4DGen: Grounded 4D Content Generation with Spatial-temporal Consistency

Yuyang Yin[1*], Dejia Xu[2*], Zhangyang Wang[2], Yao Zhao[1], Yunchao Wei[1†]
[1]Beijing Jiaotong University, [2]University of Texas at Austin
yuyangyin@bjtu.edu.cn, dejia@utexas.edu

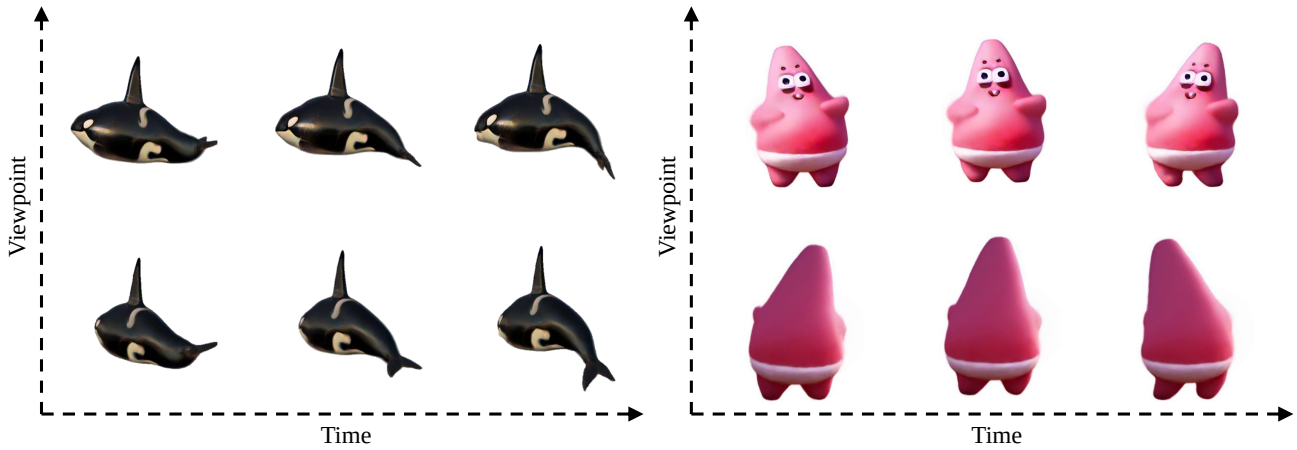

Figure 1. 4DGen introduces grounded 4D content generation. We present high-quality rendered images from diverse viewpoints at distinct timesteps. Our model exhibits rapid training and high-quality results with outstanding spatial-temporal consistency.

Abstract

Aided by text-to-image and text-to-video diffusion models, existing 4D content creation pipelines utilize score distillation sampling to optimize the entire dynamic 3D scene. However, as these pipelines generate 4D content from text or image inputs, they incur significant time and effort in prompt engineering through trial and error. This work introduces ***4DGen****, a novel, holistic framework for grounded 4D content creation that decomposes the 4D generation task into multiple stages. We identify static 3D assets and monocular video sequences as key components in constructing the 4D content. Our pipeline facilitates conditional 4D generation, enabling users to specify geometry (3D assets) and motion (monocular videos), thus offering superior control over content creation. Furthermore, we construct our 4D representation using dynamic 3D Gaussians, which permits efficient, high-resolution supervision through rendering during training, thereby facilitating high-quality 4D generation. Additionally, we employ spatial-temporal pseudo labels on anchor frames, along with seamless consistency priors implemented through 3D-aware score distillation sampling and smoothness regularizations. Compared to existing baselines, our approach yields competitive results in faithfully reconstructing input signals and realistically inferring renderings from novel viewpoints and timesteps. Most importantly, our method supports grounded generation, offering users enhanced control, a feature difficult to achieve with previous methods. Project page: https://vita-group.github.io/4DGen/*

1. Introduction

The rapid advancements of the recent text-to-image diffusion models [46, 52, 54, 55] have introduced a new generative AI era. Alongside generating images from text prompts, much attention has been put into generating more complex content, such as videos and (dynamic) 3D assets. Modern artists have previously been relying on special software tools to implement their ideas into reality. As a result, automatic content creation pipelines are in great need to effectively assist human labor and reduce time-consuming

*Equal Contribution.
†Corresponding Author.

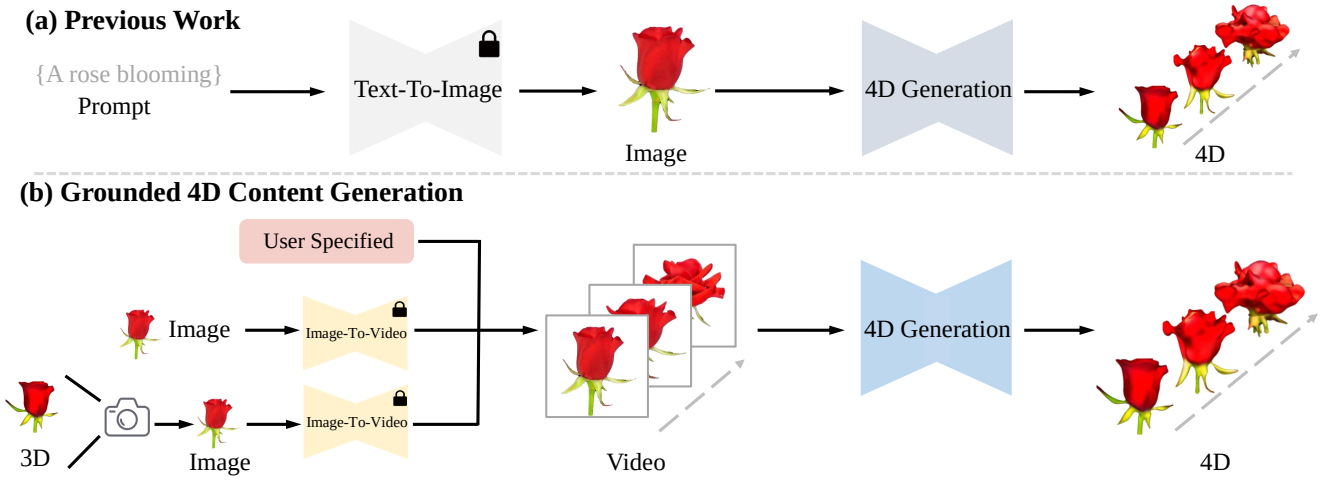

Figure 2. Previous work generates 4D content in one click. Our work introduces **Grounded 4D Content Generation**, which employs a video sequence and an optional 3D asset to specify the appearance and motion.

manual adjustments.

