DreamPhysics: Learning Physics-Based 3D Dynamics with Video Diffusion Priors

Tianyu Huang[1,2], Haoze Zhang[1], Yihan Zeng[3], Zhilu Zhang[1], Hui Li[1], Wangmeng Zuo[1,*], Rynson W. H. Lau[2,*]

[1] Harbin Institute of Technology
[2] City University of Hong Kong
[3] Huawei Noah's Ark Lab

Abstract

Dynamic 3D interaction has been attracting a lot of attention recently. However, creating such 4D content remains challenging. One solution is to animate 3D scenes with physics-based simulation, which requires manually assigning precise physical properties to the object or the simulated results would become unnatural. Another solution is to learn the deformation of 3D objects with the distillation of video generative models, which, however, tends to produce 3D videos with small and discontinuous motions due to the inappropriate extraction and application of physics priors. In this work, to combine the strengths and complementing shortcomings of the above two solutions, we propose to learn the physical properties of a material field with video diffusion priors, and then utilize a physics-based Material-Point-Method (MPM) simulator to generate 4D content with realistic motions. In particular, we propose motion distillation sampling to emphasize video motion information during distillation. In addition, to facilitate the optimization, we further propose a KAN-based material field with frame boosting. Experimental results demonstrate that our method enjoys more realistic motions than state-of-the-arts do.

Code — https://github.com/tyhuang0428/DreamPhysics

Introduction

With the development in 3D representations, *e.g.*, Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (GS) (Kerbl et al. 2023), significant progress has been made in creating 3D assets through reconstruction and generation (Poole et al. 2022; Wang et al. 2024). However, interacting with these 3D assets in a simulation environment (Savva et al. 2019; Xia et al. 2018) remains challenging, despite its importance in many applications, *e.g.*, video games (Fan et al. 2022), virtual reality (Jiang et al. 2024), and robotics (Lu et al. 2024).

Animating static 3D objects based on instructions is an important step toward this interaction goal. In the real world, object movement is intertwined with the object's internal properties (*e.g.*, material types). Hence, we can see that on the one hand, some works (Xie et al. 2023; Feng et al. 2024b) first inject physical parameters into 3D GS objects, and then perform motion predictions in a physics-based simulator. However, as all these parameters have to be manually assigned, it is difficult to set them accurately, thus producing unnatural simulation results, as demonstrated in Figure 1(a). On the other hand, pre-trained video generators (Singer et al. 2022; Khachatryan et al. 2023; Wang et al. 2023b) are trained on real-world video data, which has naturally incorporated physical phenomena and regulations. These generators should contain, to some extent, physics-based prior knowledge. Thus, some works (Singer et al. 2023; Bahmani et al. 2023; Zhao et al. 2023) directly learn time-dependent deformation with the distillation of video models. However, the generated motions tend to exhibit small and discontinuous motions across frames. We hypothesize that the main reason for this drawback is the inappropriate extraction and application of the physics prior, rather than the utilization of video models. We therefore ask this question: how can we mine and apply the physics knowledge of video generative models to achieve realistic dynamic 3D synthesis?

To this end, we rethink the usage of physics-based simulation and video generative models in this work. We propose to learn a material field, rather than a deformation field, from video diffusion models, and then deploy a physics-based simulator to animate the 3D object in this field. As such, the advantages of the above two related approaches are combined, while their shortcomings can be complemented. Learnable physical properties from video diffusion models eliminate the need for manual modulation, and the physics simulator based on reasonable properties ensures more realistic motion generation.

Specifically, we introduce a new framework named DreamPhysics. DreamPhysics takes 3D GS (Kerbl et al. 2023) as a 3D representation. It first learns the physical properties of a material field with the distillation of video diffusion priors, and then adopts a simulator based on Material Point Method (MPM) (Stomakhin et al. 2013; Jiang et al. 2016) to model the time-dependent deformation of each Gaussian kernel. During the distillation from video diffusion models, the Score Distillation Sampling (SDS) (Poole et al. 2022) may focus more on color information, and is not completely suitable for extracting motion information. Instead, we propose motion distillation sampling (MDS) to avoid the

