PhysMotion: Physics-Grounded Dynamics From a Single Image

Xiyang Tan[1,*] Ying Jiang[1,*] Xuan Li[1,*] Zeshun Zong[1] Tianyi Xie[1] Yin Yang[2]
Chenfanfu Jiang[1]
[1]University of California, Los Angeles [2]University of Utah

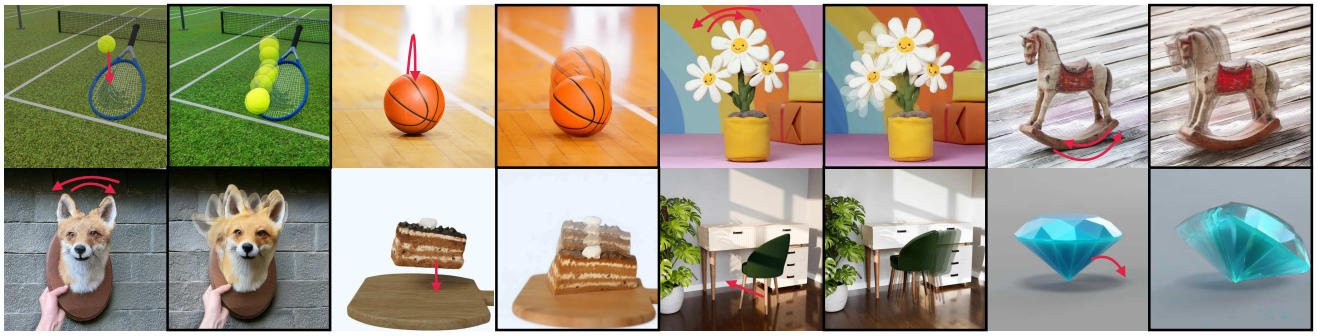

Figure 1. **PhysMotion** is a novel image-to-video framework that uses physical dynamics and 3D geometric representations to generate realistic videos, capturing rigid body motion, elastic deformation, viscoplastic flow, and fracture. Each image pair shows the input image on the left and the generated dynamics on the right.

Abstract

We introduce PhysMotion, a novel framework that leverages principled physics-based simulations to guide intermediate 3D representations generated from a single image and input conditions (e.g., applied force and torque), producing high-quality, physically plausible video generation. By utilizing continuum mechanics-based simulations as a prior knowledge, our approach addresses the limitations of traditional data-driven generative models and results in more consistent physically plausible motions. Our framework begins by reconstructing a feed-forward 3D Gaussian from a single image through geometry optimization. This representation is then time-stepped using Material Point Method (MPM) with continuum mechanics-based elastoplasticity models, which provides a strong foundation for realistic dynamics, albeit at a coarse level of detail. To enhance the geometry, appearance, and ensure spatiotemporal consistency, we refine the initial simulation using a text-to-image (T2I) diffusion model with cross-frame attention, resulting in a physically plausible video that retains intricate details comparable to the input image. We conduct comprehensive qualitative and quantitative evaluations to validate the efficacy of our method.

1. Introduction

Video generation and editing are essential in movie and game industries, where achieving high-quality content efficiently still remains challenging. Traditional workflows with commercial software—such as video recording and post-production editing—are labor-intensive and require substantial expertise. These methods often result in huge time and labor costs and restrict accessibility for novices. Recent advances in deep learning, particularly in generative models, have enabled more automated and efficient approaches to video generation. State-of-the-art image and video diffusion models can synthesize videos from text prompts or image inputs. However, the inherent sparsity of text prompts often fails to capture detailed object layouts and complex motion features, resulting in generated videos that do not fully align with user intentions [27, 85]. To support controllable video generation, additional conditions—such as motion trajectories, reference videos, poses, and camera movements—have been incorporated into diffusion models [15, 27, 29, 33, 73, 101, 109]. Despite these advancements, diffusion models still struggle to maintain physical realism, often producing outputs that violate basic physical principles [4, 14, 60].

Previous work has employed physics simulators to gen-

erate physics-based dynamics and render videos through differentiable rendering techniques [6, 31, 35, 47, 54, 92, 108]. For example, PhysGaussian [92], VR-GS [35], and PhysGen [54] utilize simulators such as MPM [34], XPBD [57], and rigid body simulators [5] to achieve realistic physics-driven video synthesis, PhysDiff [104] integrates physical constraints for human motion recovery. To further streamline physics parameter tuning, PhysDreamer [108] distills video diffusion priors via SDS to estimate the physical material properties of static 3D objects, enabling the creation of immersive, realistic interactive dynamics. PAC-NeRF [47] and GIC [6] estimate unknown geometry and physical parameters to ensure physically plausible outputs in Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS), respectively. However, these techniques often require multi-view images, videos, or 3D objects as input, limiting their generalizability. While PhysGen [54] generates dynamics from a single image, it is constrained to 2D dynamics. Our objective is to generate visually compelling, physics-grounded 3D dynamics from a single image, enhancing both visual quality and the physical plausibility of the resulting videos.

To address these challenges, we combine insights from both 3D and 2D generative models. Our goal is to generate realistic 3D dynamics, for which physical simulations are best performed on 3D geometry; therefore, we lift the image to a 3DGS representation. However, a direct application of 3DGS deformation and rendering yields low-quality videos. Inspired by Image Sculpting [100], we design a coarse-to-fine refinement stage that employs 2D image diffusion models and cross-attention mechanisms. This approach enables the final video to exhibit physically plausible motion consistent with 3D dynamics while preserving the quality and detail of the original input image. Our contributions include: (1) A novel single image-to-video framework with 3D geometry awareness and physics-grounded dynamics; (2) A video enhancement module that improves video quality, maintains spatiotemporal consistency, preserves intricate details, and aligns closely with the input image; (3) Versatile evaluation on visual coherence, physical plausibility, and generative diversity.

2. Related Work

Generative Dynamics Previous works primarily follow two paradigms to generate dynamics from images or 3D content. Data-driven methods leverage image or video generative models to create dynamic scenes from various control signals, such as text prompts [45, 93], bounding boxes [45], trajectories [48, 73, 85, 91, 110], reference videos [33, 86, 111], camera movements [85, 97], and force signals [25]. They rely on pre-trained generative models, which often lack fundamental understanding of physics [60]. Consequently, the generated motions may violate physical laws. In contrast, simulation-based methods produce generative dynamics that are physically grounded [4]. For instance, PhysGaussian [92] and PhysDreamer [108] integrate a MPM simulator with 3DGS to generate physics-based dynamics; VR-GS [35] employs Extended Position Based Dynamics (XPBD) [57] for real-time dynamics; [112] uses a spring-mass model within 3DGS to simulate elastic objects. These methods, however, require 3D contents or multi-view images as input. Recently, PhysGen [54] and PhysGen3D [11] introduced single-image rigid-body dynamics. However, their reliance on explicit mesh extraction and PBR rendering creates a disconnect between the simulation entity and the visual output. In contrast, our method leverages 3D Gaussian Splatting (3DGS) as a unified, compact representation. Inspired by [92], we integrate MPM to enable physics-based dynamics, using only a single image as input. Unlike Phy124 [51], which outputs coarse video directly from the MPM simulator, ours produces higher-quality, visually compelling dynamics through video enhancement.