Aiming at these ambitious goals, much effort has been made into creating 3D objects and videos from images or text prompts [20, 30, 36, 37, 49, 61, 65, 73]. Researchers have shown that 3D and video datasets [5, 11, 71, 76, 82] provide rich domain-specific knowledge that is beneficial in building large generative models. Optimization-based 3D generation also attracts great attention, relying more on 2D priors and reducing the need for 3D annotations. DreamFusion [49] introduces score distillation sampling (SDS) by distilling knowledge from pre-trained 2D diffusion models without the need for any 3D annotation. 3D geometry and appearance are obtained by fusing multi-view supervision into a single 3D representation, implemented with Neural Radiance Field (NeRF) [42, 57, 66, 77], Mesh [34, 72], and 3D Gaussians [10, 63].

Despite the exciting progress on either 3D or video generation, little attention has been put to the intersection of these two directions, *i.e.* the generation of 4D (dynamic 3D) assets, mainly due to the lack of high-quality data. Existing approaches mainly focus on category-specific generation [19, 24, 26, 78, 86].The most related work for general object generation is MAV3D [62], where SDS is applied on text-to-image and text-to-video diffusion models to distill a 4D representation from scratch. However, their results suffer from multiple issues. Firstly, the results are produced from a vague single image or text input, and there is no guarantee that the time-consuming optimization process will lead to the desired 4D result. This introduces extra trial and error costs on prompt engineering in practical applications. Another issue arises from the 4D representation. Though their Hexplane [7] is more efficient than vanilla NeRF, rendering high-resolution images across long frames is still memory-costly and not affordable. As a result, score distillation is conducted at a lower resolution, introducing unpleasant artifacts. Though MAV3D [62] incorporates additional super-resolution models to improve the appearance, the generation is still blurry in texture and has flickering geometry.

To overcome the above issues, we introduce **4DGen**, a novel pipeline tackling a new task of **Grounded 4D Generation**. As shown in Fig. 2, we provide users with explicit fine-grained controllability over the generation of 4D content. Our key idea is to leverage a monocular video and an optional static 3D asset as conditional signals to specify the motion and appearance of the 4D content. By conditioning on the video sequence, explicit motion information is introduced, enabling the user to directly adjust the 4D content in terms of motion. An optional static 3D asset can be provided to adjust the geometry and appearance of the final result, further improving the user experience. When the 3D asset is not specified, our pipeline performs single-image-to-3D generation as a surrogate. When the monocular video sequence is not provided, we leverage existing video generation tools to produce a video sequence.

Our framework introduces novel designs to accomplish the grounded 4D generation task. We implement our 4D asset with deformable 3D Gaussians, an efficient 4D representation that enables high-resolution rendering at the training stage. We first construct static 3D Gaussians to support deformation into motion sequences effectively. Then, we utilize spatial-temporal pseudo labels defined in 2D and 3D space to inject motion information into the 4D representation through anchor frames. Further, we leverage seamless spatial-temporal consistency priors to refine the renderings from arbitrary viewpoints and arbitrary timesteps.

In summary, our contributions are,

- We present **4DGen**, a holistic pipeline for grounded 4D

content generation. Our framework allows full control over the 3D asset's appearance and motion by specifying a static 3D asset or a monocular video sequence.
- With the help of the efficient dynamic 3D Gaussian Splatting as scene representation, our model is trained at a high resolution and long frame length, leading to visually pleasing 4D generation, as shown in Fig. 1.
- Armed with spatial-temporal pseudo labels defined in 2D and 3D space, we explicitly inject motion and appearance information at anchor frames. Moreover, we adopt seamless spatial-temporal consistency priors for arbitrary viewpoint and timestep with the help of 3D-aware score distillation sampling and unsupervised smoothness regularization.
- Experiments demonstrate our superiority against per-frame generation baselines. Our 4DGen framework delivers faithful reconstruction of the input signals while synthesizing plausible results for novel viewpoints and timesteps.

2. Related Works

2.1. 3D Representations for Content Creation

Since 3D data are not naturally stored in grids as pixels in 2D, multiple 3D representations have been studied for content creation. Polygonal meshes represent shape surfaces using vertices, edges, and faces. Previous works [16, 47, 48, 58] have achieved high quality textured 3D meshes generation. On the other hand, point cloud utilizes an unstructured set of points in 3D space to represent surfaces. Many works have explored point cloud generation by autoencoding [2, 15, 80], adversarial generation [60, 64], and diffusion model [83, 87]. However, both meshes and point clouds are known to require significant amounts of memory and slow training. NeRF [43] addresses this challenge by implicitly employing Multilayer Perceptrons (MLPs) to represent objects and scenes. While easy to optimize, rendering NeRF into high-resolution images usually requires millions of queries of the MLP network, making it hard to supervise through patch-based rendering losses. Many follow-up works aim to overcome this issue by introducing hybrid representations that leverage explicit structure to reduce the burden of NeRF's MLP. Instant NGP [45] significantly accelerates neural rendering using a compact neural network combined with a multi-resolution hash table. EG3D [8] introduces an expressive hybrid explicit-implicit network architecture for the unsupervised generation of high-quality multi-view-consistent images and 3D shapes. Recently, 3D Gaussian Splatting (3D-GS) [29] introduces an alternative 3D representation to NeRF. 3D-GS demonstrates impressive visual quality and real-time rendering, supported by a rapid visibility-aware rendering algorithm. Our work follows this inspiring direction in representing the 3D scene with 3D Gaussians, and we adopt an additional implicit network to make the 3D Gaussians deformable for dynamic scenes.

