus problem. SyncDreamer [LLZ*23] approaches multi-view diffusion by grounding features from each generated view into an explicit 3D feature space. Using a 3D-aware attention mechanism, SyncDreamer synchronizes the intermediate states between the diffusion paths across different viewpoints, establishing 3D correspondence between them.

3D-Aware Image Diffusion. Although finetuning higher-dimensional models on top of large pre-trained T2I models helps simplify the training process, the choice to build off a pre-trained model can limit control over the architectural design space. Another line of work explores training models from scratch, such that architectures may be 3D-aware or use more explicit multi-view reasoning to provide conditioning signals. GeNVS [CNC*23] tackles single-image novel-view synthesis by training a 3D-aware conditional diffusion model that incorporates geometric priors in the form of a 3D feature volume obtained from an input image (or images, when performing autoregressive generation). A feature image rendered from the feature volume at the novel viewpoint provides 3D-aware conditioning to the diffusion model. NerfDiff [GTL*23] takes a similar approach, providing PixelNeRF [YYTK21] renderings at novel views as conditioning to a 3D-aware diffusion model. SparseFusion [ZT23] and Sparse3D [ZCC*23] use an epipolar feature transformer [SESM22] to provide a conditioning signal to a finetuned image diffusion model.

In a separate vein, 3D-aware image generation has also been explored through the perpetual view generation problem introduced by InfiniteNature [LTJ*21, LWSK22]. For example, given only a text prompt, SceneScape [FAKD23] generates 3D-consistent videos depicting static scenes rendered from a specified camera trajectory. To achieve this, they leverage a pre-trained T2I model, and construct a unified mesh representation of the scene, along with the video generation process. DiffDreamer [CCP*23] trains a conditional diffusion model that takes a single input image and generates renderings of a specified camera trajectory flying into the scene. Although these tasks seem similar, flying into a scene is more challenging because this problem not only requires 3D-multi-view consistent out-painting but also super-resolution, as novel details of a previously seen structure become visible in later frames.

6.2.2. Applications

3D Editing. Following the success of InstructPix2Pix [BHE23] for instruction-based image editing, InstructNeRF2NeRF [HTE*23] achieves similar results for editing 3D scenes. Similar to SDS, the diffusion model (in this case, a text-and-image conditioned model that edits images) is queried repeatedly through optimization, and the model outputs are propagated back into a the NeRF scene. A few modifications are made to the standard SDS formulation, however. First, the diffusion model is queried not once, but rather multiple times, such that the fractionally noised image can be sampled to a clean output image (as in [ZT23]). Then, to derive a loss, instead of directly comparing the noise estimate to the injected noise, the clean image is compared to the original scene rendering. Finally, this loss is not applied for one image at a time (as in DreamFusion), but rather at a randomly shuffled set of rays from the set of captured views (which is standard in NeRF optimization). This effectively amounts to iteratively updating the dataset of images used for training the NeRF, and is similar in spirit to the method propose in SNeRF [NPLX22] for 3D style transfer.

Fig. 9 shows example edits of the same initial 3D scene with the corresponding text-based instructions. InstructNeRF2NeRF can handle a diverse set of instructions while performing multi-view consistent 3D edits. While InstructNeRF2NeRF enables diverse holistic and contextual edits with high fidelity, it can often be bottlenecked by the performance of InstructPix2Pix: If a certain edit is not possible in 2D or is too inconsistent, it will not be reflected in 3D. Similarly, one occasional failure mode of InstructPix2Pix is over-editing, i.e., when unintended parts of the scene are modified. To resolve this, and have guarantees on localized editing, one may inherit insights from methods like DreamEditor [ZWL*23] and Vox-E [SFHAE23] that can detect the region to be edited using the underlying attention maps from the diffusion model.

Scene Generation. Locally conditioned diffusion (LCD) [PW23] enables controllable 3D generation with intuitive user inputs. LCD transforms user-defined bounding boxes with corresponding text prompts and generates 3D scenes matching the desired layout. This is achieved using a modified SDS-based pipeline that takes the user input as conditioning. The method allows explicit control over the size and position of individual scene components while ensuring seamless transitions between them. Text2Room [HCO*23] generates textured 3D meshes of indoor rooms from a given text prompt. Although it also leverages a pre-trained image prior, 3D synthesis is not enabled through SDS. Instead, a pre-trained monocular depth estimator is used to provide 3D information. The method generates a textured 3D mesh by iteratively inpainting the scene at randomly sampled camera angles. During each inpainting step, a monocular depth estimator is used to create a 3D mesh of the new scene content, which is then merged with the rest of the existing geometry.

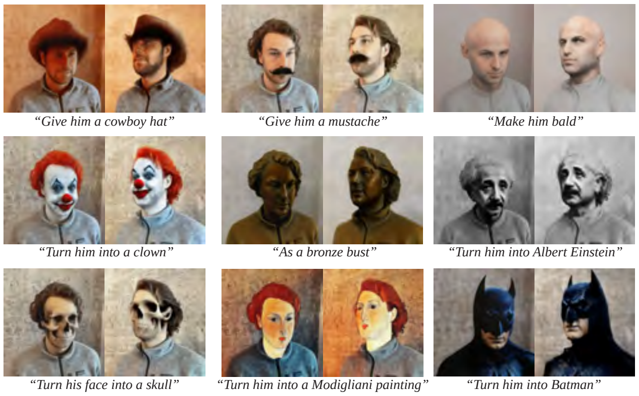


Figure 9: ***Instruction-based 3D Editing***. *Example instruction-based 3D edits from InstructNeRF2NeRF [HTE*23] with corresponding text-prompts. All edits are performed over the same input reconstructed NeRF scene. Image reproduced from [HTE*23]*

7. Towards 4D Spatio-temporal Diffusion

The capabilities of video generation and editing tools are rapidly advancing (see Sec. 5), but the underlying T2V models often lack a mechanism to model long-range temporal consistency or consistent 3D structure of their objects and scenes. 3D generative tools (see Sec. 6) are specifically designed to enable spatial reasoning. Yet, these tools are limited to modeling static scenes with rigid (camera) motion. We consider video generation and editing that have an understanding of the underlying 3D structure of a (dynamic) scene as 4D in the sense that they combine 3D structure and time, including non-rigid deformation and general motion. In this realm, we discuss articulated object and avatar generation, motion generation, and more general 4D generation and editing tools in the following.

7.1. Articulated Avatars, Animals, and Objects

Articulation refers to a class of non-rigid motions in which the deformation of an object is composed of locally rigid transformations. Many objects in everyday life are articulated, including human faces and bodies as well as animals. Articulated objects can be intuitively controlled in a manual manner or using motion capture, for example using skeleton-based or template-mesh-based handles. This set of techniques has been popular and broadly applied in the field of computer graphics for the last few decades.

Generating articulable 3D avatars from a given text prompt is a popular line of research in this area. This class of methods often generates 3D digital humans by constraining the outputs to follow a given parameterized model. For example, DreamHuman [KAZ*23] generates articulable human avatars using imGHUM [AXS21], an implicit statistical 3D human pose and shape model. Bergman et al. [BYW23] synthesize articulated 3D heads by optimizing the geometry and texture of a 3D morphoable model and TADA! [LYX*23] optimizes over a human model derived from SMPL-X [PCG*19]. DreamFace [ZQL*23] is a CLIP-based selection approach for animatable head avatar generation from text prompts. It first chooses coarse proxy geometry and then refines it for consistency with the text prompt using SDS loss and adds the hair. TECA [ZFK*23] synthesises 3D human avatars composed of mesh-based head (generated first) and other elements based on NeRF (hair, garments and accessories; added afterwards using SDS loss and guidance from segmentation) from text descriptions; the texture is inpainted with diffusion models. Animations of the generated avatars can be performed leveraging a human parametric model [PCG*19]. DreamAvatar [CCH*23] and AvatarCraft [JWZ*23] also focus on the creation of articulable avatars from text prompts and support body articulation. 4D Facial Diffusion Model [ZFY*23] generates facial expressions of a template face mesh. The AvatarStudio method [MPE*23] creates text-guided stylizations of high-quality 3D neural face avatars using a new view- and time-consistent score distillation sampling.

Diffusion-based generation of more general 3D articulated objects has been explored by NAP [LDS*23]. This system uses a tree parameterization for modeling the irregular distribution of articulated objects across different datasets and supports unconditional and conditional articulated shape generation (e.g., conditioned on object parts or joints). ARTIC3D [YRH*23] leverages a generative 2D diffusion model prior to learning articulated and animatable 3D animal shapes. Given a sparse and (not curated and without 2D or 3D annotations) image collection containing 10–30 images of an animal species, ARTIC3D estimates camera viewpoints, pose articulations, part shapes and texture for each observed instance; no shape initializations, pre-defined 3D skeletons or shape templates are required, in contrast to previous works in this area [KTEM18, STG*20, YSJ*21, YHL*22, WLJ*23].

7.2. 3D Human Motion Generation

In recent years, deep learning has made remarkable strides in advancing character animation. Thanks to the availability of large human motion datasets to the public, researchers in character animation have been training data-driven models to predict character motion, conditioned on the motion history, agent observations or various forms of user commands. However, a persistent challenge in developing predictive motion models is that the mapping between input conditions and output motions is rarely one-to-one. For example, when a character is instructed to approach a chair, they have multiple options, such as walking around it, moving it, or sitting on it. This challenge necessitates the adoption of generative models. Indeed, generative models provide a principled solution to learn a conditional data distribution, making them well-suited for modeling many-to-many mappings inherent in human motion synthesis.

A significant body of work has explored approaches based on Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to synthesize motion trajectories [BKL18] or train auto-regressive motion priors [LZCVDP20, RBH*21]. While these methods have shown varying degrees of success, the full benefits of generative models were not always realized due to issues like mode collapse, posterior collapse or training instability. However, in early 2022, diffusion-based approaches entered the field of character animation and rapidly became the preferred choice among generative models. Unlike Conditional VAEs (CVAE) and GANs, diffusion models excel at modeling multimodal distributions and offer ease of training. Additionally, diffusion models allow for flexible motion editing during inference using techniques such as "In-painting", which aligns well with the needs of digital content creation across a wide range of computer graphics applications.

Nevertheless, the successful application of diffusion in motion synthesis heavily relies on the quality and quantity of training datasets. Furthermore, for conditioned models, a demanding prerequisite is the existence of paired training datasets that establish a clear correspondence between the conditioning information and 3D human motion. These constraints can indeed limit the applicability of diffusion models. Currently, various human motion datasets exist, coupled with diverse conditions, including text [GZZ*22, GZW*20], music [LYRK21], audio [FM18], video [IPOS14, MRC*17], scene descriptions [ALV*23, GMSPM21,ZYM*22,YK19,ZMZ*22] and objects [TGBT20,FTT*23, BXP*22, LWL23]. Continuing to develop and share such large-scale, diverse, and high-quality paired motion datasets is a valuable investment in the relevant research and development communities.

Different motion representations have been used in prior work without a clear advantage of one over the other. While all assume an underlying body model, some use a minimal representation such as the 6D joint rotations [TCL23, KKC23], while others allow a redundant representation by having both positions and rotations [TRG*23, CJL*23]. It is also a common practice to include contact information as part of the motion representation to address concerns like foot sliding or surface penetration [TCL23]. Some studies have suggested employing redundant representations and using self-consistency to improve learning of diffusion models [TRG*23]. Furthermore, learning an encoded motion representation has also shown benefit in training diffusion models [CJL*23, JCL*23, AZL23].

Despite the initial success of diffusion-based approaches in the early exploration phase, several recurring issues have surfaced in motion synthesis applications. Notably, diffusion models, like most machine learning techniques, lack an inherent ability to precisely satisfy constraints, often resulting in motion artifacts such as violations of physical and geometric properties, or failure to follow the control signals or specifications provided as input. Several proposed methods aim to address this concern, including gradient guidance [RLBP*23, KPST23], learning of conditions via supervision [LWL23], direct editing during inference [TRG*23], or refining the model through reinforcement learning processes [HPD*23]. While these methods have shown varying degrees of success in controlling output motion, they cannot guarantee precise constraint satisfaction and controllability.

The remaining part of the subsection highlights a few active research directions in motion synthesis and the representative papers that advanced the areas.