Sparse-view 3D Reconstruction Recent advancements in neural rendering techniques, such as NeRF [61] and 3DGS [38], have made significant progress in 3D reconstruction. Approaches like additional depth supervision [19], depth prior loss [63, 84], CLIP-based supervision [32], pixel-aligned features [102], frequency regularizer [95], and diffusion priors [89] have been introduced to promote sparse-view reconstruction. More relatedly, to reconstruct 3DGS scenes from sparse views, methods [10, 16, 43, 64, 79, 87, 94, 113] have been developed to better constrain the optimization process. Among these, FSGS [113] and SparseGS [94] apply monocular depth estimators or diffusion models to GS under sparse-view conditions. DNGaussian [43] uses both soft and hard depth regularization to restore accurate scene geometry.

Diffusion Models [76] first introduced diffusion models to learn a reverse diffusion process for data restoration. [20] applied diffusion models to image synthesis, achieving higher visual quality in generated images than GANs. [71] proposed latent diffusion models, which first compress image data and then sample in a latent space during the denoising process. Besides image synthesis [3, 20, 23, 36, 59, 66, 71, 75, 100, 106], diffusion models have been applied to many other tasks, including 3D reconstruction and generation [21, 42, 50, 52, 55, 67, 80, 103], video generation [8, 13, 24, 26, 41, 49, 53, 54, 58, 68, 73, 85, 90, 96, 98, 107], robot policy [17, 37, 105], and scene understanding [2, 12, 56]. Our task focuses on enhancing video quality and ensuring alignment with input. Previous works utilized pre-trained diffusion models as priors, or trained video diffusion models directly, to create and edit videos [18, 26, 28, 41, 68, 99]. For example, [99] employed video diffusion models to learn space-time features for motion transfer; [41] injected temporal features into frozen

video diffusion models to edit video based on the first edited frame. However, due to the substantial computational and memory requirements, video diffusion models remain in its early stage, largely limited to short clips, or yielding low visual quality [24]. Consequently, researchers leveraged image diffusion models, adapting them for video editing tasks through spatio-temporal slices [18], reference-aware latent correction [28], interactive semantic point maps [26], or spatial-temporal attention [68]. Nonetheless, solely applying image diffusion models on video editing often causes flickering issues among frames [9]. To mitigate this, methods such as feature warping [62, 96], cross-frame attention [7, 24], and noise warping [9] have been proposed to improve video consistency.

3. Method

3.1. Preliminaries

3D Gaussian Splatting In its general format, 3D Gaussian Splatting (3DGS) [38] represents a static scene [61] with a set of anisotropic 3D Gaussian kernels $G = \{\boldsymbol{\mathcal{G}}_k : (\mathbf{x}_k, \sigma_k, \mathbf{H}_k, \mathcal{C}_k)\}_{k\in\mathcal{K}}$, where $\mathbf{x}_k$, σ_k, $\mathbf{H}_k$, and $\mathcal{C}_k$ are the centers, opacities, covariance matrices, and color representations of the Gaussians, respectively. At rendering stage, those 3D Gaussians are projected onto the 2D image space as 2D Gaussian kernels $\{\boldsymbol{\mathcal{G}}_k^{\text{proj}}\}_{k\in\mathcal{K}}$, and the color of each image pixel $\mathbf{p}$ is computed using $\mathbf{C}(\mathbf{p}) = \sum_{n\in\mathcal{N}(\mathbf{p})} \alpha_n \mathbf{c}_n \prod_{j=1}^{n-1}(1-\alpha_j)$. Here $\mathcal{N}(\mathbf{p})$ denotes the Gaussians overlapping pixel $\mathbf{p}$; α_n represents the z-depth ordered effective opacities, *i.e.*, products of the projected Gaussian weights $\boldsymbol{\mathcal{G}}_n^{\text{proj}}$ and their overall opacities σ_n; $\mathbf{c}_n$ is the decoded color for color representation $\mathcal{C}_n$. The end-to-end differentiable rendering enables efficient optimization of 3DGS parameters.

MPM Simulation for Continuum Mechanics In continuum mechanics, motion is prescribed by a continuous deformation map $\mathbf{x} = \boldsymbol{\varphi}(\mathbf{X}, t)$ between the undeformed material space $\Omega^0 \ni \mathbf{X}$ and the deformed world space $\Omega^t \ni \mathbf{x}$ at time t. The evolution of deformation map $\boldsymbol{\varphi}$ follows the conservation of mass and momentum [44]. As essentially a Lagrangian method, the Material Point Method (MPM) naturally obeys conservation of mass. The conservation of momentum can be formulated as $\rho(\mathbf{x}, t)\dot{\mathbf{v}}(\mathbf{x}, t) = \nabla \cdot \boldsymbol{\sigma}(\mathbf{x}, t) + \mathbf{f}^{\text{ext}}$, where $\mathbf{v}(\mathbf{x}, t)$ is the velocity field, $\rho(\mathbf{x}, t)$ is the density field, $\mathbf{f}^{\text{ext}}$ is the external force per unit volume. $\boldsymbol{\sigma} = \frac{1}{\det(\mathbf{F})}\frac{\partial\Psi}{\partial\mathbf{F}}\mathbf{F}\mathbf{F}^\mathsf{T}$ is the Cauchy stress associated with a energy density $\Psi(\mathbf{F})$, and $\mathbf{F}(\mathbf{X}, t) := \nabla_\mathbf{X}\boldsymbol{\varphi}(\mathbf{X}, t)$ is the local deformation gradient.