2.2. 3D Content Generation with Diffusion Prior

Generating 3D content from multi-modal input has attracted great research interest for years. The task of text-to-3D generation focuses on generating a 3D model from a textual prompt. Due to the lack of large-scale 3D data, many works try to improve 3D generation quality by improving the rendering quality from random 2D viewpoints. Early works DreamFields [25] and CLIPMesh [44] utilized CLIP [51], but the results tend to lack realism. With the help of 2D diffusion priors, the seminal work DreamFusion [49] introduces Score Distillation Sampling (SDS) and showcases encouraging results. Many follow-up works have then tried to improve SDS sampling results through better optimization [4, 21, 23, 31, 41, 72], crafting different 3D representations [10, 34, 63, 81] or constructing better diffusion priors that are suitable for 3D generation [32, 37, 50, 57]. SDS can also be adapted to image-to-3D generation [12, 37, 42, 77]. More recently, with the help of large-scale 3D asset datasets [11, 76, 82], better generative models [22, 27, 36–38, 59, 79] are built and produce outstanding 3D asset generation ability when given a single image input. Our work draws inspiration from both directions and proposes to use pseudo labels explicitly sampled from a large generative model to supervise anchor viewpoints and employ score distillation sampling to enforce consistency for arbitrary viewpoints.

2.3. Dynamic Content Generation

Unlike static content generation, producing high-quality dynamic datasets remains challenging due to the lack of data [5, 71]. Many prior works seek to leverage existing priors learned on large-scale text-image pairs [6, 9, 56] and adapt to dynamic content generations. Tune-a-Video [75] tackles one-shot video generation through subtle architecture modifications and sub-network tuning. Text2Video Zero [30] introduces a training-free method for animating a pre-trained text-to-image model. AnimateDiff [18] converts existing text-to-image diffusion models to video generators by designing a motion modeling module that is zero-shot transferrable to unseen stable diffusion variants.

Regarding 4D content generation, early works mainly focus on category-specific generation, such as digital humans [19, 24, 33, 39, 53, 67, 78, 84], animals [26] and flowers [86]. Regarding general object generation, MAV3D [62] leverages score distillation sampling from both image and video diffusion models to optimize a 4D scene. Consistent4D [28] is a concurrent method that also studies lifting a monocular video sequence to 4D. Their work leverages frame interpolation-driven consistency loss and per-frame

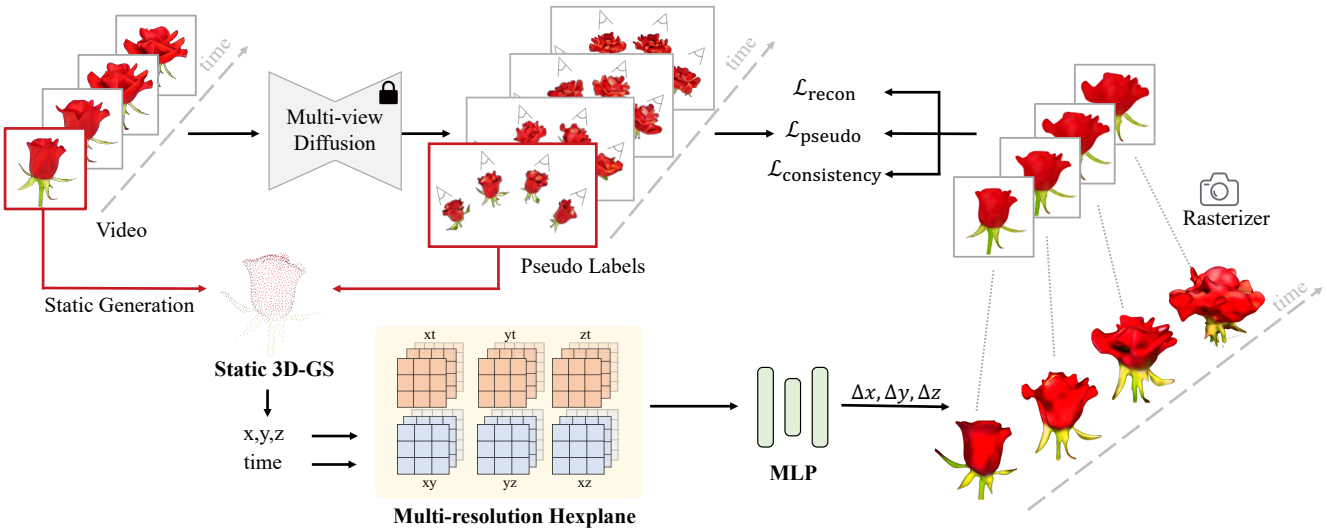

Figure 3. Overall framework of 4DGen. We conduct 4D generation grounded by a static 3D asset and a monocular video sequence. Our 4D scene is implemented by deforming a static set of 3D Gaussians. Besides reconstructing the input signals, we supervise our framework with spatial-temporal pseudo labels on anchor frames by a multi-view diffusion model. Moreover, we adopt seamless consistency priors implemented with score distillation sampling and unsupervised smoothness regularizations.

SDS from a 3D-aware diffusion model. Our work similarly studies utilizing 3D-aware diffusion priors for 4D generation, but differs in that we not only employ SDS supervision but also introduce DDIM-sampled pseudo labels for supervision at a higher resolution. Moreover, while their work adopts costly NeRF-based volumetric rendering, we utilize an efficient rasterizer brought by 3D Gaussians, supporting faster training implemented through rendering losses.

3. 4DGen: Grounded 4D Generation

Generating 4D contents without 4D annotation is highly ill-posed since multiple 4D results can all be plausible when projecting to a monocular video sequence or a static 3D asset. To this end, we adopt a pre-trained 3D-aware multi-view diffusion model to provide spatial-temporal pseudo labels. Additionally, since the spatial-temporal pseudo labels are only provided on limited 2D renderings, we include score distillation sampling from a 3D-aware diffusion model to supervise random viewpoints and unsupervised smoothness regularization to enforce temporal smoothness at random timestep in 3D explicitly.

In this section, we first introduce the background of score distillation sampling and 3D Gaussian splatting. Then, we construct our scene by generating a static scene via 3D Gaussians. Subsequently, we introduce our spatial-temporal pseudo labels on anchoring viewpoints. Finally, we propose seamless spatial and temporal consistency priors for arbitrary viewpoints and timesteps. An overview of our framework is provided in Fig. 3.

3.1. Background

Score Distillation Sampling Score Distillation Sampling (SDS) is widely adopted for distilling 2D image priors from a pre-trained diffusion model ϵ_ϕ into 3D representations through optimization on the renderings. When provided with a 3D model parameterized by θ and an image generator g, the multi-view image x can be rendered by $g(\theta)$.