Motion Generation Conditioned on Time-series. A number of recent works focus on generating motion synthesis from a time-series input, such as audio, music, or text. EDGE [TCL23] is a transformer-based diffusion model paired with a pre-trained music feature extractor Jukebox [DJP*20]—acting as cross-attention context—for realistic dance motion generation. EDGE offers versatile editing functionality and arbitrary long sequences. This is possible by replacing the known regions with forward-diffused samples of the provided constraint (through masking), which is an increasingly popular technique enabling editability at test time without the need for model retraining. Listen, Denoise, Action! [ANBH23] is a method based on a stack of Conformers [ZLC*22] for dance motion generation from audio signals (see Fig. 10). It takes acoustic feature vectors and an optional style vector as inputs. The method enables control over motion style with classifier-free guidance and dance style interpolation thanks to the product-of-experts ensembles of diffusion models (guided interpolation of different probability distributions). MoFusion [DMGT23] conditions human dance generation on Mel spectrograms. The advantage of this method is that the context embedding layer of the MoFusion architecture learns a suitable injection of the audio signal to the feature space of a U-Net, in contrast to MFCC features, for example, which offer less flexibility.

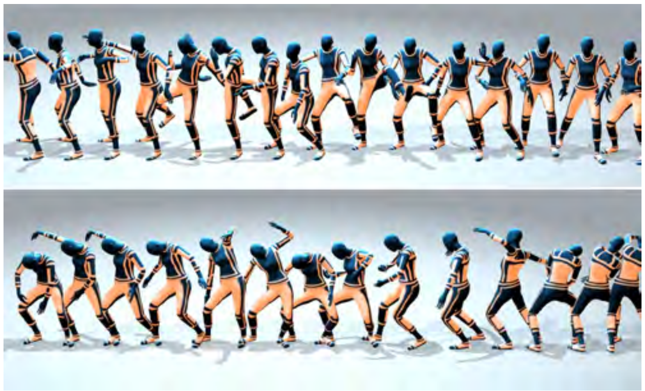


Figure 10: ***Motion Generation.*** *Dances synthesised from audios by the approach of Alexanderson et al. [ANBH23] for Locking (top row) and Krumping (bottom row) dance styles. Image reproduced from [ANBH23].*

Motion Generation with Spatial Constraints. Integration of spatial constraints into 3D human motion generation with diffusion models is another active research direction. GMD [KPST23], for example, offers two elaborate mechanisms to allow spatial conditioning including trajectory, keypoints and obstacle avoidance objectives. NIFTY's architecture [KRG*23] includes an object interaction field that guides the motion generator to enable the support of human-object interactions. AGRoL [DKP*23] is an MLP-based architecture with a conditional diffusion model for full-body motion synthesis accepting a sparse upper-body tracking signal. The method supports real-time inference rates, which places them among only a few techniques well-suitable for AR and VR applications. InterGen [LZL*23a] is a model generating interactive human motions. The paper considers the case of two virtual characters constraining each other's movements. InterGen's objective is reflected in its architecture: It contains cooperative denoisers that share weights and a mutual attention mechanism, which improves the motion generation quality. Additional interaction regularizers (losses) with a damping schedule allow modeling of complex spatial relations between the generated motions. Similarly, Shafir et al. [STKB23] train a slim communication layer to synchronize two pre-trained motion models using only a few training examples.

PhysDiff [YSI*23] goes a different way beyond geometric priors, namely physics-guided motion diffusion: It adds a physics-based motion projection step into the diffusion process of existing methods that mitigates such artefacts as floating in the air, foot-floor penetration and foot sliding. The projection step of PhysDiff is a human motion imitation policy controlling a virtual character in a physical simulator and enforcing the physical constraints.

Synthesis of Long Motion Trajectories. Being able to generate arbitrarily long motion sequences is crucial to many online applications. The TEDi approach [ZLAH23] entangles the time-axis of diffusion and temporal motion axis: In each diffusion step of their stationary motion buffer (aka pre-defined motion generation window), they remove a clean frame at the beginning of the buffer and append a new noise vector to its end. A different approach is pursued by DoubleTake, [STKB23] that extends MDM for long-term motion synthesis without additional training. Their sequential composition approach concatenates short motion sequences with a diffusion blending step (performed in a zero-shot manner): In the first "take", their method generates motion batches and the second "take" refines the transitions between individual motions. Their motion prior is trained from short 3D motion clips only, and the method allows individual control of each motion interval. MoFusion [DMGT23] can generate long motion using seed conditioning in a sliding window fashion. The provided seed motion frames are first corrupted by noise, while the remaining ones that need to be generated are initialised with random noise. At each denoising step, the seed frames are masked out, which eventually leads to generated motions complementing the seed motions. A similar policy is also applicable to MotionDiffuse [ZCP*22], MDM [TRG*23] and EDGE [TCL23]. In Alexanderson et al. [ANBH23], translation-invariant policy to encode positional information enables better generalisation to long sequences.

Motion Generation Based on Multiple Modalities. Many motion generation algorithms leverage inputs with multiple modalities. For example, gesture generation approaches use speech as a condition along with text for style; the attention mechanism is leveraged to synchronise the gestures to the speech. The approach of Deichler et al. [DMAB23] proposes a contrastive speech and motion pre-training module that learns joint embedding of speech and gesture; it learns a semantic coupling between these modalities. DiffuseStyleGesture [YWL*23] is an audio-driven co-gesture generation approach that synthesises gestures matching the music rhythm and text descriptions based on cross-local and self-attention mechanisms. It uses classifier-free guidance to manipulate the initial gestures and interpolate or extrapolate them. EMoG [YWH*23] decomposes the generation problem into two steps, i.e., joint correlation and temporal dynamics, and shows that explicitly predicting joint correlation improves generation quality. DiffMotion [ZJGL23] is a two-stage framework for speech-driven gesture synthesis. Its auto-regressive temporal encoder (LSTM-based) conditions the diffusion module on temporal dynamics extracted from the preceding human poses (gestures) and acoustic speech features.

7.3. 4D Scene Generation and Editing

4D Scene Generation. General 4D scene generation implies that no strong priors are used about types of objects and possible non-rigid movements, or weak priors are combined with stronger ones (e.g., in a compositional setting). We discuss two approaches in this category that address this most general setting among all, i.e., MAV3D and Learnable Game Engines (LGE).

MAV3D [SSP*23b] (see Fig. 1, top-right) extends DreamFusion [PJBM22] with the time dimension for non-rigid NeRF scenes generated from text. Similar to the 3D case, it is difficult to acquire large datasets of 4D scenes that could be used to train non-rigid scene generators, especially with paired textual annotations. Hence, MAV3D relies on a pre-trained text-to-video diffusion model [SPH*22] (see Sec. 5) as a non-rigid "world scene prior" that provides learnt distributions of multi-view scene projections (this is in contrast to [PJBM22] using a text-to-image model). LGE [MSL*23] is a neural model with game-engine-like features learned from annotated videos. It follows a scene composition approach with the elements of playability (in the sense of Playable Environments [MLS*22]), scene editability and learning game engine functionality from data. LGE supports the generation of compositional 4D NeRFs (e.g., of a tennis match or a Minecraft-like game) conditioned on a single RGB image and a wide spectrum of conditioning temporal signals (such as learned human actions, object locations, and textual action descriptions). In LGE, the diffusion-based animation module predicts scene states conditioned on user-provided actions, and the synthesis module renders them from desired viewpoints. Compared to MAV3D, LGE generates scenes of higher visual fidelity and with finer-grained control, at the cost of scenario-specific datasets with textual annotations.

4D Editing. Control4D [SSP*23a] can edit 4D human portraits from text inputs provided as implicit Tensor4D [SZT*23] scene representations. The method utilizes a 2D diffusion-based prior (editor) to train a 4D GAN that is applied at test time to the rendered 2D views of the dynamic (animated) portrait scenes. Note that the diffusion-based editor does not directly operate on the rendered images, and this design choice is done for the 4D GAN to ensure that the edits are temporally consistent. While the results of Control4D are also photo-realistic, it can be noticed that they preserve only a few characteristics of the original identity, i.e., the edits are dominated by the textual prompts, especially in the head area. A critical reader might ask to what extent the input portraits influence the edited results and whether they could not have been obtained by further refining the textual prompts.

This aspect is substantially improved in another work titled AvatarStudio [MPE*23] that performs editing of human head avatars in a photo-realistic and temporally consistent manner while preserving the initial identity (see Fig. 1, second row on the right). AvatarStudio accepts as input a high-resolution 360° non-rigid NeRF and uses an LDM [RBL*22] fine-tuned with viewpoints and time stamps randomly sampled from the NeRF volume as well as view-and-time aware SDS loss with a model-based guidance. The method edits the full head performance in the canonical space and propagates the edits to all time steps via a pre-trained deformation network. AvatarStudio enables personalized and photo-realistic edits that preserve the initial head identity while transferring (or mixing in) the visual appearance features associated with the prompted identity (e.g., “Vincent van Gogh” or the literature/movie character “Cruella de Vil”) or qualitative scene descriptions (e.g., “bronze bust”, “marble statue” or “blue hair”).

8. Data

In this section, we briefly review datasets commonly used for training and evaluating 3D and 4D diffusion models.

Image Datasets. Image datasets play a pivotal role in the training and validation of diffusion models for visual computing applications. High-quality, diverse, and large-scale image datasets, accompanied by rich semantic information, such as class or semantic (instance) labels [DDS*09, LMB*14], ensure that trained models can generalize across many image modalities. As the demand for labeled images grows, the scale of such datasets continues to expand. In particular, the combination of image and text data has been an important source of training data in the context of diffusion models, since such data can be crawled from the web at large scale and typically does not rely on manual annotations. Notable large-scale image datasets include the Wikipedia-based Image Text (WIT) [SRC*21], a curated set of 37.6 million rich image–text examples with 11.5 million unique images across 108 Wikipedia languages, the LAION-400M open dataset [SVB*21] of 400 million CLIP-Filtered image-text pairs, and its successor LAION-5B [SBV*22] which expands the collection to 5.85 billion.

Video Datasets. Publicly available video datasets with textual descriptions, such as WebVid-10M [BNVZ21] and HD-VILA-100M [XHZ*22], are of great importance for training video diffusion models. These datasets contain 10M and 100M text–video pairs, respectively, which is more than an order of magnitude smaller than available text–image datasets. For this reason, many T2V models are trained with both images and videos, treating images as individual video frames. Alternatively, a pre-trained T2I model can be refined with a smaller amount of training data (see Sec. 5 for more details). At the same time, these datasets can be augmented with the vast amount of unlabeled video data available on the web.

Shape Datasets. In contrast to their image-based counterparts, 3D datasets are often constrained by the paucity of training samples, primarily due to the high cost associated with obtaining 3D models. A key dataset in this domain is ShapeNet [CFG*15], which comprises 51.3k models across 55 object categories and has been enriched with part segmentation and textual annotations [AFH*19,CCS*19,KHA*22,AHS*23]. Another widely-used dataset is Amazon Berkley Objects (ABO) [CGD*22], which emphasizes texture quality and encompasses 8k 3D models from 63 classes, serving as a training ground for object-level geometry and appearance generative models. While these datasets offer synthetic models that provide abundant training data and ground-truth geometry, they also introduce a synthetic-to-real gap. To mitigate this, CO3D [RSH*21] and OmniObject3D [WZF*23] acquired large-scale real objects. While CO3D contains 19k objects, it only provides multi-view captures, whereas OmniObject3D offers 6,000 high-quality scanned 3D objects across 190 categories, complete with textured mesh and point clouds. However, the most ambitious endeavor in this space of object-level 3D data is Objaverse and Objaverse-XL [DSS*22, DLW*23], boasting over 10M 3D models sourced from Sketchfab [Ske23], accompanied by a large-scale text corpus. Despite its unprecedented scale, the dataset presents challenges due to the heterogeneous quality of 3D models and text descriptions, as well as non-uniform data distribution, thus complicating the effective utilization of these large-scale resources. For a scene-scale generation, data collection is even more challenging. While seminal datasets such as ScanNet [DCS*17], Matterport3D [CDF*17], ScanNet++ [YLND23], and RealEstate10k [ZTF*18] rely on 3D scanning real environments, others focus on building on crowd-sourced web designs for scene modeling, such as 3D Front [FCG*21]. In practice, however, many existing approaches in this space do not (yet) generate an actual 3D scene but rather a video rendering of the scene, which requires the camera trajectory and the corresponding RGB(D) images for supervision. VizDoom [KWR*16], Replica [SWM*19], Carla [DRC*17], VLN-CE [KWM*20], and ARKitScenes [BCD*21] are some examples of such datasets, where the training sequence is generated through a simulated or captured camera motion in synthetic or reconstructed static 3D scenes. However, these datasets are still severely limited in diversity and annotations. The autonomous driving industry is in an ideal position to capture diverse real-world data; however, such data is typically treated as a proprietary asset and not publicly available. We also provide references to smaller datasets that have been used in approaches mentioned in Sec. 6.1: BuildingNet [SNL*21], Pix3D [SWZ*18], Photoshape [PRFS18], CLEVR [JHVDM*17]. These datasets complement ShapeNet in terms of geometric scale and complexity, as well as rendering realism, and annotation availability.