Various works have demonstrated MPM's superb ability and compatibility to endow 3DGS with physics-based dynamics [31, 51, 69, 88, 92, 108]. In MPM, the continuum is discretized by a set of particles, each representing a small material region. At each time step t_n, we track their Lagrangian quantities such as position $\mathbf{x}_p$, velocity $\mathbf{v}_p$, and deformation gradient $\mathbf{F}_p$. To advance one time step, we discretize the momentum equation by forward Euler as

$$\frac{m_i}{\Delta t}(\mathbf{v}_i^{n+1} - \mathbf{v}_i^n) = -\sum_p V_p^0 \frac{\partial\Psi}{\partial\mathbf{F}}\mathbf{F}_p^n\mathbf{F}_p^{n\mathsf{T}}\nabla w_{ip}^n + \mathbf{f}_i^{\text{ext}}. \quad (1)$$

Here i and p represent the Eulerian grid and the Lagrangian particles, respectively; w_{ip}^n is the B-spline kernel defined on i-th grid evaluated at $\mathbf{x}_p^n$; V_p^0 is the initial representing volume, and Δt is the time step size. The updated grid velocity field $\mathbf{v}_i^{n+1}$ is transferred back to particles as $\mathbf{v}_p^{n+1}$, updating the particles' positions to $\mathbf{x}_p^{n+1} = \mathbf{x}_p^n + \Delta t\mathbf{v}_p^{n+1}$. We refer to [46, 114, 115] for more discussions on plasticity handling, and [22, 34, 78] for more details on MPM.

Physics-Integrated 3D Gaussians We follow [92] to generate physics-grounded dynamics of 3DGS. Applying a first-order approximation of the deformation map, continuum mechanics is combined with 3DGS by $\mathbf{x} \approx \tilde{\boldsymbol{\varphi}}_k(\mathbf{X}, t) := \mathbf{x}_k + \mathbf{F}_k(\mathbf{X} - \mathbf{X}_k)$, and the Gaussian distribution function for each Gaussian kernel p is augmented as $\boldsymbol{\mathcal{G}}_k(\mathbf{x}) = \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^\mathsf{T}(\mathbf{F}_k\mathbf{H}_k\mathbf{F}_k^\mathsf{T})(\mathbf{x}-\mathbf{x}_k)\right\}$. This transformation gives a time-dependent 3DGS geometrical framework as $\mathbf{x}_k(t) = \boldsymbol{\varphi}(\mathbf{X}_k, t)$, $\mathbf{h}_k(t) = \mathbf{F}_k(t)\mathbf{H}_k\mathbf{F}_k(t)^\mathsf{T}$. The world-space covariance matrix $\mathbf{h}_k$ can thus be subsequently updated as $\mathbf{h}_k^{n+1} = \mathbf{h}_k^n + \Delta t\left(\nabla\mathbf{v}_k\mathbf{h}_k^n + \mathbf{h}_k^n\nabla\mathbf{v}_k^\mathsf{T}\right)$. In this work, we use only first order sphere harmonics (SH) [38]. As color is not view-dependent with order one SH, the polar decomposition to a rotation matrix is not needed [92].

3.2. Pipeline Overview

As shown in Fig. 2, given a single image, our pipeline consists of two stages: 1) we extract and optimize a 3DGS representation of the foreground object, then generate physics-grounded dynamics using an MPM simulator, where we manually apply external forces to selected regions and modify specific particle velocities, inducing controllable physics-grounded movement; 2) we apply a diffusion-based video enhancement pipeline to generate high-fidelity video with realistic object-background interaction.

3.3. Geometry-Aware Reconstruction

Starting from a single image $\mathbf{I}^0$, using SAM [39], we segment the image as a foreground $\mathbf{I}_F^0$ containing the object of interest, and a background $\mathbf{I}_B^0$. A spatially consistent 3DGS representation of the foreground object is pivotal to high-fidelity physics-based dynamics. We use LGM [81] to obtain a coarse initial 3DGS reconstruction for the foreground object. We then optimize the coarse 3DGS with depth and color supervision. Motivated by [43], we include a hard-depth loss during the optimization stage. Given $\mathbf{I}_F^0$, as done in [81], MVDream [74] is used to generate four multi-view

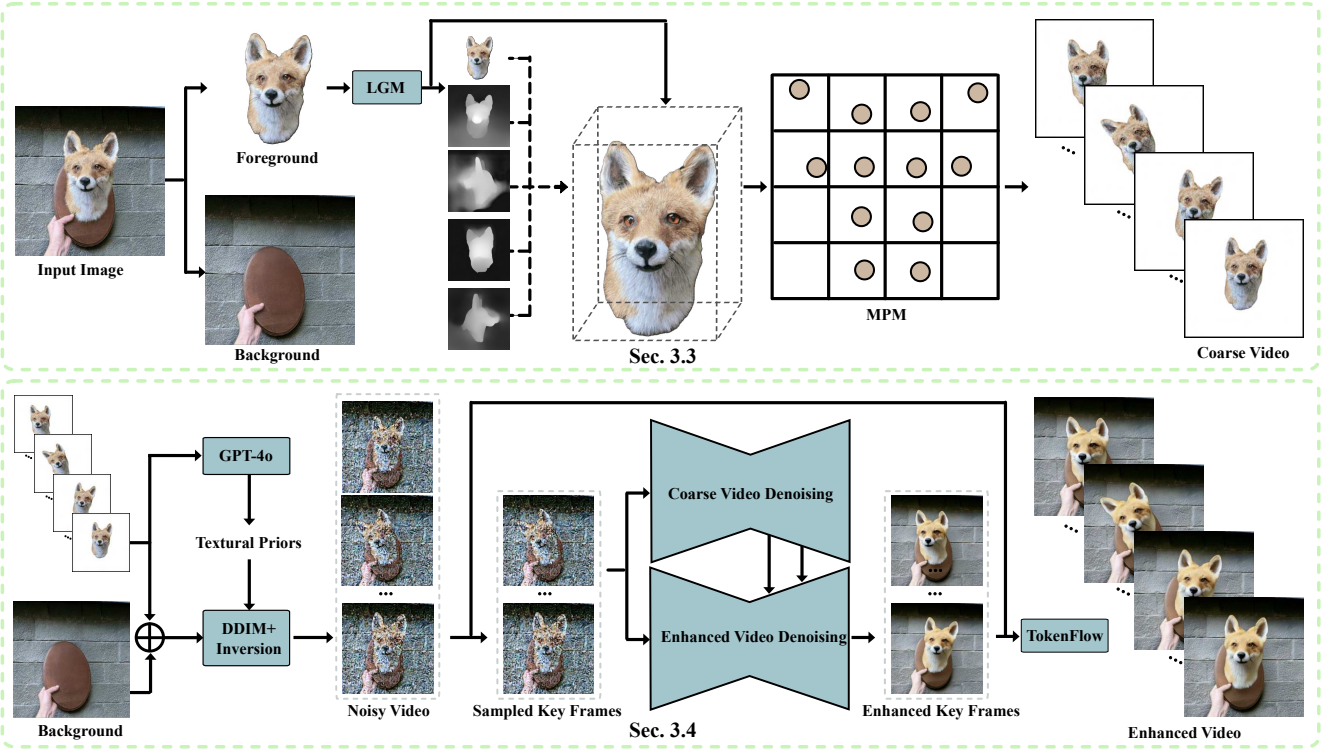

Figure 2. **Overview.** Given a single image input $\mathbf{I}^0$, we introduce a novel pipeline to generate high-fidelity, physics-grounded video with 3D understanding. Our pipeline consists of two main stages: first, we perform a single-view 3DGS reconstruction of the segmented object from the input image, then synthesize a physics-grounded coarse object dynamics $\{\mathcal{I}_j\}_{j=1}^N$ (Sec. 3.3); then, we apply a novel diffusion-based video enhancement procedure to produce the final enhanced video $\{\mathcal{I}_j^*\}_{j=1}^N$ with backgrounds (Sec. 3.4), enabling users to create visually compelling, physics-driven video from a single image with an applied conditional force or torque.

