Advances in 4D Generation: A Survey

Qiaowei Miao, Kehan Li, Jinsheng Quan, Zhiyuan Min, Shaojie Ma, Yichao Xu, Yi Yang, Yawei Luo+

(Survey Paper)

***Abstract*—Generative artificial intelligence (AI) has made significant progress across various domains in recent years. Building on the rapid advancements in 2D, video, and 3D content generation fields, 4D generation has emerged as a novel and rapidly evolving research area, attracting growing attention. 4D generation focuses on creating dynamic 3D assets with spatiotemporal consistency based on user input, offering greater creative freedom and richer immersive experiences. This paper presents a comprehensive survey of the 4D generation field, systematically summarizing its core technologies, developmental trajectory, key challenges, and practical applications, while also exploring potential future research directions. The survey begins by introducing various fundamental 4D representation models, followed by a review of 4D generation frameworks built upon these representations and the key technologies that incorporate motion and geometry priors into 4D assets. We summarize five major challenges of 4D generation: consistency, controllability, diversity, efficiency, and fidelity, accompanied by an outline of existing solutions to address these issues. We systematically analyze applications of 4D generation, spanning dynamic object generation, scene generation, digital human synthesis, 4D editing, and autonomous driving. Finally, we provide an in-depth discussion of the obstacles currently hindering the development of the 4D generation. This survey offers a clear and comprehensive overview of 4D generation, aiming to stimulate further exploration and innovation in this rapidly evolving field. Our code is publicly available at: https://github.com/MiaoQiaowei/Awesome-4D.**

***Index Terms*—4D generation, dynamic 3D generation, deep generative modeling, diffusion models**

I. Introduction

With the continuous advancement of generative models, their capabilities have seen remarkable improvements over the past decade. Initially, research in this domain primarily focused on 2D image generation, with generative methodologies such as [1], [2], [3], [4], [5], [6], [7] achieving significant milestones. Subsequently, these approaches expanded to encompass multi-view 2D image generation [8], [9], [10], [11], video generation [12], [13], [14], [15], [16], [17], and 3D content generation [8], [11], [18], [19], each contributing to the rapid evolution of generative techniques. As these methodologies continue to mature, 4D generation, which incorporates the temporal dimension into generative tasks, has emerged as a prominent and rapidly growing research focus [20], [21], [22], [23], [24], [25]. In addition to its academic significance, 4D generation has demonstrated immense potential in various commercial applications, including video games, films, digital humans, and AR/VR. By eliminating the reliance on traditional capture devices, 4D generation methods enable flexible control mechanisms for synthesizing dynamic 4D objects and scenes, paving the way for transformative advancements in human-computer interaction. As illustrated in Fig. 1, the growing interest in 4D generation stems from its ability to bridge cutting-edge research with practical applications, drawing increasing attention from the research community to this burgeoning field.

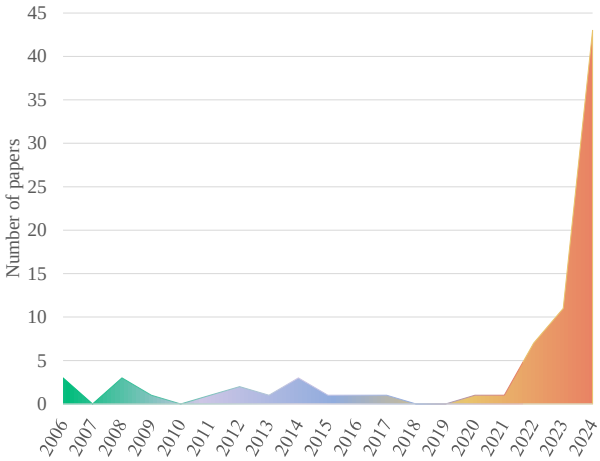

Fig. 1. Trends in the Publication of 4D Generation Research Papers.

4D generation tasks are fundamentally driven by advancements in two core technologies: **4D representation methods** and **diffusion models**. Building upon well-established 3D representation techniques—such as Neural Radiance Fields (NeRF) [26], mesh structures, 3D Gaussian functions[27], and point clouds—researchers have introduced deformation networks [28], [29], [30] to enable temporal modifications of 3D representations. These networks effectively extend static 3D representations by incorporating the temporal dimension, laying the foundation for 4D generation tasks. Complementing this, diffusion models [1], [2], [3], [4], [5], [6], [7] have gained prominence for their flexible control mechanisms and remarkable generative capabilities in image, video, and even 3D content creation. Recent studies have adapted diffusion models to 4D generation by proposing novel frameworks that transfer their generative abilities to the temporal domain, significantly enhancing the quality and realism of generated 4D assets. These advancements in 4D representation methods and diffusion models have not only accelerated the progress of 4D generation but also enabled its application across a wide range of domains. Researchers explores its potential in object generation [31], [32], scene generation [33], [34],

Qiaowei Miao, Kehan Li, Jinsheng Quan, Zhiyuan Min, Shaojie Ma, Yichao Xu, Yi Yang, and Yawei Luo are with the ReLER Lab, Zhejiang University, China.

Corresponding author: Yawei Luo.

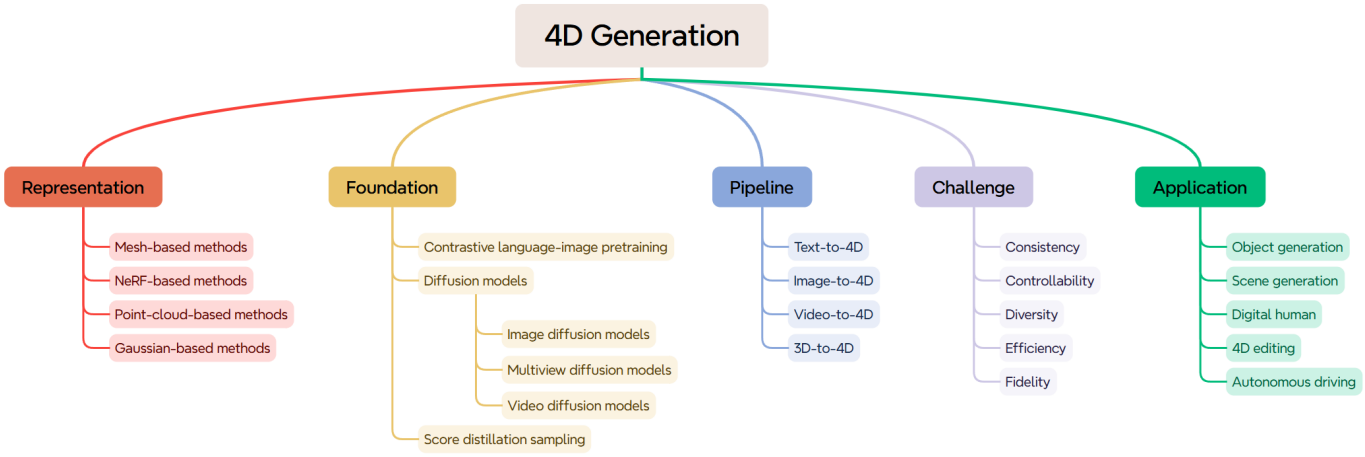

Fig. 2. We present a comprehensive survey of the rapidly evolving field of 4D generation and introduce a systematic three-tier taxonomy to organize the domain. This framework provides a clear structure for understanding key aspects of 4D generation, including representation methods, foundational techniques, pipeline categorizations, existing challenges, and current applications.

digital humans [35], [36], [37], [38], 4D editing [39], and autonomous driving [40], [41], [42]. These diverse applications highlight the versatility of 4D generation techniques, as well as their transformative potential in both research and industrial contexts.

Despite the rapid progress in 4D generation, there is currently no comprehensive survey that consolidates and systematically reviews this emerging field. **This paper aims to fill this gap by providing the first broad and in-depth survey of the latest advancements in 4D generation.** Specifically, it covers key aspects of the field, including 4D data representations, control mechanisms, generation methodologies, research directions, application domains, relevant datasets, and future development trajectories. Given the growing interest in 4D generation tasks within the research community, this survey places particular emphasis on the integration of 4D generation with technologies from related fields. By analyzing the connections and differences between 4D generation and other generative tasks, such as image, video, and 3D generation, this work aims to provide researchers with a clear and systematic understanding of the field. We believe that this survey not only offers a comprehensive perspective for researchers but also serves as a catalyst to inspire further exploration and innovation in the rapidly evolving domain of 4D generation.

Our contributions can be summarized as follows:

- **A novel categorization framework based on control conditions:** We propose a rapid categorization method designed to help researchers efficiently navigate and locate relevant works in the 4D generation field.
- **A comprehensive and timely literature review:** Recognizing the growing significance of 4D generation as a major research focus, we provide an extensive survey that systematically summarizes existing methods and related technologies. This review aims to equip readers with a thorough understanding of the current state of the field.
- **Insights into trends, challenges, and future directions:** We highlight the key development trends in 4D generation, identify the challenges faced by current methods, and offer an in-depth discussion of the opportunities and open problems in this rapidly evolving domain.

II. SCOPE OF THIS SURVEY

This survey is structured to cover the foundational techniques, generation methodologies, current challenges, diverse applications, and future directions of 4D generation, as illustrated in Fig. 2. Specifically, we begin by introducing the foundational representation models that underpin 4D generation, analyzing how existing 3D representation techniques are extended into the temporal dimension (see Sec. III). Next, we explore the technologies enabling the integration of control conditions into the 4D generation process, which is crucial for achieving flexible and accurate generation outputs (see Sec. IV). Building on this, we propose a systematic categorization of 4D generation pipelines into four major types based on different control conditions, providing a clear framework for organizing existing works and identifying research gaps (see Sec. V). We then summarize and analyze five key challenges currently faced in the field, including data acquisition, computational complexity, spatiotemporal consistency, and evaluation metrics, which represent significant barriers to further progress (see Sec. VI). Following this, we present the applications of 4D generation across a variety of domains, such as object generation, scene generation, digital humans, 4D editing, and autonomous driving, highlighting the transformative potential of these technologies (see Sec. VII). Finally, we conclude by discussing emerging trends and potential future directions, aiming to inspire further research and innovation in this rapidly evolving field (see Sec. VIII). By systematically organizing the survey into these interconnected topics, this work provides readers with a holistic understanding of the 4D generation landscape, from its theoretical foundations to its practical applications and future opportunities.

III. 4D REPRESENTATION

A. 4D Mesh

Dynamic meshes are widely used for modeling time-varying surfaces, requiring continuous updates to vertex positions

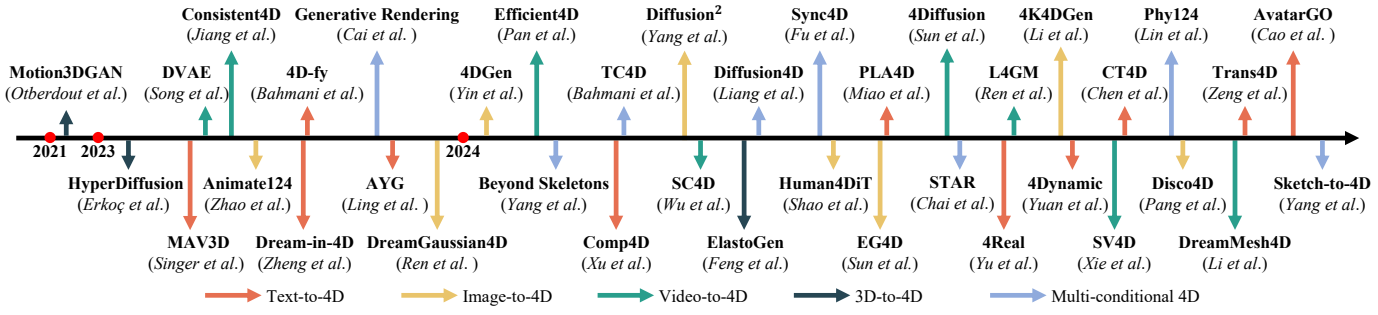

Fig. 3. Timeline of the development of 4D generation methods.

and, in some cases, changes in mesh topology. In character animation and motion capture, dynamic meshes are commonly employed to represent complex human poses and deformations. For instance, the Skinned Multi-Person Linear Model (SMPL) [43] is a widely adopted 3D human body model that represents the human body as a parameterizable triangular mesh. By leveraging linear blend skinning techniques, SMPL generates highly realistic dynamic human deformations. Building upon this, STAR [44] improves pose reconstruction accuracy through sparse optimization techniques, while GHUM [45] further enhances expressiveness by simultaneously capturing diverse body shapes and poses, excelling in marker-less dynamic human modeling.

4D meshes with physical simulations are essential for modeling soft bodies, fluids, and other dynamic phenomena. Methods such as finite element methods [46] or simulations based on elastic models enable dynamic meshes to capture physical properties like elasticity, rigidity, and viscosity. These attributes make dynamic meshes particularly valuable in domains such as medical simulation and engineering analysis. However, large-scale deformations and intricate object interactions pose significant computational challenges for traditional approaches [47]. To address these challenges, recent research has introduced deep learning-based dynamic mesh optimization techniques [48]. By integrating neural networks with physical models, these methods significantly enhance computational efficiency, enabling real-time performance in complex simulations.

Differentiable rendering further expands the capabilities of dynamic meshes by enabling the rendering process to be differentiable, allowing optimization of parameters such as shape, material, and lighting through image-based supervision. Unlike traditional rendering, differentiable rendering directly optimizes geometry and materials without relying on manual design. Recent work has explored the integration of dynamic meshes with NeRF and 3D Gaussian representations [22], [49], [50], leveraging the deformation information from dynamic meshes and the high-precision rendering capabilities of differentiable rendering. These approaches achieve high-quality dynamic scene representation, combining the strengths of explicit deformation modeling and neural rendering.

