Phys4DGen: Physics-Compliant 4D Generation with Multi-Material Composition Perception

Jiajing Lin
Xiamen University
School of Informatics
Xiamen, China
31520231154298@stu.xmu.edu.cn

Zhenzhong Wang
Xiamen University
School of Informatics
Xiamen, China
zhenzhongwang@xmu.edu.cn

Dejun Xu
Xiamen University
School of Informatics
Xiamen, China
xudejun@stu.xmu.edu.cn

Shu Jiang
Xiamen University
Institute of Artificial Intelligence
Xiamen, China
36920241153222@stu.xmu.edu.cn

Yunpeng Gong
Xiamen University
School of Informatics
Xiamen, China
fmonkey625@gmail.com

Min Jiang*
Xiamen University
School of Informatics
Xiamen, China
minjiang@xmu.edu.cn

Abstract

4D content generation aims to create dynamically evolving 3D content that responds to specific input objects such as images or 3D representations. Current approaches typically incorporate physical priors to animate 3D representations, but these methods suffer from significant limitations: they not only require users lacking physics expertise to manually specify material properties but also struggle to effectively handle the generation of multi-material composite objects. To address these challenges, we propose Phys4DGen, a novel 4D generation framework that integrates multi-material composition perception with physical simulation. The framework achieves automated, physically plausible 4D generation through three innovative modules: first, the 3D Material Grouping module partitions heterogeneous material regions on 3D representations' surfaces via semantic segmentation; second, the Internal Physical Structure Discovery module constructs the mechanical structure of object interiors; finally, we distill physical prior knowledge from multimodal large language models to enable rapid and automatic material properties identification for both objects' surfaces and interiors. Experiments on both synthetic and real-world datasets demonstrate that Phys4DGen can generate high-fidelity 4D content with physical realism in open-world scenarios, significantly outperforming state-of-the-art methods.

CCS Concepts

• **Computing methodologies → Computer vision tasks**; • **Information systems → Multimedia content creation**.

Keywords

4D generation, 3D Gaussian Splatting, Multi-modal understanding

ACM Reference Format:
Jiajing Lin, Zhenzhong Wang, Dejun Xu, Shu Jiang, Yunpeng Gong, and Min Jiang. 2025. Phys4DGen: Physics-Compliant 4D Generation with Multi-Material Composition Perception. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755647

1 Introduction

The creation of 4D content has gained substantial importance across multiple domains, including computer animation, interactive gaming, and immersive virtual reality applications [4]. In particular, recent advances in generative models have fundamentally transformed 4D generation due to their powerful visual priors [5, 29, 34, 39, 41, 50]. They leverage dynamic priors from video diffusion models [32, 37, 48, 49, 52], enabling the automated production of high-quality 4D content. However, these data-driven approaches fundamentally lack physical modeling constraints, often resulting in generated motions that violate physical laws and exhibit noticeable inconsistencies.

To ensure the generation of physically realistic 4D content, recent works [17, 44, 54] have explored the incorporation of physical priors, such as continuum mechanics, to animate 3D representations [14, 27]. PhysGaussian [44] pioneered the integration of physical properties into 3D Gaussian Splatting (3DGS) [14] representations, and introduced the Material Point Method (MPM) [36] for dynamic generation. However, it requires manually setting the material type and properties of the simulation object. Methods such as PAC-NeRF [17] and GIC [6] are capable of estimating physical properties under the supervision of multi-view videos. However, they rely heavily on multi-view video data, which is often difficult to obtain, thereby significantly hindering their practical applicability. PhysDreamer [51], DreamPhysics [12], and Physics3D [21] leverage dynamic priors from pre-trained video diffusion models to guide the estimation of material properties, based on a given 3DGS model rather than multi-view videos. This enables automatic material properties determination without strict input constraints.

[*] Corresponding author: Min Jiang, minjiang@xmu.edu.cn

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755647

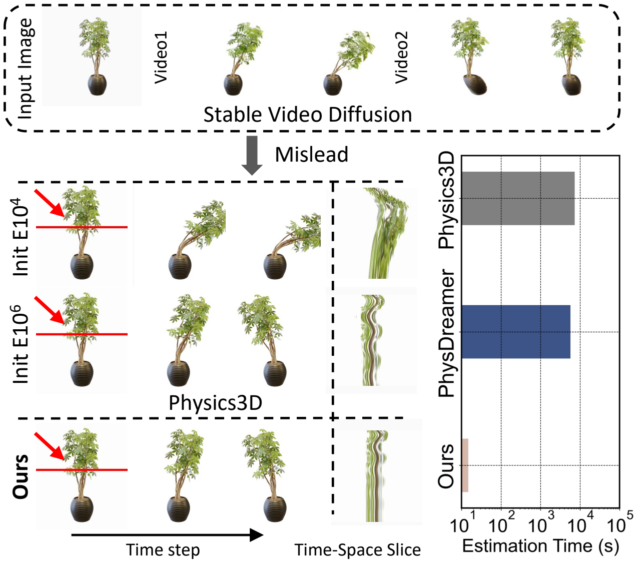


Figure 1: The red arrows indicate the direction of external forces. We use the space-time slice (right column), where the vertical axis represents time and the horizontal axis shows a spatial slice of the object (marked by red lines), to reveal motion intensity and frequency. As shown in the top, diffusion models embed unrealistic motion priors that may mislead the estimation process—e.g., Physics3D consistently overestimates the softness of the ficus, deviating from physical plausibility. Additionally, the accuracy of such approaches heavily depends on the setting of initial material properties (e.g., Init Young's modulus 10^4 vs. 10^6). In contrast, our method achieves more accurate material properties estimation within 14.88 seconds, enabling reliable simulation.

Despite these advances, existing methods still face several critical challenges. 1) First, these methods typically assume that objects are uniform entities made of a single material, whereas real-world objects often consist of multiple heterogeneous materials. Failing to distinguish between different materials hinders the accurate simulation of localized deformation behaviors, reducing physical realism. 2) Second, while 3DGS effectively captures the surface geometry of objects, it lacks the capability to model internal structures. However, object interiors often contain multiple materials, which may even differ from those on the surface. Simulations based on such structure-agnostic representations are prone to structural collapse under large deformations. 3) In addition, as shown in Fig. 1, the dynamic priors embedded in video diffusion models, which lack physical constraints, may mislead the estimation of material properties. Furthermore, due to their iterative optimization process, these methods are computationally expensive. Both their convergence speed and estimation quality are also highly sensitive to initial material properties settings, which typically require domain-specific physics knowledge. This poses a significant barrier for general users who lack such expertise in 4D content generation.