*Joint corresponding authors. Email: wmzuo@hit.edu.cn; rynson.lau@cityu.edu.hk

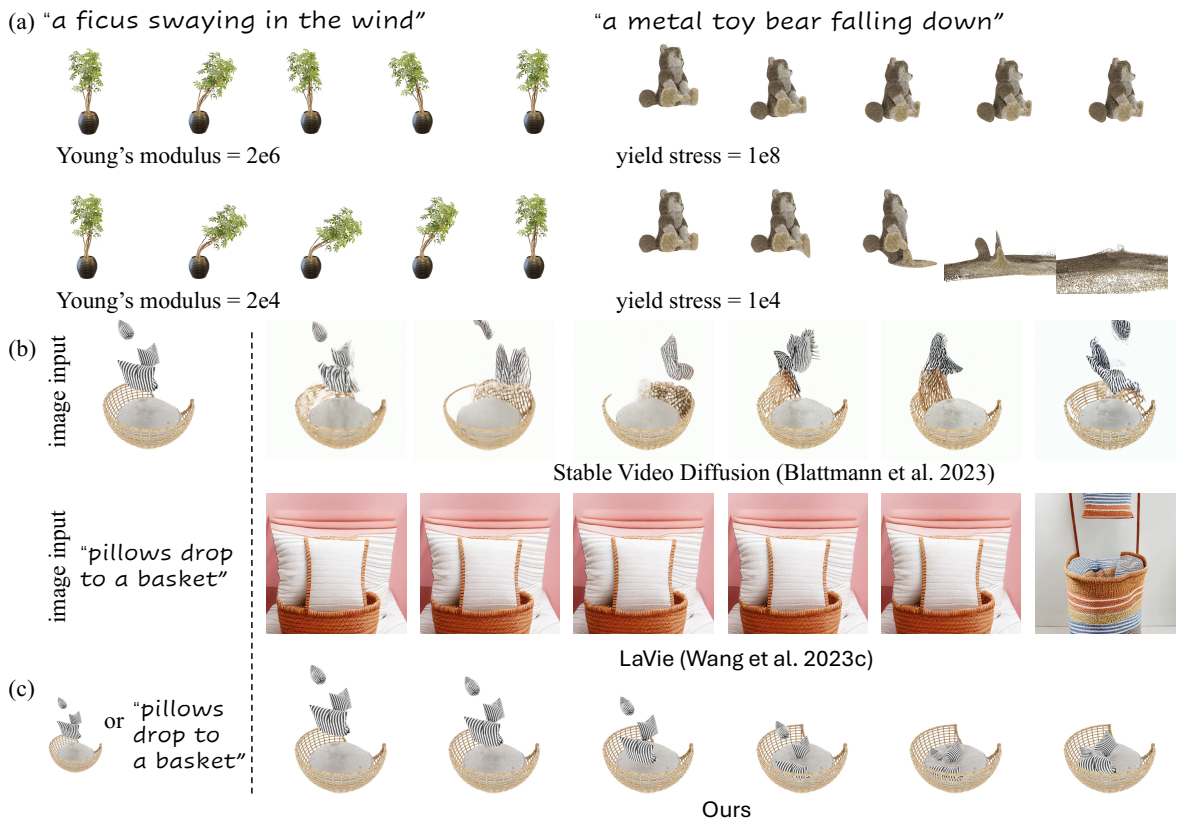

Figure 1: (a): The setting of physical properties can significantly affect the quality of the simulated videos. (b) Using state-of-the-art video diffusion models (Blattmann et al. 2023; Wang et al. 2023c) can hardly generate the desired results. (c) Our DreamPhysics can produce realistic 3D dynamic content with the distillation of video diffusion priors.

interference of color bias and emphasize the motion information in the rendered video. In addition, directly optimizing the material field can easily lead to unstable training due to the large range of possible parameter values. To facilitate the training process, we propose a KAN-based (Liu et al. 2024) material field with frame boosting.

We note that there is a concurrent work named PhysDreamer (Zhang et al. 2024), which supervises the prediction of physical properties with a ground-truth video generated by an image-to-video diffusion model. However, as shown in Figure 1(b), the video generative model can hardly produce the desired results to serve as ground truth, due to its poor motion control over the image/text condition. In contrast, our DreamPhysics supports both image-conditioned and text-conditioned optimization without the need for pre-generated ground truth, as demonstrated in Figure 1(c). Experimental results demonstrate that our method can effectively distill the video diffusion prior and assign proper values to the physical properties. Compared with state-of-the-art works, our results enjoy more realistic motion.

Our main contributions can be summarized as:

- We introduce a physics-based 3D animation framework, *i.e.*, DreamPhysics, which learns a material field for a physics simulator to support the creation of dynamic 3D content.
- We propose motion distillation sampling for the optimization of physical properties with video diffusion priors. To facilitate the optimization, we further propose a KAN-based material field with frame boosting.
- DreamPhysics can generate high-quality 4D content with either image- or text-conditioned optimization. Extensive experiments show that our results enjoy more realistic motion simulation.

Related Work

3D Generation

In recent years, 3D generation has advanced significantly, with methods broadly classified into two main categories: 3D supervised and 2D lifting approaches.

3D supervised methods (Nichol et al. 2022; Jun and Nichol 2023; Yu et al. 2023; Huang et al. 2023b; Hong et al. 2023; Tang et al. 2024) utilize text-3D data to train generators capable of directly producing 3D assets. For instance, Point-E (Nichol et al. 2022) is an early example of a text-to-3D generator that creates point clouds based on input prompts. Shap-E (Jun and Nichol 2023) and LGM (Tang et al. 2024) have expanded the scope of generated content to

include SDF (Park et al. 2019) and 3DGS (Kerbl et al. 2023) representations, respectively. Despite their efficiency in generating solid 3D content, these methods are significantly limited by the availability of 3D data. The current scale of 3D training datasets (Reizenstein et al. 2021; Deitke et al. 2023, 2024) is much smaller compared to 2D or video datasets, resulting in a constrained open-world capability relative to image or video generators. TextField3D (Huang et al. 2023b) attempts to enhance text control in 3D generators using a noisy latent space, yet it still falls short of achieving the imaginative capabilities seen in 2D generators.

Conversely, 2D lifting methods (Poole et al. 2022; Lin et al. 2023; Metzer et al. 2023; Chen et al. 2023; Wang et al. 2024) leverage the extensive prior knowledge embedded in 2D diffusion models to optimize 3D representations. DreamFusion (Poole et al. 2022) pioneered the concept of score distillation sampling (SDS), which distills 3D renderings into 2D diffusion. Although these methods produce photo-realistic results, they are prone to 3D inconsistency issues, commonly referred to as the Janus problem.