B. 4D NeRF

With the continuous development of 3D scene representation methods in computer vision, NeRF has shown limitations in modeling dynamic 3D scenes due to its static nature. To address these challenges, researchers have extended NeRF along the temporal dimension, resulting in a class of methods collectively referred to as 4D NeRF.

One of the earliest efforts to adapt NeRF for 4D tasks is Neural 3D Video [51], which introduces temporal latent variables for efficient rendering of 4D scenes. Utilizing hierarchical training and importance sampling, it achieves scalable performance, with the DyNeRF dataset serving as a benchmark for dynamic NeRF modeling. Another direction explores deformation fields, as demonstrated by Nerfies [29] and D-Nerf [28], which extend NeRF to 4D by integrating temporal variables into deformation fields. Nerfies decomposes 4D space into a static NeRF-based template and time-dependent deformation fields, while D-Nerf predicts spatial displacements over time to model dynamic scenes. Similarly, Neural Scene Flow Fields [52] encode both 3D geometry and temporal variations by training a neural scene flow field to capture spatial structure and motion dynamics.

Despite their success in extending NeRF to the temporal domain, these methods often face challenges related to computational efficiency and storage requirements due to the high complexity of 4D representations. To mitigate these issues, researchers have explored spatial feature decomposition strategies for more efficient 4D modeling. HyperReel [53] introduces techniques such as keyframes and importance sampling to decouple dynamic scenes into a ray-based sample prediction network and a compact dynamic volume representation. NeRFPlayer [54], on the other hand, decomposes spatiotemporal space into static, deforming, and newly emerging regions based on temporal characteristics, reducing computational overhead. Additionally, plane-based decomposition methods, such as HexPlane [55] and K-Planes [56], factorize the 4D space-time grid into planar representations for efficient rendering. HexPlane employs six characteristic planes, while K-Planes generalizes planar decomposition techniques to arbitrary-dimensional spaces. Building on these ideas, Tensor4D [57] further compresses the 4D tensor into nine compact feature planes, achieving both low-cost computation and expressive performance.

As 4D NeRF methods continue to evolve, neural radiance fields have become one of the most prominent representation models for dynamic 4D scenes. The growing diversity of NeRF variants offers high-quality rendering and efficient computation, paving the way for broad research opportunities and

applications in areas such as video synthesis, virtual reality, and dynamic scene understanding.

C. 4D Point Cloud

Point clouds represent three-dimensional shapes as a disordered set of discrete samples [58]. Due to their inherently unstructured nature, point clouds can be interpreted either as globally parameterized small Euclidean subsets or with an emphasis on local structures, depending on the task requirements. Dynamic point clouds have been widely applied as a representation model for 4D scenes, offering flexibility in capturing spatiotemporal variations. Early approaches to dynamic point cloud reconstruction focused on multi-view setups. Mustafa et al.[59], [60] performed 4D reconstruction of dynamic scenes using multi-view geometry, requiring wide-baseline views to adequately cover the scene. However, their method suffers from limitations such as view ambiguities and a reliance on sufficient camera coverage, which restrict its scalability to more complex scenarios. Similarly, Wand et al.[61] proposed a technique for reconstructing deforming 3D geometries from point clouds. Despite producing compelling results, their approach is constrained to smooth spatiotemporal movements, assumes dense temporal sampling of point clouds, and incurs substantial computational costs. Another line of research utilizes template models to guide the reconstruction process [62], [63], [64], [65], [66]. Template-based methods provide a valuable framework for both classical and learning-based models, enabling reconstruction with higher accuracy for specific categories such as human bodies, hands, or faces. However, the quality and availability of template models remain critical bottlenecks. Obtaining high-quality templates is often costly and domain-specific, limiting these methods to predefined shape categories [67], [43], [68], [69]. This reliance on templates restricts their generalization to diverse scenes and shapes, making them less suitable for broader applications.

D. 4D Gaussian Splatting

The concept of 4D Gaussian Splatting (4D GS) extends the static 3D GS framework by incorporating a temporal dimension, enabling more effective reconstruction of dynamic scenes. In dynamic scene reconstruction, 4D GS adopts a spatio-temporal encoding approach, where both spatial and temporal information are encoded into Gaussian features. One of the primary challenges in this domain is efficiently modeling the deformation and motion of objects within the scene. 4D GS addresses this by transforming static 3D Gaussians into dynamic counterparts through deformation fields that predict offsets over time [70].

Recent approaches based on 4D GS [71], [30], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81] can be broadly categorized into two paradigms: iterative methods and deformation-based methods. Iterative methods optimize Gaussian parameters frame by frame, gradually adjusting them to adapt to new frames. For instance, D-3DGS [80] models each frame independently after the first and constrains the motion of each Gaussian based on its prior configuration. This approach ensures coherent and interpretable dynamic representations of the scene over time. However, such methods often struggle with occluded or unseen areas in initial frames, where a lack of visibility can result in incomplete reconstructions in subsequent frames. To address this limitation, multi-camera setups are typically required for more robust results.

In contrast, deformation-based methods leverage a canonical representation shared across all frames and apply deformation fields to predict Gaussian offsets over time. This approach avoids frame-by-frame optimization by modeling the motion of Gaussians through time-dependent deformation networks. For example, Yang et al.[71] introduced a deformable 4D Gaussian representation, where a deformation MLP network predicts spatial and temporal adjustments by querying 4D coordinates. Similarly, Wu et al.[30] proposed a deformation-based framework in which a Gaussian deformation field network predicts per-Gaussian offsets at given timestamps. Their approach utilizes a spatio-temporal structure encoder and an MLP-based decoder to model dynamic scenes with high fidelity and computational efficiency.

IV. FOUNDATION TECHNOLOGIES

A. Contrastive Language–Image Pre-training

Recent advancements in multimodal learning have led to the development of models like CLIP [82] (Contrastive Language-Image Pre-training), which learns shared representations between images and text. CLIP jointly trains an image encoder and a text encoder using a symmetric InfoNCE loss [82], ensuring meaningful alignment between the two modalities in a shared embedding space. The model outputs a scalar score to measure the alignment between an image and its associated text, enabling robust cross-modal matching. A key feature of CLIP is its zero-shot learning capability, where textual descriptions embedded by the text encoder effectively synthesize a linear classifier without requiring task-specific fine-tuning. Optimization techniques further enhance its ability to achieve high-quality image-text alignment, making it versatile for various downstream tasks. In the context of 4D generation, CLIP provides a foundation for incorporating text-driven control conditions, allowing users to guide the generation process through natural language descriptions, enabling more intuitive and flexible control over complex outputs.

B. Diffusion Models

Diffusion models [1], [5], [7] have emerged as one of the most rapidly advancing generative modeling techniques, achieving remarkable success in various generation tasks. These models are probabilistic generative frameworks that learn to reverse a forward process that progressively adds noise to the training data. The training process consists of two main phases: the **forward diffusion process** and the **reverse denoising process**.

Forward Diffusion Process. In the forward process, Gaussian noise is gradually added to the input data, progressively destroying its structure until it becomes pure Gaussian noise. Let $p(x_0)$ denote the data density, where x_0 represents the original data. Given a training sample $x_0 \sim p(x_0)$, the noisy

versions $x_1, x_2, \ldots, x_T$ are produced by a Markov process defined as follows:

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

where $t \in [1, T]$, β_t is a variance schedule that controls the amount of noise added at each step, and $\mathcal{N}$ is a Gaussian transition kernel. The full forward process over all steps can be expressed as:

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

Using the reparameterization trick and defining $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, we can directly sample a noisy version x_t at any arbitrary step t conditioned on the original data x_0:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I). \tag{3}$$

This formulation allows efficient computation of noisy data at any step without iterating through all preceding steps.

Reverse Denoising Process. The reverse process is designed to reconstruct the original data x_0 by iteratively denoising a sample starting from pure Gaussian noise x_T. The goal is to model the true data distribution $q(x_0)$ by learning a parameterized approximation $p_\theta(x_0)$. The reverse process is defined as:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{4}$$

where $p_\theta(x_{t-1}|x_t)$ is a learned Gaussian distribution parameterized by a neural network $\epsilon_\theta(x_t, t)$ to predict the noise ϵ added at step t. The training objective is to minimize the difference between the predicted noise $\epsilon_\theta(x_t, t)$ and the actual noise ϵ added during the forward process. This is achieved by minimizing the following simplified loss function:

$$\mathbb{E}_{t\sim[1,T], x_0\sim q(x_0), \epsilon\sim\mathcal{N}(0,I)} \left[\lambda(t)\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \tag{5}$$

where $\lambda(t)$ is a weighting function that emphasizes different steps during optimization.

1) Image Diffusion Models: Diffusion models have achieved remarkable success in the domain of image generation, offering flexible and diverse control mechanisms to enable the synthesis of images aligned with specified conditions. Based on the nature of these control criteria, image diffusion models can be categorized into the following three types:

Text-conditioned Models. Early text-conditioned image generation models, such as Stable Diffusion [83] and Imagen [84], typically consist of a text encoder and a diffusion model. The text encoder extracts textual embeddings, which are subsequently used to guide the generative process of the diffusion model. These models are trained on large-scale text-image datasets, enabling them to learn robust mappings between textual descriptions and corresponding visual outputs. This approach allows users to generate high-quality images based on arbitrary text prompts, making text-conditioned diffusion models highly versatile in creative tasks.

Image-conditioned Models. Image-conditioned diffusion models facilitate diverse image-to-image translation tasks, where the input image serves as the primary control signal. Methods such as CtrLoRA [85] and ControlNet [86] enable transformations based on specific attributes of input images, such as Canny edges, line art, or low-light conditions. By training on paired datasets of $\langle$source image, target image$\rangle$, these models can perform tasks such as enhancing brightness, converting sketches to photorealistic images, or adapting styles. Image-conditioned diffusion models demonstrate strong capabilities in tasks requiring precise visual modifications.

Mixed-condition Models. Mixed-condition diffusion models integrate multiple control signals, such as text [83], [84], images [85], [86], [87], and skeletons [88], [89], to guide the generation process. Leveraging additional control signals enhances the alignment between generated images and user specifications, allowing for more detailed and personalized outputs. For instance, in inpainting tasks, text and masks are combined as control inputs to synthesize semantically consistent content within specified regions of an image [87], [90]. Furthermore, methods like InstructPix2Pix utilize textual instructions alongside input images to transform the source image into a target image according to user-defined guidelines, enabling fine-grained control over the generative process.

2) Multiview Diffusion Models: Multiview diffusion models aim to generate images of the same target from different viewpoints, leveraging multiview geometric principles to ensure consistency across views. Early methods such as Zero-1-to-3 [9], [91], and Zero123++ [10] train multiview diffusion models using the Objaverse dataset series [91], [92]. These models take a reference view and its camera parameters as input and generate images corresponding to different camera poses. In contrast, MVDream [8] introduces a text-driven approach for multiview image generation. By combining multiview training data from Objaverse [92] with 2D text-to-image data from LAION [93], MVDream uses the Stable Diffusion 2.1 framework [94] to generate multiview images directly from textual input, enabling text-guided multiview synthesis.

Despite their effectiveness, methods like Zero-1-to-3 and MVDream face limitations in the number of generated viewpoints and inconsistencies between views. To address these challenges, ConsistNet [95] extends Zero-123 by integrating a multiview consistency module based on geometric multiview principles. This module facilitates information exchange across single-view diffusion processes, increasing the number of generated views to eight while improving cross-view consistency. Furthermore, Liu et al. [96] proposed a synchronized multiview diffusion approach that achieves early consensus between viewpoint diffusion processes, ensuring consistent textures and coherent geometry across views.

3) Video Diffusion Models: With the advancement of generative models, diffusion models have achieved significant progress in video generation tasks. Based on the control conditions, video diffusion models can be categorized into two main types: **text-to-video diffusion models** and **image-to-video diffusion models**.

Text-to-video Diffusion Models. Text-to-video diffusion models generate a continuous sequence of video frames based on textual input. These models are typically trained on large-scale text-video datasets, such as those used by VideoCrafter [13], [12] and Sora [97]. By understanding

scenes, objects, and actions described in the text, these models translate textual descriptions into coherent video sequences, ensuring both logical and visual consistency across frames. Furthermore, dynamic objects within these videos often exhibit viewpoint variations, capturing realistic temporal dynamics. Text-to-video diffusion models have wide applications in creative content generation, video synthesis, and storytelling.

Image-to-video Diffusion Models. Image-to-video diffusion models aim to generate video sequences conditioned on a single image or a series of input images. These models take the input image as a reference to guide the generation of temporally coherent video frames, extending both the spatial and semantic content of the input. For instance, models like Gen-1 [98] and SVD [99] leverage novel diffusion architectures and are trained on large-scale video datasets, using the first frame as the conditioning input. By incorporating motion priors and temporal consistency modules, image-to-video diffusion models can produce dynamic scenes that preserve spatial details and maintain coherence across frames. This approach is particularly valuable for applications such as video extrapolation, creative content generation, and video editing, where a static reference image serves as the foundation for generating temporally consistent video sequences.

C. Score Distillation Sampling

Methods for 4D generation using diffusion models typically rely on SDS (Score Distillation Sampling) [100] optimization or related variants. The central idea of this approach is to align the features of the rendered images with the prior knowledge encoded in the diffusion model, ensuring that the diffusion model accurately predicts the random noise added to the features of the rendered images. The SDS method operates as follows: A 4D model θ renders an image $x = g(\theta)$ at specified camera poses, where g represents the rendering process. The UNet in the diffusion model functions as a denoising network $\epsilon_\phi(x; y, t)$, with y serving as control conditions (e.g., text, skeleton, lighting). By calculating the SDS loss between the predicted and true noise, the gradient direction for optimizing the 4D model parameter θ can be effectively determined:

$$\bigtriangledown_\theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\left(w(t)\epsilon_\phi(x; y, t) - \epsilon)\frac{\partial x}{\partial \theta}\right), \tag{6}$$

where $w(t)$ is a timestep-related weight function of diffusion model. Through extensive optimization, SDS (Score Distillation Sampling) aligns the rendered image x with the diffusion model's prior. This alignment indirectly enhances the geometry and dynamics of the underlying 4D model by ensuring that its representations are consistent with the probabilistic features captured by the diffusion prior. This process facilitates the improvement of the 4D model's structural fidelity and temporal coherence.