4D and Human Motion Datasets. Another important set of datasets are those focusing on 3D capture in motion [LTT*21, BZTN20]. Although the human body spans many possible motions, 3D human motion is expensive to acquire, and hence datasets remain scarce. In the context of humans, most current sizable datasets are based on AMASS [MGT*19], which is a collection of other datasets acquired by different technologies, all aligned to a uniform representation [LMR*15]. In total, these recordings contain over 40 hours and comprise 11k sequences; however, this alignment does not include any labeling. Hence, other datasets [PMA16, PCA*21, GZZ*22] have sought to textually annotate it. These datasets include richer and cleaner data, with more elaborate textual descriptions. Although already large, these datasets still do not capture the full expressiveness of human motion. A more elaborate dataset has recently been released [LZL*23b], consisting of an order of magnitude more data, and more expressive annotations that include hand motions and facial expressions. As literature exploring the capabilities of this dataset is still in its infancy, it is too soon to predict the impact it will have on the field. At the same time, higher-fidelity animation sequences from 3D scans or multi-view capture setups have been released in the context of human faces [GKG*23, KQG*23] and bodies [PZX*21, IRG*23, LHR*21, HLX*21, HXZ*20, XCZ*18]. Although these datasets are of high quality for each captured instance, the costly recording process limits the largest of such datasets to several hundred captured people. A new dataset with audio and high-quality 3D motion capture with various dance genres for 3D human motion generation tasks was introduced in Alexanderson et al. [ANBH23]. AIST++ [LYRK21] is another dataset of people dancing widely used in 3D human motion generation.

The CIRCLE dataset [ALV*23] utilizes optical motion capture and virtual reality to capture 3D human motions in cluttered indoor environments, paired with an egocentric view of the scene. A very large dataset of speech-aligned 3D face, body and hand motion extracted from talk show host videos is presented in [HXM*21]. Finally, the OMOMO dataset [LYC*23] contains more than 10 hours of full-body manipulation of various objects, including pushing furniture, carrying household objects, manipulating cleaning equipment and more.

9. Metrics

This section briefly discusses the metrics used for evaluating the reviewed methods with diffusion models.

Image Quality and Diversity. Robust evaluation of image diffusion models requires the use of specific metrics that can effectively capture both the quality and diversity of the generated samples. Widely adopted metrics for gauging diversity and fidelity of image diffusion models include Inception Score (IS) [SGZ*16], Fréchet Inception Distance (FID) [HRU*17] and Kernel Inception Distance (KID) [BSAG21]. IS aims at capturing quality and diversity under a single metric by analyzing the distribution of labels obtained from a pre-trained classifier [SVI*15]. FID computes the Fréchet distance between inception features [SVI*15] derived from a set of real and synthesised images under the assumption that the feature vectors follow a Gaussian distribution. KID aims to improve on FID by relaxing the Gaussian assumption, directly measuring the Maximum Mean Discrepancy (MMD) between the two feature sets. Despite their popularity, inception-based metrics face several fundamental limitations. For instance, such metrics are reliant on the Inception-v3 pre-trained model, which could inherit biases depending on how the model was trained.

With the introduction of large-scale T2I diffusion models, generalization capabilities must also be taken into consideration. Zero-Shot FID [SCS*22] provides a solution to this, evaluating FID for images generated from a subset (30k in Saharia et al. [SCS*22]) of unseen prompts taken from the validation set and comparing them with reference images from the full validation set.

Conventional metrics for evaluating image quality include PSNR, SSIM, and LPIPs. PSNR measures the peak signal-to-noise ratio, SSIM measures the structural similarity, and LPIPs captures the perceived similarity based on learned features between two images. Recent methods also provide mid and high-level metrics that compare images, such as DreamSim [FTS*23] and CLIP [RKH*21]. However, such metrics are only applicable when ground truth images are available for comparison, i.e., reconstruction tasks. In the context of evaluating generative models, these metrics are rarely used directly.

Video Quality and Diversity. Naturally, the aforementioned image metrics have been extended to video, most notably using the Fréchet video distance (FVD) [UVSK*18] (e.g., implemented by the I3D network [CZ17]). FVD is often reported in conjunction with (video) IS. T2V datasets that are smaller than the large training sets, such as UCF101 [SZS12] and MSR-VTT [XMYR16], are often used to evaluate IS and FVD scores.

As noted in several works, e.g., [BHA*22,BRL*23], FVD is sensitive to the realism of individual frames and motion over short segments, but it does not capture long-term realism. Unrealistic repetitions over time, for example, are not penalized. Moreover, as noted by [STE22], FVD is highly sensitive to small implementation differences, implying that reported results between papers are not always directly comparable. To address these challenges, many T2V approaches use human evaluation in addition to IS and FVD, as discussed below.

Evaluating 3D Object Fidelity. The first option is to use FID and KID on the multiple renderings of the 3D models. Given a suitable encoder, FID can be applied to the latent space of the 3D models to directly measure the quality and diversity. P-FID, for example, uses PointNet++ [QYSG17] as the encoder to obtain the latent representation from pre-sampled points on the surface of the 3D models. For conditioned generation, such as single-view 3D reconstruction, ground truth multi-view images are often available. In this case, PSNR, LPIPS and other conventional image quality metrics are applied for the evaluation of novel view synthesis, whereas shape reconstruction metrics such as Chamfer Distance, F-Score, and Intersection of Union (IoU) are used to measure geometry accuracy. In case the text or image condition does not have ground truth, which is the more general case, prompt fidelity becomes the main metric. It measures the alignment of the generation with the conditional input, e.g., text and image. Similar to FID, one can fall back to measure image-prompt fidelity (see the **Prompt Fidelity** paragraph below) by using the renderings of the generated 3D models as a proxy. Recently, with the development of joint shape-text-image embedding space, e.g., in [ZLC*23] and [XGX*23], it is possible to directly assess the shape-prompt fidelity,e.g., Shape-Image Score (SI-S) and Shape-Text Score (ST-S), by comparing the embedding of shape and the conditional inputs on a pre-trained shape-text-image embedding space.

Evaluating Animated and Articulated Objects. Lei et al. [LDS*23] introduced the Instantiation Distance (ID) for measuring the similarity between a pair of articulated objects. This metric considers the part-level geometry and the object motion patterns. The physical realism of generated motion can be evaluated by physics-based metrics. For this purpose, EDGE [TCL23] proposed a metric to measure the dynamic realism of the motion. Diversity (Div), Multi-Modality (MM) and Beat Alignment Score (BAS) [LYRK21] are widely used metrics for the evaluation of 3D human dance motions. The core idea of BAS is to assess how close the kinematic beats (i.e., directional changes of the per-joint velocity vectors) coincide with the music beats. Note, however, that in many dancing styles, kinematic beats deliberately do not coincide with music beats. Hence, BAS should be used in combination with other metrics (FID, Div and MM).

Human Evaluation. The standard metrics discussed above, including FID and FVD, are unreliable at best and should be considered proxy metrics. For this reason, many papers describing new diffusion models also perform human evaluation by running user studies. Oftentimes, the users are asked to subjectively evaluate image or video "quality" and "faithfulness" (e.g., with respect to a prompt) or they are asked to directly compare and rank two or more images based on some metric. These types of user user studies are difficult to replicate and resource intense to conduct, as discussed in more detail in the next section on open challenges.

Prompt Fidelity. As conditional generation using text prompts is one of the most popular applications of the diffusion model, it is important to evaluate the faithfulness of the generated content with respect to the text prompt. To measure the faithfulness of a generated image with respect to the text prompt condition, the average cosine similarity between prompt and image CLIP [RKH*21] embeddings is often computed. Similarly, CLIPSIM [WLJ*22] measures the average CLIP similarity between generated video frames and text with T2V models. Similar metrics have also been proposed to measure alignment of edited images and human instructions [BHE23].

Identity Preservation. To assess multi-view facial identity (ID) consistency for generated 3D faces, the mean ArcFace [DGXZ19] cosine similarity score between pairs of views of the same synthesized face rendered from random camera poses is the de-facto standard metric. When reference images are available for generated identities—as is the case when finetuning diffusion models on a few images of a specific subject—the average pairwise cosine similarity between CLIP embeddings of generated and real (reference) images of the same subject can be used to evaluate the ID consistency. Note, however, that this approach is not constructed to distinguish between different subjects that could have highly similar text descriptions (e.g., two different yellow clocks). For this reason, [RLJ*22] proposed to also evaluate ID consistency using the average pairwise cosine similarity between ViTS/16 DINO embeddings of generated and real images. The advantage of this metric is that DINO [CTM*21] is not trained to ignore differences between subjects of the same class by design.

10. Open Challenges

Despite the immense progress made in generative models in recent years, there remain a large number of unsolved problems. In this section, we detail some prominent ones, although many more exist and are described in greater detail in the references cited throughout this report.

Evaluation Metrics. As discussed before, standard image and video quality metrics, such as FID and FVD, are not always well aligned with human judgement, and often make undesirable assumptions about the distributional similarity across datasets. Despite this, available alternatives are not much better: comparative metrics like PSNR and LPIPs require matching ground-truth pairs, and conducting a user study can be costly, time-consuming, and often not much more informative. One direction for obvious improvement in this domain is the creation of better metrics that reliably and automatically assess the quality and diversity of generated content. These metrics should be applicable to a variety of data types, such 2D images, video, 3D, and dynamic 3D scenes. Furthermore, they should be aligned with human preferences, to enable faster, automated progress on method development without regular human intervention.

Training Data. Captioned image data is available in abundance, but labeled training data for 3D, video, and 4D generation is scarce. This makes it difficult to train higher-dimensional generative models. Open problems remain in the collection of large-scale datasets, whether they be explicitly in higher dimensions (e.g., a large set of 3D models), or lower-dimensional projections of this data that can be used for learning high-dimensional priors (e.g., a large dataset of multi-view images). Both options require additional research in their corresponding training protocols—large-scale 3D datasets will undoubtedly have domain gaps with real-world scenes (since it will be difficult to match the diversity of real-world scenes), and large-scale multi-view datasets may not encode the same distribution of 3D-consistent scenes. The ideal training protocol and dataset may even be a combination of these options, or trained in stages, as is common with inflated models.

Another way to frame this decision is as an option of quality versus quantity. Can diffusion models leverage weaker supervision from a larger amount of training samples more effectively than stronger supervision from fewer samples? For instance, large language models trained on text leverage trillions of training samples, while image generation models are trained with only billions of images, and video models with even fewer. Multi-view datasets and 3D datasets provide more supervision than monocular video captures, but available datasets are at least yet another order of magnitude smaller in size. One may consider opting for a smaller, more informative dataset and instead choose to tackle the problem of data *efficiency*: the ability of a generative model to generalize in low-data regimes.

Efficiency. Inference-time sampling speed continues to be a concern for diffusion-based generative models. Where other generative methods (e.g., GANs) only require a single forward pass of a neural network, diffusion models can require up to thousands [DN21, HJA20] of network evaluations to produce a single generated result. The sequential nature of the forward and reverse diffusion processes poses as a fundamental bottleneck for efficiency. A straightforward method for increasing sampling speed is by designing newer, lightweight efficient architectures, such that fewer time is spent on each denoising step [LWJ*23]. Distillation is another category of technique that aims to achieve the same result. Salimans et al. [SH22] and Meng et al. [MRG*23] distill a pre-trained diffusion model into a model that requires fewer sampling steps. Consistency Models [SDCS23], TRACT [BAL*23], and BOOT [GZZ*23] take this to the extreme, proposing single-step generation distilled from pre-trained diffusion samplers. The question of how to generate the highest-quality output in the most efficient manner, however, remains an unsolved problem.