To address this issue, recent works (Liu et al. 2023; Shi et al. 2023; Long et al. 2023) have explored the synthesis of multi-view images of 3D objects. Zero-1-to-3 (Liu et al. 2023) generates images of the same object from different viewpoints based on a given image and viewpoint angles. MVDream (Shi et al. 2023) enhances consistency by generating orthogonal multi-view images of the same object. Wonder3D (Long et al. 2023) supports depth generation to achieve a more precise reconstruction. In this work, we collect static 3D scenes from both reconstruction data and 3D generation methods, providing more available assets.

3D Animation

3D animation creation has significantly increased demand across various applications, such as video games, virtual reality, and robotic simulation. However, manually creating such 4D content is a time-consuming process that necessitates a high level of expertise. To animate a 3D object, the common practice is to bind the object with a template skeleton, also known as rigging. TADA (Liao et al. 2023) produces 3D assets based on SMPL-X (Pavlakos et al. 2019), which is a human-body 3D template that supports animation. DreamControl (Huang et al. 2024a) proposes to generate 3D assets conditioned by input skeletons, which can be rigged easily for animation.

As the success of video generative models (Wang et al. 2023c,b; Blattmann et al. 2023; Zhang et al. 2023), some methods (Zhao et al. 2023) attempt to leverage video diffusion models to guide the prediction of the 3D deformation. DreamGuassian4D (Ren et al. 2023) uses a pre-generated video to supervise the deformation of static scenes. Animate124 (Zhao et al. 2023) proposes to distill the priors of video diffusion models to its deformation fields.

The deformation prediction in these methods is not accurate. Recent works (Xie et al. 2023; Feng et al. 2024a; Zhang et al. 2024) introduce physics simulation to the 3D deformation. PhysGaussian (Xie et al. 2023) deploys the finite element method to model the deformation of elastic objects like collision and shaking. Feng et al. further supports the simulation of liquid. However, these methods require manually setting the physical properties for objects before simulation. PhysDreamer (Zhang et al. 2024) attempts to optimize these properties with pre-generated videos, but the quality of generated videos can hardly be ensured. In this work, we propose to distill the priors of video models to simulation environments, enabling automatic setting of physical properties.

Preliminaries

Point-Based Representation

Point cloud (Guo et al. 2020) is an explicit 3D representation, which generally consists of the coordinates for all points. Normal and color information (Dai et al. 2017; Qi et al. 2017) can also be considered to further enrich the feature space of the point cloud. Despite the succinct representation, its rendering quality is heavily restricted by the number of points (Huang et al. 2023a). Derived from NeRF (Mildenhall et al. 2021), 3D Gaussian Splatting (GS) (Kerbl et al. 2023) introduces a point-based explicit radiance field. Points are modeled as a set of Gaussian kernels $\{\mathcal{G}_i\} = \{x_i, \sigma_i, \Sigma_i, C_i\}$, where x_i, σ_i, Σ_i, and C_i denote the center coordinate, opacity, covariance matrix, and spherical harmonic coefficient of the i-th kernel $\mathcal{G}_i$. To render a 3D GS scene at a specific viewpoint $\mathbf{r}$, color can be formulated as:

$$\mathbf{C} = \sum_{i=1}^{N} T_i \alpha_i C_i, \text{ with } \mathrm{T_i} = \prod_{\mathrm{j}=1}^{\mathrm{i}-1} (1 - \alpha_\mathrm{j}), \tag{1}$$

where N is the set of sorted Gaussian kernels related to the pixel and the viewpoint. α_i is the effective opacity given by evaluating a 2D Gaussian with Σ and σ. 3D GS can reconstruct high-fidelity views by real-time rendering, and support explicit interaction and editing.

Material Point Method

The material point method (MPM) (Stomakhin et al. 2013; Jiang et al. 2016) is a numerical simulation mechanic for the analysis of continuum forces. In MPM, the continuum is represented by a set of particles placed in a grid-based space. Different from mesh-based numerical mechanics, MPM can be naturally applied to point-based representation 3D GS. Following PhysGaussian (Xie et al. 2023), we have a time-dependent state for each Gaussian kernel as:

$$x_i(t) = \Delta(x_i, t), \ \Sigma_i(t) = F_i(t) \Sigma_i F_i(t)^T, \tag{2}$$

where $\Delta(\cdot, t)$ and $F_i(t)$ are the coordinate deformation and the deformation gradient at timestep t. Considering the continuum rotation $\Omega_i(t)$, the rendering viewpoint also requires adjustment to satisfy the view direction of spherical harmonic coefficient C_i.

Score Distillation Sampling

The score distillation sampling (SDS) (Poole et al. 2022; Wang et al. 2023a) distills pre-trained 2D diffusion models to the parameters of the 3D representation, widely used in 3D generation methods. Recently, SDS has had various extensions. Variational score Distillation (VSD) (Wang et al.