V. 4D Generation

Building upon the aforementioned foundational technologies, 4D generation has emerged as a promising direction in generative modeling, enabling the synthesis of dynamic spatio-temporal content. Recent advancements have significantly lowered the barriers to 4D generation, with various methods supporting generation under diverse control conditions. This section provides an overview of current approaches to 4D generation, categorizing them based on their control conditions: text, image, video, 3D, and multi-condition inputs. We summarize the current pipelines of 4D generation as illustrated in Fig. 5.

A. Text-to-4D Generation

Utilizing text as a control condition for 4D generation is both conceptually intuitive and technically demanding. This requires precise alignment between the 4D target's geometry and textual semantics, as well as accurate synchronization of its motion dynamics with the textual description. Building upon advancements in text-to-3D and text-to-video generation, which address geometric and temporal aspects respectively, integrating these methodologies into 4D representation models for text-driven 4D content generation represents a natural and strategic extension of existing approaches.

MAV3D [107] was among the first to explore text-to-4D generation. Building upon models developed for motion-image (MAV) generation, it introduced a three-stage generation method: Stage 1 applies SDS optimization using a text-to-image (T2I) diffusion model, Stage 2 incorporates a text-to-video (T2V) diffusion model with SDS optimization, and Stage 3 further applies SDS optimization at higher resolutions. MAV3D simultaneously utilizes SDS optimization from both T2V and T2I diffusion models to provide priors for texture, geometry, and motion in 4D object generation, while leveraging the text-to-image generation task based on 4D NeRF. Dream-in-4D [22] adopts a similar approach, proposing a two-stage generation method: In the static stage, a motion-view (MV) diffusion model with SDS optimization is used to generate a 3D NeRF, followed by the incorporation of a video diffusion model with SDS distillation to obtain a 4D NeRF model.

With the proposal of the 4D Gaussian Splatting technique [108], [71], [30], Ling et al. introduced a text-to-4D generation method based on the 4D Gaussian model, known as AYG [24]. This method comprises two stages: static generation and dynamic generation. In the initial stage, AYG employs a 3D generation method based on SDS to produce a 4D initialization state from text. During the dynamic generation phase, a text-to-video diffusion model is introduced to provide motion priors, thereby enriching the spatiotemporal dimensions. As AYG utilizes the superior video diffusion model AYL [109] compared to MAV [110], the 4D objects generated by AYG exhibit enhanced dynamic effects. Furthermore, Comp4D employs large language models to generate multiple objects independently and uses LLMs to determine the motion trajectories of these objects based on text, integrating them into a single scene. Comp4D [111] leverages GPT-4 to extract entities from user input text and generates them using 3D generation methods. It then utilizes GPT-4's text-to-image understanding to create Newtonian physics-based motion trajectories, guiding object movement. Finally, multiple dynamic 3D objects are integrated into a single scene.

Recent advancements in 4D generation have shifted from merely emphasizing geometric and temporal consistency to fo-

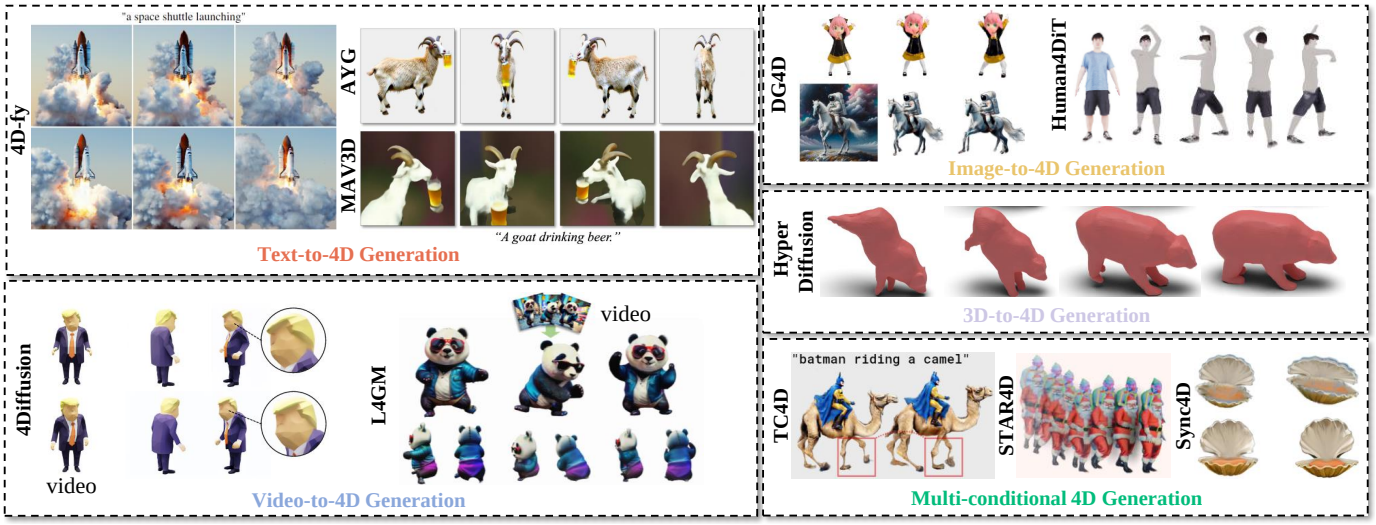

Fig. 4. Representative directions in 4D generation. Based on different control modalities, 4D generation tasks are categorized into five key domains: (1) **Text-to-4D Generation**, where methods such as 4D-fy [21], MAV3D [20], and AYG [24] enable the generation of diverse 4D assets using text as the control condition; (2) **Image-to-4D Generation**, exemplified by DreamGaussian4D [25] (DG4D) and Human4DiT [101], which focuses on faithfully reconstructing 4D assets from input images; (3) **Video-to-4D Generation**, as demonstrated by 4Diffusion [102] and L4GM [32], emphasizes maintaining spatial consistency over time in generated 4D sequences; (4) **3D-to-4D Generation**, like HyperDiffusion [103], extends static 3D assets into the temporal dimension to create dynamic 4D outputs; (5) **Multi-conditional 4D Generation**, showcased by TC4D [104], STAR4D [105] and Sync4D [106], integrates multiple control conditions to achieve precise and controllable 4D generation.

cusing on enhancing control conditions and improving generative outcomes. PLA4D [31] introduces a pixel-aligned pipeline aimed at using text as a control condition, incorporating video generation as an explicit 4D alignment target. This approach allows for the generation of dynamic 4D Gaussian representations akin to video within 15 minutes, using text as the control condition. Similarly, 4Dynamic [49] recognizes that video generation models can not only provide prior guidance through SDS optimization but also use generated videos as reference inputs for 4D generation. 4Real [33] focuses on generating photorealistic 4D scenes, leveraging modifications to the Snap Video Model, a text-to-video diffusion model. These modifications enable the model to generate multiview video frames, which, when combined with existing 4D reconstruction methods, achieve scene-level generative effects. CT4D [112] proposes a three-stage Generate-Refine-Animate (GRA) method for text-driven 4D mesh generation. GRA initially generates a 3D mesh based on text and refines the 3D textures, finally driving the transformation from 3D mesh to 4D mesh by animating the movement of points within the mesh. AvatarGO [113] introduces a zero-shot generation approach in the domain of 4D full-body human-object interactions (HOI), generating initial human models through text-to-3D processes and driving them using the linear blend skinning function from SMPL-X [43], [114].

B. Image-to-4D Generation

Using images as conditions to generate 3D dynamic objects has become a prominent research direction in 4D generation. The task of image-to-4D generation requires not only multiview geometric generation based on the input image but also the expansion of temporal dimensions to create dynamic spatio-temporal content.

Animate124 [115] represents a pioneering study in animating a single in-the-wild image into a 3D video. It adopts a two-stage generation approach: In the coarse stage, a static 3D NeRF is generated using image-to-3D techniques; in the refinement stage, a video diffusion model optimized by Score Distillation Sampling (SDS) is employed to learn subject-specific motion dynamics. DreamGaussian4D [25] is the first image-to-4D generation model based on dynamic Gaussian representation. This method initially generates static Gaussian splats, followed by dynamic generation through Gaussian deformation using HexPlane. To further refine the generated textures, dynamic mesh UV-space texture maps are enhanced with a pre-trained image-to-video diffusion model. The reference image is used as a control condition, while SDS optimization improves texture mapping and temporal consistency. Diffusion2^2 [116] leverages priors from both video and multiview diffusion models, introducing the Variance-Reducing Sampling (VRS) technique to reconcile heterogeneous scores during denoising. This technique mitigates potential conflicts between priors and facilitates the generation of seamless results with high consistency across multiple views and frames. Using these generated image matrices with strong spatiotemporal consistency, existing 4D reconstruction methods can be employed to accomplish the task of image-to-4D representation generation.

Shao et al. proposed Human4DiT [101], an innovative approach for generating high-quality, spatiotemporally coherent 360-degree human videos from a single image. This method employs a hierarchical 4D transformer architecture, efficiently modeling 4D spaces by integrating human features, camera parameters, and temporal signals for precise conditioning. A multidimensional dataset was compiled, and a tailored training strategy was adopted to overcome the limitations of generative adversarial networks and conventional diffusion

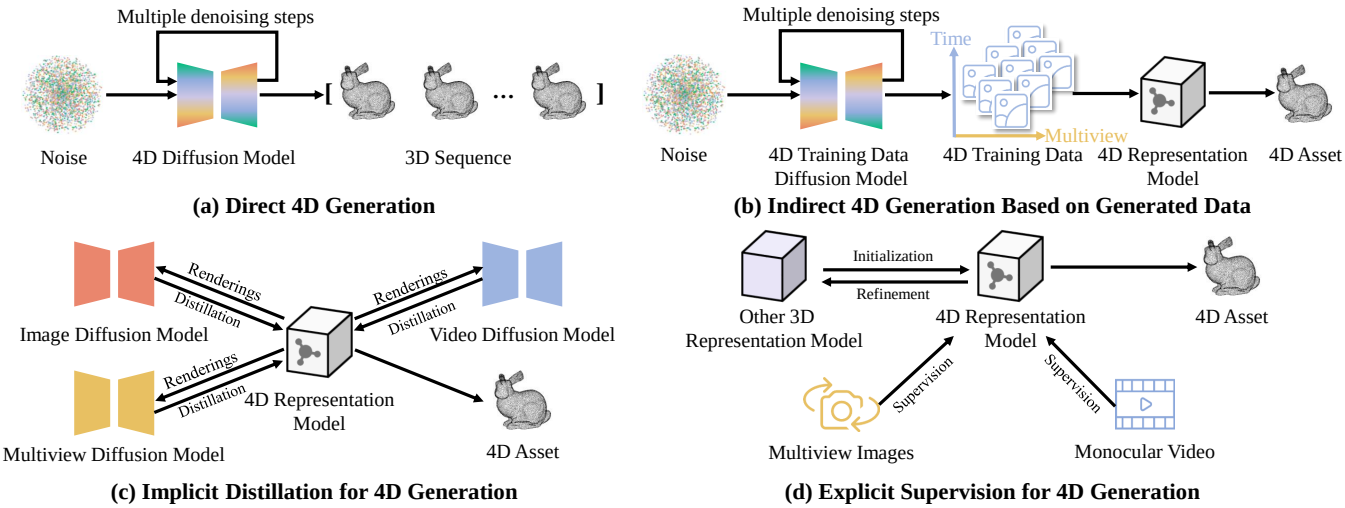

Fig. 5. We summarize the pipelines for generating 4D assets using two approaches: inference-based and optimization-based methods. These pipelines include: (a) directly generating 4D assets conditioned on input parameters; (b) leveraging diffusion models to generate multi-temporal and multi-view training data, enabling indirect 4D generation; (c) combining multiple diffusion models to provide generative priors through implicit distillation, achieving 4D targets via multi-stage training; and (d) utilizing multi-modal data to provide explicit supervisory signals for 4D generation.

models in handling complex motions and viewpoint variations. To address the prolonged training times associated with SDS optimization methods, Sun et al. proposed EG4D [117], an explicit image-to-4D generation method that bypasses score distillation. EG4D utilizes SVD, an image-to-video generation model, to expand temporal dimensions and subsequently employs SV3D, an image-to-multiview diffusion model, to generate multiview images for each frame, forming a temporal and angular image matrix. This matrix is then used in 4D reconstruction methods to build dynamic 4D targets.

4K4DGen [34] presents a method for transforming panoramic images into 4D immersive environments. By adapting 2D diffusion priors to animate 360-degree images, it generates panoramic videos with dynamic scenes. These videos are further enhanced via dynamic panoramic refinement, enabling the creation of 4D immersive environments while maintaining spatial and temporal consistency. Disco4D [118] introduces an innovative Gaussian splatting framework designed to generate vivid, dynamic 4D human animations from a single image. It uniquely separates the processing of clothing and the human body by using Gaussian models for clothing and the SMPL-X model for the human body. This separation enhances the level of detail and flexibility in generation. Disco4D further incorporates technical innovations such as efficient fitting of clothing Gaussians, diffusion-based enhancement of 3D generation, and identity encoding for clothing Gaussians.

C. Video-to-4D Generation

The primary goal of video-to-4D generation is to expand multiview generation from each frame of a video while maintaining temporal consistency across newly generated viewpoints. This task bridges the gap between video content and dynamic 4D representations, enabling the synthesis of spatio-temporal models with coherent geometry and motion.