Training efficiency also poses a concern. The majority of existing models are currently trained at scale by corporations with large computational resource pools. Exploring solutions for reducing the training compute requirements remains a very valuable open problem, as training specialized models for targeted applications is currently out of reach for most researchers.

Controllability. As discussed in Maneesh Agrawala's blog post [Agr23], most diffusion models are unpredictable black boxes. Most models are conditioned only on text prompts, and therefore require extensive prompt engineering to generate a desired image. Furthermore, text input alone often does not offer sufficient control to specify a particular image's exact appearance. This poor interface typically results in a lengthy trial-and-error process. A conversational interface is a suitable alternative. Explored briefly in InstructPix2Pix [BHE23] and further developed in Dall-E3 [Ope23b], conversations and instructions allow for relative and intuitive adjustments to a current image. This process of aligning generations with human intentions largely remains an open problem.

Other existing forms of control include matched prompt editing [HMT*22], conditioning (Sec. 3.3), customization (Sec. 3.4), and guidance [EJP*23]. These are all effective strategies to make diffusion models more controllable and predictable, better enabling a user to achieve a desired image, but are far from a polished solution. Designing better and more intuitive interfaces around diffusion models that provide generalized control and predictable outputs remains an important open challenge.

Physical Grounding. Controlling, constraining, or guiding diffusion models to adhere to the rules of physics is also a promising direction. Certain modalities, such as 3D geometry and motion are heavily constrained by the physics of the underlying scene they model. In practice, these physical properties and rules can be embedded in the training process or network design to encourage generations to be more plausible. GANs, for example, have greatly benefited from being constrained by the physics of projective geometry, non-rigid object deformation, and physically based lighting to enable the training of 3D-aware models from single-view 2D image datasets [NPLT*19, CMK*21] to generate articulated 3D characters [WLJ*23, BKY*22, YSI*23] or re-lightable digital humans [DWW23, WHZZ23], respectively. Exploring similar uses of diffusion models, e.g., by leveraging a model's emergent geometry or correspondences [TJW*23, LDP*23], is an encouraging open challenge.

Robustness and Reproducibility. Despite the meteoric advancements in the field, catalyzed by seminal works such as DALL-E 2 [Ope23a] and Imagen [SCS*22], a disconcerting discrepancy persists between officially publicized results and those obtained through independent reproduction, even when employing the officially released code. This gap is often bridged through manual prompt manipulation, hyper-parameter optimization, and random seed manipulation—practices that have regrettably become tacit yet universally accepted methods for achieving favorable outcomes. This inflation of success is especially detrimental in optimization-based paradigms, including text-to-3D generation, and is exacerbated by the prevailing competitive ethos within the research community. Consequently, the imperative to develop robust and reproducible algorithms that fundamentally ameliorate these issues cannot be overstated.

11. Social Implications and Ethical Concerns

Distribution of Harmful Content. Generative AI tools automate the process of content creation with photorealistic quality. This ability could be used to generate fake photos or videos of real people (DeepFakes). DeepFakes pose a societal threat that could be used for harm, either intentional or unintentional. Anyone with unrestricted access to generative AI tools could, for example, create an image or video of a celebrity with the intention of tarnishing their reputation.

A number of measures can be put in place to prevent the distribution of harmful content. First, the ability of a user to generate harmful content should be prevented as best as possible. Most image generation models today, for example, prevent the generation of content depicting violence, gore, harassment, drugs, adult content, and generally offensive topics. Second, forensic techniques to detect DeepFakes are being developed by the AI community (e.g., [AFG*19, RCV*19, FLK*21]). As the quality of generative AI tools advances, these types of efforts are becoming increasingly important but also challenging.

Copyright, Legal Exposure, and Privacy Concerns. Foundation models are trained on billions of images, including content that may have been scraped without the consent of the creator, that may have been legally protected otherwise, or that may contain personally identifiable or sensitive information. Indeed, copyright infringement lawsuits have already been filed by artists against some of the companies behind foundation models for visual computing.

Bias and Fairness. Similar to most machine learning methods, diffusion models can inadvertently learn and perpetuate biases present in their training data. This is a significant concern in terms of fairness and ethical considerations; further research to develop means to mitigate such biases is needed.

Environmental Concerns. Training foundation models requires substantial computational resources. For example, the relatively small StableDiffusion model was reportedly trained on 2.3 Billion images using 256 Nvidia A100 GPUs on Amazon Web Services for a total of 150,000 GPU-hours [Mos22]. According to unverified sources, OpenAI's large GPT-4 model was trained on about 25,000 Nvidia A100 GPUs for 90–100 days, costing more than $100 Million. The carbon footprint of training GPT-4 is estimated to be between 12,456 and 14,994 metric tons of carbon dioxide equivalent [Lud23]. For comparison, the average yearly carbon footprint across humans on Earth is 4 tons. These considerations lead to significant concerns about the environmental impact of training foundation models.

Economic Impacts. As generative AI models become more integrated into various sectors, there are concerns about their impact on jobs and economic structures. For example, the ability to generate high-quality digital doubles or, more generally, digital humans as well as 3D scenes is expected to have a significant impact on video production, the visual effects, among many other industries. Some jobs in this creative industry may be displaced, but at the same time new creative job profiles will arise and new ways to monetize creative work will be arise through Generative AI. Generative AI also automates many tasks that workers in other areas do today. Although this is the case for many technologies, generative AI raises concerns about worker displacements at an accelerated rate.

Explainability, Trust, and Accountability. Machine learning and generative AI models learn correlations within their training data, not causality. Therefore, it may be challenging to understand why a model gave the answer it did. Moreover, generative AI models can synthesize new data, which may not always be truthful, or the models could be involuntarily trained on data that contains factual errors. These issues call trustworthiness into question. Some outcomes may even have legal consequences, raising questions of accountability.

Researchers working in the field of generative AI must be cognizant of these issues. Moreover, policy and lawmakers, industry, and the research community must engage in a constructive dialogue to set meaningful legal boundaries for the unprecedented capabilities and dangers of emerging generative AI.

12. Discussion and Conclusion

In this state-of-the-art report, we have reviewed the theory and practice of emerging diffusion models for visual computing. We have introduced the basic mathematical concepts, implementation details and design choices of popular diffusion models, and important strategies for finetuning, sampling, conditioning and inversion, among others. Moreover, we have given a comprehensive overview of the rapidly growing literature in this space, categorized by the type of generated medium, and discussed available datasets, metrics, open challenges, and social implications. Yet, this is an exploding field with papers and commercial models being released on a weekly or even daily basis. Thus, we hope that this STAR provides an intuitive starting point for the interested reader—artist, practitioner, and researcher alike.

Although most of the numerous papers discussed in this STAR have been published in the last (few) year(s) and all important aspects of diffusion models have seemingly been addressed, many open challenges remain. Perhaps one of the highest-level goals of the field of visual computing is to amplify the creative potential of novice and advanced users alike and empower them to intuitively convert their imagination into an image, video, or 3D scene. The generative AI tools discussed in this STAR are a big step forward, but the community still has much work ahead to achieving this goal.

Acknowledgements

We thank the Stanford Graphics lab, the Snap Research team, and Guandao Yang for fruitful discussions. R.P. was supported by the Stanford Graduate Fellowship. W.Y. was supported by the SNF Postdoc Mobility fund. V.G. and C.T. were supported by the ERC Consolidator Grant *4DReply* (770784) and by the Saarbrücken Research Center for Visual Computing, Interaction and Artificial Intelligence (VIA Center, via-center.science). T.D. was supported by the Israeli Science Foundation (grant 2303/20). A.K. was supported by BAIR and BDD sponsors, as well as DARPA Fiddler. M.N. was supported by an ERC Starting Grant Scan2CAD (804724) and the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing". B.O. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project 421703927 and the bidt project KLIMA-MEMES. G.W. was supported by a PECASE from the ARO, by Samsung, and by the Stanford Institute for Human-Centered AI (HAI).

References

[AAF*23] Avrahami O., Aberman K., Fried O., Cohen-Or D., Lischinski D.: Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311* (2023). 7

[AFG*19] Agarwal S., Farid H., Gu Y., He M., Nagano K., Li H.: Protecting world leaders against deep fakes. In *CVPR workshops* (2019), vol. 1, p. 38. 22

[AFH*19] Achlioptas P., Fan J., Hawkins R., Goodman N., Guibas L. J.: Shapeglot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8938–8947. 13, 19

[Agr23] Agrawala M.: Unpredictable black boxes are terrible interfaces. https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible, 2023. 22

[AHS*23] Achlioptas P., Huang I., Sung M., Tulyakov S., Guibas L.: Shapetalk: A language dataset and framework for 3d shape edits and deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12685–12694. 13, 19

[ALF22] Avrahami O., Lischinski D., Fried O.: Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18208–18218. 6

[ALV*23] Araújo J. P., Li J., Vetrivel K., Agarwal R., Wu J., Gopinath D., Clegg A. W., Liu K.: Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 21211–21221. 17, 20

[ANBH23] Alexanderson S., Nagy R., Beskow J., Henter G. E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph. 42*, 4 (2023), 44:1–44:20. 17, 18, 20

[And82] Anderson B. D.: Reverse-time diffusion equation models. *Stochastic Processes and their Applications 12*, 3 (1982), 313–326. 3

[Arn23] Arnautu E.: 30 best midjourney prompts to get amazing results. https://mspoweruser.com/best-midjourney-prompts/, 2023. Accessed: 2023-09-30. 1

[AXF*23] Anciukevičius T., Xu Z., Fisher M., Henderson P., Bilen H., Mitra N. J., Guerrero P.: Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12608–12618. 11, 12, 13

[AXS21] Alldieck T., Xu H., Sminchisescu C.: imghum: Implicit generative models of 3d human shape and articulated pose, 2021. arXiv:2108.10842. 16

[AZL23] Ao T., Zhang Z., Liu L.: Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv preprint arXiv:2303.14613* (2023). 17

[BAL*23] Berthelot D., Autef A., Lin J., Yap D. A., Zhai S., Hu S., Zheng D., Talbott W., Gu E.: Tract: Denoising diffusion models with transitive closure time-distillation, 2023. arXiv:2303.04248. 21

[BCD*21] Baruch G., Chen Z., Dehghan A., Dimry T., Feigin Y., Fu P., Gebauer T., Joffe B., Kurz D., Schwartz A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897* (2021). 19

[BGA*22] Bautista M. A., Guo P., Abnar S., Talbott W., Toshev A., Chen Z., Dinh L., Zhai S., Goh H., Ulbricht D., ET AL.: Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems 35* (2022), 25102–25116. 11, 12, 13

[BHA*21] BOMMASANI R., HUDSON D. A., ADELI E., ALTMAN R., ARORA S., VON ARX S., BERNSTEIN M. S., BOHG J., BOSSELUT A., BRUNSKILL E., ET AL.: On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021). 2

[BHA*22] BROOKS T., HELLSTEN J., AITTALA M., WANG T.-C., AILA T., LEHTINEN J., LIU M.-Y., EFROS A., KARRAS T.: Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems 35* (2022), 31769–31781. 20

[BHE23] BROOKS T., HOLYNSKI A., EFROS A. A.: Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18392–18402. 7, 15, 21, 22

[BKL18] BARSOUM E., KENDER J., LIU Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2018), pp. 1418–1427. 16

[BKY*22] BERGMAN A., KELLNHOFER P., YIFAN W., CHAN E., LINDELL D., WETZSTEIN G.: Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems 35* (2022), 19900–19916. 22

[BNVZ21] BAIN M., NAGRANI A., VAROL G., ZISSERMAN A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision* (2021). 19

[BRL*23] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 22563–22575. 8, 9, 20

[BRO*22] BLATTMANN A., ROMBACH R., OKTAY K., MÜLLER J., OMMER B.: Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems 35* (2022), 15309–15324. 5