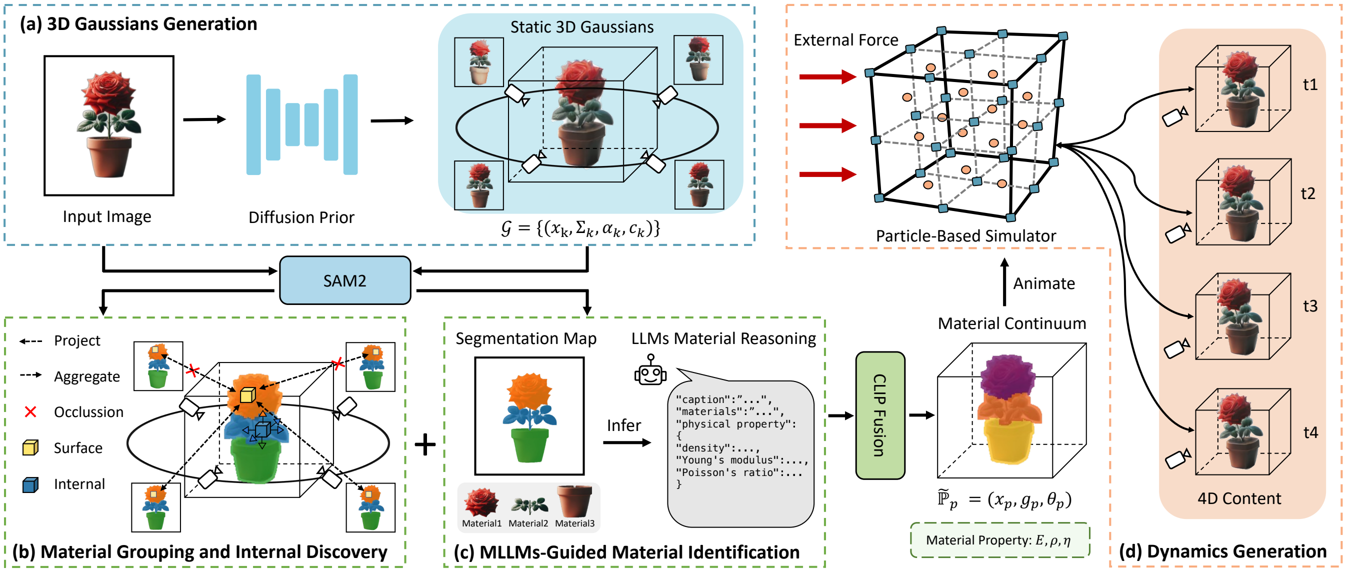
To overcome the limitations of previous approaches, we introduce *Phys4DGen*, a novel physics-driven 4D generation framework that integrates multi-material composition perception into the 4D generation pipeline for the first time, enabling fast, user-friendly, and physically plausible 4D content generation from a single image or 3D representation. Our framework addresses three key challenges: (1) the 3D Material Grouping module, which extends the segmentation semantics of large vision models (e.g., SAM2[31]) from 2D to 3D space for accurate surface material grouping; (2) the Internal Physical Structure Discovery module, which models the mechanical structure of object interiors; and (3) a multimodal physics expert that leverages physical knowledge embedded in large language models to automatically and rapidly identify material properties for both surfaces and internal structures. By unifying these components, *Phys4DGen* enables more complete and physically realistic 4D generation. Our key contributions include:

- We propose a 4D generation framework that distills prior knowledge from a foundation model to enable the multi-material composite perception, while incorporating physical simulations to achieve user-friendly and physically realistic 4D generation.
- 3D Material Grouping is introduced to partition object surfaces into distinct material regions, and Physical Internal Structure Discovery for modeling internal structures, together enabling the handling of multi-material composition.
- We are the first to leverage physical prior knowledge from a multimodal large language model to enable automatic and efficient material identification.
- Extensive qualitative and quantitative comparisons on both synthetic and real-world datasets demonstrate that our method can generate physically consistent and high-fidelity 4D content across various materials.

2 Related Work

2.1 4D Generation

4D generation aims to generate dynamic 3D content that aligns with user input conditions such as text, images, and videos. Unlike 3D generation [16, 20, 26, 30, 40, 42], which primarily focuses on producing spatially consistent geometry and appearance, 4D generation must additionally ensure temporal realism and consistency across frames, making the task significantly more challenging. Based on the input conditions, 4D generation can be categorized into three types: text-to-4D [2, 3, 19, 35, 53], video-to-4D [8, 11, 28, 47, 49], and image-to-4D [37, 45, 46, 52]. MAV3D [35] first employs temporal SDS from the text-to-video diffusion model to optimize HexPlane [7] representation. 4D-fy [3] introduces hybrid score distillation during training for high-quality text-to-4D generation. Instead of text input, Consistent4D [13] uses SDS for geometry optimization and interpolation loss for spatiotemporal consistency in 4D generation from monocular video. Animate124 [52] pioneered an image-to-4D framework using a coarse-to-fine strategy that combines different diffusion priors [24, 25]. Generating 4D content from an image in DreamGaussian4D [32] avoids using temporal SDS and instead performs optimization based on reference videos generated by a video diffusion model. Our framework can generate physically plausible and temporally coherent 4D content efficiently, without extra optimization steps.



```
"caption":"...",
"materials":"...",
"physical property":
{
"density":...,
"Young's modulus":...,
"Poisson's ratio":...
}
```

Figure 2: Framework of *Phys4DGen*. (a) 3D Gaussians Generation: Given an input image, a static 3D Gaussians is generated under the guidance of the diffusion model. (b) Material Grouping and Internal Discover: 3D Material Grouping is applied to partition the 3D Gaussians into distinct material groups. Concurrently, Internal Physical Structure Discovery is used to fill internal particles and determine their corresponding material groups. (c) MLLMs-Guided Material Identification: Surface and internal material properties are visually inferred by MLLMs. These inferred results are then integrated into the 3D representation $\mathbb{P}$ through the CLIP Fusion module, forming a material continuum representation $\tilde{\mathbb{P}}$. (d) 4D Dynamics Generation: Given external forces, MPM simulation is performed to animate the material continuum, thereby generating 4D content.