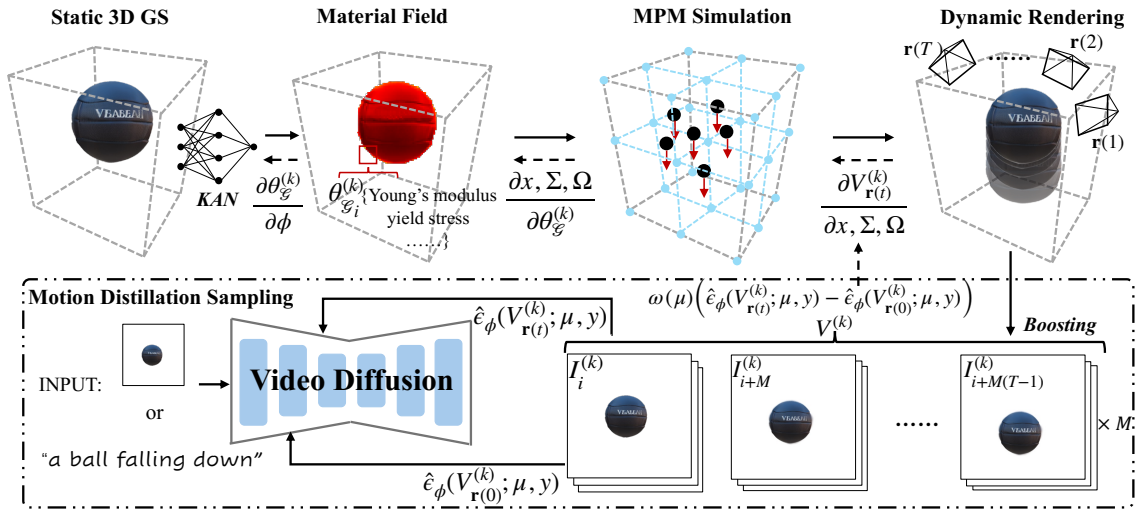

Figure 2: Overview of DreamPhysics. First, a set of physical parameters is initialized with a KAN-based material field for a static 3D GS. Then, it is fed to an MPM simulator to render a 3D video. Finally, we leverage motion distillation sampling to optimize the rendered video, and the distillation gradients are back-propagated to refine the physical parameters.

2024) proposes an additional LoRA term ϵ_θ to learn the distribution of current 3D scenes, which is attached to the score as:

$$\nabla_\theta \mathcal{L}_{\text{VSD}}(\theta) \triangleq \mathbb{E}\left[\omega(t)\left(\hat{\epsilon}_{\text{2D}}(\boldsymbol{x}_t, t, y) - \hat{\epsilon}_\theta(\boldsymbol{x}_t, t, c, y)\right)\frac{\partial \boldsymbol{x}}{\partial \theta}\right], \tag{3}$$

where t is the noise timestep and y is the input condition. $\hat{\epsilon}_{\text{2D}}$ and $\hat{\epsilon}_\theta$ are noises predicted by a pre-trained 2D diffusion model and the LoRA. Another extension, SDS-T, is for dynamic 3D generation, where video diffusion models are deployed to supervise the time-dependent deformation of static 3D objects. Specifically, given a camera trajectory $\mathbf{r}(t)$, SDS-T optimizes the rendered 3D video $V_{\mathbf{r}(t)}$ with predicted noise $\hat{\epsilon}_{\text{V}}$, as:

$$\nabla_\theta \mathcal{L}_{\text{SDS-T}}(\theta) \triangleq \mathbb{E}\left[\omega(\mu)\left(\hat{\epsilon}_{\text{V}}(V_{\mathbf{r}(t)}; \mu, y) - \epsilon\right)\frac{\partial V_{\mathbf{r}(t)}}{\partial \theta}\right], \tag{4}$$

where μ is noise timestep and θ is target deformation.

DreamPhysics

Method Overview

As shown in Figure 2, given a generated object or a reconstructed scene $\{\mathcal{G}_i\}$ represented by 3D GS (Kerbl et al. 2023), DreamPhysics aims to estimate the corresponding physical parameters $\{\theta_{\mathcal{G}_i}\}$ for the MPM-based simulator. For each Gaussian kernel $\mathcal{G}_i$, we initialize its parameters $\theta^{(0)}_{\mathcal{G}_i} = \boldsymbol{\phi}(x_i)$ with a KAN-based (Liu et al. 2024) material field $\boldsymbol{\phi}$ and then simulate a time-dependent state $\{x_i(t), \Sigma_i(t), \Omega_i(t)\}$, which can be rendered as a L-length video $V^{(0)} = \{I^{(0)}_1, I^{(0)}_2, ..., I^{(0)}_L\}$. The rendered video may look unrealistic due to the inaccurate initialization of $\theta^{(0)}_{\mathcal{G}}$. Therefore, we propose motion distillation sampling (MDS), which distills video diffusion's motion priors while weakening its color bias. The distillation gradient is then propagated backward to the material field, updating corresponding parameters to $\theta^{(1)}_{\mathcal{G}}$. Similarly, for each training iteration k, we can obtain an optimized $\theta^{(k+1)}_{\mathcal{G}}$ via the distillation of $V^{(k)}$. Considering current video diffusion models' low frame rate, we further propose a frame-boosting strategy to supervise more simulation frames. After several rounds of optimization, the final physical parameters $\hat{\theta}_{\mathcal{G}}$ can converge to a reasonable range.