SDS optimizes the 3D parameter θ as follows:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, x) = w(t)\left[(\hat{\epsilon_\phi}(x_t; y, t) - \epsilon)\right]\frac{\partial x}{\partial \theta}, \quad (1)$$

where $w(t)$ is a weighting function, x_t is obtained by perturbing x with randomly sampled Gaussian noise, and y is the conditional signal. In practice, $\hat{\epsilon_\phi}(z_t; y, t)$ is implemented by classifier-free guidance, which combines the scores of a conditional model $\epsilon_\phi(z_t; y, t)$ and an unconditional model $\epsilon_\phi(z_t; , t)$ to achieve better generation conditioned on y.

3D Gaussian Splatting 3D Gaussian Splatting(3D-GS) [29] is an explicit 3D scene representation utilizing a set of differentiable 3D Gaussians. Each Gaussian is defined with center position $\mu \in \mathbb{R}^3$, covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$, color $c \in \mathbb{R}^3$ and opacity $\alpha \in \mathbb{R}^1$. It can be formulated as:

$$G(x) = e^{-\frac{1}{2}(x)^T\Sigma^{-1}(x)}, \quad (2)$$

where x means the distance between the center position μ and the query point. Color can be calculated as:

$$C(r) = \sum_{i \in N} c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \tag{3}$$

where $\sigma_i = \alpha_i G(x_i)$ and N means the number of sample Gaussian points.

3.2. Static Gaussian Generation

The vanilla 3D Gaussians are only designed for reconstructing static 3D scenes, and many follow-up works [40, 74] extend the framework to dynamic scenes by constructing a static scene first and then learning the point deformations to support scene motion. We draw inspiration from 4D Gaussian Splatting [74] to use a set of static Gaussians that are shared across different timesteps and utilize an additional HexPlane [7] representation to express the deformations for point attributes.

When the user specifies a 3D asset, we can readily construct the static scene by converting the specified 3D asset into a point cloud and subsequently initializing 3D Gaussians from this point cloud. However, in the absence of a static 3D asset, our framework must generate the static Gaussian initialization to enable scene deformation. Owing to the ill-posed and data-limited nature of our generative task, additional design considerations are necessary to construct a plausible static scene.

To achieve this, we utilize a multi-view diffusion model, pre-trained on 3D datasets, to infer the scene structure. We perform DDIM sampling using SyncDreamer [38] to obtain multi-view predictions of the scene, and then optimize a set of static 3D Gaussians using the sampled images. Our framework facilitates both random initialization and initialization from pre-processed point clouds for the 3D Gaussians. Specifically, for the initial frame, off-the-shelf multi-view reconstruction algorithms [68, 70] are employed to estimate coarse geometry and extract point clouds.

It should be noted that our static 3D Gaussians are not kept fixed but jointly optimized with the deformation field once the static 3D Gaussians are properly constructed. During the joint optimization phases, the static 3D Gaussians are not compelled to reconstruct the initial frame of the 4D scene. Rather, the deformation field can modify the static 3D Gaussians at each timestep. This approach additionally addresses the issue of poor geometric quality arising from inconsistencies in multi-view predictions.

3.3. Spatial-temporal Anchor Frame Pseudo-labels

MAV3D [62] demonstrates outstanding performance by jointly distilling video and image diffusion priors. However, the lack of powerful open-source video diffusion models hinders the wide application of this direction. Possibly due to the lack of large-scale, high-quality dynamic video dataset, open-source video diffusion models usually leverage image-video joint training [30] or transfer learning [18, 75] to benefit from the knowledge learned during the foundational training of Stable diffusion [54]. As a result, the video generation quality suffers from limited realism and, therefore, is not ideal enough for distillation-based 4D content generation [62] directly. To address this challenge, we propose a novel lightweight solution by introducing anchor frames and explicitly specifying the 3D motion information at anchor frames for the 4D scene without injecting video diffusion priors.

Spatial-temporal 2D Pseudo Labels Owing to the scarcity of data, we endeavor to transfer knowledge from external priors. Given the monocular video sequence, we utilize SyncDreamer [38] to synthesize multi-view images at each timestamp. SyncDreamer is a diffusion model pre-trained on multi-view image datasets and aims to generate 3D consistent multi-view results. Consequently, we aim to distill this geometry information into the 4D content we generate.

However, while effective, SyncDreamer processes each frame individually, and for each frame, multiple 3D objects can be plausible due to the ill-posed nature of image-to-3D generation. Consequently, there is no consistency assurance in generation across frames, with this issue being particularly pronounced in synthesized back views. Furthermore, despite SyncDreamer's 3D-aware architecture aiming to ensure geometric consistency in novel view predictions, 3D consistency across viewpoints from DDIM-sampled images remains unguaranteed. Therefore, it is impractical to train a visually pleasing 4D representation using these supervisions directly. Instead, they serve as pseudo labels to facilitate the initial warming up of the 4D representation at anchor frames.

3D Deformation Pseudo Label Since these supervisions are defined on the 2D renderings, we rely on the differentiable rendering to propagate essential knowledge to the 3D points. However, it is observed that when Gaussians are presented in incorrect locations, they tend to reduce their visibility by changing their appearance instead of moving to the desired locations. Though ideally, this issue can be alleviated through scene flow supervision, estimating these dense 3D point correspondences from monocular video is highly challenging and requires a time-consuming optimization process [69].