2.2 Physics-Based Dynamic Generation

Some recent methods have attempted to leverage physical simulation to create visual dynamics with physical realism. PhysGen [23] introduces rigid-body physics simulation to achieve physically grounded video synthesis from a single image. PAC-NeRF [17] recovers object geometry and physical properties from multi-view videos by combining NeRF with differentiable physics, without requiring known shapes. However, NeRF's implicit representation is not ideal for physical simulation. The explicit representation of 3DGS [14, 43] through a set of anisotropic Gaussian kernels, which can be interpreted as particles in space, enables the many applications of physical simulations [10, 18, 44, 54]. PhysGaussian [44] is the first to apply MPM to 3D Gaussian representations, enabling the simulation of realistic physical dynamics. Spring-Mass [54] introduces a novel integration of a spring-mass system into the 3DGS framework for elastic material simulation. However, they require manual specification of material types, material properties, and simulation regions. PhysDreamer [51] utilizes reference videos from video diffusion models for supervision to estimate material properties. Likewise, DreamPhysics and Physics3D [12, 21] use score distillation sampling from video diffusion models to optimize physical properties. The stochastic nature of video diffusion models and their non-physical dynamic priors can distort material property estimation. Our method does not require an optimization process and can efficiently infer material properties. Notably, we uniquely consider the possibility that a single object may comprise multiple materials, including differing internal and external compositions.

3 Methodology

The proposed *Phys4DGen* is illustrated in Fig 2. Given a single input image, a static 3D Gaussians representation is generated under the guidance of a diffusion model (Sec. 3.1). We then employ **3D Material Grouping** (Sec. 3.2) to assign suitable material groups to different regions of the 3D Gaussians. To further explore internal details, **Internal Physical Structure Discovery** (Sec. 3.3) is employed to model internal geometries and assign the internal material groupings. Subsequently, **MLLM-Guided Material Identification** (Sec. 3.4) infers both surface and internal material properties for each material region from the input image. These are fused into the 3D representation through CLIP Fusion, producing a material continuum representation. Finally, given external forces and boundary conditions, the 4D dynamics are simulated using MPM (Sec. 3.5) based on the material continuum.

3.1 Static 3D Gaussians Generation

Given the recent advances in image-to-3D generation [38, 39], which have demonstrated the ability to produce high-quality 3D content, we directly employ these methods to generate static 3D Gaussians. We choose 3DGS as a representation for its explicit representation nature. 3DGS represents 3D objects using a collection of anisotropic Gaussian kernels [14]. Thus, 3D Gaussians can be viewed as a discretization of the continuum, which is highly beneficial for integrating particle-based physical simulation algorithms. In this phase, we obtain a static 3D Gaussians $\mathcal{G}$ for subsequent simulation. Each Gaussian kernel can be represented as

$\mathcal{G}_{\boldsymbol{k}} = (\mathbf{x}_k, \Sigma_k, \alpha_k, \mathbf{c}_k)$. Here, $\mathbf{x}_k$, Σ_k, α_k, and $\mathbf{c}_k$ represent the center position, covariance matrix, opacity, and color of the Gaussian kernel k, respectively.

3.2 3D Material Grouping

In practice, many objects are composites composed of different materials. Prior to material identification, it is essential to group objects into distinct material components. Thus, we propose 3D Material Grouping, which lifts the segmentation semantics from the vision foundation model from 2D to 3D space, enabling the assignment of a unique material group to each Gaussian kernel.

3.2.1 Pre-Process. Given the static 3D Gaussians generated in the previous phase, we render the scene from multiple views, producing an image sequence and its corresponding depth maps. SAM2 [15, 31] is a powerful vision foundation model, supports accurate video segmentation. To ensure consistency of 2D mask maps across views—i.e., that the same material region receives the same mask index from different views—we consider the multi-view image sequence as a video input and apply SAM2 for segmentation to generate the associated mask maps $\mathbf{M} = \{\mathbf{M}_o\}_{o=1}^{N}$.

3.2.2 Projection and Aggregation. We treat each mask index as a material group label g, resulting in a sequence of mask maps with consistent material groupings across views. For a given Gaussian kernel $\mathcal{G}_k \in \mathcal{G}$ and a mask map $\mathbf{M}_o \in \mathbf{M}$, we use the camera's intrinsic and extrinsic parameters to project the Gaussian kernel into 2D space, obtaining its 2D coordinates $\mathbf{x}_p^{2d}$ on the corresponding mask map. This process can be expressed as:

$$\mathbf{x}_k^{2d} = \mathbf{K}[\mathbf{R}_o|\mathbf{T}_o]\mathbf{x}_k, \tag{1}$$

where $\mathbf{K}$ and $[\mathbf{R}_o|\mathbf{T}_o]$ represent the camera's intrinsic and extrinsic parameters, respectively. We use the 3DGS-estimated depth to check if the Gaussian kernel is visible in the segmentation map $\mathbf{M}_o$. If it is visible, we include the material group from this view in the voting process. After processing the segmentation maps for all views, we perform majority voting, assigning the material group g_k that appears most frequently across all views to Gaussian kernel $\mathcal{G}_{\boldsymbol{k}} = (\mathbf{x}_k, \Sigma_k, \alpha_k, \mathbf{c}_k, g_k)$. Repeating these steps allows us to determine the material groups for all Gaussian kernels $\mathcal{G}$.

3.3 Physical Internal Structure Discovery

Due to an inherent limitation of the 3DGS representation, a large number of Gaussian kernels are distributed only on the surface, leaving the interior empty. This leads to reduced fidelity in physical simulations, especially under large deformations. To address this limitation, we propose a strategy termed Physical Internal Structure Discovery, which enables internal material filling and grouping.

Concretely, grid particles are initialized by uniform sampling within the bounding box of the 3D Gaussians $\mathcal{G}$. These particles are then projected onto multi-view mask images and depth maps. Each particle is filtered by comparing the projected depth with the rendered depth and checking whether it lies within the foreground mask. Based on the 3D object represented by the valid grid particles, we further classify them into surface and internal particles.