Parameter Optimization with MDS

Video generative models are trained with real-world captured videos that cover kinds of physical phenomena. As a result, given a simulated video V, we can assess whether it is natural and realistic based on the judgement of video models. To this end, one direct solution is to treat videos generated by video models as ground truth, supervising V with reconstruction loss (Zhang et al. 2024). However, limited by the control capability, existing video generators can hardly produce desired ground-truth videos. We consider exploring distillation methods to optimize simulated results. Motion distillation sample is thus proposed to enhance the distillation of video diffusion's motion priors.

Motion Distillation Sample. With the simulation of MPM, a time-dependent state $\{x_i(t), \Sigma_i(t), \Omega_i(t)\}$ is predicted according to Eq. (2), representing a motion in the 3D space. Our intention is to optimize this simulated motion. However, the information of a video can be divided into two terms, *i.e.*, color and motion, where color biases between video diffusion models and the simulated video should be dismissed. In VSD (Wang et al. 2024), a LoRA term pushes the distribution of the target object away from the gradient direction of the current state. Similarly, we can adopt an additional term to omit the information in the color space. We suppose

that the first frame can represent the color for a whole video, so our motion distillation sample s_{MDS} is formulated as,

$$\boldsymbol{s}_{\text{MDS}} = \omega(\mu)\left(\hat{\epsilon}_{\text{V}}(V_{\mathbf{r}(t)}; \mu, y) - \hat{\epsilon}_{\text{V}}(V_{\mathbf{r}(0)}; \mu, y)\right), \quad (5)$$

where $\mathbf{r}(0)$ is the camera viewpoint in the first frame.

Note that the gradient of s_{MDS} cannot be directly propagated to the target physical parameters $\theta_{\mathcal{G}}$, and it needs to go through the differentiable MPM. Thus, our training objective can be written as:

$$\nabla_{\theta_{\mathcal{G}}} \mathcal{L}_{\text{MDS}}(\theta_{\mathcal{G}}, \mathbf{r}(t)) \triangleq \mathbb{E}\left[\boldsymbol{s}_{\text{MDS}} \frac{\partial V_{\mathbf{r}(t)}}{\partial x, \Sigma, \Omega} \frac{\partial x, \Sigma, \Omega}{\partial \theta_{\mathcal{G}}}\right]. \quad (6)$$

Parameter Estimation with Material Field

The value range for physical properties $\theta_{\mathcal{G}}$ can be very large, *e.g.*, the reasonable values for Young's modulus can vary from $1e4$ to $1e8$. However, during gradient updates, the same gradient can result in varying update granularity across different magnitudes, causing parameters to get stuck within a specific magnitude range. To enable parameters to converge more quickly to a reasonable range, we propose to perform a KAN-based tri-plane representation to model the material field and conduct frame boosting to further facilitate the training process.

KAN-Based Triplane. Tri-plane is widely used to encode spatial information. Given a 3D coordinate x, the tri-plane extractor projects it onto three orthogonal planes, *i.e.*, the front view, side view, and top view. These projections match x with 2D features that represent different perspectives of the 3D space. We extract features with KAN (Liu et al. 2024), which integrates kernel methods and attention mechanisms to offer superior modeling capabilities for physics-based tasks compared to traditional MLPs. Extracted features are then combined to form a unified representation, which constitutes our physical parameters $\theta_{\mathcal{G}}$. The gradient is propagated as:

$$\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\phi}}(x, \mathbf{r}(t)) \triangleq \mathbb{E}\left[\mathcal{L}_{\text{MDS}}(\boldsymbol{\phi}(x), \mathbf{r}(t)) \frac{\partial \theta_{\mathcal{G}}}{\partial \boldsymbol{\phi}}\right]. \quad (7)$$

Frame Boosting. The MPM simulator is a sequential model, which can easily lead to gradient vanishing or exploding like RNN (Rumelhart, Hinton, and Williams 1986). We have to conduct truncated back-propagation through time (BPTT), preserving the gradient of key frame simulation only. Truncated BPTT can effectively prevent gradient issues, but the supervision could be limited to specific frames. To ensure that our supervision covers as many video frames as possible, we further suggest a frame-boosting strategy. Specifically, given a total number of frames $M \times T$, we can separate them into M groups of frames with equal intervals, *i.e.*, $V_{t_i} = \{I_i, I_{i+M}, ..., I_{i+M(T-1)}\}$ for the i-th group. These groups formulate different videos, which are fed into the supervision process alternately. Finally, the boosted motion distillation can be formulated as:

$$\mathcal{L}_{\hat{\boldsymbol{\phi}}}(x) = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_{\boldsymbol{\phi}}(x, \mathbf{r}(t_i)), \quad (8)$$

where $\hat{\boldsymbol{\phi}}$ is the boosted material field.

Experiments

In this section, we show our 4D generation content on both text-conditioned and image-conditioned optimizations and compare it with previous state-of-the-art methods. Extensive ablation studies are then conducted to demonstrate the effectiveness of our newly proposed components.