To this end, we propose constructing point clouds from SyncDreamer predictions and treating them as pseudo labels for scene deformation. Note that point clouds at different frames are unrelated to 3D point correspondences because the points constructing the point clouds are stored as unordered sets and constructed individually. Therefore, we

use Chamfer distance for the 3D Gaussian locations, which measures the difference between two point sets without the need of point correspondences. It instead searches for the nearest neighbor as an alternate, defined as follows,

$$\begin{aligned}\mathcal{L}_{\text{chamfer}} = &\frac{w_1}{|P_1|}\sum_{p_{1i}\in P_1}\min_{p_{2j}\in P_2}\left(\|p_{1i}-p_{2j}\|_2^2\right)\\ &+\frac{w_2}{|P_2|}\sum_{p_{2j}\in P_2}\min_{p_{1i}\in P_1}\left(\|p_{2j}-p_{1i}\|_2^2\right),\end{aligned} \tag{4}$$

where P_1 is our 3D Gaussians 3D mean and P_2 refers to the pre-processed point clouds at each timeframe. This novel design effectively prevents the scene from mode-collapsing into static, allowing us to integrate the framework with score distillation sampling [49], which is widely observed to be prone to mode collapse. The loss functions for pseudo labels are summarized as follows,

$$\mathcal{L}_{\text{pseudo}} = \mathcal{L}_{\text{sync}} + \omega_1\mathcal{L}_{\text{chamfer}}, \tag{5}$$

where ω_1 is a weighting factor and $\mathcal{L}_{\text{sync}}$ refers to the LPIPS [85] distance between the rendered output and the pseudo label.

3.4. Seamless Spatial-temporal Consistency Priors

While efficient, dealing with anchor frames only will lead to flickering results if the model is inference with a different frame rate than training time. Inspired by MAV3D [62], we utilize a random frame sampling rate at training time to provide temporal augmentation of the model. By randomly selecting a frame rate and a random starting time step, our model can render a random short sequence of continuous time frames for refinement. Since we do not have access to ground truth signals for intermediate frames and arbitrary viewpoints, we adopt external priors and unsupervised regularization as alternates.

Arbitrary Viewpoint Spatial Consistency through SDS Motivated by DreamFusion [49], we adopt score distillation sampling (SDS) to overcome the multi-view inconsistency issue produced by the pseudo labels. Unlike regressing the inconsistent DDIM-sampled images, SDS optimizes the scene representation such that our rendered images maintain a high likelihood as evaluated by the pre-trained diffusion prior. With the help of recently released large-scale 3D-aware diffusion prior [11], we implement our image-conditioned SDS through conditioning on the front view of each timestep. SDS not only overcomes the aforementioned inconsistent issue introduced by the SyncDreamer predictions but also introduces dense supervision coverage on arbitrary viewpoints, supporting better visual appearance.

Arbitrary Timestep Temporal Consistency via Smoothness Regularization Thanks to the explicit nature of 3D Gaussians, we obtain the point locations for each timestep during rendering. We thus enforce unsupervised smoothness priors in intermediate frame renderings. We first implement spatial total variance that is widely used in prior works [7, 14],

$$\begin{aligned}\mathcal{L}_{\text{TV}}(\mathbf{P}) = \frac{1}{|C|n^2}\sum_{c,i,j}&\left(\left\|\mathbf{P}_c^{i,j}-\mathbf{P}_c^{i-1,j}\right\|_2^2\right.\\ &\left.+\left\|\mathbf{P}_c^{i,j}-\mathbf{P}_c^{i,j-1}\right\|_2^2\right),\end{aligned} \tag{6}$$

where i, j are indices on the current plane $\mathbf{P}$. We further utilize temporal smoothness prior that penalizes the acceleration of individual 3D points over time,

$$\mathcal{L}_{\text{smooth}}(\mathbf{P}) = \frac{1}{|C|n^2}\sum_{c,i,t}\left\|\mathbf{P}_c^{i,t-1}-2\mathbf{P}_c^{i,t}+\mathbf{P}_c^{i,t+1}\right\|_2^2. \tag{7}$$

To further encourage temporal smoothness, we initialize the space-time planes used in our scene representation as a constant one so that the space-time features are not necessarily changed if the corresponding spatial content remains static across different timesteps. The consistency priors can be formulated as follows,

$$\mathcal{L}_{\text{consistency}} = \mathcal{L}_{\text{smooth}} + \omega_2\mathcal{L}_{\text{TV}} + \omega_3\mathcal{L}_{\text{SDS}}, \tag{8}$$

leading to the overall loss function,

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \omega_4\mathcal{L}_{\text{pseudo}} + \omega_5\mathcal{L}_{\text{consistency}}, \tag{9}$$

where ω_i are weighting factors and $\mathcal{L}_{\text{recon}}$ refers to LPIPS [85] loss enforced to reconstruct the original input.

4. Experiments

4.1. Implementation Details

Architecture We implement our 4D representation using 4D Gaussian Spaltting [74]. We adopt a multi-resolution HexPlane voxel module to encode the deformation of 3D Gaussians. The six planes represent the combinations of spatial-temporal dimensions: $(x, y), (x, z), (y, z), (x, t), (y, t), (z, t)$, with the last three planes preserving the temporal information. Each queried feature is interpolated based on the nearby voxel grid features at each resolution. The initial voxel resolution is set to $[64, 64]$. Four levels of upsampled resolutions are employed, each doubling the initial resolution. An additional MLP head is utilized to decode interpolated features into the final predictions of position deformation $\Delta\mathbf{X}$.

Training Our training process comprises static, coarse, and fine stages. We start by constructing the static 3D Gaussians to support further deformation. If the user specifies a static 3D asset, we can initialize the static Gaussians with