Song et al. [119] propose a 4D generation method that extracts dynamic information from videos to construct a dynamic knowledge graph (DKG) and develop a variational autoencoder (DVAE) based on DKG. By leveraging kinematic knowledge extracted from video frames, this approach generates animations and automatically produces plausible 3D dynamic objects. With the advancement of 4D NeRF, Consistent4D introduces a method for generating 4D dynamic objects viewable from 360 degrees using monocular videos. To densify sparse data in monocular videos, Consistent4D generates additional video frames through interpolation, achieving multiview interpolation at fixed moments and multi-moment interpolation at fixed views. Explicit supervision with the densified data is combined with SDS optimization from a diffusion model to jointly generate 4D NeRF representations. Efficient4D[120] further accelerates the video-to-4D generation process. Efficient4D employs SyncDreamer [121], a multiview image generation network, to expand single-view videos into a time-by-view image matrix, which is then integrated with a 4D reconstruction network for rapid 4D object generation. SC4D proposes a two-stage generation approach: In the coarse stage, SC4D learns suitable shape and motion initialization using sparse control Gaussians. In the fine stage, these sparse control Gaussians act as implicit control points, enabling Linear Binding Skinning to drive dense Gaussian splats, resulting in detailed 4D representations.

Unlike approaches that directly utilize pre-trained models, 4Diffusion [102] introduces a learnable motion module within the diffusion model, specifically designed for multiview video generation. After training on a carefully curated dataset, 4Diffusion transforms monocular videos into multiview video sequences. Coupled with SDS optimization, it generates 4D representations parameterized by 4D NeRF. Similarly, SV4D [122] focuses on constructing a robust diffusion model capable of multi-moment and multiview generation. Using a filtered 4D rendering dataset, SV4D incorporates a main-view video and initial-moment multiview video as control conditions to train the diffusion model for completing

video frame matrices across non-main views for subsequent moments. By leveraging SV4D's powerful generation capabilities, dynamic representations can be directly synthesized with SDS optimization. DreamMesh4D [108] begins with a coarse mesh obtained through an image-to-3D generation process. Sparse points are uniformly sampled across the mesh surface to build a deformation graph, which drives the motion of the 3D object. This approach enhances computational efficiency while providing additional constraints for dynamic object generation. To optimize the performance of dynamic meshes, DreamMesh4D introduces a skinning algorithm that combines LBS (Linear Blending Skinning) and DQS (Dual-Quaternion Skinning). This technique achieves superior spatial-temporal consistency compared to 4D NeRF and 4D gaussian splatting representation methods.

D. 3D-to-4D Generation

The 3D-to-4D generation task involves transforming a single static 3D representation model into a sequence of 3D models, collectively forming a dynamic representation model. This process expands the spatial geometry of 3D objects into the temporal dimension, enabling the synthesis of 4D content.

A notable example of such an approach is Motion3DGAN and Sparse2Dense Mesh Decoder proposed by Otberdout et al. [123]. Their method takes a 3D facial model and an expression label as input to Motion3DGAN, which predicts the displacement of 3D facial landmarks based on motion dynamics. The Sparse2Dense Mesh Decoder then refines the deformation of the input 3D face by modifying its mesh structure according to these landmark displacements, generating a sequence of expressive meshes. By leveraging the GAN framework, this approach incorporates temporal deformation information into static 3D facial models, thereby accomplishing the 4D generation task. HyperDiffusion [124] introduces a novel approach that directly models neural field representations for dynamic 4D generation. In the initial stage, a set of neural field MLPs is fitted to a dataset of 3D or 4D shapes through an overfitting process. Subsequently, a transformer-based architecture is employed to model a diffusion process directly on the optimized MLP weights, predicting denoised weights and biases as flattened vectors which enables the synthesis of new NeRFs, from which dynamic meshes can be extracted using the Marching Cubes algorithm.

ElastoGen [125] represents a knowledge-driven AI model designed for generating physically accurate 4D elastodynamics. The process begins by rasterizing the input 3D model, including boundary conditions, to derive parameters for the NeuralMTL module. A recurrent neural network (RNN) is then utilized to iteratively relax local stresses using a 3D convolutional approach, while another RNN integrates these local deformations into the overall displacement field of the object. ElastoGen automatically verifies the accuracy of predictions from both RNN loops and outputs dynamic 4D results once the prediction error falls below a specified threshold. This approach ensures the physical plausibility and temporal coherence of the generated 4D elastodynamic models.

E. Multi-conditional Generation

To enhance the controllability of 4D generation results, recent approaches have explored the use of multiple control conditions. These methods aim to enable fine-grained manipulation of the generated 4D outputs by integrating diverse inputs such as text, images, videos, and sketches. This integration allows for greater flexibility and precision in generating dynamic spatio-temporal content, addressing challenges in producing high-quality, customizable results.

Generative Rendering combines the controllability of dynamic 3D meshes with the expressiveness and editability of emerging diffusion models. By injecting real correspondence information from animated, low-fidelity rendered meshes into various stages of pre-trained text-to-image generative models, this method produces high-quality, temporally consistent frames. These frames can then be integrated with sparse 4D reconstruction methods to achieve high-fidelity 4D results. Similarly, TC4D [104] introduces a trajectory-conditioned text-to-4D generation framework, enabling objects to move along user-defined trajectories. Motion is decomposed into global and local components: global motion is represented by rigid transformations along spline-parameterized trajectories, while local deformations are learned under the supervision of text-to-video models to align with these trajectories. This hierarchical motion modeling ensures coherent motion dynamics and temporal consistency. Diffusion4D [126] extends these ideas by leveraging a carefully curated dynamic 3D dataset to develop a 4D-aware video diffusion model. By fine-tuning text-to-3D and image-to-3D models on this dataset, Diffusion4D expands their capabilities into text-conditioned and image-conditioned orbital view generation models, addressing diverse generation tasks and enabling multi-modal control.

Beyond Skeletons [127] focuses on generating coherent 4D sequences by animating 3D shapes with dynamically evolving forms and colors, conditioned on both text and images. To address multiview inconsistencies and artifacts (e.g., translation errors and misalignment) commonly encountered in static 3D human animations, STAR [105] introduces a skeleton-aware text-driven 4D avatar generation framework. It employs in-network motion retargeting techniques to correct mismatched source motions, accounting for geometric and skeletal differences between the template mesh and the target avatar. Furthermore, STAR incorporates a skeleton-conditioned Score Distillation Sampling (SDS) optimization approach to refine surface details and ensure consistent motions in dynamic human models. This combination of skeleton-aware optimization and motion retargeting enables STAR to produce high-fidelity 4D representations with superior temporal and spatial coherence.

Sync4D [106] introduces a physics-driven methodology for creating controllable dynamics within generated 3D Gaussian models using casually captured reference videos. The method transfers motion from reference videos to generated 3D Gaussians across various categories, ensuring precise and customizable motion transfer. Sync4D begins by generating 3D Gaussians from text or image prompts and uses a triplane representation to create a delta velocity field for the Gaussians.

TABLE I
LIST OF THE DIRECTION OF 4D GENERATION METHODS.

Methods	Representation	Optimization	Condition	Main Motivation
Motion3DGAN [123]	mesh	L1	mesh	Diversity
HyperDiffusion [124]	Implicit neural field	BCE	MLPs	Fidelity
ElastoGen [125]	Grid	Fitting error	Physical parameters	Consistency
Consistent4D [23]	NeRF	SDS	Video	Consistency
SV4D [122]	NeRF	MSE	Video	Consistency
4Diffusion [102]	NeRF	SDS	Video	Consistency&Fidelity
SC4D [130]	Gaussian	SDS	Video	Consistency&Efficiency&Fidelity
L4GM [32]	Gaussian	MSE	Video	Consistency
Efficient4D [120]	Gaussian	SDS+MSE	Video	Consistency&Efficiency
DreamMesh4D [108]	Gaussian+Mesh	SDS+MSE	Video	Efficiency&Controllability
Sync4D [106]	Gaussian+Material Point	Displacement loss	Video+Text/Image	Consistency&Controllable&Fidelity
DKG4D [131]	VAE	MSE	Image	Consistency&Controllability
Diffusion^2 [116]	Gaussian	MSE	Image	Consistency&Efficiency&Fidelity
Disco4D [118]	Gaussian	SDS+MSE	Image	Consistency&Fidelity
DreamGaussian4D [25]	Gaussian	SDS+MSE	Image	Consistency&Controllability&Diversity
4K4DGen [34]	Gaussian	L1	Panorama	Consistency
Human4DiT [101]	Video	Sample loss	Image+Dynamic SMPL sequences	Consistency&Controllability&Efficiency
STAR [105]	3D model+Deformation net	SDS	Text	Consistency&Diversity&Fidelity
Dream-in-4D [132]	NeRF	SDS	Text	Consistency&Fidelity
MAV3D [107]	NeRF	SDS	Text	Consistency
4D-fy [133]	NeRF	SDS+MSE	Text	Fidelity
CT4D [112]	NeRF+Mesh	SDS+MSE	Text	Consistency
AYG [24]	Gaussian	SDS	Text	Consistency
Trans4D [134]	Gaussian	SDS	Text	Efficiency
Comp4D [111]	Gaussian	SDS	Text	Controllability&Fidelity
AvatarGO [113]	Gaussian	SDS	Text	Fidelity
4Real [33]	Gaussian	SDS+MSE	Text	Consistency&Efficiency
AvatarCLIP [135]	mesh	CLIP-guided loss	Text	Consistency&Fidelity
GAvatar	Gaussian+mesh	SDS	Text	Consistency&Efficiency&Fidelity
PLA4D [31]	Gaussian+Mesh	SDS+MSE	Text	Consistency&Controllability&Diversity&Efficiency&Fidelity
Animate124 [115]	NeRF	SDS+MSE	Text+Image	Fidelity
Beyond Skeletons [127]	Mesh	L1	Text+Image	Consistency&Controllability&Fidelity
EG4D [117]	Gaussian	L1	Text+Image	Consistency
Phy124 [128]	Gaussian+Material Point	SDS	Text+Image	Consistency&Controllability&Efficiency&Fidelity
Generative Rendering [136]	Video	-	Text+Mesh	Consistency&Controllability
Sketch-2-4D [129]	NeRF	SDS	Text+Sketch	Consistency&Controllability
TC4D [104]	NeRF	SDS+MSE	Text+Trajectory	Consistency&Controllability
Diffusion4D [126]	Gaussian	L1	Text/Image	Consistency&Efficiency
4DGen [137]	Gaussian	SDS+MSE	Text/Image	Consistency&Controllability
4Dynamic [138]	Grid	SDS+MSE	Text/Video	Consistency&Diversity

This motion sequence is then optimized using differentiable Material Point Method (MPM) simulation, enabling the transfer of motion dynamics from the reference video to the generated 3D Gaussian. By integrating these motion dynamics, Sync4D produces high-fidelity 4D results with consistent temporal and spatial behavior. Similarly, Phy124 addresses the absence of physical priors in diffusion models and the extended optimization time by proposing a fast, physics-driven 4D content generation method. Phy124 [128] treats each 3D Gaussian kernel as a particle within a continuum, assigning physical properties such as density and mass to the particles. By employing MPM simulation and incorporating external forces derived from user-provided text inputs, Phy124 ensures that the resulting 4D content adheres to natural physical laws while achieving superior computational efficiency.

Sketch-to-4D [129] focuses on enabling user-specified control through the simultaneous use of text and sketches in 4D generation. This method introduces a novel controlled Score Distillation Sampling approach, allowing for precise control over the dynamics of generated 4D scenes. To address spatial and temporal inconsistencies introduced by sketch-based control, Sketch-to-4D incorporates spatial consistency and temporal consistency modules into its framework. These modules ensure that the generated results maintain high-fidelity 4D representations while adhering to the user-specified constraints. By integrating text and sketches, Sketch-to-4D provides a flexible and intuitive interface for controlling 4D content generation.

VI. 4D GENERATION CHALLENGES

While several works currently exist on 4D generation tasks, allowing for the creation of object-level to scene-level 4D assets with various control conditions, these methods are primarily motivated by addressing five core challenges: consistency, controllability, diversity, efficiency, and fidelity (see Tab. I).

A. Consistency

Consistency is a critical challenge in 4D generation, encompassing both geometric and temporal aspects. Geometric consistency ensures that observations of objects or scenes from different viewpoints at a single moment remain coherent, preserving stable geometric relationships. Temporal consistency, on the other hand, refers to the visual continuity of object motion across different time points from a fixed viewpoint. For 4D generation, maintaining both geometric and temporal consistency is essential to deliver a satisfactory visual experience when viewpoints and timestamps change simultaneously.

To address consistency issues, some methods [139] incorporate physics-based simulations to enhance motion plausibility during dynamic processes. These approaches bind static 3D keypoints to particles within a physics engine, allowing

particle motion to be governed by physical principles such as mass and momentum conservation. This motion drives the movement of the static 3D keypoints, enabling the dynamic evolution of the generated 4D assets. By ensuring that particle dynamics align with physical laws, these methods improve geometric consistency across the temporal dimension, resulting in motion states that are both realistic and visually coherent.

Other methods focus on optimizing deformation networks and modifying diffusion models to directly enhance consistency. For instance, 4Diffusion [102] modifies ImageDream [140] by fine-tuning it to learn temporal patterns, while 4K4DGen [34] integrates pre-trained diffusion models with depth information to ensure geometric consistency across varying viewpoints and timestamps. Generative Rendering strengthens visual consistency by enhancing self-attention layers in image generation models, introducing correspondence-aware feature blending between input and output features. Moreover, it employs UV-space noise initialization combined with correspondence-aware attention to improve cross-frame generation consistency. Similarly, L4GM [32] incorporates an additional cross-view self-attention layer into the U-Net architecture, learning consistency patterns from data to directly generate 4D targets from input images.