images $\{\hat{\mathbf{I}}_F^0, \hat{\mathbf{I}}_F^1, \hat{\mathbf{I}}_F^2, \hat{\mathbf{I}}_F^3\}$, where $\hat{\mathbf{I}}_F^a$ is the image generated at viewpoint with azimuth $\theta = \pi a/2$ for $a = 0, 1, 2,$ and 3. Each generated image $\hat{\mathbf{I}}_F^a$ is decomposed as a set of pixel patches $\hat{\mathbf{I}}_F^a = \bigcup_i \mathbf{P}_i^a$. The hard-depth loss $\mathcal{L}_{\text{hard}}$ for each pixel patch is formulated as

$$\mathcal{L}_{\text{hard}}(\mathbf{P}_i^a) = \left\| \mathrm{N}\left(\mathbf{D}_{\text{hard}}(\mathbf{P}_i^a) - \hat{\mathbf{D}}(\mathbf{P}_i^a)\right) \right\|_2, \quad (2)$$

where hard-depth $\mathbf{D}_{\text{hard}}(\mathbf{P}_i^a)$ is defined as

$$\mathbf{D}_{\text{hard}}(\mathbf{P}_i^a) = \sum_{\mathbf{p}\in\mathbf{P}_i^a} \sum_{n\in\mathcal{N}(\mathbf{p})} \|\mathbf{x}_n - \mathbf{o}_a\|_2 \cdot \left(\delta(1-\delta)^{p-1}\boldsymbol{\mathcal{G}}_n^{\text{proj}}(\mathbf{p})\right). \quad (3)$$

$\boldsymbol{\mathcal{G}}_n^{\text{proj}}(\mathbf{p})$ represents the projected 2D Gaussians overlapping pixel $\mathbf{p}$; $\mathbf{o}_a$ is the camera center of view a; $\delta \in (0, 1)$ is a scalar closer to 1 so that the camera only sees the nearest Gaussians. $\hat{\mathbf{D}}(\cdot)$ is the ground truth monocular depth extracted using MiDaS [77] from the training image pixel patch; $\mathrm{N}(\cdot)$ is a balanced local and global normalization operator as in [43]. The final hard-depth loss can be written as

$$\mathcal{L}_{\text{hard}} = \sum_{a=1}^{3} \sum_{\mathbf{P}_i^a \subset \hat{\mathbf{I}}_F^a} \mathcal{L}_{\text{hard}}(\mathbf{P}_i^a) + \sum_{\mathbf{P}_i^0 \subset \mathbf{I}_F^0} \mathcal{L}_{\text{hard}}(\mathbf{P}_i^0), \quad (4)$$

where we replace $\hat{\mathbf{I}}_F^0$ with the available ground truth image $\mathbf{I}_F^0$. Following [43], we freeze all parameters except $\mathbf{x}_k, k \in \mathcal{K}$ for depth guidance, ensuring the optimization of geometrical information.

For color supervision, we observe that using generated multi-view images results in poor spatial consistency, due to the inaccurate camera poses. Hence, color supervision is only applied on the input view image $\mathbf{I}_F^0$. Following [38], we write the color supervision as

$$\mathcal{L}_{\text{color}} = \mathcal{L}_1\left(\mathbf{I}(G, 0), \mathbf{I}_F^0\right) + \lambda \mathcal{L}_{\text{D-SSIM}}\left(\mathbf{I}(G, 0), \mathbf{I}_F^0\right), \quad (5)$$

where $\mathbf{I}(G, 0)$ is the rendered image at view with azimuth $\theta = 0$ from the Gaussians $G = \{\boldsymbol{\mathcal{G}}_k\}_{k\in\mathcal{K}}$. The output of the optimized Gaussians are send to a MPM simulator, resulting in a coarse object dynamics $\{\mathcal{I}_j\}_{j=1}^N$ consisting of N frames. We adopt the open-sourced MPM solver [114].

3.4. Generative Video Enhancement

We present a diffusion-based video enhancement pipeline to generate enhanced video $\{\mathcal{I}_j^*\}_{j=1}^N$ from the coarse frames $\{\mathcal{I}_j\}_{j=1}^N$, which integrates object dynamics with the background $\mathbf{I}_B^0$ for realistic interactions, and encourages generative dynamics to maintain temporal consistency. Inspired by [24], given the coarse frames $\{\mathcal{I}_j\}_{j=1}^N$, a subset of key-

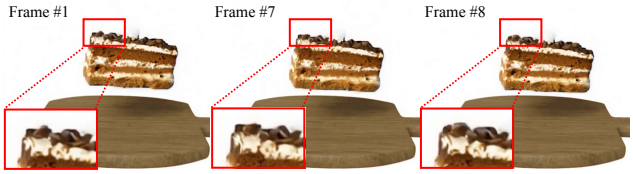

Figure 3. Applying blending in latent space causes temporal inconsistency at the cake's edge.

frames $\{\mathcal{I}_{j_k}\}_{k=1}^{K} \subset \{\mathcal{I}_j\}_{j=1}^{N}$, with $K < N$, is selected for key-frame enhancement. The key-frames $\{\mathcal{I}_{j_k}\}_{k=1}^{K}$ are first jointly enhanced with shared attention features, then the enhanced features of key-frames are propagated through the whole video to ensure temporal consistency. In our setting, the first frame is always selected as a key frame, *i.e.* $j_1 \equiv 1$, as it contains the most faithful information to the input $\mathbf{I}^0$. We introduce our video enhancement pipeline in two stages: inversion stage (Sec. 3.4.1) and sampling stage (Sec. 3.4.2).