[BSAG21] BIŃKOWSKI M., SUTHERLAND D. J., ARBEL M., GRETTON A.: Demystifying mmd gans, 2021. arXiv:1801.01401. 20

[BTLLW21] BOND-TAYLOR S., LEACH A., LONG Y., WILLCOCKS C. G.: Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence* (2021). 2

[BTOAF*22] BAR-TAL O., OFRI-AMAR D., FRIDMAN R., KASTEN Y., DEKEL T.: Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision* (2022), Springer, pp. 707–723. 1, 10

[BXP*22] BHATNAGAR B. L., XIE X., PETROV I., SMINCHISESCU C., THEOBALT C., PONS-MOLL G.: Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (jun 2022), IEEE. 17

[BYW23] BERGMAN A. W., YIFAN W., WETZSTEIN G.: Articulated 3d head avatar generation using text-to-image diffusion models. *arXiv preprint arXiv:2307.04859* (2023). 16

[BZTN20] BOZIC A., ZOLLHOFER M., THEOBALT C., NIESSNER M.: Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 7002–7012. 19

[CCH*23] CAO Y., CAO Y.-P., HAN K., SHAN Y., WONG K.-Y. K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916* (2023). 16

[CCP*23] CAI S., CHAN E. R., PENG S., SHAHBAZI M., OBUKHOV A., VAN GOOL L., WETZSTEIN G.: Diffdreamer: Consistent single-view perpetual view generation with conditional diffusion models. In *ICCV* (2023). 15

[CCS*19] CHEN K., CHOY C. B., SAVVA M., CHANG A. X., FUNKHOUSER T., SAVARESE S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14* (2019), Springer, pp. 100–116. 13, 19

[CDF*17] CHANG A., DAI A., FUNKHOUSER T., HALBER M., NIESSNER M., SAVVA M., SONG S., ZENG A., ZHANG Y.: Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)* (2017). 19

[CFG*15] CHANG A. X., FUNKHOUSER T., GUIBAS L., HANRAHAN P., HUANG Q., LI Z., SAVARESE S., SAVVA M., SONG S., SU H., ET AL.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015). 12, 19

[CGC*23] CHEN H., GU J., CHEN A., TIAN W., TU Z., LIU L., SU H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714* (2023). 11, 12, 13

[CGD*22] COLLINS J., GOEL S., DENG K., LUTHRA A., XU L., GUNDOGDU E., ZHANG X., VICENTE T. F. Y., DIDERIKSEN T., ARORA H., ET AL.: Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 21126–21136. 19

[CHL*23] CHEN W., HU H., LI Y., RUIZ N., JIA X., CHANG M.-W., COHEN W. W.: Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186* (2023). 7

[CHM23] CEYLAN D., HUANG C.-H., MITRA N. J.: Pix2video: Video editing using image diffusion. In *ICCV* (2023). 9, 10

[CJ23] CAO A., JOHNSON J.: Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 130–141. 10

[CJL*23] CHEN X., JIANG B., LIU W., HUANG Z., FU B., CHEN T., YU G.: Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18000–18010. 17

[CLC*22] CHAN E. R., LIN C. Z., CHAN M. A., NAGANO K., PAN B., DE MELLO S., GALLO O., GUIBAS L. J., TREMBLAY J., KHAMIS S., ET AL.: Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16123–16133. 12, 13

[CLT*23] CHENG Y.-C., LEE H.-Y., TULYAKOV S., SCHWING A. G., GUI L.-Y.: SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4456–4465. 11, 12

[CMK*21] CHAN E. R., MONTEIRO M., KELLNHOFER P., WU J., WETZSTEIN G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 5799–5809. 22

[CNC*23] CHAN E. R., NAGANO K., CHAN M. A., BERGMAN A. W., PARK J. J., LEVY A., AITTALA M., MELLO S. D., KARRAS T., WETZSTEIN G.: GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *ICCV* (2023). 15

[CTM*21] CARON M., TOUVRON H., MISRA I., JÉGOU H., MAIRAL J., BOJANOWSKI P., JOULIN A.: Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)* (2021). 21

[CWQ*23] CAO M., WANG X., QI Z., SHAN Y., QIE X., ZHENG Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465* (2023). 6

[CWX*23] CHEN W., WU J., XIE P., WU H., LI J., XIA X., XIAO X., LIN L.: Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840* (2023). 9

[CZ17] CARREIRA J., ZISSERMAN A.: Quo vadis, action recognition?

a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 6299–6308. 20

[DCLT18] DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* (2018). 7

[DCS*17] DAI A., CHANG A. X., SAVVA M., HALBER M., FUNKHOUSER T., NIESSNER M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 5828–5839. 19

[DDS*09] DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L.: Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255. 19

[DGXZ19] DENG J., GUO J., XUE N., ZAFEIRIOU S.: Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4690–4699. 21

[DJP*20] DHARIWAL P., JUN H., PAYNE C., KIM J. W., RADFORD A., SUTSKEVER I.: Jukebox: A generative model for music. *arXiv e-prints* (2020). 17

[DKP*23] DU Y., KIPS R., PUMAROLA A., STARKE S., THABET A., SANAKOYEU A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. *arXiv e-prints* (2023). 17

[DLW*23] DEITKE M., LIU R., WALLINGFORD M., NGO H., MICHEL O., KUSUPATI A., FAN A., LAFORTE C., VOLETI V. S., GADRE S. Y., VANDERBILT E., KEMBHAVI A., VONDRICK C., GKIOXARI G., EHSANI K., SCHMIDT L., FARHADI A.: Objaverse-xl: A universe of 10m+ 3d objects. *ArXiv abs/2307.05663* (2023). 12, 13, 14, 15, 19

[DMAB23] DEICHLER A., MEHTA S., ALEXANDERSON S., BESKOW J.: Diffusion-based co-speech gesture generation using joint text and audio representation. *arXiv e-prints* (2023). 18

[DMGT23] DABRAL R., MUGHAL M. H., GOLYANIK V., THEOBALT C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)* (2023). 1, 17, 18

[DN21] DHARIWAL P., NICHOL A.: Diffusion models beat gans on image synthesis, 2021. arXiv:2105.05233. 6, 21

[DRC*17] DOSOVITSKIY A., ROS G., CODEVILLA F., LOPEZ A., KOLTUN V.: Carla: An open urban driving simulator. In *Conference on robot learning* (2017), PMLR, pp. 1–16. 19

[DSS*22] DEITKE M., SCHWENK D., SALVADOR J., WEIHS L., MICHEL O., VANDERBILT E., SCHMIDT L., EHSANI K., KEMBHAVI A., FARHADI A.: Objaverse: A universe of annotated 3d objects. *arXiv:2212.08051* (2022). 12, 13, 14, 15, 19

[DVK22] DOCKHORN T., VAHDAT A., KREIS K.: Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems 35* (2022), 30150–30166. 4

[DWW23] DENG B., WANG Y., WETZSTEIN G.: Lumigan: Unconditional generation of relightable 3d human faces. *arXiv preprint arXiv:2304.13153* (2023). 22

[ECA*23] ESSER P., CHIU J., ATIGHEHCHIAN P., GRANSKOG J., GERMANIDIS A.: Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011* (2023). 8, 9

[EJP*23] EPSTEIN D., JABRI A., POOLE B., EFROS A. A., HOLYNSKI A.: Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986* (2023). 6, 22

[EMS*23] ERKOÇ Z., MA F., SHAN Q., NIESSNER M., DAI A.: Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. *arXiv preprint arXiv:2303.17015* (2023). URL: https://arxiv.org/abs/2303.17015. 11, 12

[ERO21] ESSER P., ROMBACH R., OMMER B.: Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 12873–12883. 5

[FAKD23] FRIDMAN R., ABECASIS A., KASTEN Y., DEKEL T.: Scenescape: Text-driven consistent scene generation. *ArXiv abs/2302.01133* (2023). 9, 15

[FCG*21] FU H., CAI B., GAO L., ZHANG L.-X., WANG J., LI C., ZENG Q., SUN C., JIA R., ZHAO B., ET AL.: 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10933–10942. 19

[FKYT*22] FRIDOVICH-KEIL S., YU A., TANCIK M., CHEN Q., RECHT B., KANAZAWA A.: Plenoxels: Radiance fields without neural networks. In *CVPR* (2022). 12

[FLK*21] FOX G., LIU W., KIM H., SEIDEL H.-P., ELGHARIB M., THEOBALT C.: VideoForensicsHQ: Detecting high-quality manipulated face videos. In *IEEE International Conference on Multimedia and Expo (ICME 2021)* (Shenzhen, China (Virtual), 2021), IEEE. doi:10.1109/ICME51207.2021.9428101. 22

[FM18] FERSTL Y., MCDONNELL R.: Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (2018), pp. 93–98. 17

[FTS*23] FU S., TAMIR N., SUNDARAM S., CHAI L., ZHANG R., DEKEL T., ISOLA P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344* (2023). 20

[FTT*23] FAN Z., TAHERI O., TZIONAS D., KOCABAS M., KAUFMANN M., BLACK M. J., HILLIGES O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 17

[GAA*22] 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 preprint arXiv:2208.01618* (2022). 7

[GAA*23a] 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 preprint arXiv:2302.12228* (2023). 7

[GAA*23b] GAL R., ARAR M., ATZMON Y., BERMANO A. H., CHECHIK G., COHEN-OR D.: Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–13. 7

[GBTBD23] GEYER M., BAR-TAL O., BAGON S., DEKEL T.: Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373* (2023). 9, 10

[GGZ*23] GU J., GAO Q., ZHAI S., CHEN B., LIU L., SUSSKIND J.: Learning controllable 3d diffusion models from single-view images, 2023. arXiv:2304.06700. 13

[GKG*23] GIEBENHAIN S., KIRSCHSTEIN T., GEORGOPOULOS M., RÜNZ M., AGAPITO L., NIESSNER M.: Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 21003–21012. 19

[GMSPM21] GUZOV V., MIR A., SATTLER T., PONS-MOLL G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 4318–4329. 17

[GSW*21] GUI J., SUN Z., WEN Y., TAO D., YE J.: A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering 35*, 4 (2021), 3313–3332. 2

[GTL*23] Gu J., Trevithick A., Lin K.-E., Susskind J., Theobalt C., Liu L., Ramamoorthi R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion, 2023. arXiv:2302.10109. 15

[GXN*23] Gupta A., Xiong W., Nie Y., Jones I., Oğuz B.: 3dgen: Triplane latent diffusion for textured mesh generation, 2023. arXiv:2303.05371[cs]. 11, 12

[GYR*23] Guo Y., Yang C., Rao A., Wang Y., Qiao Y., Lin D., Dai B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2023. 9

[GZW*20] Guo C., Zuo X., Wang S., Zou S., Sun Q., Deng A., Gong M., Cheng L.: Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2021–2029. 17

[GZZ*22] Guo C., Zou S., Zuo X., Wang S., Ji W., Li X., Cheng L.: Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 5152–5161. 17, 19

[GZZ*23] Gu J., Zhai S., Zhang Y., Liu L., Susskind J.: Boot: Data-free distillation of denoising diffusion models with bootstrapping. *arXiv preprint arXiv:2306.05544* (2023). 21

[HACO23] Hertz A., Aberman K., Cohen-Or D.: Delta denoising score. *arXiv preprint arXiv:2304.07090* (2023). 7, 14

[HCL*23] Huang L., Chen D., Liu Y., Shen Y., Zhao D., Zhou J.: Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778* (2023). 9

[HCO*23] Höllein L., Cao A., Owens A., Johnson J., Niessner M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. *arXiv preprint arXiv:2303.11989* (2023). 15

[HCS*22] Ho J., Chan W., Saharia C., Whang J., Gao R., Gritsenko A., Kingma D. P., Poole B., Norouzi M., Fleet D. J., et al.: Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022). 8

[HDZ*22] Hong W., Ding M., Zheng W., Liu X., Tang J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022). 8

[HGJ*19] Houlsby N., Giurgiu A., Jastrzebski S., Morrone B., de Laroussilhe Q., Gesmundo A., Attariyan M., Gelly S.: Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751* (2019). 7

[HJA20] Ho J., Jain A., Abbeel P.: Denoising diffusion probabilistic models, 2020. arXiv:2006.11239. 5, 21