To distinguish surface particles, we analyze the spatial relationship between each Gaussian kernel and the grid particles using the Mahalanobis distance [9]. Specifically, for a Gaussian kernel with mean μ and covariance Σ, the Mahalanobis distance from a particle located at position $\mathbf{x}$ is computed as:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}, \tag{2}$$

If this distance is less than a predefined threshold, the grid particle is considered to be covered by the Gaussian kernel. All such particles are labeled as surface particles $\mathcal{P}_S$, while those not included by any Gaussian are classified as internal particles $\mathcal{P}_I$.

Since the internal structure of an object can only be inferred from surface information available in the input image, we establish a surface-to-interior correspondence by assigning material groups from surface particles to internal ones. To ensure closure of the boundary set, we additionally designate the outermost layer of grid particles as surface particles. Each surface particle is then assigned to its nearest Gaussian kernel via a nearest-neighbor search and inherits its material group accordingly. For each internal particle, rays are cast along the six principal axes to collect material groups from the intersected surface particles. The most frequently observed material group g_i is then assigned to the internal particle. Subsequently, we merge 3D Gaussians $\mathcal{G}$ with the filled internal particles $\mathcal{P}_I$ to obtain a unified continuum particles representation with material groups: $\mathbb{P} = \{\mathcal{G}|\mathcal{P}_I\} = \{\{(\mathbf{x}_k, g_k)\}|\{(\mathbf{x}_i, g_i)\}\}$. Based on this representation, we can easily assign material properties to different material regions of the 3D object.

3.4 MLLMs-Guided Material Identification

3.4.1 Material Information Reasoning. In the real world, objects are typically composed of different materials. In 4D generation, users often lack the necessary physical knowledge to provide reasonable material properties for simulation. This greatly limits the physical realism of the simulation results. Recently, multimodal large language models have advanced rapidly, exhibiting knowledge far beyond that of humans, including rich physical prior knowledge. Inspired by this, we introduce GPT-4o [1], which reasons the material properties (e.g., Density ρ, Young's modulus E, Poisson's ratio v) of the internal and external parts of objects through vision.

Specifically, the user-input image is treated as the canonical view $\mathbf{I}_c$ for reasoning with the MLLMs. Sub-images representing different material components are extracted from this view based on the segmentation mask map. Following this, the canonical view and its segmented sub-images are fed into GPT-4o, prompting it to reason the internal and external material types and properties for the object described in each segmented sub-image. Detailed prompts are provided in the appendix.

3.4.2 CLIP Fusion. To integrate the inferred surface and internal material properties into the 3D representation. It is necessary to align the material groupings of the continuum particles—obtained in Sec. 3.2 and Sec. 3.3—with those derived from the canonical view. Specifically, we first extract the CLIP embedding for all segmented sub-images in the canonical view and the rendered image $\mathbf{I}_f$ from the front view, which contains the same object information as the canonical view. This is represented as:

$$\mathbf{L}_c(\cdot) = \mathbf{V}(\mathbf{I}_c \odot \mathbf{M}_c(\cdot)), \tag{3}$$

$$\mathbf{L}_f(\cdot) = \mathbf{V}(\mathbf{I}_f \odot \mathbf{M}_f(\cdot)), \tag{4}$$

where $\mathbf{V}$ is the CLIP image encoder and $\mathbf{L}$ represents the mask CLIP embedding. $\mathbf{M}(\cdot)$ denotes the sub-region of the mask map labeled by the specified mask index.

Next, we calculate the similarity between the CLIP features of the canonical view and those of the rendered image. By selecting the matching pairs with the highest similarity, we establish a correspondence between the material groupings. Based on this, we can assign the surface and internal material information to the corresponding continuum particles. Finally, we construct a complete material continuum representation $\tilde{\mathbb{P}} = (\mathbf{x}_p, g_p, \theta_p)$, where θ_p denotes the material properties, for use in subsequent physical simulation to generate 4D dynamics.

3.5 4D Dynamics Generation

Phys4DGen can integrate any particle-based physical simulation algorithm. In this paper, we use the Material Point Method (MPM) [44] to simulate the dynamics of 4D content, which enables the modeling of motion and deformation behavior of continuum under external forces. For details on MPM and external forces, please refer to the appendix. For physical simulation, we further assign temporal properties t to the material continuum, along with other physical attributes involved in the simulation process, such as mass m, deformation gradient $\mathbf{F}$, and velocity $\mathbf{v}$. Then, we employ MPM to perform physical simulations on the material continuum $\tilde{\mathbb{P}}^t$. This allows us to track the position and local deformation of each particle at every time step:

$$\mathbf{x}^{t+1}, \mathbf{F}^{t+1}, \mathbf{v}^{t+1} = \text{MPMSimulator}(\tilde{\mathbb{P}}^t), \tag{5}$$

where $\mathbf{x}^{t+1} = \{\mathbf{x}_p^{t+1}\}_{p=1}^P$ denotes the positions of all particles at time step t+1. $\mathbf{F}^{t+1} = \{\mathbf{F}_p^{t+1}\}_{p=1}^P$ represents deformation gradients, which describe the local deformation of each particle at time step $t+1$. To reconstruct the 3D Gaussian Splatting (3DGS) representation at time step t from the simulation results, we isolate the 3DGS-relevant components from the simulated material continuum. To incorporate the local deformation behavior of each GS kernel, we interpret the deformation gradient as a local affine transformation applied to the Gaussian kernel. Consequently, we can derive the covariance matrix of the Gaussian kernel k in step $t+1$:

$$\Sigma_k^{t+1} = (\mathbf{F}_k^{t+1})\Sigma_k^t(\mathbf{F}_k^{t+1})^T. \tag{6}$$

At each step of the MPM simulation, we obtain the deformed 3DGS representation. The sequence of 3DGS representations across all time steps collectively forms the 4D content. This enables the generation of physically plausible 4D dynamics.

4 Experiments

4.1 Experimental Setup

4.1.1 Implementation Details. Phys4DGen supports both a single image and a 3D model as input. Given a single image, we use LGM [38] to generate static 3D Gaussians. For material grouping, we render multiview images from the generated 3D Gaussians and apply SAM2 [31] to obtain cross-view consistent material mask maps. We used GPT-4o to identify the material for each material region in the image. For 4D dynamics generation, we perform physical simulation using MPM [36]. For each example, the simulation environment is configured based on the material information inferred by GPT-4o, and different external forces are applied according to the specific case, allowing the generation of physically plausible dynamic 4D sequences. All experiments were conducted on NVIDIA A40(48GB) GPU. For more detailed information on the experimental settings, please refer to the appendix.