3.4.1 Inversion Stage

We use the $\boldsymbol{\alpha}$ values from the optimized foreground 3DGS to obtain blended image $\bar{\mathcal{I}}_j = \boldsymbol{\alpha}\mathcal{I}_j + (\mathbf{1}-\boldsymbol{\alpha})\mathbf{I}_B^0, 1 \leq j \leq N$. A DDIM+ inversion [100] is applied on blended images to obtain the noisy latents $\{\tilde{\mathcal{I}}_j\}_{j=1}^{N}$ as $\{\bar{\mathcal{I}}_j\}_{j=1}^{N} \xrightarrow[\text{Inversion}]{\text{DDIM+}} \{\tilde{\mathcal{I}}_j\}_{j=1}^{N}$. They capture the intricate appearance of the object and spatial information, enhancing the realistic generative dynamics with high-quality texture and geometric details.

The DDIM+ procedure is outlined as the following. We fine-tune a pre-trained image diffusion model [72] based on the segmented object and generated multi-view images. Leveraging image diffusion priors, this process enhances the consistency of texture and appearance of the object with input image. Additionally, we utilize the depth map $\mathbf{D}(\mathcal{I})$ rendered from 3DGS and the canny edge $\mathbf{C}(\mathcal{I})$ as control signals for ControlNet [106] to maintain geometric consistency between the foreground and the background within each frame. Moreover, we query GPT-4o to retrieve detailed text descriptions of the object's appearance and background context. These descriptions are used in the DDIM+ inversion process to enrich the prompt conditioning, ensuring that intricate material textures, environmental details, and motion semantics are faithfully preserved in the inverted latents.

Unlike [100], we blend the foreground and background in image space before DDIM+ inversion rather than in latent space at sampling stage. Although blending in latent space preserves the original background and ensures output consistency with the input image, it heavily depends on accuracy of the mask. Misalignments during mask encoding can lead to boundary inconsistencies in the final video, as indicated in Fig. 3, making it unsuitable for video enhancement tasks.

3.4.2 Sampling Stage

During DDIM+ sampling stage, we perform coarse and enhanced sampling processes simultaneously. Following [82], we switch the output of residual blocks and self-attention blocks in the enhanced sampling stage with corresponding outputs from the coarse sampling stage as $\forall j, 1 \leq j \leq N$:

$$\mathbf{f}_e(\tilde{\mathcal{I}}_j) \leftarrow \mathbf{f}_c(\tilde{\mathcal{I}}_j), \quad (\mathbf{Q}_e(\tilde{\mathcal{I}}_j), \mathbf{K}_e(\tilde{\mathcal{I}}_j)) \leftarrow (\mathbf{Q}_c(\tilde{\mathcal{I}}_j), \mathbf{K}_c(\tilde{\mathcal{I}}_j)).$$

The subscripts e and c stand for the features in **e**nhanced stage and **c**oarse stage; $\mathbf{f}_*(\tilde{\mathcal{I}}_j), * = e, c$ indicates the output of residual blocks for the latent frame $\tilde{\mathcal{I}}_j$; $\mathbf{Q}_*$ and $\mathbf{K}_*$, $* = e, c$ are queries and keys within the transformer attention blocks. Feature injection is applied to all upsampling layers (*i.e.* the decoding stage) in the UNet. The timesteps for feature and attention injection is controlled by two hyperparameters, τ_f and $\tau_A \in (0, 1)$.

For latent key-frames $\{\tilde{\mathcal{I}}_{j_k}\}_{k=1}^{K}$ during enhanced sampling, the self-attention features $\{\mathbf{Q}_{j_k}\}_{k=1}^{K}$ (queries), $\{\mathbf{K}_{j_k}\}_{k=1}^{K}$ (keys), $\{\mathbf{V}_{j_k}\}_{k=1}^{K}$ (values) are concatenated and shared to form the extended attention, with the queries and keys being replaced by the corresponding values in the coarse stage (for simplicity, we denote the $\mathbf{Q}_*(\tilde{\mathcal{I}}_j)$ as $\mathbf{Q}_{j,*}$, where $* = e, c$, same as $\mathbf{K}$) as, for $k = 1, 2, ..., K$,

$$\mathcal{A}_{j_k,e} = \text{Softmax}\left(\frac{\mathbf{Q}_{j_k,c}[\mathbf{K}_{j_1,c}, \cdots, \mathbf{K}_{j_K,c}]^{\mathsf{T}}}{\sqrt{d}}\right). \quad (6)$$

Here d is the dimension of embedded vectors within $\mathbf{Q}, \mathbf{K}$ [83]. For each latent key-frame $\tilde{\mathcal{I}}_{j_k}$, its output of attention block during enhanced sampling $\boldsymbol{\phi}_e(\tilde{\mathcal{I}}_{j_k})$ is thus

$$\boldsymbol{\phi}_e(\tilde{\mathcal{I}}_{j_k}) = \mathcal{A}_{j_k,e} \cdot [\mathbf{V}_{j_1,e}, \cdots, \mathbf{V}_{j_K,e}], \ \forall 1 \leq k \leq K. \quad (7)$$

For latent non-key-frames, following [24], we propagate the enhanced key-frames to them, through the extracted Nearest-Neighbor correspondences from the coarse blended video $\{\bar{\mathcal{I}}_j\}_{j=1}^{N}$. For a spatial location $\mathbf{q}$ in the feature map of frame $j \in \{1, \cdots, N\} \backslash \{j_1, \cdots, j_K\}$, we replace its attention block output by a weighted average between neighboring key-frames

$$\boldsymbol{\phi}_e(\tilde{\mathcal{I}}_j, \mathbf{q}) = w_j \boldsymbol{\phi}_e(\tilde{\mathcal{I}}_{j+})\,[\nu_{j+}[\mathbf{q}]] + (1-w_j)\boldsymbol{\phi}_e(\tilde{\mathcal{I}}_{j-})\,[\nu_{j-}[\mathbf{q}]]\,. \quad (8)$$

Here, $j\pm$ denotes the index of the closest future (+) and past (-) key frames; $w_j \in (0, 1)$ is a scalar proportional to the distance between frame j and its neighboring key frames; $\nu_{j\pm}$ is defined over coarse stage on the original blended video:

$$\nu_{j\pm}[\mathbf{q}] = \arg\min_{\mathbf{q}^*} \mathcal{D}\left(\boldsymbol{\phi}_c(\bar{\mathcal{I}}_j)[\mathbf{q}], \boldsymbol{\phi}_c(\bar{\mathcal{I}}_{j\pm})[\mathbf{q}^*]\right), \quad (9)$$

where $\mathcal{D}$ is the cosine distance. The entire enhanced sampling will output a enhanced video:

$$\{\tilde{\mathcal{I}}_j\}_{j=1}^{N} \rightarrow \{\boldsymbol{\phi}_{j,e}\}_{j=1}^{N}; \{\mathbf{f}_{j,e}\}_{j=1}^{N} \rightarrow \{\mathcal{I}_j^*\}_{j=1}^{N}. \quad (10)$$