Experimental Setup

Implementation Details. The simulation is based on the warp (Macklin 2022) implementation of MPM (Stomakhin et al. 2013; Jiang et al. 2016). For most simulation scenes, we set the simulation duration as 5×10^{-5} second and the frame duration as 4×10^{-2} second. Thus, we simulate 800 steps between every two renderings and include the simulation gradient of the last step in the optimization. We leverage a text-to-video diffusion model ModelScope (Wang et al. 2023b) and an image-to-video diffusion model Stable Video Diffusion (SVD) (Blattmann et al. 2023) to conduct text-conditioned and image-conditioned optimization, respectively. The numbers of their generated video frames T are 16 and 25, respectively. For frame boosting, we set $M = 5$, boosting the video slices to 5 groups. The setting of MDS follows SDS, where CFG value is set to 100. We stop the training if optimized parameter values stabilize within one order of magnitude. The training process requires around 30 iterations. The iteration time highly depends on the number of input Gaussian kernels, and it is within 30 seconds for most cases.

Dataset. We collect seven 3D static scenes or objects from previous works (Xie et al. 2023; Zhang et al. 2024) and 3D GS generative models (Tang et al. 2024). The content includes three plants, a beanie hat, a telephone cord, a sofa with pillows, and a ball, where two motions (rotation and collision) are involved in the simulator.

Evaluation Metric. We use the aesthetic quality from VBench (Huang et al. 2024b), grading the artistic score from 0 to 10 using the LAION aesthetic predictor (LAION-AI 2022). This metric can reflect aesthetic aspects such as the naturalness of the video, which exactly meets our evaluation requirements. In addition, we will add user study results in the supplementary materials.

Compared Methods. Since physics-based 4D generation is still under development, we compare three existing methods PhysGaussian (Xie et al. 2023), PhysDreamer (Zhang et al. 2024), and DreamGaussian4D (Ren et al. 2023). PhysGaussian is a pioneer work that manually sets all the physical properties in a physics-based simulator. PhysDreamer is a concurrent work that supervises physical parameters with ground-truth videos. DreamGaussian4D predicts the deformation of 3D GS without physical constraint, which is different from the above two works.

3D Dynamics Generation

Text Condition. In Figure 3(a), we select the ficus scene in PhysGaussian (Xie et al. 2023) and input a text prompt "*ficus swaying in the wind*" to simulate the rotation motion. The ficus would excessively tilt to one side and have difficulty returning to its original position if its Young's modulus

Figure 3: (a) Text-conditioned optimization; (b) Image-conditioned optimization. Right images are the space-time (X-t) slices, one axis represents time and the other axis shows a space slice (red line) of the object.

DreamGaussian4D	PhysGaussian	PhysDreamer	Ours	GT
4.61	4.98	4.84	**5.03**	5.13

Table 1: Quantitative results for the comparison with previous works on 4 scenes from Figure 4. The higher aesthetic quality score indicates better generation quality.

is set too low. After the optimization by our DreamPhysics, Young's modulus falls within a normal range, and the swaying looks more natural. From the space-time slices, the optimized motion trajectory looks more realistic.

Image Condition. For image-conditioned optimization, the first frame is regarded as the input image. We select a generated ball and try to optimize its dropping process, which is an example of collision motion, as shown in Figure 3(b). When hitting the ground, the ball would exhibit excessive deformation if the physical properties are not initialized accurately. Our method can effectively adjust these properties to a reasonable range after the optimization.

Comparison with State-of-the-art Works. We report the quantitative results of all the compared methods in Table 1. Since PhysDreamer hasn't released its training implementation, we can only compare four evaluation scenes, where the corresponding ground-truth videos are provided in the video demo. Considering that other methods don't have extra text inputs, we use the first frame as the image condition to conduct the optimization. According to the evaluation of aesthetic quality, our results are the closest to the ground truth. PhysDreamer has a lower score compared with PhysGaussian, which indicates that pre-generated videos may not be a proper ground truth for supervision. The generation quality of DreamGaussian4D is the worst because its deformation prediction didn't consider physical constraints.

We also provide the visualization of space-time slices in Figure 4. Since all the physical properties in PhysGaussian are manually set, its generated motions often look too extreme. DreamGaussian4D generates the most consistent motions but appears less natural, as its prediction lacks physical constraint. PhysDreamer can exhibit energy dissipation to some extent, while our results look more similar to the ground-truth visualization, in terms of amplitude and frequency of the simulated motions.

Method	+KAN	+$\mathcal{L}_{\text{MDS}}$	+Boost	***Score***↑	***Iter***↓
(a)				4.86	36.86
(b)	✓			4.89	34.29
(c)	✓	✓		**4.94**	33.86
Ours	✓	✓	✓	4.93	**29.71**

Table 2: Quantitative results of ablation study on 7 scenes. ***Score*** denotes the average aesthetic quality score, and ***Iter*** denotes the average training iterations.

Ablation Study

To evaluate the effectiveness of our newly proposed modules, we conduct ablation studies on all 7 scenes. Our baseline uses a vanilla SDS-T loss (Eq. (4)), where gradients are propagated to the physical parameters without KAN. Based on this, we attach our KAN-based material field, motion distillation sampling, and frame boosting step by step.