4.1.2 Datasets. To thoroughly assess the effectiveness of our approach across varying input types, we establish separate datasets for the Image-to-4D and 3D-to-4D tasks. For the image-to-4D task, we use a total of 11 samples, including 8 synthetic image samples (4 sourced from Zero123[22] and Animate124[52], and 4 created by ourselves), as well as 3 image samples collected from the real world. The datasets feature a range of materials, including elastic, elastoplastic, granular media, and snow. For the 3D-to-4D task, we choose four real-world static scenes from PhysDreamer [51] (alocasia, carnations, telephone and hat), along with additional scenes: ficus from PhysGaussain[44] and basketball from Physics3D[21].

4.1.3 Baselines. We compare our method both qualitatively and quantitatively with existing SOTA 4D generation methods. For the image-to-4D task, we compare our approach with several image-to-4D generation methods, including DG4D [32], STAG4D [48], and L4GM [33], with a primary focus on assessing their performance in terms of spatiotemporal consistency and physical realism. For the 3D-to-4D task, we compare our method with PhysGaussian [44], PhysDreamer [51], and Physics3D [21]—approaches that incorporate physical priors—using the PhysDreamer datasets.

4.1.4 Metrics. Following previous works [47, 52], we use CLIP-T score which calculates the average cosine similarity between the CLIP embeddings of every two adjacent frames in rendered video from a given view. To further assess the spatiotemporal consistency, we render videos from the right, back, and left views to calculate the CLIP-T-other score. We also conduct a user preference study to evaluate the physical realism (PR) and overall quality (OQ) of the generated 4D content, with both metrics rated on a 10-point scale. More details on the user study can be found in the appendix.

4.2 Showcase of 4D Generation Results

Fig. 3 visualizes the 4D content generated by Phys4DGen. For each example, we present the perceived material properties, including the density and Young's modulus of both the object's surface and internal regions. As shown, Phys4DGen effectively discerns the material compositions of 3D objects, such as differentiating the distinct physical attributes of carnations' petals and stems, and recognizing the material differences in a doll's fabric surface and polyester fiberfill interior. Furthermore, we render the corresponding 4D content from dynamically changing viewpoints. The results demonstrate Phys4DGen's capability to generate physically realistic 4D content from a single image or 3D content. Please refer to the appendix for additional visual results.

4.3 Comparison in Image-to-4D Generation

Fig. 4 presents a qualitative comparison of our Phys4DGen method with state-of-the-art (SOTA) image-to-4D generation methods. Generating 4D content using STAG4D and DG4D heavily depends on the quality of reference videos produced by video diffusion models. However, due to the inherently stochastic and data-driven nature of

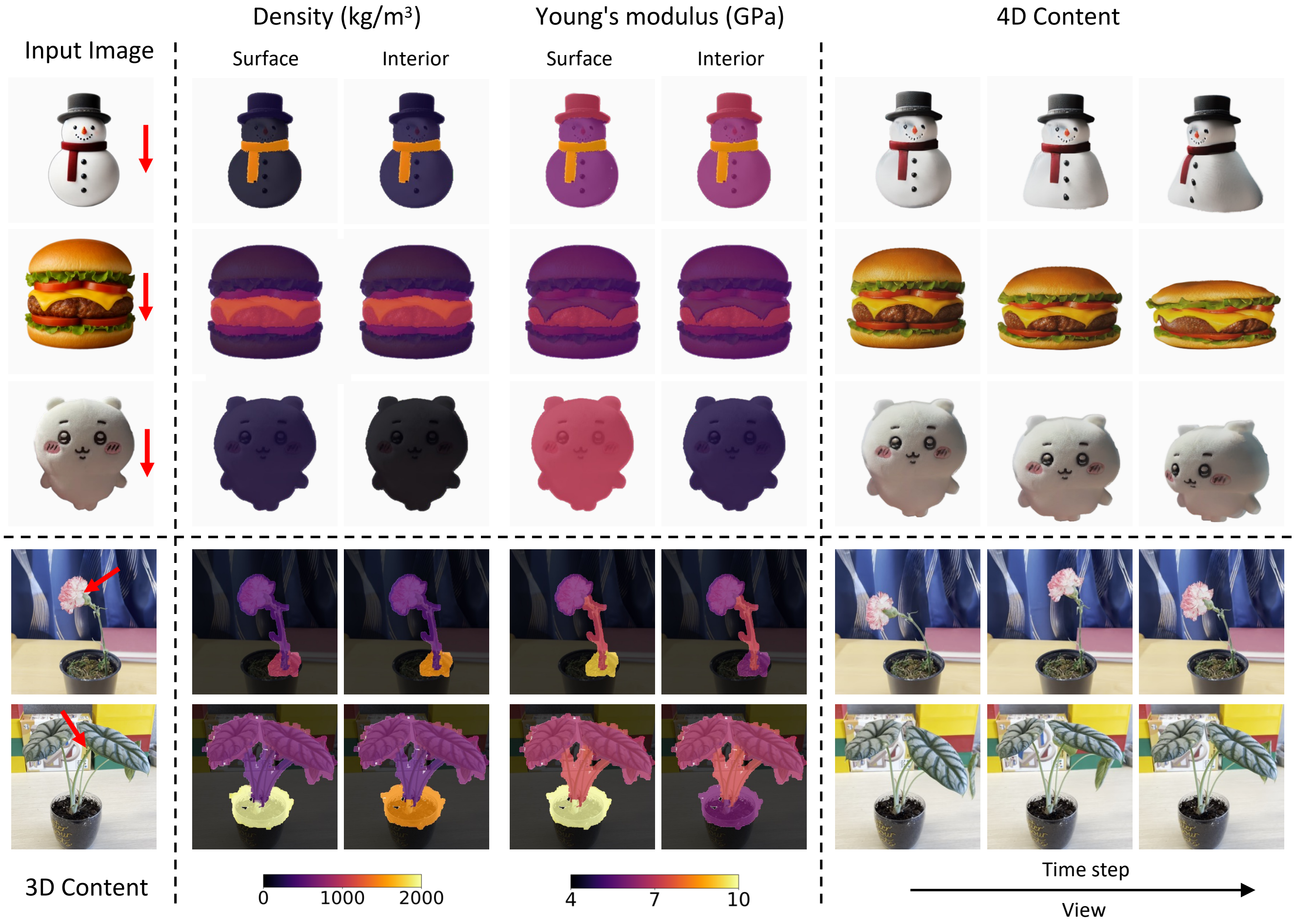


Figure 3: Visual results of *Phys4DGen*. *Phys4DGen* is capable of perceiving the multi-material composition of 3D objects and generating physically realistic 4D content under given external forces (red arrows).