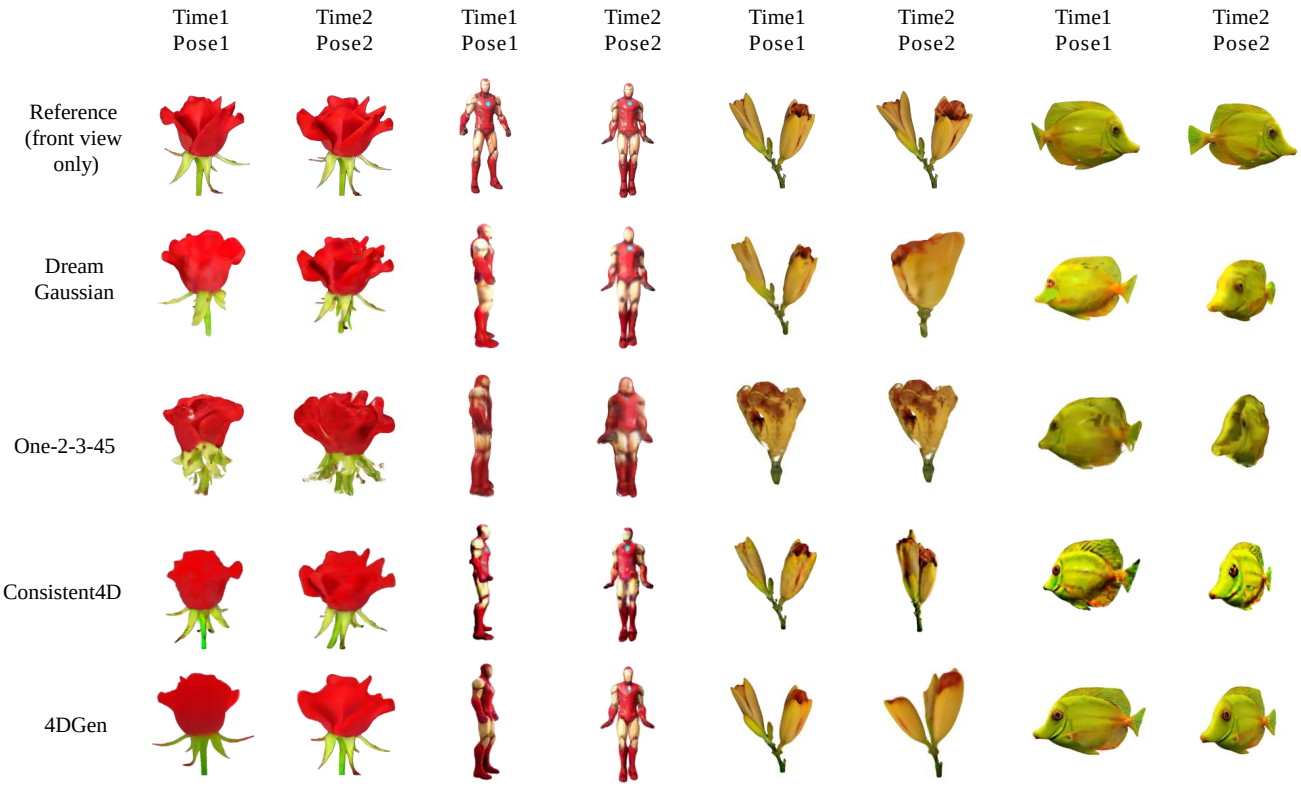

Figure 4. Comparison with baseline methods. We provide **video comparisons** in supplementary. The reference only contains front-view images, while our 4DGen model is capable of rendering at any timestep from any viewpoint. Our method maintains greater consistency and higher quality in both spatial and temporal consistency. In contrast, other methods result in significant artifacts or blurring, attributed to inaccurate feature estimation from different views. Moreover, these baselines generate individual frames separately and cannot be rendered at intermediate timesteps.

the point cloud converted from the 3D asset. Otherwise, we randomly initialize points in a unit sphere and spend 1,000 iterations to optimize the 3D Gaussians. We exclusively densify and prune 3D Gaussians in static stage. In the coarse and fine stage, we jointly optimize the static 3D Gaussians and the deformation field. Our coarse stage involves 1,000 iterations of optimization using $\mathcal{L}_{\text{pseudo}}$. As for the fine stage, we aim to enhance the visual quality while maintaining consistency in both spatial and temporal domains. To address the limitation of optimizing against inconsistent spatial-temporal pseudo labels, which tends to produce blurred results rather than accurately capturing the position and color of 3D Gaussian points, we thus introduce our seamless consistency priors to enhance overall consistency. We randomly use 10% of iterations to sample arbitrary frame rate and starting timestep to enforce temporal consistency for arbitrary timestep sequences. For the rest 90% iterations we randomly sample a timestep among the anchor frames and supervise the rendering of a random viewpoint through SDS implemented with Zero123 [37]. The fine stage contains 3,000 iterations.

All experiments are conducted using a single RTX 3090 GPU. The training time is highly effected by the number of 3D Gaussians. The time is about 45 minutes with around 25,000 3D Gaussians and 120 minutes with about 90,000 3D Gaussians. Densification introduces more 3D Gaussians, increasing the training time and also bringing better details.

4.2. Baseline Methods

MAV3D [62] is the sole work preceding 4D generation for 4D content creation. A direct comparison is not feasible since their code has not been released. To evaluate our method, we instead establish two baseline approaches following the practice in MAV3D. DreamGaussian [63] is a state-of-the-art image-to-3D method utilizing 3D Gaussian splatting. DreamGaussian is provided with individual video frames to generate a total of T 3D models, where T is the total number of frames. Subsequently, we can render images from every view at each time. For a fair comparison, we only use their first stage, which optimizes 3D Gaussians directly. The second baseline involves One-2-3-45 [36], a