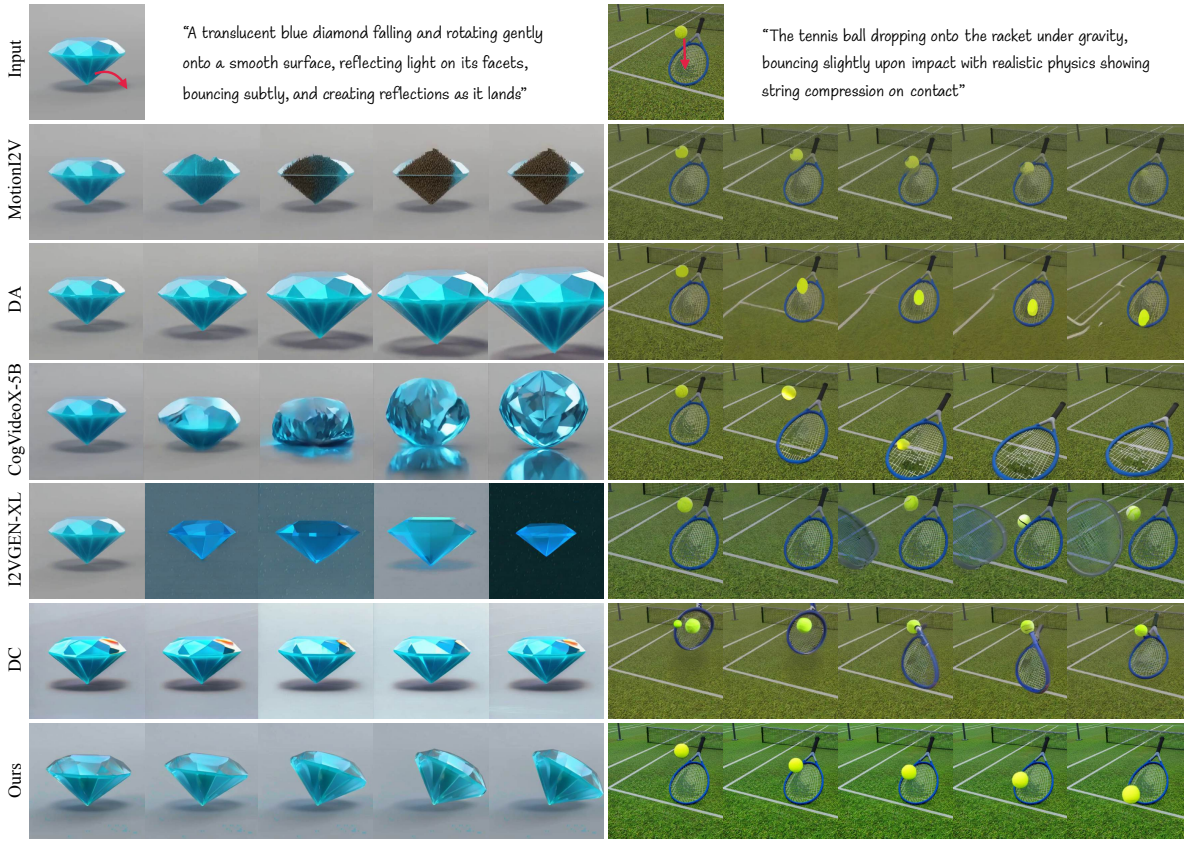

Figure 4. **Qualitative Comparison**. We use compare our results against MotionI2V [73], DragAnything (DA) [91], CogVideoX-5B [98], DynamiCrafter (DC) [93] and I2VGen-XL [73]. Text prompts for CogVideoX-5B, I2VGen-XL and DynamiCrafter are generated using ChatGPT-4o, while trajectories are used for DragAnything and Motion-I2V.

4. Experiment

4.1. Results

Implementation Details We use MiDaS [77] for monocular depth extraction for 3DGS optimization. For diffusion-based video enhancement, we leverage a pre-trained Stable-Diffusion [70] and ControlNet [106] weights, and we fine-tune them using LoRA [30] trained for $1,500$ steps with learning rate 10^{-6}. Physical material parameters (mass, Young's modulus, Poisson's ratio and density) are manually assigned based on semantic category of the object.

Showcases We showcase various 3D-generated dynamics produced by PhysMotion (Fig. 5), where *cake* and *chair* are generated from synthetic data and others are from are from real-world images. Our method creates visually compelling, temporal-consistent videos with physics-grounded dynamics. PhysMotion supports material properties such as elasticity, plasticity, rigid bodies, granular particles, and fracture effects. Additionally, the method effectively handles collisions and multi-object interactions.

Table 1. **Quantitative Comparisons.** We use VideoPhy [4] to compare the average normalized Physical Commonsense (PC) and Semantic Adherence (SA) scores, along with user study results showing preferences for our method over baselines in temporal consistency (TC) and overall preference.

Methods	**PC ↑**	**SA ↑**	**TC**	**PP**	**Overall**
I2VGen-XL	0.1853	-0.8446	89%	90%	90%
CogVideoX-5B	0.0692	0.3480	76%	76%	78%
DynamiCrafter	-0.2337	-0.3091	88%	92%	92%
MotionI2V	-0.5896	0.1530	92%	94%	91%
DragAnything	0.0545	-0.0703	92%	90%	93%
Ours	0.5142	0.7230	—	—	—

4.2. Quantitative Evaluation

Baselines and Metrics For quantitative evaluation, we compare PhysMotion with several open-sourced state-of-the-art image-to-video generation models: I2VGen-XL[107], CogVideoX-5B[98], DynamiCrafter[93], MotionI2V [73], and DragAnything [91]. The first three methods use text-based conditions, while the latter two use trajectory-based conditions to generate dynamics from a