these models, it remains challenging to obtain reference videos that simultaneously adhere to physical laws and faithfully align with the user's intent. This limitation is evident in Fig. 4, where the 4D content generated by STAG4D and DG4D struggles with controllability and often violates fundamental physical principles. For example, in Fig. 4, the 4D content generated by DG4D [32] for a balloon tied to a wooden block moves upwards, violating gravity. In stark contrast, Fig. 4 clearly demonstrates that our generated 4D content achieves high fidelity and adheres physical laws, significantly outperforming baseline methods. This qualitative observation is further supported by the quantitative results presented in Tab. 1. As shown in the table, our method achieves the highest CLIP-T and CLIP-T-other scores, indicating that it is capable of generating spatiotemporally consistent 4D content. The user study results on physical realism (PR) and overall quality (OQ) indicate that participants found our generated 4D content more physically plausible and expressed a stronger preference for our results over the baselines. Additionally, our method offers superior controllability, enabling users to adjust external forces according to their specific needs—an aspect where baseline approaches fall short.

4.4 Comparison in 3D-to-4D Generation

Following PhysDreamer[51], we compare our results with real captured videos and simulations from other methods using space-time slices, where the vertical axis represents time and the horizontal axis shows a spatial slice of the object, as indicated by the red lines in the "object" column. We use 90-frame video sequences to generate space-time slices. This representation enables an effective comparison of the magnitude and frequencies of the oscillatory motions generated by different methods. Fig. 5 demonstrates that the dynamics generated by our method more closely resemble those of real-world videos. As shown in Tab. 2, our method achieves the highest average scores in physical realism (PR), overall quality (OQ), and CLIP-T, significantly outperforming baseline methods. These results validate the effectiveness of our proposed multi-material composition perception capability, which enables relatively accurate material property estimation within only 14.88 seconds, in

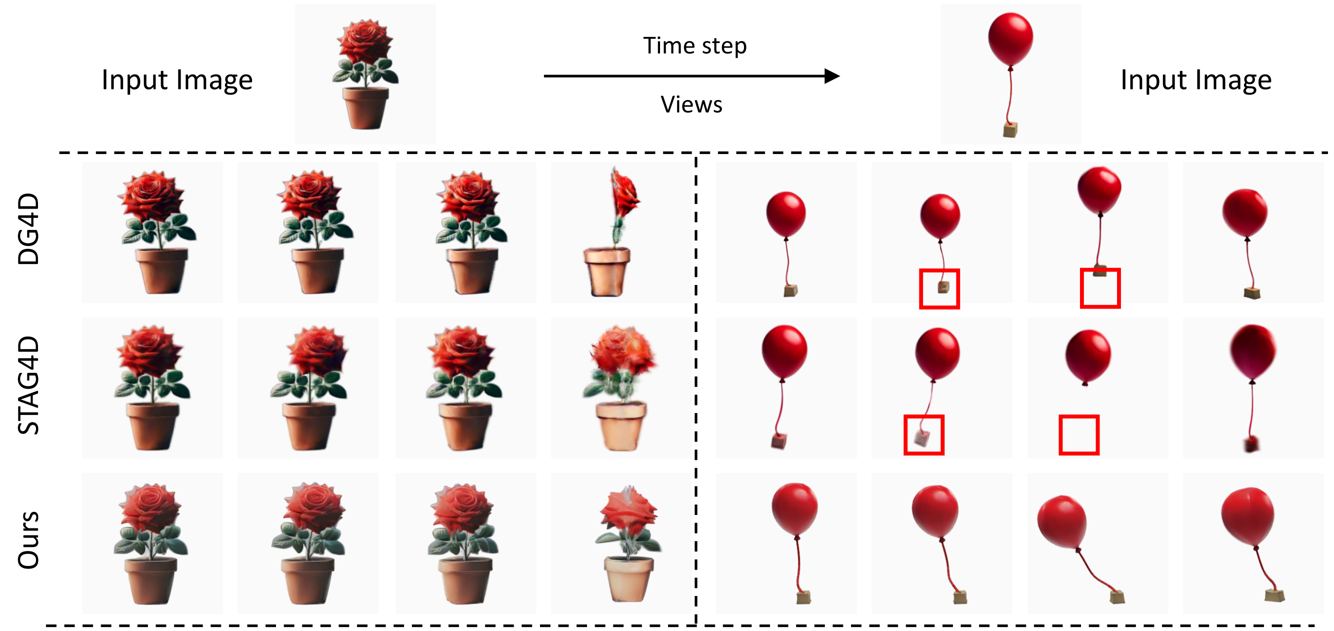


Figure 4: Qualitative comparison in image-to-4D generation. To compare the spatiotemporal consistency, the rendering view changes with each time step. The red box highlights regions exhibiting physically implausible behavior for further observation. The dynamics generated by our method are more consistent with physical laws compared to the baseline method.

Method	PR ↑	OQ ↑	CLIP-T ↑	CLIP-T other ↑
DG4D [32]	5.90	6.25	0.98983	0.98536
STAG4D [48]	5.55	5.97	0.98813	0.98492
L4GM [33]	6.30	6.70	0.99242	0.99275
Ours	**7.50**	**7.72**	**0.99459**	**0.99409**

Table 1: Quantitative comparison in image-to-4D generation. PR evaluates the physical realism of the 4D content, OQ measures its overall quality, and CLIP and CLIP-T-other assess its spatiotemporal consistency.