[HLHF22] Hui K.-H., Li R., Hu J., Fu C.-W.: Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. 11, 12, 13

[HLX*21] Habermann M., Liu L., Xu W., Zollhoefer M., Pons-Moll G., Theobalt C.: Real-time deep dynamic characters. *ACM Transactions on Graphics 40*, 4 (aug 2021). 19

[HLZ*23] Han L., Li Y., Zhang H., Milanfar P., Metaxas D., Yang F.: Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305* (2023). 7

[HLZH21] Hong Y., Li Q., Zhu S.-C., Huang S.: Vlgrammar: Grounded grammar induction of vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 1665–1674. 13

[HMT*22] Hertz A., Mokady R., Tenenbaum J., Aberman K., Pritch Y., Cohen-Or D.: Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022). 4, 6, 10, 22

[HPD*23] Han B., Peng H., Dong M., Xu C., Ren Y., Shen Y., Li Y.: Amd autoregressive motion diffusion. *arXiv preprint arXiv:2305.09381* (2023). 17

[HRU*17] 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. *Advances in neural information processing systems 30* (2017). 20

[HS22] Ho J., Salimans T.: Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022). 6, 7

[HSG*22] Ho J., Salimans T., Gritsenko A., Chan W., Norouzi M., Fleet D. J.: Video diffusion models. *arXiv:2204.03458* (2022). 8

[HSW*21] Hu E. J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W.: Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021). 7, 14

[HTE*23] Haque A., Tancik M., Efros A. A., Holynski A., Kanazawa A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789* (2023). 1, 15, 16

[HXM*21] Habibie I., Xu W., Mehta D., Liu L., Seidel H.-P., Pons-Moll G., Elgharib M., Theobalt C.: Learning speech-driven 3d conversational gestures from video. In *ACM International Conference on Intelligent Virtual Agents (IVA)* (2021). 20

[HXZ*20] Habermann M., Xu W., Zollhoefer M., Pons-Moll G., Theobalt C.: Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (jun 2020), IEEE. 19

[HYZ*22] He Y., Yang T., Zhang Y., Shan Y., Chen Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221* (2022). 8

[IPOS14] Ionescu C., Papava D., Olaru V., Sminchisescu C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*, 7 (jul 2014), 1325–1339. 17

[IRG*23] Işik M., Rünz M., Georgopoulos M., Khakhulin T., Starck J., Agapito L., Niessner M.: Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–12. doi:10.1145/3592415. 19

[Itô50] Itô K.: Stochastic differential equations in a differentiable manifold. *Nagoya Mathematical Journal 1* (1950), 35 – 47. 3

[Itô51] Itô K.: On a formula concerning stochastic differentials. *Nagoya Mathematical Journal 3* (1951), 55 – 65. 3

[IZZE18] Isola P., Zhu J.-Y., Zhou T., Efros A. A.: Image-to-image translation with conditional adversarial networks, 2018. arXiv:1611.07004. 5

[JCL*23] Jiang B., Chen X., Liu W., Yu J., Yu G., Chen T.: Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795* (2023). 17

[JHVDM*17] Johnson J., Hariharan B., Van Der Maaten L., Fei-Fei L., Lawrence Zitnick C., Girshick R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 2901–2910. 19

[JMPTdCM20] Jolicoeur-Martineau A., Piché-Taillefer R., des Combes R. T., Mitliagkas I.: Adversarial score matching and improved sampling for image generation, 2020. arXiv:2009.05475. 5

[JN23] Jun H., Nichol A.: Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463* (2023). 11, 12, 13

[JWZ*23] Jiang R., Wang C., Zhang J., Chai M., He M., Chen D., Liao J.: Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. *arXiv preprint arXiv:2303.17606* (2023). 16

[JZC*23] Jia X., Zhao Y., Chan K. C., Li Y., Zhang H., Gong B., Hou T., Wang H., Su Y.-C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642* (2023). 7

[KAAL22] KARRAS T., AITTALA M., AILA T., LAINE S.: Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems 35* (2022), 26565–26577. 4, 5

[KAZ*23] KOLOTOUROS N., ALLDIECK T., ZANFIR A., BAZAVAN E. G., FIERARU M., SMINCHISESCU C.: Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329* (2023). 16

[KBY*23] KIM S. W., BROWN B., YIN K., KREIS K., SCHWARZ K., LI D., ROMBACH R., TORRALBA A., FIDLER S.: Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 8496–8506. 11, 12, 13

[KDSD23] KOCHANOWICZ J., DOMAGAŁA M., STACHOWIAK D., DZIEDZIC K.: Diffusion models in practice. part 1: The tools of the trade. https://deepsense.ai/diffusion-models-in-practice-part-1-the-tools-of-the-trade/, 2023. 8

[KHA*22] KOO J., HUANG I., ACHLIOPTAS P., GUIBAS L. J., SUNG M.: Partglot: Learning shape part segmentation from language reference games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16505–16514. 19

[KHWKS23] KARRAS J., HOLYNSKI A., WANG T.-C., KEMELMACHER-SHLIZERMAN I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025* (2023). 9

[KKC23] KIM J., KIM J., CHOI S.: Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 8255–8263. 17

[KKY22] KIM G., KWON T., YE J. C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 2426–2435. 6

[KMT*23] KHACHATRYAN L., MOVSISYAN A., TADEVOSYAN V., HENSCHEL R., WANG Z., NAVASARDYAN S., SHI H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439* (2023). 9, 10

[KMVN23] KARNEWAR A., MITRA N. J., VEDALDI A., NOVOTNY D.: Holofusion: Towards photo-realistic 3d generative modeling. *arXiv preprint arXiv:2308.14244* (2023). 11, 12, 13

[KOWD21] KASTEN Y., OFRI D., WANG O., DEKEL T.: Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG) 40*, 6 (2021), 1–12. 10

[KPST23] KARUNRATANAKUL K., PREECHAKUL K., SUWAJANAKORN S., TANG S.: Gmd: Controllable human motion synthesis via guided diffusion models. *arXiv preprint arXiv:2305.12577* (2023). 17

[KQG*23] KIRSCHSTEIN T., QIAN S., GIEBENHAIN S., WALTER T., NIESSNER M.: Nersemble: Multi-view radiance field reconstruction of human heads, 2023. arXiv:2305.03027, doi:10.48550/arXiv.2305.03027. 19

[KRG*23] KULKARNI N., REMPE D., GENOVA K., KUNDU A., JOHNSON J., FOUHEY D., GUIBAS L.: Nifty: Neural object interaction fields for guided human motion synthesis, 2023. arXiv:2307.07511. 17

[KTEM18] KANAZAWA A., TULSIANI S., EFROS A. A., MALIK J.: Learning category-specific mesh reconstruction from image collections. In *ECCV* (2018). 16

[KVNM23] KARNEWAR A., VEDALDI A., NOVOTNY D., MITRA N. J.: Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18423–18433. 11, 12, 13

[KWM*20] KRANTZ J., WIJMANS E., MAJUMDAR A., BATRA D., LEE S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16* (2020), Springer, pp. 104–120. 19

[KWR*16] KEMPKA M., WYDMUCH M., RUNC G., TOCZEK J., JAŚKOWSKI W.: Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *2016 IEEE conference on computational intelligence and games (CIG)* (2016), IEEE, pp. 1–8. 19

[KZL*23] KAWAR B., ZADA S., LANG O., TOV O., CHANG H., DEKEL T., MOSSERI I., IRANI M.: Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 6007–6017. 7

[KZZ*23] KUMARI N., ZHANG B., ZHANG R., SHECHTMAN E., ZHU J.-Y.: Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1931–1941. 7

[Lab23] LAB D.: Deepfloyd if. https://github.com/deep-floyd/IF, 2023. 1

[LCW*23] LI X., CHU W., WU Y., YUAN W., LIU F., ZHANG Q., LI F., FENG H., DING E., WANG J.: Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation, 2023. arXiv:2309.00398. 8

[LCZ*23] LUO Z., CHEN D., ZHANG Y., HUANG Y., WANG L., SHEN Y., ZHAO D., ZHOU J., TAN T.: Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 10209–10218. 9

[LDP*23] LUO G., DUNLAP L., PARK D. H., HOLYNSKI A., DARRELL T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *arXiv preprint arXiv:2305.14334* (2023). 22

[LDS*23] LEI J., DENG C., SHEN B., GUIBAS L., DANIILIDIS K.: Nap: Neural 3d articulation prior. *arXiv e-prints* (2023). 16, 20

[LDZL23] LI M., DUAN Y., ZHOU J., LU J.: Diffusion-SDF: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12642–12651. 11, 12

[LGT*23] LIN C.-H., GAO J., TANG L., TAKIKAWA T., ZENG X., HUANG X., KREIS K., FIDLER S., LIU M.-Y., LIN T.-Y.: Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 300–309. 14

[LH21] LUO S., HU W.: Diffusion probabilistic models for 3d point cloud generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2837–2845. 11, 12

[LHR*21] LIU L., HABERMANN M., RUDNEV V., SARKAR K., GU J., THEOBALT C.: Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)* (2021). 19

[LJC*23] LEE Y.-C., JANG J.-Z. G., CHEN Y.-T., QIU E., HUANG J.-B.: Shape-aware text-driven layered video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 14317–14326. 10

[LLZ*23] LIU Y., LIN C., ZENG Z., LONG X., LIU L., KOMURA T., WANG W.: Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453* (2023). 15

[LMB*14] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (2014), Springer, pp. 740–755. 19

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 34*, 6 (Oct. 2015), 248:1–248:16. 19

[LTJ*21] LIU A., TUCKER R., JAMPANI V., MAKADIA A., SNAVELY N., KANAZAWA A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021). 15

[LTLH19] Liu Z., Tang H., Lin Y., Han S.: Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems 32* (2019). 13

[LTSH23] Li Z., Tucker R., Snavely N., Holynski A.: Generative image dynamics. *arXiv preprint arXiv:2309.07906* (2023). 9

[LTT*21] Li Y., Takehara H., Taketomi T., Zheng B., Niessner M.: 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12706–12716. 19

[Lud23] Ludvigsen K. G. A.: The carbon footprint of gpt-4. Towards Data Science, 2023. https://medium.com/towards-data-science/the-carbon-footprint-of-gpt-4-d6c676eb21ae. 22

[LWH*23] Liu R., Wu R., Hoorick B. V., Tokmakov P., Zakharov S., Vondrick C.: Zero-1-to-3: Zero-shot one image to 3d object. *ArXiv abs/2303.11328* (2023). 14

[LWJ*23] Li Y., Wang H., Jin Q., Hu J., Chemerys P., Fu Y., Wang Y., Tulyakov S., Ren J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *arXiv preprint arXiv:2306.00980* (2023). 21

[LWL23] Li J., Wu J., Liu C. K.: Object motion guided human motion synthesis. *ACM Transactions on Graphics (SIGGRAPH Asia)* (2023). 17

[LWSK22] Li Z., Wang Q., Snavely N., Kanazawa A.: Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *ECCV* (2022). 15

[LYC*23] Lu C., Yin F., Chen X., Liu W., Chen T., Yu G., Fan J.: A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 7557–7567. 20

[LYRK21] Li R., Yang S., Ross D. A., Kanazawa A.: Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 17, 20

[LYX*23] Liao T., Yi H., Xiu Y., Tang J., Huang Y., Thies J., Black M. J.: Tada! text to animatable digital avatars. *ArXiv* (Aug 2023). 16

[LYZ*23] Liew J. H., Yan H., Zhang J., Xu Z., Feng J.: Magicedit: High-fidelity and temporally coherent video editing. In *arXiv* (2023). 9

[LZCVDP20] Ling H. Y., Zinno F., Cheng G., Van De Panne M.: Character controllers using motion vaes. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 40–1. 16

[LZL*23a] Liang H., Zhang W., Li W., Yu J., Xu L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv e-prints* (2023). 17

[LZL*23b] Lin J., Zeng A., Lu S., Cai Y., Zhang R., Wang H., Zhang L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset, 2023. arXiv:2307.00818. 19

[LZL*23c] Liu S., Zhang Y., Li W., Lin Z., Jia J.: Video-p2p: Video editing with cross-attention control. *arXiv:2303.04761* (2023). 9

[LZW*23] Li C., Zhang C., Waghwase A., Lee L.-H., Rameau F., Yang Y., Bae S.-H., Hong C. S.: Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131* (2023). 2