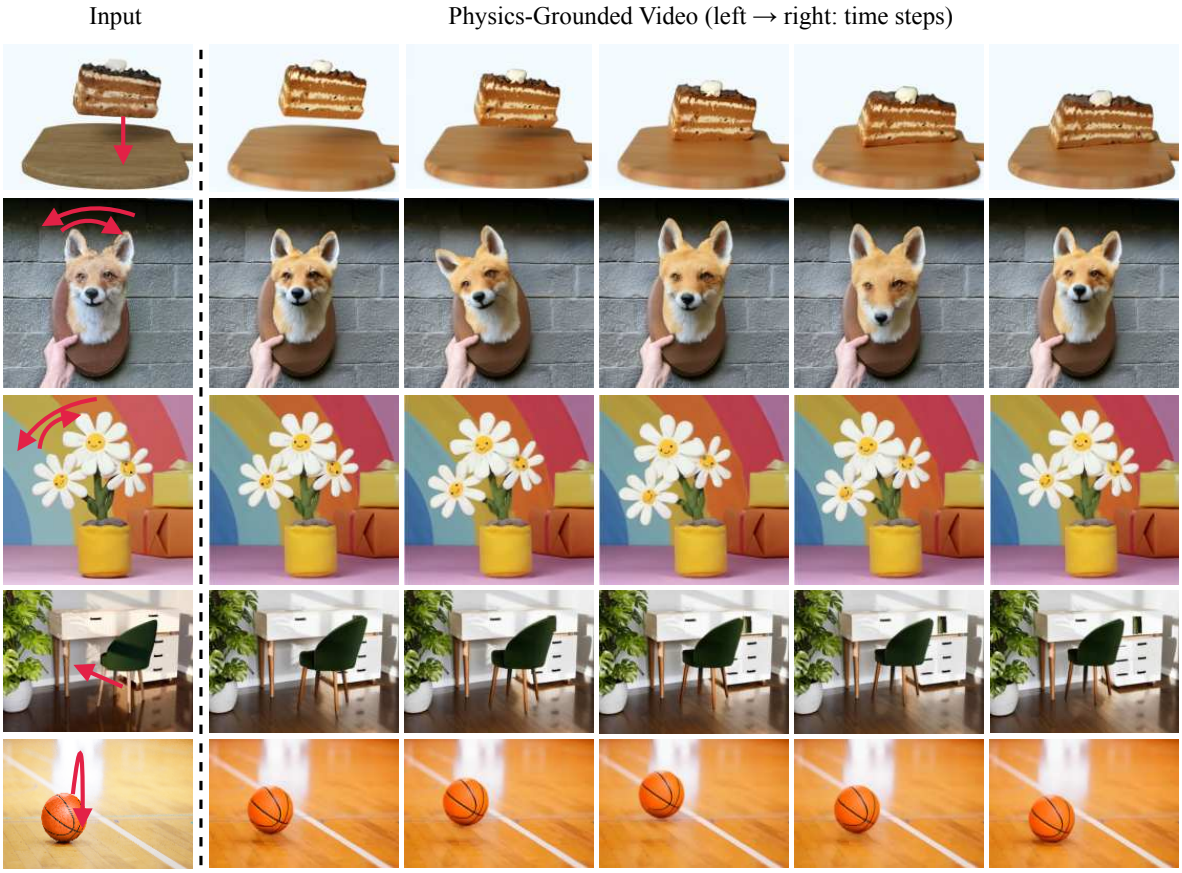

Figure 5. We demonstrate exceptional versatility of our approach across a wide variety of examples.

single image. Following [73], we build a test set that includes 13 images of different scenes covering various physical materials. We use ChatGPT-4o [1] to generate the prompts for each image on possible dynamics for video generation. For a fair comparison, input images are resized to the desired height-width ratio of each model. The final output videos are resized to 768×768, same as the output of our pipeline. We adopt three different metrics: (1) the Physical Commonsense (PC) of generated videos reported by VideoPhy [4]; (2) the Semantic Adherence (SA) between videos and text prompt reported by VideoPhy; and (3) the temporal consistency (TC), physics plausibility (PP), and user preference (UP), assessed through a user study. As VideoPhy gives scores at different scales on different scenes, we apply a standard z-score normalization and report the z-score averaged over all scenes for a meaningful comparison. The normalized z-scores for each scene have mean zero and variance of one.

For user study, we adopt a two-alternative forced choice (2AFC)) method as suggested by [24, 40, 65], where participants are asked to select which video demonstrates greater temporal consistency, visual quality, and adherence to physics in motion among two results generated from ours and a baseline respectively. The survey includes approximately $1,000$ judgments per baseline, provided by 20 users.

Comparison As indicated in Tab. 1, based on the Physics Commonsense (PC) and Semantic Adherence (SA) evaluation, PhysMotion achieved the highest z-scores in both directions, indicating superior performance across various physics scenes. These results validate the effectiveness of PhysMotion in generating accurate dynamics with excellent semantic adherence across diverse scenarios. Furthermore, our proposed method consistently outperforms all baselines in all evaluation criteria in the user preference results (in both TC, PP, and overall). These results demonstrate PhysMotion's capability to synthesize realistic physical dynamics while maintaining temporal consistency.

4.3. Qualitative Evaluation

We show the qualitative comparison results in Fig. 4. MotionI2V [73] tends to shift textures without synthesizing true geometric motion, resulting in shape inconsistencies. DragAnything [91], often misinterpreting the motion relationship between the camera and object, and moves the camera instead of the object. Image-to-video generative

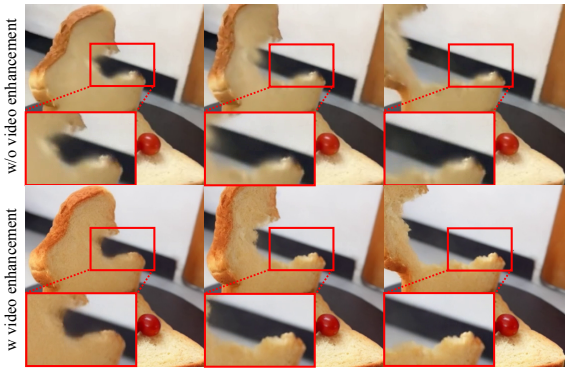

Figure 6. **Video Enhancement.** Our generative video enhancement captures the realistic texture of bread in the torn area.

methods with text conditions, such as DynamiCrafter [93], I2VGen-XL [107], and CogVideoX-5B [98], frequently produce inconsistent motion or unintended object deformations. For example, in I2VGen-XL and CogVideoX-5B, the tennis ball disappears, and in DynamiCrafter, the tennis racket deforms and incorrectly embeds the tennis ball. In contrast, our physics-guided approach ensures physically plausible dynamics and high visual quality, preserving both the geometry and texture details of the input image.

4.4. Ablation Studies

We conduct ablation studies to justify the necessity of our pipeline components. We first evaluate the effectiveness of our geometry-aware reconstruction stage. The reconstructed coarse 3DGS of LGM often appears blurry and exhibits incorrect textures. To address this, we refine the 3DGS using both depth and color supervision, leading to improved reconstruction quality. As shown in Fig. 7, while the addition of color supervision enhances texture details, it introduces floating artifacts around the object due to geometric ambiguities. The use of hard-depth loss mitigates this issue by incorporating depth information, ensuring that the refined front view aligns accurately with the input image and better preserves the object's shape. We then examine the image blending stage. As discussed in Sec. 3.4.1, applying alpha-blending in the latent space compromises temporal consistency, especially at object edges. Fig. 8 shows that inaccuracies in the latent space mask lead to the loss of high-frequency details, such as around the legs of the chair. Lastly, we evaluate the full enhancement pipeline. Prior works [92, 108] use fully reconstructed 3D scenes to generate physics-based dynamics, serving as coarse dynamics in our pipeline. However, due to challenges in filling internal 3DGS without internal information, [92] fails to show realistic dynamics for exposed internal particles, leading to artifacts and blurriness. As shown in Fig. 6, our method enhances blurred sections, e.g. in the bread-tearing video, rendering realistic internal textures.