Method	PR ↑	OQ ↑	CLIP-T ↑	Time ↓
PhysGaussian [44]	5.67	6.30	0.99855	-
PhysDreamer [51]	6.43	6.90	0.99862	5688.79s
Physics3D [21]	7.17	7.53	0.99852	7273.32s
Ours	**7.87**	**7.97**	**0.99925**	**723.14s+14.88s**

Table 2: Quantitative comparison in 3D-to-4D generation. Since PhysGaussians lacks a material property estimation stage, we exclude its runtime from our evaluation. Our method requires only 14.88s for property estimation.

contrast to the hour-level computation time required by baseline methods. This efficient perception contributes to the generation of 4D content that is not only physically plausible but also more preferred by human observers. More comparative results are provided in the appendix.

4.5 Ablation Analysis

To validate the effectiveness of Internal Physical Structure Discovery (IPSD) and multi-material partitioning through Material Grouping, we conduct the ablation study illustrated in Fig. 6. As shown in the figure, the complete model (top-left) demonstrates ideal simulation performance. With both Material Grouping and IPSD enabled, the bun and the beef patty exhibit different deformation behaviors under the same external force: the bun, being softer, undergoes larger deformation, while the beef, being relatively stiffer, deforms less. Meanwhile, the overall structure remains stable, reflecting strong physical plausibility and internal consistency. When the Material Grouping module is removed (top-right), internal structural cues are still present, but the model fails to distinguish between materials effectively. As a result, the bun and beef respond with similar levels of deformation under force, which clearly contradicts physical reality. In contrast, when the IPSD module is removed (bottom-left), the model still captures material differences in deformation response. However, the lack of internal structural support leads to a collapse of the overall geometry under larger external forces, revealing significant issues in maintaining structural stability. The worst performance occurs when both modules are disabled (bottom-right). The simulation results in chaotic material interactions and complete structural failure. In summary, both Material Grouping and IPSD are essential for enhancing the physical realism of 4D content generation. They complement each other—Material Grouping ensures that the behavioral differences between materials are properly captured, while IPSD

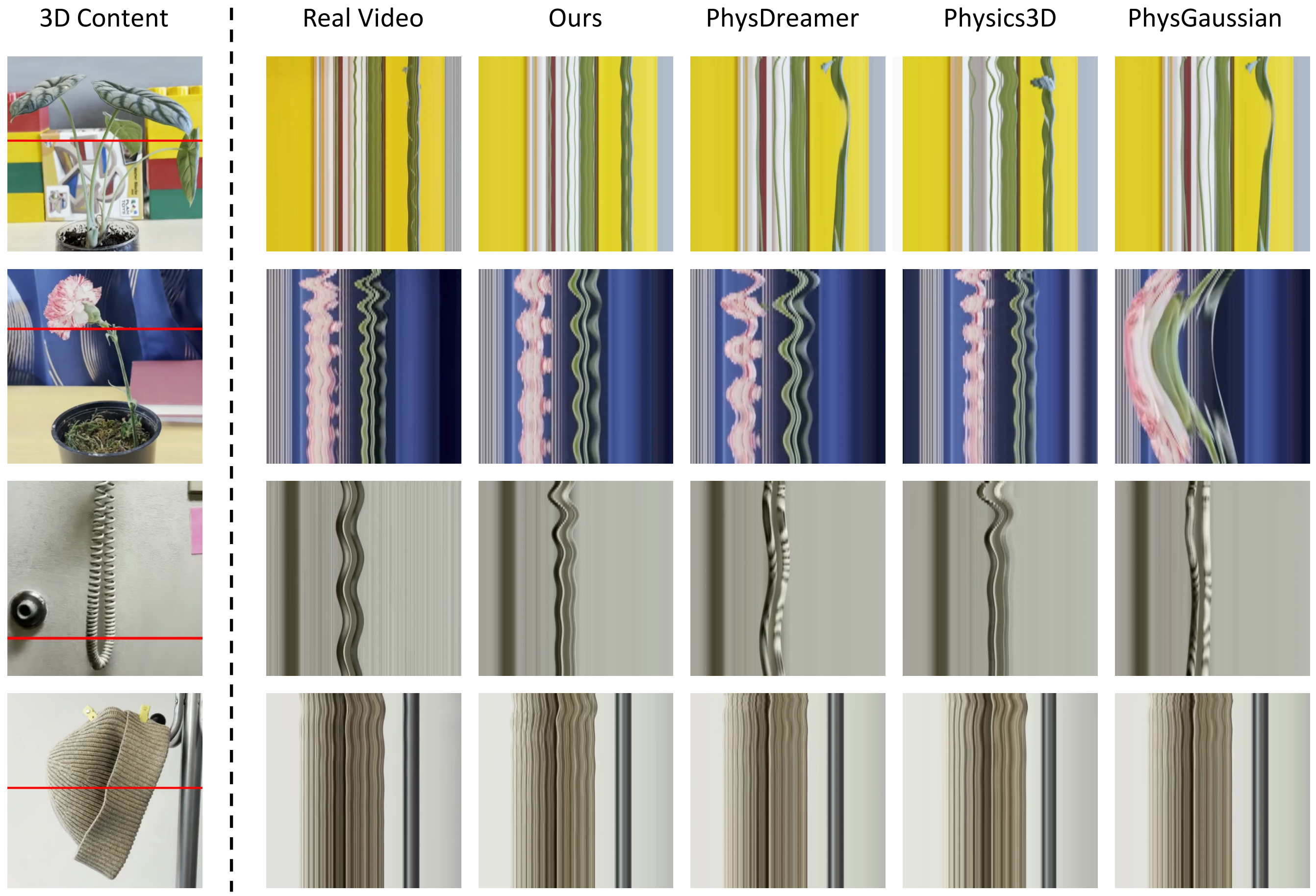


Figure 5: Qualitative comparison for 3D-to-4D generation. We compare our results with real videos and baselines using space-time slices, These slices reveal the motion's intensity and frequency. Our results more closely match the ground truth.