Another class of methods generates multi-view, multi-temporal image grids as intermediate 4D datasets, which are subsequently reconstructed into 4D assets. Diffusion^2 [116] introduces the Variable Reconciliation Strategy (VRS) to harmonize heterogeneous scores during denoising, mitigating conflicts and producing smoother results. Diffusion4D [126], Human4DiT [141], SV4D [122], and 4DGen [137] further retrain diffusion models on curated 4D datasets to generate consistent 4D data. Hybrid approaches, such as Consistent4D [23], and 4Real [33], combine models like SVD [99] and SV3D [142] to infer and generate 4D image grids. These models also employ advanced image interpolation techniques to ensure grid smoothness and consistency under varying viewpoints and across temporal changes.

To further enhance consistency, some researchers integrate multiple 4D representation models to leverage complementary strengths. For instance, CT4D [112] and DreamGaussian4D combine Gaussian and Mesh representations. Gaussian representations excel at driving smooth and consistent geometry, which complements the detailed structure modeling capabilities of Mesh-based approaches. This fusion delivers 4D assets with superior consistency compared to using standalone representation methods, demonstrating the benefits of multi-representation integration in generating coherent 4D results.

B. Controllability

Controllability for 4D refers to the ability to enable users to finely manipulate both the process and outcomes of 4D generation. To enhance the controllability , current approaches primarily follow two directions: leveraging multi-representations or integrating diverse control conditions.

Multi-representation methods aim to combine the strengths of different representations to achieve more flexible and controllable 4D generation. For instance, DreamMesh4D [108] fuses Gaussian and mesh representations by binding Gaussian points with mesh surface patches. This design allows Gaussian models to capture motion information while leveraging mesh structures for high-quality rendering, thus balancing temporal flexibility and visual fidelity. Similarly, Sync4D [106] and Phy124 [128] integrate Gaussian representations with the Material Point Method (MPM). In these methods, physically simulated material points guide the motion of Gaussian points, ensuring that the generated 4D outputs adhere to physical principles and exhibit physically plausible motion dynamics.

Condition integration methods focus on injecting external guidance into the generation process to enhance controllability. For example, Human4DiT [101] incorporates human identity, temporal information, and camera parameters into its network modules, enabling precise control over 4D video generation. Generative Rendering [136] uses dynamic meshes as input to guide pre-trained text-to-image generative models, enabling direct control over rendered 4D outputs. To control motion, TC4D [104] leverages input motion trajectories, while Beyond Skeletons [127] incorporates human skeletons to refine motion dynamics. Similarly, DG4D [25] and PLA4D [31] utilize video supervision to guide the motion of generated 4D assets. For appearance control, Sketch-2-4D [129] generates outputs based on user-provided sketches, and 4DGen [137] expands control capabilities by allowing textual or visual inputs to guide the generation process.

These methods collectively advance the field of controllable 4D generation by introducing diverse representations and integrating external control mechanisms, making the generation process more flexible, precise, and user-friendly.

C. Diversity

In the context of 4D generation, diversity refers to the ability of models to produce a wide range of variations in 4D outputs, encompassing differences in motion, appearance, shape, textures, and other attributes, while maintaining coherence and realism. Diversity is a critical aspect of 4D generation as it reflects the model's capacity for versatility and adaptability in addressing complex generative tasks. STAR [44] departs from traditional two-stage methods that first learn static representations and then inject motion into static targets. Instead, STAR directly updates geometry, texture, and motion jointly through text guidance, enabling the generation of a broader range of dynamic human motion outputs. This approach significantly enhances diversity by integrating multiple attributes in a unified framework. Similarly, DG4D [25] and PLA4D [31] leverage diverse videos as intermediate control conditions. The diversity of their generated outputs is closely tied to the variety of input videos, allowing the models to produce outputs with a wide range of motions and appearances. 4Dynamic [49] further explores diversity by employing the Score Distillation Sampling (SDS) method based on 2D diffusion models. This approach transfers the inherent diversity of 2D image generation into the 4D domain, enabling the synthesis of 4D outputs with rich variations in shape, motion, and textures. By leveraging the strengths of diffusion-based

methods, 4Dynamic expands the generative potential of 4D tasks to include broader variations while maintaining temporal and spatial coherence.

Despite these advancements, discussions on the diversity of 4D generation remain limited. As generative models continue to evolve, diversity is expected to emerge as a critical metric for evaluating their performance. Future research will likely focus on developing more advanced techniques to enhance diversity across multiple dimensions of 4D outputs, addressing the growing demand for versatile and realistic 4D content.

D. Efficiency

Efficiency in 4D generation is a critical challenge, as most current methods rely on SDS optimization to introduce priors, which is computationally expensive and time-consuming. To address these limitations, various approaches have been proposed to improve the efficiency of 4D generation by avoiding extensive optimization, leveraging efficient representations, or utilizing intermediate outputs.

Efficient4D [120] avoids optimization-based generation by directly creating dense multi-frame, multi-view training data. This approach leverages existing advanced 4D reconstruction methods to efficiently construct 4D outputs, significantly reducing computational overhead. Similarly, Diffusion^2 [116] extends input images into multi-frame, multi-view data matrices using image-to-multi-view and image-to-video models, which are subsequently used with 4D reconstruction pipelines. These methods focus on directly creating 4D training datasets, bypassing the need for iterative optimization. Diffusion4D [126] further advances this concept by building a unified diffusion model that maps text or images to 4D datasets. By training on carefully curated datasets, Diffusion4D utilizes the model's generalization capability to directly generate 4D training data based on user-provided text or images, improving both efficiency and usability. Other methods focus on efficient representations or intermediate guidance to accelerate the 4D generation process. Phy124 [128] employs Gaussian point-based representations, demonstrating that explicit models such as Gaussian representations can effectively reduce training time and computational cost. PLA4D [31] reduces reliance on SDS optimization by utilizing intermediate outputs, such as images or videos, as pixel-level guidance targets during the 4D generation process. This strategy not only accelerates training but also improves the stability of the generation process. Trans4D [134] also adopts Gaussian point-based representations, emphasizing the control of Gaussian point quantity to enhance both training and rendering efficiency. Similarly, 4Real [33] introduces a thresholding mechanism to filter Gaussian points, striking a balance between computational efficiency and visual quality.

These methods collectively aim to reduce computational overhead and accelerate the 4D generation process, while ensuring that the generated content maintains fidelity and consistency. By addressing the efficiency bottleneck, these approaches enable scalable and practical applications of 4D generation in real-world scenarios.

E. Fidelity

Fidelity in 4D generation refers to the accuracy and realism of the generated content in capturing detailed geometry, motion, and textures while adhering to input conditions and maintaining temporal consistency. Several methods have been developed to enhance fidelity in 4D generation, focusing on different aspects of input adherence and output realism.

4Diffusion [102] aims to capture multi-view spatial-temporal correlations for multi-view video generation, thereby creating 4D data. By leveraging the generalization capability of generative models, it ensures that the generated 4D data faithfully represent the input images. SC4D [130] employs mean squared error (MSE) loss between real and generated videos to maintain fidelity in both shape and motion, ensuring that the output closely matches the input characteristics. Animate124 [115] incorporates ControlNet to impose additional control constraints, thereby enhancing the fidelity of the generated 4D outputs to the input images. This approach allows for more precise adherence to the original input conditions, improving overall output quality. Similarly, Disco4D [118] isolates Gaussian representations of clothing to optimize them separately, ensuring that changes in attire do not affect the underlying human body structure, thus maintaining consistent fidelity. Dream-in-4D [132] and 4D-fy [21] utilize guidance from multiple diffusion models to enhance the fidelity of the generated 4D results to the input text. These models leverage diverse diffusion processes to ensure that the semantic and stylistic elements of the input text are accurately reflected in the 4D outputs. Finally, Comp4D [143] addresses the challenge of multi-object generation by independently generating and optimizing each object. This approach ensures that each target maintains high fidelity individually before being combined into a cohesive scene, thus preserving the integrity and realism of each component within the larger context.

These methods collectively advance the field of 4D generation by focusing on fidelity, ensuring that generated content is both accurate and realistic in accordance with the specified input conditions.

VII. 4D Generation Applications

A. Object Generation

The generation of 4D objects refers to the creation of specified single objects in 4D space based on given conditions. Currently, 4D object generation faces challenges in terms of consistency, controllability, fidelity, and efficiency.

Temporal and spatial consistency remain major challenges in 4D object generation. Temporal consistency issues often arise from reliance on pre-trained video diffusion models, which suffer from motion discontinuity and limited generation durations. Spatial consistency, on the other hand, is hindered by the limitations of 2D prior diffusion models, which lack 3D awareness. This often leads to inconsistent or repetitive 3D structures when observing objects from different viewpoints, a phenomenon known as the Janus problem. To address these issues, one approach is to use 3D-aware diffusion models, which generate images from different camera viewpoints.

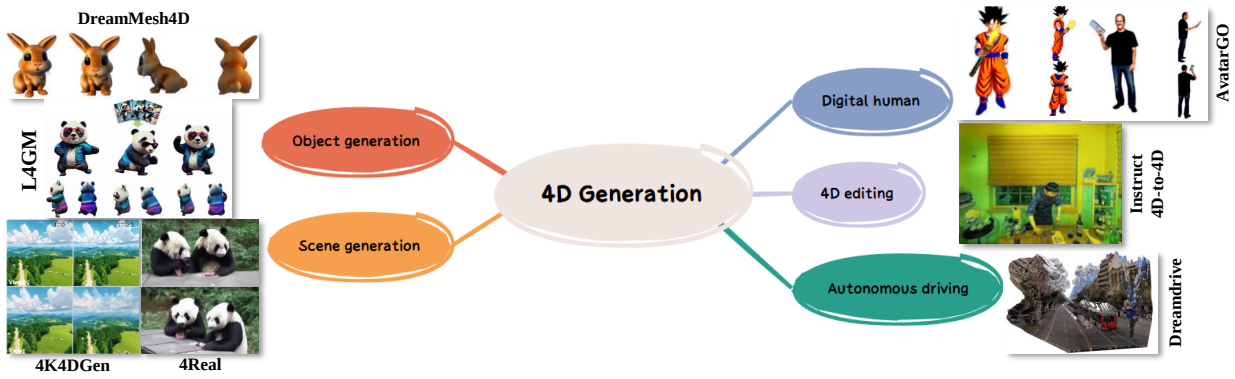

Fig. 6. Applications of 4D generation across five critical domains: (1) **Object Generation**, which focuses on capturing and synthesizing dynamic, time-evolving objects with fine-grained geometric and appearance details, enabling advancements in robotics, virtual simulations, and object recognition; (2) **Scene Generation**, aimed at reconstructing complex spatiotemporal environments with high semantic fidelity and geometric accuracy, supporting immersive experiences in virtual/augmented reality and digital twins; (3) **Digital Humans**, emphasizing the creation of lifelike human avatars with temporal coherence, realistic motion, and detailed expressions, applicable in gaming, telepresence, and virtual assistants; (4) **4D Editing**, facilitating intuitive spatiotemporal manipulation of dynamic content, supporting tasks such as motion retargeting, shape deformation, and appearance editing over time; and (5) **Autonomous Driving**, utilizing 4D generation to enhance perception, prediction, and decision-making pipelines by modeling dynamic objects and scenes, ultimately improving safety and efficiency in real-world driving scenarios. These applications collectively underscore the transformative potential of 4D generation in bridging spatial and temporal dimensions across a wide range of practical and research contexts.

Methods such as [24], [133], [22] combine pre-trained diffusion models (e.g., text-to-image, text-to-3D, text-to-video) with SDS optimization to improve 3D structure. Alternatively, multi-view video diffusion models can generate images across multiple viewpoints and time instances, which are then used to produce 4D content from sparse images [144], [122]. These methods advance spatial and temporal consistency but face challenges in scalability and computational cost.

Controllability and fidelity are critical for creating realistic and precise 4D objects. Current research focuses on text-driven, image-driven, and video-driven approaches [138], [34], [104]. However, textual descriptions often lack sufficient spatial detail for complex scenarios, images are inherently static, and videos struggle to generate new viewpoints or dynamic content. Furthermore, most methods rely on volume or neural rendering, which are less compatible with rasterization pipelines, while polygonal representations often result in overly complex models with excessive faces [108], [145]. Addressing these limitations requires more flexible and precise input conditions to improve both controllability and fidelity, especially in high-complexity scenarios.

Efficiency is another pressing challenge in 4D object generation. SDS-based methods are computationally expensive, and NeRF-based approaches suffer from low rendering efficiency compared to GS-based methods. While some works achieve high-quality 4D generation by combining multiple pre-trained diffusion models, generating a single model can take several hours [115], [22]. To address these inefficiencies, DreamGaussian4D introduces 3D Gaussian splatter and HaxPlan techniques, which explicitly represent 4D content and learn motion from video. These innovations reduce generation time to just a few minutes, significantly accelerating the optimization process [25]. Such advancements highlight the potential of explicit representations for improving efficiency in 4D generation.

Through advancements in consistency, controllability, fidelity, and efficiency, 4D object generation is evolving to address the challenges of dynamic, high-fidelity content creation. However, further development is needed to balance computational efficiency with high-quality outputs in complex scenarios.