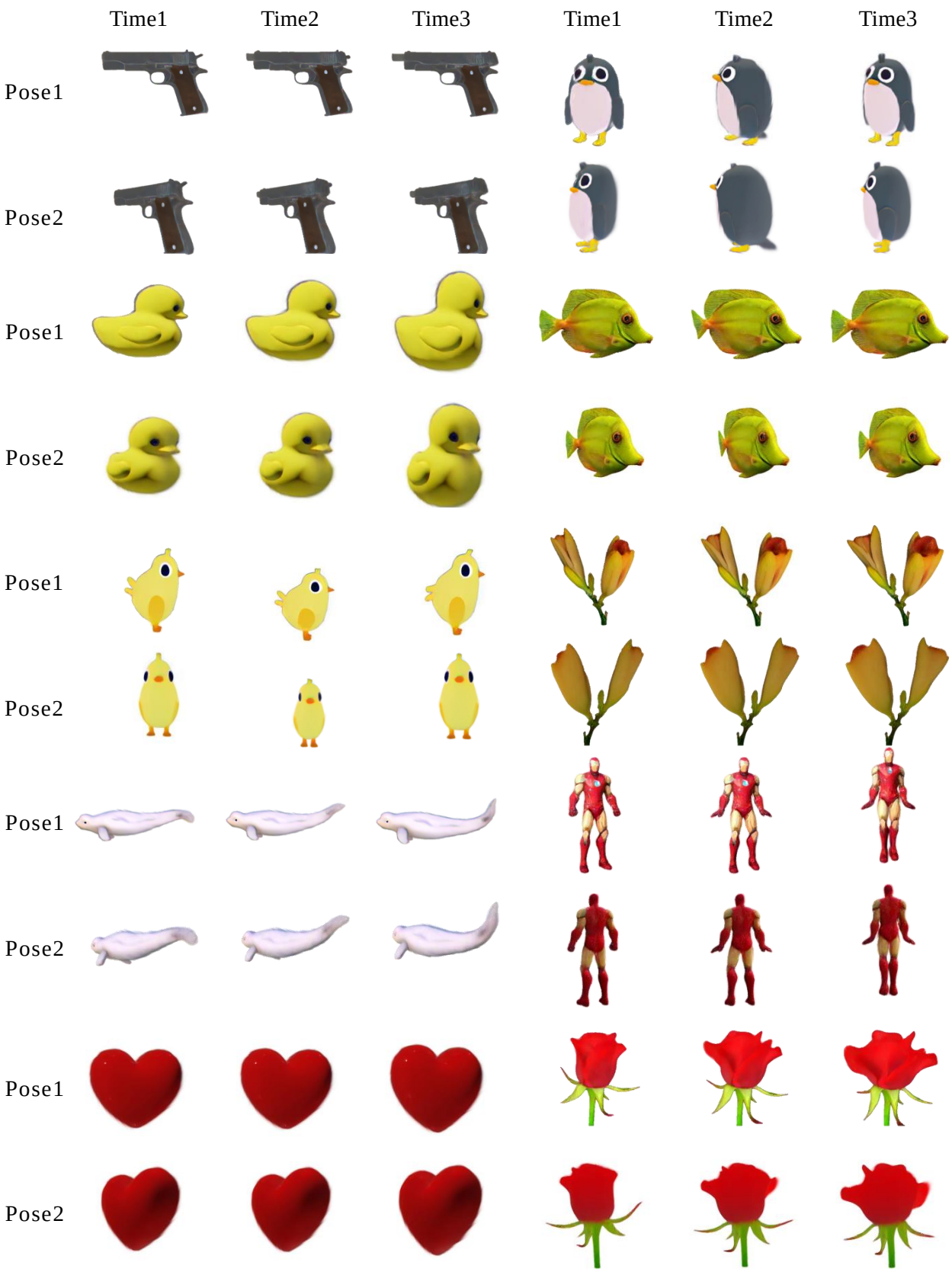

Figure 5. More cases about results of 4DGen. Our method achieves high-quality detail from arbitrary viewpoints and timesteps, meanwhile maintains seamless spatial-temporal consistency.

w/o $\mathcal{L}_{\text{sds}}$ | w/o $\mathcal{L}_{pseudo}$ | w/o $\mathcal{L}_{recon}$ (less points) | Full model

Figure 6. Ablation studies of our proposed components. Models lacking either $\mathcal{L}_{\text{SDS}}$ or $\mathcal{L}_{\text{pseudo}}$ fail to generate high-quality details.

high-quality multi-view diffusion model. Given an image at each time, One-2-3-45 can generate multi-view images and fuse them into a mesh. For these two baselines, we concatenate the T results to construct a 4D model. We also provide comparisons against a concurrent method Consistent4D [28]. The datasets consist of three parts, with each test case containing only a reference video sequence from the front view. The first part consists of a video dataset released by Consistent4D [28]. The second part of the dataset is collected either in the wild or from Sketchfab [1] by ourselves. The final part is generated using a pre-trained image-to-video diffusion model [3].

4.3. Evaluation Metric

Currently, there is no well-established evaluation metric for assessing the quality of 4D generations. Our assessment focuses on the quality of 4D content generation from the perspectives of space and time. In other words, images rendered from any viewpoint for any timestep should all enjoy high quality, while image sequences should demonstrate realistic consistency across frames and viewpoints.

We use the widely adopted CLIP distance as in previous 3D content creation works [63, 77] to measure the view synthesis quality. For temporal smoothness, we employ CLIP-T as in prior works [13, 17], which is the average CLIP distance between adjacent frames. Since novel view synthesis gets more challenging as novel-view camera gets more distant, we choose to also measure CLIP-T distance at different viewpoints, not only for the frontal view but also for the back and side views, denoted as CLIP-T-f, CLIP-T-b, and CLIP-T-s, respectively.

4.4. Qualitative and Quantitative Comparisons

We visualize our comparisons against baselines in Fig. 4, DreamGaussian [63] struggles to generate fine-grained details, possibly due to the limited number of Gaussian points. One-2-3-45 [35] adopts multi-view reconstruction from Zero123's noisy predictions. Possibly due to the limited generalization ability of Zero123, their results usually contain inaccurate geometry predictions and white texture artifacts. As each frame is generated individually, adapting these frameworks' results to 4D requires concatenating images from corresponding views at each time to produce a video. This results in noticeable temporal discontinuity, characterized by frequent flickering and texture inconsistencies over time.Consistent4D exhibits color distortion and artifacts, likely stemming from the limited capabilities of the NeRF model in low resolution and the post-processing stage of the super-resolution network.

Input Images | SVD Outputs | Ours View1 | Ours View2

Figure 7. Image-to-4D Results. Our pipeline supports the Image-to-4D task by a video sequence generated by an image-to-video generative model [3].

Since the code of Consistent4D [28]is only partially released, we are unable to conduct quantitative comparisons with Consistent4D under the same settings. Our 4DGen significantly outperforms baselines in spatial and temporal consistency, evident in enhanced details and smooth transitions between frames. Furthermore, previous methods only support the generation of a fixed number of frames, which is equal to the number of input images. Our method, on the other hand, benefits from the seamless consistency priors, and supports sampling videos of arbitrary length from arbitrary viewpoints. As shown in Table. 1, Our 4DGen highly surpasses the compared methods in term of CLIP-T-f, CLIP-T-b, and CLIP-T-s and achieves competitive performances regarding CLIP metric. The better spatio-temporal continuity reflected in the CLIP-T metric aligns with observations made in the videos.

4.5. Ablation Study

We conduct ablation studies to validate the effectiveness of our novel components. As shown in Tab. 2 and Fig. 6, when training without $\mathcal{L}_{\text{SDS}}$, $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{sync}}$, $\mathcal{L}_{\text{TV}}$ or $\mathcal{L}_{\text{chamfer}}$, the output video suffer from distorted geometric shapes or unsatisfactory flickering artifacts. In comparison, our full model delivers the best performance.

Table 1. Quantitative comparisons on the four datasets.