We report the aesthetic quality score and training iterations in Table 2. In (a), the physical parameters can hardly converge to a reasonable range, with the evaluation score and required iterations being the worst. Equipped with a KAN-based material field, (b) can facilitate the optimization and improve the generation quality. Then, we use motion distillation sampling $\mathcal{L}_{\text{MDS}}$ in (c), where the aesthetic score is further improved. In (d), our final method enjoys a faster optimization speed within 30 training iterations, demonstrating that our frame boosting can fasten the parameter conver-

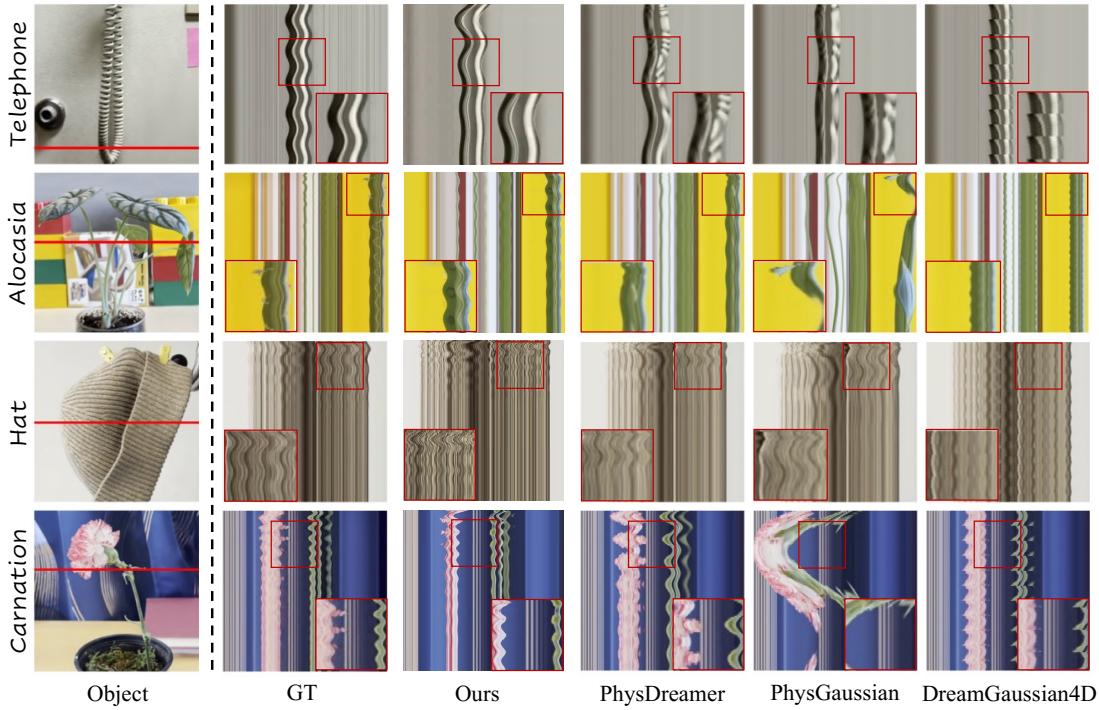

Figure 4: Viualization of space-time slices. Compared with previous works, our results are more close to the ground truth.

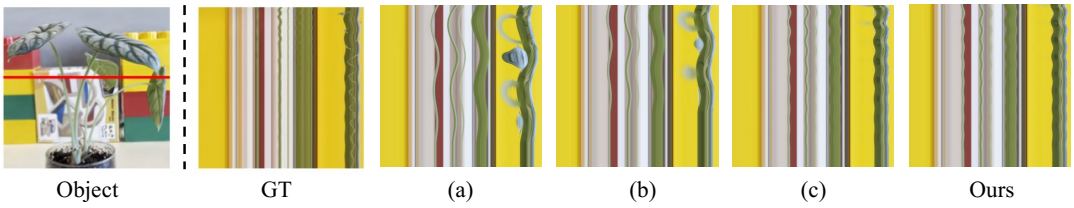

Figure 5: Visualization of space-time slices for ablation study. (a) and (b) are not quite consistent with the ground truth. (c) and our method can generate closer content compared with the ground truth.

gence. Note that, frame boosting is not designed for optimization quality, so our final score is similar to (c).

We provide the visualization of Alocasia in Figure 5. The space-time slices of (a) and (b) are not quite consistent with the ground truth, while (c) and our final method can produce 4D content that is competitive to real-captured videos. These results are consistent with our quantitative results in Table 2.

Conclusion

In this work, we introduced a new framework DreamPhysics, which learns the physical properties of 3D Gaussian Splatting with video diffusion priors. Based on the physics-based simulation, DreamPhysics distills the motion priors to physical parameters with motion distillation sampling. To facilitate that process, we further propose a KAN-based material field with frame boosting. Extensive experiments demonstrate that our method can produce high-quality 4D content with both text and image conditions.

Albeit the improvement compared with previous works, the physics-based 3D dynamics research still faces two problems, *i.e.*, simulated motions and scene-level interaction. Each kind of motion depends on independent physical constraints. Current frameworks can hardly combine all the motions into one simulator. Moreover, simulators can only handle the interactions of a few target objects, but environments are dismissed. For example, in the simulation of the telephone (Figure 4), shadows on the wall cannot change with the movement of the telephone cord. We will explore these problems for future work.

Acknowledgments

This work is in part supported by the National Key R&D Program of China (2021YFF0900500), the National Natural Science Foundation of China (NSFC) under grants 62441202, and two GRF grants from the Research Grants Council of Hong Kong (RGC No.: 11211223 and 11220724).