[MCST22] Mittal P., Cheng Y.-C., Singh M., Tulsiani S.: AutoSDF: Shape priors for 3d completion, reconstruction and generation. In *CVPR* (2022). 12

[MGT*19] Mahmood N., Ghorbani N., Troje N. F., Pons-Moll G., Black M. J.: Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 5442–5451. 19

[MHA*23] Mokady R., Hertz A., Aberman K., Pritch Y., Cohen-Or D.: Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 6038–6047. 4, 7

[MHS*21] Meng C., He Y., Song Y., Song J., Wu J., Zhu J.-Y., Ermon S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021). 6

[Mid23] Midjourney: Midjourney. https://www.midjourney.com/, 2023. 2

[MLS*22] Menapace W., Lathuilière S., Siarohin A., Theobalt C., Tulyakov S., Golyanik V., Ricci E.: Playable environments: Video manipulation in space and time. In *Computer Vision and Pattern Recognition* (2022). 18

[Mos22] Mostaque E.: Twitter post, 2022. URL: https://twitter.com/emostaque/status/1563870674111832066. 22

[MPE*23] Mendiratta M., Pan X., Elgharib M., Teotia K., R M. B., Tewari A., Golyanik V., Kortylewski A., Theobalt C.: Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM ToG (SIGGRAPH Asia)* (2023). 1, 16, 18

[MRC*17] Mehta D., Rhodin H., Casas D., Fua P., Sotnychenko O., Xu W., Theobalt C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on* (2017), IEEE. 17

[MRG*23] Meng C., Rombach R., Gao R., Kingma D. P., Ermon S., Ho J., Salimans T.: On distillation of guided diffusion models, 2023. arXiv:2210.03142. 21

[MRP*23] Metzer G., Richardson E., Patashnik O., Giryes R., Cohen-Or D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12663–12673. 14

[MSL*23] Menapace W., Siarohin A., Lathuilière S., Achlioptas P., Golyanik V., Ricci E., Tulyakov S.: Plotting behind the scenes: Towards learnable game engines. *arXiv e-prints* (2023). 18

[MSP*23] Müller N., Siddiqui Y., Porzi L., Bulo S. R., Kontschieder P., Niessner M.: Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4328–4338. 1, 11

[MWS*23] Mou C., Wang X., Song J., Shan Y., Zhang J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421* (2023). 7

[MWX*23] Mou C., Wang X., Xie L., Zhang J., Qi Z., Shan Y., Qie X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023). 6

[ND21] Nichol A. Q., Dhariwal P.: Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning* (2021), PMLR, pp. 8162–8171. 4

[NDR*21] 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 preprint arXiv:2112.10741* (2021). 6

[NJD*22] Nichol A., Jun H., Dhariwal P., Mishkin P., Chen M.: Point-e: A system for generating 3d point clouds from complex prompts, 2022. arXiv:2212.08751[cs]. 11, 12, 13

[NPLT*19] Nguyen-Phuoc T., Li C., Theis L., Richardt C., Yang Y.-L.: Hologan: Unsupervised learning of 3d representations from natural images. In *The IEEE International Conference on Computer Vision (ICCV)* (Nov 2019). 22

[NPLX22] Nguyen-Phuoc T., Liu F., Xiao L.: Snerf: Stylized neural implicit representations for 3d scenes. *ACM Trans. Graph. 41*, 4 (jul 2022). URL: https://doi.org/10.1145/3528223.3530107, doi:10.1145/3528223.3530107. 15

[Ope23a] OpenAI: DALL·E 2 — openai.com. https://openai.com/dall-e-2/, 2023. [Accessed 26-09-2023]. 2, 22

[Ope23b] OpenAI: DALL·E 3 — openai.com. https://openai.com/dall-e-3, 2023. [Accessed 05-10-2023]. 2, 22

[OWX*23] OUYANG H., WANG Q., XIAO Y., BAI Q., ZHANG J., ZHENG K., ZHOU X., CHEN Q., SHEN Y.: Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926* (2023). 10

[PCA*21] PUNNAKKAL A. R., CHANDRASEKARAN A., ATHANASIOU N., QUIROS-RAMIREZ A., BLACK M. J.: BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 722–731. 19

[PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10975–10985. 16

[PJBM22] POOLE B., JAIN A., BARRON J. T., MILDENHALL B.: Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022). 1, 14, 18

[PKSZ*23] PARMAR G., KUMAR SINGH K., ZHANG R., LI Y., LU J., ZHU J.-Y.: Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023), pp. 1–11. 7

[PLWZ19] PARK T., LIU M.-Y., WANG T.-C., ZHU J.-Y.: Semantic image synthesis with spatially-adaptive normalization, 2019. arXiv:1903.07291. 5

[PMA16] PLAPPERT M., MANDERY C., ASFOUR T.: The kit motion-language dataset. *Big data 4*, 4 (2016), 236–252. 19

[PNM*20] PENG S., NIEMEYER M., MESCHEDER L., POLLEFEYS M., GEIGER A.: Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (2020), Springer, pp. 523–540. 12

[PRFS18] PARK K., REMATAS K., FARHADI A., SEITZ S. M.: Photoshape: Photorealistic materials for large-scale shape collections. *arXiv preprint arXiv:1809.09761* (2018). 19

[PTL*23] PAN X., TEWARI A., LEIMKÜHLER T., LIU L., MEKA A., THEOBALT C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023). 7

[PW23] PO R., WETZSTEIN G.: Compositional 3d scene generation using locally conditioned diffusion. *ArXiv abs/2303.12218* (2023). 1, 15

[PZX*21] PENG S., ZHANG Y., XU Y., WANG Q., SHUAI Q., BAO H., ZHOU X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR* (2021). 19

[QCZ*23] QI C., CUN X., ZHANG Y., LEI C., WANG X., SHAN Y., CHEN Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535* (2023). 9, 10

[QYSG17] QI C. R., YI L., SU H., GUIBAS L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems 30* (2017). 20

[RAY*16] REED S., AKATA Z., YAN X., LOGESWARAN L., SCHIELE B., LEE H.: Generative adversarial text to image synthesis, 2016. arXiv:1605.05396. 5

[RBH*21] REMPE D., BIRDAL T., HERTZMANN A., YANG J., SRIDHAR S., GUIBAS L. J.: Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 11488–11499. 16

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695. 2, 5, 6, 8, 18

[RCV*19] ROSSLER A., COZZOLINO D., VERDOLIVA L., RIESS C., THIES J., NIESSNER M.: Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 1–11. 22

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. *MICCAI* (2015). 5

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. *ICML* (2021). 2, 7, 20, 21

[RLBP*23] REMPE D., LUO Z., BIN PENG X., YUAN Y., KITANI K., KREIS K., FIDLER S., LITANY O.: Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 13756–13766. 17

[RLJ*22] RUIZ N., LI Y., JAMPANI V., PRITCH Y., RUBINSTEIN M., ABERMAN K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 22500–22510. 1, 7, 21

[RLJ*23] RUIZ N., LI Y., JAMPANI V., WEI W., HOU T., PRITCH Y., WADHWA N., RUBINSTEIN M., ABERMAN K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949* (2023). 7

[RSH*21] REIZENSTEIN J., SHAPOVALOV R., HENZLER P., SBORDONE L., LABATUT P., NOVOTNY D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10901–10911. 12, 19

[SBV*22] 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, 2022. arXiv:2210.08402. 14, 19

[SCC*22] SAHARIA C., CHAN W., CHANG H., LEE C. A., HO J., SALIMANS T., FLEET D. J., NOROUZI M.: Palette: Image-to-image diffusion models, 2022. arXiv:2111.05826. 5

[SCL*22] SANGHI A., CHU H., LAMBOURNE J. G., WANG Y., CHENG C.-Y., FUMERO M., MALEKSHAN K. R.: Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18603–18613. 13

[SCP*23] SHUE J. R., CHAN E. R., PO R., ANKNER Z., WU J., WETZSTEIN G.: 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 20875–20886. 11, 12

[SCS*22] SAHARIA C., CHAN W., SAXENA S., LI L., WHANG J., DENTON E. L., GHASEMIPOUR K., GONTIJO LOPES R., KARAGOL AYAN B., SALIMANS T., ET AL.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems 35* (2022), 36479–36494. 2, 20, 22

[SDCS23] SONG Y., DHARIWAL P., CHEN M., SUTSKEVER I.: Consistency models. *arXiv preprint arXiv:2303.01469* (2023). 21

[SESM22] SUHAIL M., ESTEVES C., SIGAL L., MAKADIA A.: Generalizable patch-based neural rendering. In *European Conference on Computer Vision* (2022), Springer. 15

[SFHAE23] SELLA E., FIEBELMAN G., HEDMAN P., AVERBUCH-ELOR H.: Vox-e: Text-guided voxel editing of 3d objects, 2023. 15

[SFL*23] SANGHI A., FU R., LIU V., WILLIS K. D., SHAYANI H., KHASAHMADI A. H., SRIDHAR S., RITCHIE D.: Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18339–18348. 13

[SGZ*16] SALIMANS T., GOODFELLOW I., ZAREMBA W., CHEUNG V., RADFORD A., CHEN X.: Improved techniques for training gans. *Advances in neural information processing systems 29* (2016). 20

[SH22] Salimans T., Ho J.: Progressive distillation for fast sampling of diffusion models, 2022. arXiv:2202.00512. 21

[Ske23] Sketchfab: Sketchfab — sketchfab.com. https://sketchfab.com/, 2023. [Accessed 25-09-2023]. 19

[SME20] Song J., Meng C., Ermon S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020). 4, 7

[SNL*21] Selvaraju P., Nabail M., Loizou M., Maslioukova M., Averkiou M., Andreou A., Chaudhuri S., Kalogerakis E.: Buildingnet: Learning to label 3d buildings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10397–10407. 19

[SPH*22] Singer U., Polyak A., Hayes T., Yin X., An J., Zhang S., Hu Q., Yang H., Ashual O., Gafni O., et al.: Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022). 8, 18

[SPX*22] Shi Z., Peng S., Xu Y., Liao Y., Shen Y.: Deep generative models on 3d representations: A survey, 2022. 2

[SRC*21] Srinivasan K., Raman K., Chen J., Bendersky M., Najork M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). 19

[SRL*23] Sohn K., Ruiz N., Lee K., Chin D. C., Blok I., Chang H., Barber J., Jiang L., Entis G., Li Y., Hao Y., Essa I., Rubinstein M., Krishnan D.: Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983* (2023). 7

[SRV23] Szymanowicz S., Rupprecht C., Vedaldi A.: Viewset diffusion: (0-)image-conditioned 3d generative models from 2d data. *arXiv e-prints* (2023). 13

[SSDK*20] Song Y., Sohl-Dickstein J., Kingma D. P., Kumar A., Ermon S., Poole B.: Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020). 3, 4

[SSME22] Su X., Song J., Meng C., Ermon S.: Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382* (2022). 4

[SSP*23a] Shao R., Sun J., Peng C., Zheng Z., Zhou B., Zhang H., Liu Y.: Control4d: Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *ArXiv abs/2305.20082* (2023). 18

[SSP*23b] Singer U., Sheynin S., Polyak A., Ashual O., Makarov I., Kokkinos F., Goyal N., Vedaldi A., Parikh D., Johnson J., Taigman Y.: Text-to-4d dynamic scene generation. *ArXiv abs/2301.11280* (2023). 1, 10, 18

[STE22] Skorokhodov I., Tulyakov S., Elhoseiny M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 3626–3636. 20

[STG*20] Sidhu V., Tretschk E., Golyanik V., Agudo A., Theobalt C.: Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision (ECCV)* (2020). 16

[STKB23] Shafir Y., Tevet G., Kapon R., Bermano A. H.: Human Motion Diffusion as a Generative Prior. *arXiv e-prints* (2023). 17, 18

[SVB*21] Schuhmann C., Vencu R., Beaumont R., Kaczmarczyk R., Mullis C., Katta A., Coombes T., Jitsev J., Komatsuzaki A.: LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR abs/2111.02114* (2021). URL: https://arxiv.org/abs/2111.02114, arXiv:2111.02114. 19

[SVI*15] Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.: Rethinking the inception architecture for computer vision, 2015. arXiv:1512.00567. 20