B. Scene Generation

4D scene generation focuses on modeling dynamic interactions among multiple targets, global spatial organization, and the overall spatiotemporal consistency of the scene. The generated outputs consist of dynamic scenes that include multiple targets and backgrounds, emphasizing the relationships between targets and their coupling with the environment. In contrast, 4D object generation prioritizes the geometry, texture, and temporal motion changes of a single target, with a technical focus on localized dynamic modeling and detail fidelity. The outputs in 4D object generation are typically independent dynamic targets, without considering interactions with other targets or the background. Several methods have been proposed to tackle the challenges of 4D scene generation. For instance, 4K4DGen [34] supports the generation of large-scale 4D panoramic scenes, where users can adjust viewpoints using a mouse and observe dynamic landscapes from new perspectives. Generative Rendering [136] produces 4D datasets containing multiple targets and complex backgrounds. These datasets can be directly integrated with 4D reconstruction models to generate complete dynamic scenes, ensuring a high level of spatiotemporal consistency. Similarly, 4Real [33] generates reference and freeze-time videos based on input text and uses these intermediate outputs to guide 4D generation. This approach enables the inclusion of video-like rich backgrounds, resulting in more visually coherent and immersive 4D scenes. However, some methods [129], [111], [20] that claim to perform 4D scene generation often lack realistic backgrounds and primarily focus on individual targets with limited environmental interactions. Consequently, such

methods are better suited for object-centric tasks rather than true scene-generation tasks, which require the synthesis of realistic dynamic interactions within a cohesive environment.

C. Digital Human

Digital human refers to a highly detailed, realistic, and interactive virtual representation of a human being, typically generated through advanced construction and generation technologies. It is significant to distinguish digital human generation from human body reconstruction. The digital human introduced in this section refers to a dynamic 3D human avatar generated from text, images, or other types of prompts like skeleton, excluding dynamic human individuals reconstructed using multi-view data [36], [146].

AvatarCLIP [135] is among the first methods to generate dynamic 3D avatar models, which enables users to generate a customized 3D avatar with mesh-based representation and animate it using natural language-described motions. AvatarCLIP is capable of generating 3D avatar models with high-quality textures and vivid motions. GAvatar [147] introduces a Gaussian-based representation and leverages neural implicit fields to predict Gaussian attributes, enhancing animation stability and learning efficiency. In another way, Human4DiT [141] trains a hierarchical 4D transformer-based model that factorizes self-attention spatiotemporally to generate high-quality coherent 360-degree human videos from a single image, collecting a multi-dimensional dataset of images, videos, multi-view data, and limited 4D footage. AvatarGO [113] proposes a novel framework that generates animatable 4D human-object interaction(HOI) scenes through LLM-guided contact retargeting and correspondence-aware motion optimization, outperforming existing methods in generating robust and realistic 4D HOI animations. As for other types of prompts, STAR [44] introduces a skeleton-aware text-driven 4D avatar generation framework that integrates in-network motion retargeting and skeleton-conditioned SDS optimization to correct motion mismatches and refine surface details, achieving high-fidelity 4D representations with superior temporal and spatial coherence.

In conclusion, 4D generation technology empowers the creation of digital humans effectively and efficiently. As 4D generation techniques continue to evolve and become more sophisticated, they enable the generation of digital humans with greater detail, diversity, and realism. This advancement opens up new possibilities for creating highly refined and varied digital representations, paving the way for more immersive and dynamic human models.

D. 4D Editing

4D editing refers to generating or modifying 4D content within a specified region based on user instructions or prompts. Depending on the scope of the editing region, it can be classified into local editing and global editing. Based on functionality, 4D editing can be further categorized into object addition and removal, attribute modification, and stylization tasks.

Several methods have been proposed to tackle the challenges of 4D editing. Instruct 4D-to-4D [39], as the first method for pseudo-3D editing, leverages Instructpix2pix [148] to edit anchor images and propagates changes across the scene using optical flow. While effective for stylization tasks, it struggles with object addition, removal, and attribute modification due to its reliance on image-based editing techniques. 4D-Editor [149] introduces hybrid radiance fields and semantic fields distilled from the DINO teacher model, enabling users to interactively edit objects in dynamic scenes by selecting a reference view and marking on it. This method excels in object addition and removal but is less effective for attribute modification and stylization. Alternatively, Control4D [150] proposes GaussianPlanes for human portrait editing based solely on text instructions. By integrating GaussianPlanes with a GAN-based generator and a 2D diffusion-based editor, Control4D ensures consistent and high-quality temporal editing for dynamic human portraits.

Despite these advancements, 4D editing still faces significant challenges, primarily due to the intricate requirements for spatiotemporal consistency. Temporal continuity is critical for ensuring smooth transitions across adjacent frames, while spatial coherence across multiple viewpoints adds another layer of complexity. These dual demands require advanced algorithms capable of seamlessly integrating dynamic elements while preserving both temporal and spatial fidelity. Consequently, 4D editing remains a highly challenging area, necessitating further research and innovation to address these issues and unlock its full potential for future applications.

E. Autonomous Driving

The field of autonomous driving has made significant strides in recent years, with advances in perception, planning, and control. However, one of the key challenges lies in enhancing the vehicle's ability to understand and predict dynamic environments. Traditional approaches often struggle to generate realistic and temporally consistent visual data for autonomous vehicles. To address these challenges, researchers are increasingly leveraging the capabilities of 4D generation technologies, which integrate spatial and temporal dimensions to simulate dynamic driving environments.

Advancements in 4D visual data generation have significantly impacted autonomous driving. Methods like MagicDrive3D [40] use diffusion priors but face challenges in novel view synthesis. DreamDrive [41] addresses this by employing video diffusion models and a hybrid Gaussian representation to enhance temporal consistency, generalization, and visual quality. Similarly, Stag-1 [151] integrates sparse point cloud reconstruction with video diffusion to achieve both cross-view and temporal consistency in autonomous driving simulations.

The emergence of world models [152] has opened new possibilities for vehicles to better perceive, predict, and plan in autonomous driving environments. World models rely heavily on 4D generation to simulate dynamic driving scenarios, enabling systems to learn driving behaviors and strategies by integrating spatial and temporal information. For instance, DriveDreamer [42] constructs a world model exclusively

from real-world driving scenes, facilitating a comprehensive understanding of structured traffic constraints and enabling precise video generation. To efficiently generate long video sequences, OccSora [153] introduces diffusion-based techniques to directly synthesize 4D scene representations, establishing a dynamic 4D occupancy world model.

While 4D generation highlights its transformative potential in advancing autonomous driving, significant challenges remain. Achieving high visual fidelity in novel viewpoints and maintaining consistency across long temporal sequences are critical hurdles. Overcoming these limitations will be essential for the widespread adoption and real-world deployment of these technologies.

VIII. Discussion

The field of 4D generation has witnessed remarkable progress in recent years, driven by advances in modeling dynamic spatiotemporal phenomena and diverse application scenarios. However, several critical challenges remain unresolved, and addressing these issues is essential for unlocking the full potential of 4D generation technologies in both research and practical applications. In this section, we identify and discuss four core challenges that must be addressed to advance the field: (1) developing comprehensive and standardized evaluation metrics, (2) improving model design and optimization strategies, (3) enhancing generalization across diverse application domains, and (4) understanding and mitigating the broader social impacts of 4D generation technologies.

A. 4D Dataset

Building on the experience of 3D generation, a multi-view diffusion model can be trained on 3D representation models using the Score Distillation Sampling (SDS) technique to provide priors. Extending this approach to 4D generation involves training a diffusion model that incorporates both temporal and view variations, resulting in a 4D representation model via SDS. However, the primary bottleneck in training such a model is the lack of comprehensive 4D datasets. A 4D dataset refers to a collection of multi-view images of a target captured at multiple time instances, accompanied by their corresponding camera parameters.

To address this limitation, previous methods [122], [126] have proposed constructing datasets that disentangle viewpoint and temporal variations, enabling the training of 4D diffusion models. For instance, Diffusion4D [126] curates a high-quality dynamic 3D dataset from the large-scale 3D repositories Objaverse-1.0 and Objaverse-XL [92], [91]. Since Objaverse primarily consists of static 3D assets, many of which exhibit low quality (e.g., partial scans or missing textures), Diffusion4D applies a series of empirical rules to filter and refine the dataset. This includes employing the Structural Similarity Index Measure (SSIM) to evaluate the temporal dynamics of assets, removing those with either overly subtle or extreme motions, as well as those extending beyond scene boundaries. Ultimately, Diffusion4D collects 54,000 high-quality dynamic assets, providing a robust resource for downstream 4D generation tasks.

Similarly, SV4D [122] utilizes Objaverse [92] to construct its own 4D dataset, named ObjaverseDy. To ensure high-quality data, SV4D measures motion magnitude by subsampling keyframes from videos and applying a thresholding approach based on the maximum L1 distance between consecutive frames. During rendering, they dynamically adjust the camera distance to ensure that objects remain fully visible in all frames. Starting from an initial base distance, the camera distance is incrementally increased until the object fits entirely within each frame. Additionally, the temporal sampling step is dynamically adapted, starting with an initial value and progressively increasing until the L1 distance between consecutive keyframes exceeds a predefined threshold. These steps enable SV4D to create a high-quality and dynamic 4D dataset.

Despite these advancements, current 4D datasets are primarily constructed by selecting subsets from large-scale 3D datasets, resulting in a noticeable gap in data scale compared to 3D datasets. Furthermore, existing 4D datasets are typically limited to rendered images and lack corresponding textual descriptions. Building large-scale text-4D datasets is a foreseeable and important direction for future development. Such datasets would enable the research community to construct more multimodal 4D generation models, thereby unlocking new possibilities in 4D content creation and applications.

B. Model Design and Optimization

In current 4D generation tasks, many approaches [21], [31], [24], [137], [22] adopt multi-stage processing pipelines that leverage 2D diffusion models, multi-view diffusion models, and video diffusion models, optimized via Score Distillation Sampling (SDS), to generate 4D assets. While effective to some extent, this multi-stage optimization strategy suffers from several limitations, including challenges in consistency, controllability, efficiency, diversity, and fidelity.

Consistency remains a critical challenge in 4D generation. Current methods lack a unified diffusion model capable of simultaneously ensuring spatial and temporal consistency during 4D asset generation via SDS. Instead, existing approaches rely on multi-view diffusion models and video diffusion models separately to maintain spatial and spatiotemporal consistency. However, this separation introduces inherent limitations: multi-view diffusion models lack temporal consistency priors, while video diffusion models lack multi-view priors. This mismatch often results in inconsistencies in the generated 4D assets, such as appearance discrepancies during viewpoint transitions or spatiotemporal transformations [109], [13], [110]. These issues significantly degrade the visual coherence and quality of the outputs.

Moreover, the multi-diffusion model optimization process incurs substantial computational overhead. The generation of 4D assets requires extensive optimization time and consumes significant memory and GPU resources. This inefficiency is further compounded by the implicit nature of the SDS distillation process, which makes the generated 4D assets highly sensitive to both the control conditions and the pre-trained priors of the diffusion models [8], [9]. Even under identical control conditions, different diffusion model priors can lead

to markedly different 4D outputs. As a result, the quality and diversity of 4D generation are heavily constrained by the generative capacity and diversity of the underlying diffusion models.

Fidelity in 4D generation, particularly for image-to-4D, video-to-4D, and 3D-to-4D methods, fundamentally depends on the generalization capabilities of the diffusion models. These models must extrapolate spatial and viewpoint information from input images or videos to generate 4D assets faithful to the input. However, achieving such fidelity remains a significant challenge, as current diffusion-based pipelines often struggle to preserve fine-grained details and ensure accurate extrapolation in complex scenes. This highlights the need for improved model architectures and optimization strategies to enhance the fidelity and accuracy of 4D generation.

Overall, addressing these challenges requires the development of unified diffusion models that can seamlessly integrate spatial and temporal consistency, as well as more efficient optimization techniques to reduce computational costs. Additionally, improving the generalization capabilities of diffusion models will be critical for achieving high fidelity and diversity in 4D generation, paving the way for broader applications across diverse domains.

C. Benchmark

4D generative models have garnered significant attention due to their diverse control mechanisms and the variety of assets they can produce. However, this diversity also introduces substantial challenges in evaluating the generative capabilities of these models. Without a fair, objective, and human-aligned evaluation standard, it is difficult to determine which methods among the rapidly emerging 4D generative models are truly effective and which aspects require further improvement. Current approaches lack reliable automated metrics and primarily rely on generating outputs under identical control conditions, followed by user studies for comparison. While these methods provide some insights, they often suffer from significant subjectivity. To address this issue, methodologies from image and video generation benchmarks offer valuable inspiration.

We can evaluate the renderings of generated 4D assets to perform quantitative assessments of 4D generation models using image generation benchmarks. For instance, T2I-CompBench [154] introduces a comprehensive benchmark for open-world compositional text-to-image generation, categorizing prompts into three main categories (attribute binding, object relationships, and complex compositions) and six subcategories (color binding, shape binding, texture binding, spatial relationships, non-spatial relationships, and complex compositions). These categories can be adapted to evaluate text-to-4D generative models, enabling a systematic assessment of the generated 4D outputs. Similarly, DreamBench++ [155] employs advanced multimodal GPT models to automate human-aligned benchmarks, which can be leveraged to evaluate the semantic consistency between textual prompts and the generated 4D assets under text-controlled generation tasks.

In video generation benchmarks, the evaluation of 4D generative models can be indirectly performed by analyzing multi-view and multi-temporal rendered videos of 4D outputs. EvalCrafter [156] provides valuable references for text-to-video generation evaluation, focusing on four key aspects: visual quality, content quality, motion quality, and text-video alignment, using 17 carefully designed objective metrics. These evaluation criteria can be extended to text-to-4D generation. Similarly, VBench [157] refines "video generation quality" into specific, hierarchical, and disentangled dimensions, with each dimension accompanied by tailored prompts and evaluation methods. It evaluates videos across 16 dimensions, including subject identity inconsistency, motion smoothness, temporal flickering, and spatial relationships. By batch-evaluating rendered videos from multiple 4D generative models under fixed input text and rendering parameters, VBench offers a viable approach for comparative analysis of 4D generation quality. Furthermore, FETV [158] categorizes text prompts into three orthogonal dimensions: major content, controllable attributes, and prompt complexity. It incorporates temporal awareness by introducing several time-specific categories tailored for video generation, revealing the strengths and weaknesses of text-to-video models across different categories. These methodologies can be adapted to assess and improve 4D generative models by incorporating both temporal and categorical evaluation criteria.