References

Bahmani, S.; Skorokhodov, I.; Rong, V.; Wetzstein, G.; Guibas, L.; Wonka, P.; Tulyakov, S.; Park, J. J.; Tagliasacchi, A.; and Lindell, D. B. 2023. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv:2311.17984*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv:2311.15127*.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22246–22256.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and pattern Recognition*, 5828–5839.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Fan, L.; Wang, G.; Jiang, Y.; Mandlekar, A.; Yang, Y.; Zhu, H.; Tang, A.; Huang, D.-A.; Zhu, Y.; and Anandkumar, A. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35: 18343–18362.

Feng, Y.; Feng, X.; Shang, Y.; Jiang, Y.; Yu, C.; Zong, Z.; Shao, T.; Wu, H.; Zhou, K.; Jiang, C.; and Yang, Y. 2024a. Gaussian Splashing: Unified Particles for Versatile Motion Synthesis and Rendering. *arXiv:2401.15318*.

Feng, Y.; Feng, X.; Shang, Y.; Jiang, Y.; Yu, C.; Zong, Z.; Shao, T.; Wu, H.; Zhou, K.; Jiang, C.; et al. 2024b. Gaussian Splashing: Dynamic Fluid Synthesis with Gaussian Splatting. *arXiv:2401.15318*.

Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; and Bennamoun, M. 2020. Deep learning for 3D point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv:2311.04400*.

Huang, T.; Dong, B.; Yang, Y.; Huang, X.; Lau, R. W.; Ouyang, W.; and Zuo, W. 2023a. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22157–22167.

Huang, T.; Zeng, Y.; Dong, B.; Xu, H.; Xu, S.; Lau, R. W.; and Zuo, W. 2023b. TextField3D: Towards Enhancing Open-Vocabulary 3D Generation with Noisy Text Fields. *arXiv:2309.17175*.

Huang, T.; Zeng, Y.; Zhang, Z.; Xu, W.; Xu, H.; Xu, S.; Lau, R. W.; and Zuo, W. 2024a. Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5364–5373.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024b. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiang, C.; Schroeder, C.; Teran, J.; Stomakhin, A.; and Selle, A. 2016. The material point method for simulating continuum materials. In *ACM SIGGRAPH 2016 Courses*, 1–52.

Jiang, Y.; Yu, C.; Xie, T.; Li, X.; Feng, Y.; Wang, H.; Li, M.; Lau, H.; Gao, F.; Yang, Y.; et al. 2024. VR-GS: a physical dynamics-aware interactive gaussian splatting system in virtual reality. In *Proceedings of the ACM SIGGRAPH 2024 Conference*.

Jun, H.; and Nichol, A. 2023. Shap-E: Generating Conditional 3D Implicit Functions. *arXiv:2305.02463*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15954–15964.

LAION-AI. 2022. aesthetic-predictor. https://github.com/LAION-AI/aesthetic-predictor.

Liao, T.; Yi, H.; Xiu, Y.; Tang, J.; Huang, Y.; Thies, J.; and Black, M. 2023. Tada! text to animatable digital avatars. *arXiv:2308.10899*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9298–9309.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024. Kan: Kolmogorov-arnold networks. *arXiv:2404.19756*.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2023. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. *arXiv:2310.15008*.

Lu, G.; Zhang, S.; Wang, Z.; Liu, C.; Lu, J.; and Tang, Y. 2024. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. *arXiv:2403.08321*.

Macklin, M. 2022. Warp: A High-performance Python Framework for GPU Simulation and Graphics. https://github.com/nvidia/warp. NVIDIA GPU Technology Conference (GTC).

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

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv:2212.08751*.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 165–174.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. 2019. 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*, 10975–10985.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv:2209.14988*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Reizenstein, J.; Shapovalov, R.; Henzler, P.; Sbordone, L.; Labatut, P.; and Novotny, D. 2021. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10901–10911.

Ren, J.; Pan, L.; Tang, J.; Zhang, C.; Cao, A.; Zeng, G.; and Liu, Z. 2023. DreamGaussian4D: Generative 4D Gaussian Splatting. *arXiv:2312.17142*.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088): 533–536.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9339–9347.

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

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

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

Stomakhin, A.; Schroeder, C.; Chai, L.; Teran, J.; and Selle, A. 2013. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4): 1–10.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv:2402.05054*.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023a. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12619–12629.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023b. Modelscope text-to-video technical report. *arXiv:2308.06571*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023c. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv:2309.15103*.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Xia, F.; Zamir, A. R.; He, Z.; Sax, A.; Malik, J.; and Savarese, S. 2018. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9068–9079.

Xie, T.; Zong, Z.; Qiu, Y.; Li, X.; Feng, Y.; Yang, Y.; and Jiang, C. 2023. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv:2311.12198*.

Yu, C.; Lu, G.; Zeng, Y.; Sun, J.; Liang, X.; Li, H.; Xu, Z.; Xu, S.; Zhang, W.; and Xu, H. 2023. Towards High-Fidelity Text-Guided 3D Face Generation and Manipulation Using only Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15326–15337.

Zhang, T.; Yu, H.-X.; Wu, R.; Feng, B. Y.; Zheng, C.; Snavely, N.; Wu, J.; and Freeman, W. T. 2024. PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation. *arXiv:2404.13026*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv:2305.13077*.

Zhao, Y.; Yan, Z.; Xie, E.; Hong, L.; Li, Z.; and Lee, G. H. 2023. Animate124: Animating one image to 4d dynamic scene. *arXiv:2311.14603*.