[SWM*19] Straub J., Whelan T., Ma L., Chen Y., Wijmans E., Green S., Engel J. J., Mur-Artal R., Ren C., Verma S., et al.: The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019). 19

[SWY*23] Shi Y., Wang P., Ye J., Mai L., Li K., Yang X.: Mv-dream: Multi-view diffusion for 3d generation. *arXiv:2308.16512* (2023). 15

[SWZ*18] Sun X., Wu J., Zhang X., Zhang Z., Zhang C., Xue T., Tenenbaum J. B., Freeman W. T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2974–2983. 19

[SXLJ23] Shi J., Xiong W., Lin Z., Jung H. J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411* (2023). 7

[SXP*23] Shi Y., Xue C., Pan J., Zhang W., Tan V. Y., Bai S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435* (2023). 7

[SZS12] Soomro K., Zamir A. R., Shah M.: A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision 2*, 11 (2012). 20

[SZT*23] Shao R., Zheng Z., Tu H., Liu B., Zhang H., Liu Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Computer Vision and Pattern Recognition (CVPR)* (2023). 18

[TCL23] Tseng J., Castellon R., Liu K.: Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 448–458. 17, 18, 20

[TGBD23] Tumanyan N., Geyer M., Bagon S., Dekel T.: Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1921–1930. 6, 10

[TGBT20] Taheri O., Ghorbani N., Black M. J., Tzionas D.: GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)* (2020). 17

[TJW*23] Tang L., Jia M., Wang Q., Phoo C. P., Hariharan B.: Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881* (2023). 22

[TRC*23] Tang L., Ruiz N., Chu Q., Li Y., Holynski A., Jacobs D. E., Hariharan B., Pritch Y., Wadhwa N., Aberman K., et al.: Realfill: Reference-driven generation for authentic image completion. *arXiv preprint arXiv:2309.16668* (2023). 7

[TRG*23] Tevet G., Raab S., Gordon B., Shafir Y., Cohen-or D., Bermano A. H.: Human motion diffusion model. In *International Conference on Learning Representations (ICLR)* (2023). 17, 18

[TSB*22] Thomason J., Shridhar M., Bisk Y., Paxton C., Zettlemoyer L.: Language grounding with 3d objects. In *Conference on Robot Learning* (2022), PMLR, pp. 1691–1701. 13

[TYC*23] Tewari A., Yin T., Cazenavette G., Rezchikov S., Tenenbaum J. B., Durand F., Freeman W. T., Sitzmann V.: Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *arXiv* (2023). 13

[UVSK*18] Unterthiner T., Van Steenkiste S., Kurach K., Marinier R., Michalski M., Gelly S.: Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018). 20

[VACO23] Voynov A., Aberman K., Cohen-Or D.: Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023), pp. 1–11. 5, 6

[VCCOA23] Voynov A., Chu Q., Cohen-Or D., Aberman K.: *p*+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522* (2023). 7

[Vin11] VINCENT P.: A connection between score matching and denoising autoencoders. *Neural computation 23*, 7 (2011), 1661–1674. 4

[VJMP22] VOLETI V., JOLICOEUR-MARTINEAU A., PAL C.: Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *(NeurIPS) Advances in Neural Information Processing Systems* (2022). URL: https://arxiv.org/abs/2205.09853. 9

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. *NIPS* (2017). 5, 7

[WCMB*22] WATSON D., CHAN W., MARTIN-BRUALLA R., HO J., TAGLIASACCHI A., NOROUZI M.: Novel view synthesis with diffusion models, 2022. arXiv:2210.04628. 15

[WDL*22] WANG H., DU X., LI J., YEH R. A., SHAKHNAROVICH G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022. arXiv:2212.00774. 14

[WDIT22] WU C. H., DE LA TORRE F.: Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559* (2022). 7

[WGN23] WALLACE B., GOKUL A., NAIK N.: Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 22532–22541. 7

[WGW*22] WU J. Z., GE Y., WANG X., LEI W., GU Y., HSU W., SHAN Y., QIE X., SHOU M. Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565* (2022). 9, 10

[WHZZ23] WANG Y., HOLYNSKI A., ZHANG X., ZHANG X.: Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 20792–20802. 22

[WLJ*22] WU C., LIANG J., JI L., YANG F., FANG Y., JIANG D., DUAN N.: Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision* (2022), Springer, pp. 720–736. 21

[WLJ*23] WU S., LI R., JAKAB T., RUPPRECHT C., VEDALDI A.: MagicPony: Learning articulated 3d animals in the wild. 16, 22

[WLW*23] WANG Z., LU C., WANG Y., BAO F., LI C., SU H., ZHU J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213* (2023). 14

[WYC*23] WANG J., YUAN H., CHEN D., ZHANG Y., WANG X., ZHANG S.: Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023). 8

[WYZ*23] WANG X., YUAN H., ZHANG S., CHEN D., WANG J., ZHANG Y., SHEN Y., ZHAO D., ZHOU J.: Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018* (2023). 9

[WZF*23] WU T., ZHANG J., FU X., WANG Y., REN J., PAN L., WU W., YANG L., WANG J., QIAN C., ET AL.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 803–814. 19

[WZJ*23] WEI Y., ZHANG Y., JI Z., BAI J., ZHANG L., ZUO W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848* (2023). 7

[WZZ*23] WANG T., ZHANG B., ZHANG T., GU S., BAO J., BALTRUSAITIS T., SHEN J., CHEN D., WEN F., CHEN Q., GUO B.: RODIN: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4563–4573. 12

[XCZ*18] XU W., CHATTERJEE A., ZOLLHÖFER M., RHODIN H., MEHTA D., SEIDEL H., THEOBALT C.: Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph. 37*, 2 (2018), 27. URL: https://doi.org/10.1145/3181973, doi:10.1145/3181973. 19

[XGX*23] XUE L., GAO M., XING C., MARTÍN-MARTÍN R., WU J., XIONG C., XU R., NIEBLES J. C., SAVARESE S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1179–1189. 20

[XHZ*22] XUE H., HANG T., ZENG Y., SUN Y., LIU B., YANG H., FU J., GUO B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 19

[XMYR16] XU J., MEI T., YAO T., RUI Y.: Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 5288–5296. 20

[XYF*23] XIAO G., YIN T., FREEMAN W. T., DURAND F., HAN S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431* (2023). 7

[YHL*22] YAO C.-H., HUNG W.-C., LI Y., RUBINSTEIN M., YANG M.-H., JAMPANI V.: Lassie: Learning articulated shape from sparse image ensemble via 3d part discovery. In *NeurIPS* (2022). 16

[YK19] YUAN Y., KITANI K.: Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019), pp. 10082–10092. 17

[YLND23] YESHWANTH C., LIU Y.-C., NIESSNER M., DAI A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. *arXiv preprint arXiv:2308.11417* (2023). 19

[YLT*22] YE V., LI Z., TUCKER R., KANAZAWA A., SNAVELY N.: Deformable sprites for unsupervised video decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022). 10

[YRH*23] YAO C.-H., RAJ A., HUNG W.-C., LI Y., RUBINSTEIN M., YANG M.-H., JAMPANI V.: Artic3d: Learning robust articulated 3d shapes from noisy web image collections. *arXiv preprint arXiv:2306.04619* (2023). 16

[YSI*23] YUAN Y., SONG J., IQBAL U., VAHDAT A., KAUTZ J.: Physdiff: Physics-guided human motion diffusion model. In *ICCV* (2023). 18, 22

[YSJ*21] YANG G., SUN D., JAMPANI V., VLASIC D., COLE F., CHANG H., RAMANAN D., FREEMAN W. T., LIU C.: Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR* (2021). 16

[YSKS23] YU S., SOHN K., KIM S., SHIN J.: Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18456–18466. 8

[YWH*23] YIN L., WANG Y., HE T., LIU J., ZHAO W., LI B., JIN X., LIN J.: Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. *arXiv e-prints* (2023). 18

[YWL*23] YANG S., WU Z., LI M., ZHANG Z., HAO L., BAO W., CHENG M., XIAO L.: Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919* (2023). 18

[YYTK21] YU A., YE V., TANCIK M., KANAZAWA A.: pixelNeRF: Neural radiance fields from one or few images. In *CVPR* (2021). 15

[YZLL23] YANG S., ZHOU Y., LIU Z., LOY C. C.: Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954* (2023). 10

[ZA23] ZHANG L., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023). 6, 9, 10

[ZCC*23] ZOU Z.-X., CHENG W., CAO Y.-P., HUANG S.-S., SHAN Y., ZHANG S.-H.: Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views, 2023. arXiv:2308.14078. 15

[ZCP*22] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022). 18

[ZDW21] ZHOU L., DU Y., WU J.: 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5826–5835. 11, 12

[ZFK*23] ZHANG H., FENG Y., KULITS P., WEN Y., THIES J., BLACK M. J.: Teca: Text-guided generation and editing of compositional 3d avatars. *arXiv* (2023). 16

[ZFY*23] ZOU K., FAISAN S., YU B., VALETTE S., SEO H.: 4D Facial Expression Diffusion Model. *arXiv e-prints* (2023). 16

[ZJGL23] ZHANG F., JI N., GAO F., LI Y.: Diffmotion: Speech-driven gesture synthesis using denoising diffusion model. *arXiv e-prints* (2023). 18

[ZLAH23] ZHANG Z., LIU R., ABERMAN K., HANOCKA R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis. *arXiv preprint arXiv:2307.15042* (2023). 18

[ZLC*22] ZHANG M., LIU C., CHEN Y., LEI Z., WANG M.: Music-to-dance generation with multiple conformer. In *International Conference on Multimedia Retrieval* (2022), p. 34–38. 17

[ZLC*23] ZHAO Z., LIU W., CHEN X., ZENG X., WANG R., CHENG P., FU B., CHEN T., YU G., GAO S.: Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *arXiv preprint arXiv:2306.17115* (2023). 11, 12, 13, 20

[ZMZ*22] ZHANG S., MA Q., ZHANG Y., QIAN Z., KWON T., POLLEFEYS M., BOGO F., TANG S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision (ECCV)* (Oct. 2022). 17

[ZPW*23] 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. Graph. 42*, 4 (2023), 91:1–91:13. doi:10.1145/3592103. 11, 12, 13

[ZQL*23] ZHANG L., QIU Q., LIN H., ZHANG Q., SHI C., YANG W., SHI Y., YANG S., XU L., YU J.: Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv e-prints* (2023). 16

[ZT23] ZHOU Z., TULSIANI S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR* (2023). 15

[ZTF*18] ZHOU T., TUCKER R., FLYNN J., FYFFE G., SNAVELY N.: Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH) 37* (2018). URL: https://arxiv.org/abs/1805.09817. 19

[ZTNW23] ZHANG B., TANG J., NIESSNER M., WONKA P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph. 42*, 4 (jul 2023). doi:10.1145/3592442. 11, 12, 13

[ZVW*22] ZENG X., VAHDAT A., WILLIAMS F., GOJCIC Z., LITANY O., FIDLER S., KREIS K.: Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978* (2022). 11, 12, 13

[ZWL*23] ZHUANG J., WANG C., LIU L., LIN L., LI G.: Dreameditor: Text-driven 3d scene editing with neural fields. *SIGGRAPH Asia* (2023). 15

[ZWY*22] ZHOU D., WANG W., YAN H., LV W., ZHU Y., FENG J.: Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018* (2022). 8

[ZYM*22] ZHENG Y., YANG Y., MO K., LI J., YU T., LIU Y., LIU K., GUIBAS L. J.: Gimo: Gaze-informed human motion prediction in context. *arXiv preprint arXiv:2204.09443* (2022). 17

[ZYW*23] ZHAN F., YU Y., WU R., ZHANG J., LU S., LIU L., KORTYLEWSKI A., THEOBALT C., XING E.: Multimodal image synthesis and editing: The generative ai era. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). 2, 5

[ZZL*23] ZHAO W. X., ZHOU K., LI J., TANG T., WANG X., HOU Y., MIN Y., ZHANG B., ZHANG J., DONG Z., DU Y., YANG C., CHEN Y., CHEN Z., JIANG J., REN R., LI Y., TANG X., LIU Z., LIU P., NIE J., RONG WEN J.: A survey of large language models. *ArXiv abs/2303.18223* (2023). 2