Currently, dedicated benchmarks for 4D generative models remain scarce, with existing references limited to low-dimensional generative tasks, primarily controlled by textual inputs. However, as benchmarks for image and video generation continue to mature, it is foreseeable that tailored benchmarks for 4D generation will emerge and become a central topic of methodological discussions. By advancing towards benchmarks that are objective, flexible, and human-aligned, the research community will establish a solid foundation for the ongoing development and refinement of 4D generative models, enabling the systematic evaluation of their capabilities and contributions across diverse applications.

D. Social Impact

The societal impact of 4D generation is multifaceted, offering significant opportunities while also presenting notable challenges. On the positive side, advancements in 4D generation have the potential to drive innovation across diverse industries, including entertainment, education, healthcare, and virtual reality. For example, 4D content can enhance the immersive experience in gaming and filmmaking, improve the accuracy and realism of medical simulations, and transform e-learning platforms by delivering dynamic, interactive educational materials. Furthermore, 4D technology holds great promise for cultural heritage preservation, enabling the dynamic reconstruction and visualization of historical artifacts, thus fostering a deeper understanding and appreciation of cultural history.

Despite these benefits, 4D generation technology also raises several societal concerns. The ability to generate highly realistic and dynamic content introduces ethical challenges, such as the misuse of 4D technology for creating deepfakes, spreading misinformation, or producing inappropriate or harmful content. Moreover, the large-scale collection and generation of

4D data pose privacy risks, spark debates over data ownership, and exacerbate environmental concerns due to the significant computational resources required for content processing and rendering.

To ensure that 4D generation technology contributes positively to society, researchers and practitioners must establish clear ethical guidelines and implement robust protective measures throughout the technology's lifecycle. These measures include promoting transparency in 4D generation processes, ensuring accountability mechanisms to address misuse, and encouraging the responsible adoption of 4D technology. By fostering responsible innovation, the goal is to maximize the societal benefits of 4D generation while minimizing its potential risks and challenges.

IX. Conclusion

This survey provides a comprehensive review of the advancements in 4D generation, an emerging field with significant potential across diverse domains. We systematically introduce the representation methods, design, and training strategies of generative models, highlighting key challenges and summarizing corresponding solutions. Additionally, we explore open problems and potential research directions, offering insights into future opportunities for innovation in this area. By addressing the foundational aspects of 4D generation and its associated challenges, we aim to provide new researchers and practitioners with a detailed and accessible overview of the field. This survey serves not only as a convenient reference for understanding the current landscape of 4D generation but also as a catalyst for advancing its broad range of downstream applications, including entertainment, education, healthcare, autonomous driving, and cultural preservation. Ultimately, we hope this survey will inspire further research and foster innovation, driving the progress of 4D generation technologies and unlocking their transformative potential across scientific and practical domains.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," in *International Conference on Learning Representations*, 2022.

[3] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv:2010.02502*, October 2020. [Online]. Available: https://arxiv.org/abs/2010.02502

[4] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[6] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

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

[8] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," *arXiv preprint arXiv:2308.16512*, 2023.

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

[10] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," *arXiv preprint arXiv:2310.15110*, 2023.

[11] J. Lin, "OneTo3D: One Image to Re-editable Dynamic 3D Model and Video Generation," *CoRR*, vol. abs/2405.06547, 2024, arXiv: 2405.06547. [Online]. Available: https://doi.org/10.48550/arXiv.2405.06547

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

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

[14] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[15] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," *Entropy*, vol. 25, no. 10, p. 1469, 2023.

[16] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi, "Diffusion models for video prediction and infilling," *arXiv preprint arXiv:2206.07696*, 2022.

[17] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, "Flexible diffusion modeling of long videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 953–27 965, 2022.

[18] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich, "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 619–12 629.

[19] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron *et al.*, "Dreambooth3d: Subject-driven text-to-3d generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2349–2359.

[20] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson *et al.*, "Text-to-4d dynamic scene generation," *arXiv preprint arXiv:2301.11280*, 2023.

[21] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell, "4d-fy: Text-to-4d generation using hybrid score distillation sampling," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[22] Y. Zheng, X. Li, K. Nagano, S. Liu, O. Hilliges, and S. D. Mello, "A unified approach for text- and image-guided 4d scene generation," in *CVPR*, 2024.

[23] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao, "Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video," *arXiv preprint arXiv:2311.02848*, 2023.

[24] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis, "Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8576–8588.

[25] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu, "Dreamgaussian4d: Generative 4d gaussian splatting," *arXiv preprint arXiv:2312.17142*, 2023.

[26] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[27] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[28] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327.

[29] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874.

[30] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, "4d gaussian splatting for real-time dynamic scene rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20 310–20 320.

[31] Q. Miao, Y. Luo, and Y. Yang, "PLA4D: Pixel-Level Alignments for Text-to-4D Gaussian Splatting," *CoRR*, vol. abs/2405.19957, 2024, arXiv: 2405.19957. [Online]. Available: https://doi.org/10.48550/arXiv.2405.19957

[32] J. Ren, K. Xie, A. Mirzaei, H. Liang, X. Zeng, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim *et al.*, "L4gm: Large 4d gaussian reconstruction model," *arXiv preprint arXiv:2406.10324*, 2024.

[33] H. Yu, C. Wang, P. Zhuang, W. Menapace, A. Siarohin, J. Cao, L. A. Jeni, S. Tulyakov, and H.-Y. Lee, "4Real: Towards Photorealistic 4D Scene Generation via Video Diffusion Models," *CoRR*, vol. abs/2406.07472, 2024, arXiv: 2406.07472. [Online]. Available: https://doi.org/10.48550/arXiv.2406.07472

[34] R. Li, P. Pan, B. Yang, D. Xu, S. Zhou, X. Zhang, Z. Li, A. Kadambi, Z. Wang, and Z. Fan, "4K4DGen: Panoramic 4D Generation at 4K Resolution," *CoRR*, vol. abs/2406.13527, 2024, arXiv: 2406.13527. [Online]. Available: https://doi.org/10.48550/arXiv.2406.13527

[35] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges, "Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 858–12 868.

[36] F. Zhao, Y. Jiang, K. Yao, J. Zhang, L. Wang, H. Dai, Y. Zhong, Y. Zhang, M. Wu, L. Xu *et al.*, "Human performance modeling and rendering via neural animated mesh," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–17, 2022.

[37] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "Livecap: Real-time human performance capture from monocular video," *ACM Transactions On Graphics (TOG)*, vol. 38, no. 2, pp. 1–17, 2019.

[38] M. Habermann, W. Xu, M. Zollhofer, G. Pons-Moll, and C. Theobalt, "Deepcap: Monocular human performance capture using weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5052–5063.

[39] L. Mou, J.-K. Chen, and Y.-X. Wang, "Instruct 4d-to-4d: Editing 4d scenes as pseudo-3d scenes using 2d diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 176–20 185.

[40] R. Gao, K. Chen, Z. Li, L. Hong, Z. Li, and Q. Xu, "Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes," 2024. [Online]. Available: https://arxiv.org/abs/2405.14475

[41] J. Mao, B. Li, B. Ivanovic, Y. Chen, Y. Wang, Y. You, C. Xiao, D. Xu, M. Pavone, and Y. Wang, "Dreamdrive: Generative 4d scene modeling from street view images," 2025. [Online]. Available: https://arxiv.org/abs/2501.00601

[42] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," 2023. [Online]. Available: https://arxiv.org/abs/2309.09777

[43] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[44] A. A. Osman, T. Bolkart, and M. J. Black, "Star: Sparse trained articulated human body regressor," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 598–613.

[45] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "Ghum & ghuml: Generative 3d human shape and articulated pose models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6184–6193.

[46] J. N. Reddy, "An introduction to the finite element method," *New York*, vol. 27, no. 14, 1993.

[47] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P. Battaglia, "Learning mesh-based simulation with graph networks," in *International conference on learning representations*.

[48] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia, "Learning to simulate complex physics with graph networks," in *International conference on machine learning*. PMLR, 2020, pp. 8459–8468.

[49] I. Liu, H. Su, and X. Wang, "Dynamic gaussians mesh: Consistent mesh reconstruction from monocular videos," *arXiv preprint arXiv:2404.12379*, 2024.

[50] Z. Li, Y. Chen, and P. Liu, "Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[51] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe *et al.*, "Neural 3d video synthesis from multi-view video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5521–5531.

[52] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6498–6508.

[53] B. Attal, J.-B. Huang, C. Richardt, M. Zollhoefer, J. Kopf, M. O'Toole, and C. Kim, "Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 610–16 620.

[54] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger, "Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2732–2742, 2023.

[55] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141.

[56] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 479–12 488.

[57] R. Shao, Z. Zheng, H. Tu, B. Liu, H. Zhang, and Y. Liu, "Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 632–16 642.

[58] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[59] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton, "General dynamic scene reconstruction from multiple view video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 900–908.

[60] ——, "Temporally coherent 4d reconstruction of complex dynamic scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4660–4669.

[61] M. Wand, P. Jenke, Q. Huang, M. Bokeloh, L. Guibas, and A. Schilling, "Reconstruction of deforming geometry from time-varying point clouds," in *Symposium on Geometry processing*, 2007, pp. 49–58.

[62] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8387–8397.

[63] J. Dong, J. G. Burnham, B. Boots, G. Rains, and F. Dellaert, "4d crop monitoring: Spatio-temporal reconstruction for agriculture," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3878–3885.

[64] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5614–5623.

[65] Q. Zheng, X. Fan, M. Gong, A. Sharf, O. Deussen, and H. Huang, "4d reconstruction of blooming flowers," in *Computer Graphics Forum*, vol. 36, no. 6. Wiley Online Library, 2017, pp. 405–417.

[66] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," *Advances in neural information processing systems*, vol. 30, 2017.

[67] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 157–164.

[68] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele, "Building statistical shape spaces for 3d human modeling," *Pattern Recognition*, vol. 67, pp. 276–286, 2017.

[69] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.

[70] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao, "Recent advances in 3d gaussian splatting," *Computational Visual Media*, vol. 10, no. 4, pp. 613–642, 2024.

[71] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 331–20 341.

[72] X. Zhou, Z. Lin, X. Shan, Y. Wang, D. Sun, and M.-H. Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 634–21 643.

[73] Y.-H. Huang, Y.-T. Sun, Z. Yang, X. Lyu, Y.-P. Cao, and X. Qi, "Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4220–4230.

[74] Y. Lin, Z. Dai, S. Zhu, and Y. Yao, "Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2024, pp. 21 136–21 145.
[75] Z. Li, Z. Chen, Z. Li, and Y. Xu, "Spacetime gaussian feature splatting for real-time dynamic view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8508–8520.
[76] T. Xie, Z. Zong, Y. Qiu, X. Li, Y. Feng, Y. Yang, and C. Jiang, "Physgaussian: Physics-integrated 3d gaussians for generative dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4389–4398.
[77] A. Kratimenos, J. Lei, and K. Daniilidis, "Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting," *arXiv preprint arXiv:2312.00112*, 2023.
[78] M. You and J. Hou, "Decoupling dynamic monocular videos for dynamic view synthesis," *IEEE Transactions on Visualization and Computer Graphics*, 2024.
[79] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation," *arXiv preprint arXiv:2403.08321*, 2024.
[80] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, "Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis," *arXiv preprint arXiv:2308.09713*, 2023.
[81] Z. Guo, W. Zhou, L. Li, M. Wang, and H. Li, "Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction," *arXiv preprint arXiv:2403.11447*, 2024.
[82] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445
[83] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
[84] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.
[85] Y. Xu, Z. He, S. Shan, and X. Chen, "Ctrlora: An extensible and efficient framework for controllable image generation," *arXiv preprint arXiv:2410.09400*, 2024.
[86] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.
[87] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.
[88] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *Advances in Neural Information Processing Systems*, 2023.
[89] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu, "Humansd: A native skeleton-guided diffusion model for human image generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 988–15 998.
[90] C. Corneanu, R. Gadde, and A. M. Martinez, "Latentpaint: Image inpainting in latent space with diffusion models," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 4334–4343.
[91] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[92] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 142–13 153.
[93] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in neural information processing systems*, vol. 35, pp. 25 278–25 294, 2022.
[94] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.
[95] J. Yang, Z. Cheng, Y. Duan, P. Ji, and H. Li, "Consistnet: Enforcing 3d consistency for multi-view images diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7079–7088.
[96] Y. Liu, M. Xie, H. Liu, and T.-T. Wong, "Text-guided texturing by synchronized multi-view diffusion," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.
[97] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, "Video generation models as world simulators," 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators
[98] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 7346–7356.
[99] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.
[100] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
[101] R. Shao, Y. Pang, Z. Zheng, J. Sun, and Y. Liu, "Human4DiT: Free-view Human Video Generation with 4D Diffusion Transformer," *CoRR*, vol. abs/2405.17405, 2024, arXiv: 2405.17405. [Online]. Available: https://doi.org/10.48550/arXiv.2405.17405
[102] H. Zhang, X. Chen, Y. Wang, X. Liu, Y. Wang, and Y. Qiao, "4Diffusion: Multi-view Video Diffusion Model for 4D Generation," *CoRR*, vol. abs/2405.20674, 2024, arXiv: 2405.20674. [Online]. Available: https://doi.org/10.48550/arXiv.2405.20674
[103] Z. Dou, "Dynamic Realms: 4D Content Analysis, Recovery and Generation with Geometric, Topological and Physical Priors," *CoRR*, vol. abs/2409.14692, 2024, arXiv: 2409.14692. [Online]. Available: https://doi.org/10.48550/arXiv.2409.14692
[104] S. Bahmani, X. Liu, W. Yifan, I. Skorokhodov, V. Rong, Z. Liu, X. Liu, J. J. Park, S. Tulyakov, G. Wetzstein, A. Tagliasacchi, and D. B. Lindell, "TC4D: Trajectory-Conditioned Text-to-4D Generation," in *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., vol. 15104. Springer, 2024, pp. 53–72. [Online]. Available: https://doi.org/10.1007/978-3-031-72952-2_4
[105] Z. Chai, C. Tang, Y. Wong, and M. S. Kankanhalli, "STAR: Skeleton-aware Text-based 4D Avatar Generation with In-Network Motion Retargeting," *CoRR*, vol. abs/2406.04629, 2024, arXiv: 2406.04629. [Online]. Available: https://doi.org/10.48550/arXiv.2406.04629
[106] Z. Fu, J. Wei, W. Shen, C. Song, X. Yang, F. Liu, X. Yang, and G. Lin, "Sync4D: Video Guided Controllable Dynamics for Physics-Based 4D Generation," *CoRR*, vol. abs/2405.16849, 2024, arXiv: 2405.16849. [Online]. Available: https://doi.org/10.48550/arXiv.2405.16849
[107] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson, and Y. Taigman, "Text-To-4D Dynamic Scene Generation," *CoRR*, vol. abs/2301.11280, 2023, arXiv: 2301.11280. [Online]. Available: https://doi.org/10.48550/arXiv.2301.11280
[108] Z. Li, Y. Chen, and P. Liu, "DreamMesh4D: Video-to-4D Generation with Sparse-Controlled Gaussian-Mesh Hybrid Representation," *CoRR*, vol. abs/2410.06756, 2024, arXiv: 2410.06756. [Online]. Available: https://doi.org/10.48550/arXiv.2410.06756
[109] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "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. 22 563–22 575.
[110] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.
[111] D. Xu, H. Liang, N. P. Bhatt, H. Hu, H. Liang, K. N. Plataniotis, and Z. Wang, "Comp4d: Llm-guided compositional 4d scene generation," *arXiv preprint arXiv:2403.16993*, 2024.
[112] C. Chen, S. Huang, X. Chen, G. Chen, X. Han, K. Zhang, and M. Gong, "CT4D: Consistent Text-to-4D Generation with Animatable Meshes," *CoRR*, vol. abs/2408.08342, 2024, arXiv: 2408.08342. [Online]. Available: https://doi.org/10.48550/arXiv.2408.08342