Dataset	Metrics	DreamGaussian	One-2-3-45	Ours
Rose	CLIP↓	**0.2161**	0.3177	0.2231
	CLIP-T-f↓	0.0310	0.0770	**0.0124**
	CLIP-T-b↓	0.0463	0.0866	**0.0082**
	CLIP-T-s↓	0.0754	0.0807	**0.0169**
Fish	CLIP↓	0.3227	0.3501	**0.2491**
	CLIP-T-f↓	0.0347	0.0553	**0.0095**
	CLIP-T-b↓	0.0441	0.0516	**0.0135**
	CLIP-T-s↓	0.0788	0.0455	**0.0142**
Tulip	CLIP↓	0.2987	0.4050	**0.2362**
	CLIP-T-f↓	0.0291	0.0936	**0.0088**
	CLIP-T-b↓	0.0529	0.1062	**0.0118**
	CLIP-T-s↓	0.0694	0.1389	**0.0105**
Ironman	CLIP↓	**0.2637**	0.3234	0.2853
	CLIP-T-f↓	0.0264	0.0416	**0.0132**
	CLIP-T-b↓	0.0468	0.0593	**0.0107**
	CLIP-T-s↓	0.0566	0.0583	**0.0235**

Table 2. Ablation study for our proposed components.

Method	CLIP↓	CLIP-T-f↓	CLIP-T-b↓	CLIP-T-s↓
w/o $\mathcal{L}_{\text{SDS}}$	0.3004	0.0157	0.0191	0.0724
w/o $\mathcal{L}_{\text{sync}}$	0.3096	0.0137	0.0312	0.0729
w/o $\mathcal{L}_{\text{recon}}$	0.2550	0.0099	0.0163	0.0143
w/o $\mathcal{L}_{\text{TV}}$	0.2540	0.0078	**0.0113**	0.0144
w/o $\mathcal{L}_{\text{chamfer}}$	0.2493	0.0096	0.0142	0.0153
Full model	**0.2491**	**0.0095**	0.0135	**0.0142**

4.6. Limitations

Despite our exciting 4D generation results, 4DGen still has some limitations. Multi-object generation is beyond our assumption since our diffusion prior is pre-trained on object-centric 3D datasets. We will extend our framework to compositional and scene-level generation in the future.

5. Conclusion

In this work, we construct a novel 4DGen pipeline for grounded 4D content creation. We empower users to fully control the appearance and motion by specifying a static 3D asset or a monocular video sequence. Armed with our spatial-temporal pseudo labels and seamless consistency regularization, we effectively construct deformable 3D Gaussians that can be rendered from any viewpoint at any timestep. Our framework presents a faithful reconstruction of the input signals and visually pleasing novel view synthesis results at arbitrary timesteps.

References

[1] Sketchfab. https://sketchfab.com/. 9

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 3

[3] Stability AI. Stable video diffusion: Scaling latent video diffusion models to large datasets. https://stability.ai/research/stable-video-diffusion-scaling-latent-video-diffusion-models-to-large-datasets, 2023. 9

[4] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 3

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 2, 3

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 3

[7] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2, 5, 6

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[10] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*, 2023. 2, 3

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3, 6

[12] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *arXiv preprint arXiv:2212.03267*, 2022. 3

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 9

[14] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12479–12488, 2023. 6

[15] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 3

[16] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 3

[17] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 9

[18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3, 5

[19] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 2, 3

[20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[21] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023. 3

[22] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[23] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3

[24] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 2, 3

[25] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3

[26] Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3d: Learning articulated 3d animals by distilling 2d diffusion. *arXiv preprint arXiv:2304.10535*, 2023. 2, 3

[27] Yifan Jiang, Hao Tang, Jen-Hao Rick Chang, Liangchen Song, Zhangyang Wang, and Liangliang Cao. Efficient-3dim: Learning a generalizable single-image novel-view synthesizer in one day. *arXiv preprint arXiv:2310.03015*, 2023. 3

[28] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360° dynamic object generation from monocular video. *arxiv*, 2023. 3, 9

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 3, 4

[30] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2, 3, 5

[31] Subin Kim, Kyungmin Lee, June Suk Choi, Jongheon Jeong, Kihyuk Sohn, and Jinwoo Shin. Collaborative score distillation for consistent visual synthesis. *arXiv preprint arXiv:2307.04787*, 2023. 3

[32] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 3

[33] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 3

[34] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2, 3

[35] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 9

[36] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 2, 3, 7

[37] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 2, 3, 7

[38] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3, 5

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[40] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 5

[41] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems

with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023. 3

[42] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 2, 3

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[44] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 3

[45] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[46] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[47] Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, and Aurelien Lucchi. Convolutional generation of textured 3d meshes. *Advances in Neural Information Processing Systems*, 33:870–882, 2020. 3

[48] Dario Pavllo, Jonas Kohler, Thomas Hofmann, and Aurelien Lucchi. Learning generative models of textured 3d meshes from real-world images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13879–13889, 2021. 3

[49] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3, 6

[50] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[53] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13756–13766, 2023. 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 5

[55] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. 1

[56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3

[57] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 2, 3

[58] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 3

[59] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[60] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3859–3868, 2019. 3

[61] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[62] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 2, 3, 5, 6, 7

[63] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2, 3, 7, 9

[64] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Learning localized generative models for 3d point clouds via graph convolution. In *International conference on learning representations*, 2018. 3

[65] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[66] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg 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*, pages 12619–12629, 2023. 2

[67] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20796–20806, 2023. 3

[68] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 5

[69] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 5

[70] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 5

[71] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 2, 3

[72] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3

[73] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[74] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 5, 6

[75] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3, 5

[76] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2, 3

[77] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 {\deg} views. *arXiv preprint arXiv:2211.16431*, 2022. 2, 3, 9

[78] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37 (2):1–15, 2018. 2, 3

[79] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 3

[80] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 3

[81] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 3

[82] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9150–9161, 2023. 2, 3

[83] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 3

[84] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions On Graphics (TOG)*, 42(4):1–14, 2023. 3

[85] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[86] Yunzhi Zhang, Shangzhe Wu, Noah Snavely, and Jiajun Wu. Seeing a rose in five thousand ways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 962–971, 2023. 2, 3

[87] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 3