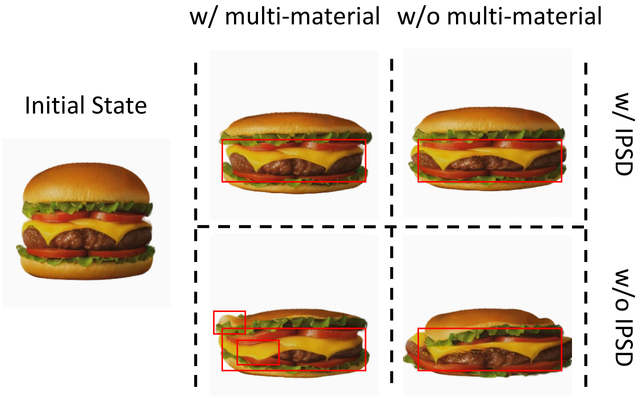


Figure 6: Ablation study on Internal Physical Structure Discover and multi-material partitions through Material Grouping. All examples are tested under the same external forces. The red box is used to assist in observing the deformation.

ensures structural integrity—making them indispensable components for achieving faithful and stable physical effects. Additional experimental analyses can be found in the appendix.

5 Conclusion

In this paper, we propose Phys4DGen, a physics-compliant 4D generation framework that effectively perceives complex multi-material compositions. By seamlessly integrating these perceptual capabilities with physical simulation, our approach enables intuitive and physically plausible 4D generation from a single image or a 3D input. To handle multi-material compositions, we propose the 3D Material Grouping module, which segments an object surface, represented by 3D Gaussians, into distinct material regions. Furthermore, the internal structure of the object is modeled through Physical Internal Structure Discovery. We distill extensive physical priors from GPT-4o to identify surface and internal material properties, which are then assigned to the 3D representation to construct a complete simulation object. Extensive experiments on synthetic and real-world datasets show that our approach generates physically realistic 4D content.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62276222.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. 2024. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*. Springer, 53–72.

[3] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 2024. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7996–8006.

[4] Jason Bailey. 2020. The tools of generative art, from flash to neural networks. *Art in America* 8 (2020), 1.

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[6] Junhao Cai, Yuji Yang, Weihao Yuan, Yisheng He, Zilong Dong, Liefeng Bo, Hui Cheng, and Qifeng Chen. 2024. GIC: Gaussian-Informed Continuum for Physical Property Identification and Simulation. *arXiv preprint arXiv:2406.14927* (2024).

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

[8] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. 2024. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280* (2024).

[9] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems* 50, 1 (2000), 1–18.

[10] Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, et al. 2024. Gaussian splashing: Dynamic fluid synthesis with gaussian splatting. *arXiv preprint arXiv:2401.15318* (2024).

[11] Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann. 2024. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365* (2024).

[12] Tianyu Huang, Yihan Zeng, Hui Li, Wangmeng Zuo, and Rynson WH Lau. 2025. DreamPhysics: Learning Physical Properties of Dynamic 3D Gaussians with Video Diffusion Priors. *AAAI* (2025).

[13] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. 2024. Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video. *Int. Conf. Learn. Represent.* (2024).

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[16] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2024. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *Int. Conf. Learn. Represent.* (2024).

[17] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. 2023. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. *Int. Conf. Learn. Represent.* (2023).

[18] Jiajing Lin, Zhenzhong Wang, Yongjie Hou, Yuzhou Tang, and Min Jiang. 2024. Phy124: Fast physics-driven 4d content generation from a single image. *arXiv preprint arXiv:2409.07179* (2024).

[19] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. 2024. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8576–8588.

[20] Fangfu Liu, Hanyang Wang, Weiliang Chen, Haowen Sun, and Yueqi Duan. 2024. Make-your-3d: Fast and consistent subject-driven 3d content generation. In *European Conference on Computer Vision*. Springer, 389–406.

[21] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. 2024. Physics3D: Learning Physical Properties of 3D Gaussians via Video Diffusion. *arXiv preprint arXiv:2406.04338* (2024).

[22] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 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.

[23] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. 2024. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*. Springer, 360–378.

[24] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2294–2305.

[25] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6430–6440.

[26] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8446–8455.

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

[28] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. 2024. Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742* (2024).

[29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2023. Dreamfusion: Text-to-3d using 2d diffusion. *Int. Conf. Learn. Represent.* (2023).

[30] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. 2024. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *Int. Conf. Learn. Represent.* (2024).

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

[32] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. 2023. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142* (2023).

[33] Jiawei Ren, Cheng Xie, Ashkan Mirzaei, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al. 2024. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems* 37 (2024), 56828–56858.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[35] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. 2023. Text-to-4d dynamic scene generation. *Proc. Int. Conf. Mach. Learn.* (2023).

[36] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. 2013. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 1–10.

[37] Qi Sun, Zhiyang Guo, Ziyu Wan, Jing Nathan Yan, Shengming Yin, Wengang Zhou, Jing Liao, and Houqiang Li. 2024. EG4D: Explicit Generation of 4D Object without Score Distillation. *Eur. Conf. Comput. Vis.* (2024).

[38] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *Eur. Conf. Comput. Vis.* (2024).

[39] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2024. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *Int. Conf. Learn. Represent.* (2024).

[40] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF international conference on computer vision*. 22819–22829.

[41] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2024. Lavie: High-quality video generation with cascaded latent diffusion models. *Int. J. Comput. Vis.* (2024).

[42] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. 2024. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[43] Zhicong Wu, Hongbin Xu, Gang Xu, Ping Nie, Zhixin Yan, Jinkai Zheng, Liangqiong Qu, Ming Li, and Liqiang Nie. 2025. TextSplat: Text-Guided Semantic Fusion for Generalizable Gaussian Splatting. *arXiv preprint arXiv:2504.09588* (2025).

[44] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. 2024. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4389–4398.

[45] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. 2024. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470* (2024).

[46] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. 2024. Diffusion ^{2}: Dynamic 3D Content Generation via Score Composition of Orthogonal Diffusion Models. *arXiv e-prints* (2024), Adv. Neural Inform. Process. Syst.

[47] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 2023. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225* (2023).

[48] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. 2024. Stag4d: Spatial-temporal anchored generative 4d gaussians. *Eur. Conf. Comput. Vis.* (2024).

[49] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 2024. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems* 37 (2024), 15272–15295.

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[51] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. 2024. Physdreamer: Physics-based interaction with 3d objects via video generation. *Eur. Conf. Comput. Vis.* (2024).

[52] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. 2023. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603* (2023).

[53] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. 2024. A unified approach for text-and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7300–7309.

[54] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. 2024. Reconstruction and Simulation of Elastic Objects with Spring-Mass 3D Gaussians. *Eur. Conf. Comput. Vis.* (2024).