[113] Y. Cao, L. Pan, K. Han, K.-Y. K. Wong, and Z. Liu, "AvatarGO: Zero-shot 4D Human-Object Interaction Generation and Animation," *CoRR*, vol. abs/2410.07164, 2024, arXiv: 2410.07164. [Online]. Available: https://doi.org/10.48550/arXiv.2410.07164

[114] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[115] Y. Zhao, Z. Yan, E. Xie, L. Hong, Z. Li, and G. H. Lee, "Animate124: Animating One Image to 4D Dynamic Scene," *CoRR*, vol. abs/2311.14603, 2023, arXiv: 2311.14603. [Online]. Available: https://doi.org/10.48550/arXiv.2311.14603

[116] Z. Yang, Z. Pan, C. Gu, and L. Zhang, "Diffusion\(^\mbox2\): Dynamic 3D Content Generation via Score Composition of Orthogonal Diffusion Models," *CoRR*, vol. abs/2404.02148, 2024, arXiv: 2404.02148. [Online]. Available: https://doi.org/10.48550/arXiv.2404.02148

[117] Q. Sun, Z. Guo, Z. Wan, J. N. Yan, S. Yin, W. Zhou, J. Liao, and H. Li, "EG4D: Explicit Generation of 4D Object without Score Distillation," *CoRR*, vol. abs/2405.18132, 2024, arXiv: 2405.18132. [Online]. Available: https://doi.org/10.48550/arXiv.2405.18132

[118] H. E. Pang, S. Liu, Z. Cai, L. Yang, T. Zhang, and Z. Liu, "Disco4D: Disentangled 4D Human Generation and Animation from a Single Image," *CoRR*, vol. abs/2409.17280, 2024, arXiv: 2409.17280. [Online]. Available: https://doi.org/10.48550/arXiv.2409.17280

[119] R. K. Shah, N. Dawood, and S. Castro, "Automatic generation of progress profiles for earthwork operations using 4D visualisation model," *J. Inf. Technol. Constr.*, vol. 13, pp. 491–506, 2008. [Online]. Available: https://www.itcon.org/paper/2008/29

[120] Z. Pan, Z. Yang, X. Zhu, and L. Zhang, "Efficient4d: Fast dynamic 3d object generation from a single-view video," 2024. [Online]. Available: https://arxiv.org/abs/2401.08742

[121] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," *arXiv preprint arXiv:2309.03453*, 2023.

[122] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang, and V. Jampani, "SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency," *CoRR*, vol. abs/2407.17470, 2024, arXiv: 2407.17470. [Online]. Available: https://doi.org/10.48550/arXiv.2407.17470

[123] N. Otberdout, C. Ferrari, M. Daoudi, S. Berretti, and A. D. Bimbo, "3D to 4D Facial Expressions Generation Guided by Landmarks," *CoRR*, vol. abs/2105.07463, 2021, arXiv: 2105.07463. [Online]. Available: https://arxiv.org/abs/2105.07463

[124] Z. Erkoç, F. Ma, Q. Shan, M. Nießner, and A. Dai, "Hyperdiffusion: Generating implicit neural fields with weight-space diffusion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 14 300–14 310.

[125] Y. Feng, Y. Shang, X. Feng, L. Lan, S. Zhe, T. Shao, H. Wu, K. Zhou, H. Su, C. Jiang, and Y. Yang, "Elastogen: 4d generative elastodynamics," 2024. [Online]. Available: https://arxiv.org/abs/2405.15056

[126] H. Liang, Y. Yin, D. Xu, H. Liang, Z. Wang, K. N. Plataniotis, Y. Zhao, and Y. Wei, "Diffusion4D: Fast Spatial-temporal Consistent 4D Generation via Video Diffusion Models," *CoRR*, vol. abs/2405.16645, 2024, arXiv: 2405.16645. [Online]. Available: https://doi.org/10.48550/arXiv.2405.16645

[127] Q. Yang, M. Feng, Z. Wu, S. Sun, W. Dong, Y. Wang, and A. Mian, "Beyond Skeletons: Integrative Latent Mapping for Coherent 4D Sequence Generation," *CoRR*, vol. abs/2403.13238, 2024, arXiv: 2403.13238. [Online]. Available: https://doi.org/10.48550/arXiv.2403.13238

[128] J. Lin, Z. Wang, Y. Hou, Y. Tang, and M. Jiang, "Phy124: Fast Physics-Driven 4D Content Generation from a Single Image," *CoRR*, vol. abs/2409.07179, 2024, arXiv: 2409.07179. [Online]. Available: https://doi.org/10.48550/arXiv.2409.07179

[129] G.-W. Yang, D.-Y. Chen, and T.-J. Mu, "Sketch-2-4D: Sketch driven dynamic 3D scene generation," *Graph. Model.*, vol. 136, p. 101231, 2024. [Online]. Available: https://doi.org/10.1016/j.gmod.2024.101231

[130] Z. Wu, C. Yu, Y. Jiang, C. Cao, F. Wang, and X. Bai, "SC4D: Sparse-Controlled Video-to-4D Generation and Motion Transfer," in *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XIII*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., vol. 15071. Springer, 2024, pp. 361–379. [Online]. Available: https://doi.org/10.1007/978-3-031-72624-8_21

[131] W. Song, X. Zhang, Y. Guo, S. Li, A. Hao, and H. Qin, "Automatic Generation of 3D Scene Animation Based on Dynamic Knowledge Graphs and Contextual Encoding," *Int. J. Comput. Vis.*, vol. 131, no. 11, pp. 2816–2844, 2023. [Online]. Available: https://doi.org/10.1007/s11263-023-01839-1

[132] Y. Zheng, X. Li, K. Nagano, S. Liu, K. Kreis, O. Hilliges, and S. D. Mello, "A Unified Approach for Text- and Image-guided 4D Scene Generation," *CoRR*, vol. abs/2311.16854, 2023, arXiv: 2311.16854. [Online]. Available: https://doi.org/10.48550/arXiv.2311.16854

[133] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. J. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell, "4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 7996–8006. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.00764

[134] B. Zeng, L. Yang, S. Li, J. Liu, Z. Zhang, J. Tian, K. Zhu, Y. Guo, F.-Y. Wang, M. Xu, S. Ermon, and W. Zhang, "Trans4D: Realistic Geometry-Aware Transition for Compositional Text-to-4D Synthesis," *CoRR*, vol. abs/2410.07155, 2024, arXiv: 2410.07155. [Online]. Available: https://doi.org/10.48550/arXiv.2410.07155

[135] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu, "Avatarclip: Zero-shot text-driven generation and animation of 3d avatars," *arXiv preprint arXiv:2205.08535*, 2022.

[136] S. Cai, D. Ceylan, M. Gadelha, C.-H. P. Huang, T. Y. Wang, and G. Wetzstein, "Generative Rendering: Controllable 4D-Guided Video Generation with 2D Diffusion Models," *CoRR*, vol. abs/2312.01409, 2023, arXiv: 2312.01409. [Online]. Available: https://doi.org/10.48550/arXiv.2312.01409

[137] Y. Yin, D. Xu, Z. Wang, Y. Zhao, and Y. Wei, "4DGen: Grounded 4D Content Generation with Spatial-temporal Consistency," *CoRR*, vol. abs/2312.17225, 2023, arXiv: 2312.17225. [Online]. Available: https://doi.org/10.48550/arXiv.2312.17225

[138] Y.-J. Yuan, L. Kobbelt, J. Liu, Y. Zhang, P. Wan, Y.-K. Lai, and L. Gao, "4Dynamic: Text-to-4D Generation with Hybrid Priors," *CoRR*, vol. abs/2407.12684, 2024, arXiv: 2407.12684. [Online]. Available: https://doi.org/10.48550/arXiv.2407.12684

[139] S. Meng, Y. Luo, and P. Liu, "Grounding creativity in physics: A brief survey of physical priors in aigc," *arXiv preprint arXiv:2502.07007*, 2025.

[140] P. Wang and Y. Shi, "Imagedream: Image-prompt multi-view diffusion for 3d generation," *arXiv preprint arXiv:2312.02201*, 2023.

[141] R. Shao, Y. Pang, Z. Zheng, J. Sun, and Y. Liu, "Human4dit: 360-degree human video generation with 4d diffusion transformer," *arXiv preprint arXiv:2405.17405*, 2024.

[142] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion," in *European Conference on Computer Vision*. Springer, 2024, pp. 439–457.

[143] D. Xu, H. Liang, N. P. Bhatt, H. Hu, H. Liang, K. N. Plataniotis, and Z. Wang, "Comp4D: LLM-Guided Compositional 4D Scene Generation," *CoRR*, vol. abs/2403.16993, 2024, arXiv: 2403.16993. [Online]. Available: https://doi.org/10.48550/arXiv.2403.16993

[144] R. Wu, R. Gao, B. Poole, A. Trevithick, C. Zheng, J. T. Barron, and A. Holynski, "Cat4d: Create anything in 4d with multi-view video diffusion models," *arXiv preprint arXiv:2411.18613*, 2024.

[145] Y. Chen, Y. Wang, Y. Luo, Z. Wang, Z. Chen, J. Zhu, C. Zhang, and G. Lin, "Meshanything v2: Artist-created mesh generation with adjacent mesh tokenization," *arXiv preprint arXiv:2408.02555*, 2024.

[146] M. Işık, M. Rünz, M. Georgopoulos, T. Khakhulin, J. Starck, L. Agapito, and M. Nießner, "Humanrf: High-fidelity neural radiance fields for humans in motion," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–12, 2023.

[147] Y. Yuan, X. Li, Y. Huang, S. De Mello, K. Nagano, J. Kautz, and U. Iqbal, "Gavatar: Animatable 3d gaussian avatars with implicit mesh learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 896–905.

[148] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 392–18 402.

[149] D. Jiang, Z. Ke, X. Zhou, and X. Shi, "4d-editor: Interactive object-level editing in dynamic neural radiance fields via 4d semantic segmentation," *arXiv preprint arXiv:2310.16858*, 2023.

[150] R. Shao, J. Sun, C. Peng, Z. Zheng, B. Zhou, H. Zhang, and Y. Liu, "Control4d: Efficient 4d portrait editing with text," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4556–4567.

[151] L. Wang, W. Zheng, D. Du, Y. Zhang, Y. Ren, H. Jiang, Z. Cui, H. Yu, J. Zhou, J. Lu, and S. Zhang, "Stag-1: Towards realistic 4d driving simulation with video generation model," 2024. [Online]. Available: https://arxiv.org/abs/2412.05280

[152] D. Ha and J. Schmidhuber, "World models," 2018. [Online]. Available: https://zenodo.org/record/1207631

[153] L. Wang, W. Zheng, Y. Ren, H. Jiang, Z. Cui, H. Yu, and J. Lu, "Occsora: 4d occupancy generation models as world simulators for autonomous driving," 2024. [Online]. Available: https://arxiv.org/abs/2405.20337

[154] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 723–78 747, 2023.

[155] Y. Peng, Y. Cui, H. Tang, Z. Qi, R. Dong, J. Bai, C. Han, Z. Ge, X. Zhang, and S.-T. Xia, "Dreambench++: A human-aligned benchmark for personalized image generation," *CoRR*, vol. abs/2406.16855, 2024.

[156] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan, "Evalcrafter: Benchmarking and evaluating large video generation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 139–22 149.

[157] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818.

[158] Y. Liu, L. Li, S. Ren, R. Gao, S. Li, S. Chen, X. Sun, and L. Hou, "Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 62 352–62 387, 2023.