Figure 7. Coarse 3DGS shows blur and texture errors; without hard-depth (HD) supervision, floaters appeared; with hard-depth (HD) supervision, the reconstruction displays better appearance and correct shape.

Figure 8. Background blending in latent space yields poor results due to inaccurate masking compared to blending in image space.

5. Conclusion and Future Work

We introduce PhysMotion, a novel framework that combines physics-integrated 3DGS with image diffusion models to generate physics-grounded dynamics from a single image. Compared with text-guided or trajectory-guided image-to-video methods, PhysMotion achieves more visually realistic and compelling motion synthesis, by enhancing coarse videos obtained from physics simulators through image diffusion models. Nevertheless, the introduction of image diffusion models may introduce additional artifacts, such as slight color distortions, affecting the overall color consistency. How color fidelity can be maintained in video enhancement is still an unsolved problem. Future research can also investigate the integration of video diffusion models with physics-based simulators.

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. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 7

[2] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv:2112.00390*, 2021. 2

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Ait-

tala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv:2211.01324*, 2022. 2

[4] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv:2406.03520*, 2024. 1, 2, 6, 7

[5] Victor Blomqvist. Pymunk. https://pymunk.org, 2023. 2

[6] Junhao Cai, Yuji Yang, Weihao Yuan, Yisheng He, Zilong Dong, Liefeng Bo, Hui Cheng, and Qifeng Chen. Gaussian-informed continuum for physical property identification and simulation. *arXiv:2406.14927*, 2024. 2

[7] Duygu Ceylan, Chun-Hao Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[8] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *International Conference on Computer Vision (ICCV)*, pages 23040–23050, 2023. 2

[9] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[10] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 19457–19467, 2024. 2

[11] Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. *CVPR*, 2025. 2

[12] Honghua Chen, Chen Change Loy, and Xingang Pan. Mvip-nerf: Multi-view 3d inpainting on nerf scenes via diffusion prior. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5344–5353, 2024. 2

[13] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7310–7320, 2024. 2

[14] Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. *arXiv preprint arXiv:2508.10858*, 2025. 1

[15] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv:2305.13840*, 2023. 1

[16] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv:2403.14627*, 2024. 2

[17] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 2

[18] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. *arXiv:2405.12211*, 2024. 2, 3

[19] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, 2022. 2

[20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[21] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv:2405.10314*, 2024. 2

[22] Theodore F Gast, Craig Schroeder, Alexey Stomakhin, Chenfanfu Jiang, and Joseph M Teran. Optimization integrator for large time steps. *IEEE transactions on visualization and computer graphics*, 21(10):1103–1115, 2015. 3

[23] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023. 2

[24] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv:2307.10373*, 2023. 2, 3, 4, 5, 7

[25] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *arXiv preprint arXiv:2505.19386*, 2025. 2

[26] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7621–7630, 2024. 2, 3

[27] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025. 1

[28] Sai Sree Harsha, Ambareesh Revanur, Dhwanit Agarwal, and Shradha Agrawal. Genvideo: One-shot target-image and shape aware video editing using t2i diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7559–7568, 2024. 2, 3

[29] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with

retrieval-augmented video generation. *arXiv:2307.06940*, 2023. 1

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[31] Tianyu Huang, Yihan Zeng, Hui Li, Wangmeng Zuo, and Rynson WH Lau. Dreamphysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors. *arXiv:2406.01476*, 2024. 2, 3

[32] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[33] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9212–9221, 2024. 1, 2

[34] Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *Acm siggraph 2016 courses*, pages 1–52. 2016. 2, 3

[35] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, et al. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–1, 2024. 2

[36] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023. 2

[37] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv:2402.10885*, 2024. 2

[38] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 3, 4

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[40] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10051–10060, 2019. 7

[41] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv:2403.14468*, 2024. 2

[42] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv:2311.06214*, 2023. 2

[43] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *arXiv:2403.06912*, 2024. 2, 3, 4

[44] Minchen Li, Chenfanfu Jiang, and Zhaofeng Luo. *Physics-Based Simulation*. 2024. 3

[45] Mingxiao Li, Bo Wan, Marie-Francine Moens, and Tinne Tuytelaars. Animate your motion: Turning still images into dynamic videos. *arXiv:2403.10179*, 2024. 2

[46] Xuan Li, Minchen Li, and Chenfanfu Jiang. Energetically consistent inelasticity for optimization time integration. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 3

[47] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. *arXiv:2303.05512*, 2023. 2

[48] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Computer Vision and Pattern Recognition (CVPR)*, pages 24142–24153, 2024. 2

[49] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv:2308.14749*, 2023. 2

[50] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. 2

[51] Jiajing Lin, Zhenzhong Wang, Yongjie Hou, Yuzhou Tang, and Min Jiang. Phy124: Fast physics-driven 4d content generation from a single image. *arXiv:2409.07179*, 2024. 2, 3

[52] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[53] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8599–8608, 2024. 2

[54] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2025. 2

[55] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9970–9980, 2024. 2

[56] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022. 2

[57] Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, pages 49–54, 2016. 2

[58] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7038–7048, 2024. 2

[59] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14297–14306, 2023. 2

[60] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv:2410.05363*, 2024. 1, 2

[61] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[62] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18444–18455, 2023. 3

[63] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5480–5490, 2022. 2

[64] Avinash Paliwal, Wei Ye, Jinhui Xiong, Dmytro Kotovenko, Rakesh Ranjan, Vikas Chandra, and Nima Khademi Kalantari. Coherentgs: Sparse novel view synthesis with coherent 3d gaussians. *arXiv:2403.19495*, 2, 2024. 2

[65] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020. 7

[66] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023. 2

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

[68] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 2, 3

[69] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv:2312.17142*, 2023. 3

[70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 6

[71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[72] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv:2208.12242*, 2022. 5

[73] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 6, 7

[74] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[75] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8839–8849, 2024. 2

[76] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[77] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv:2305.10853*, 2023. 4, 6

[78] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Trans. Graph.*, 32(4), 2013. 3

[79] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

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

[81] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv:2402.05054*, 2024. 3

[82] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven

image-to-image translation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 5

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 5

[84] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 2

[85] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2

[86] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6537–6549, 2024. 2

[87] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv:2403.16292*, 2024. 2

[88] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 3

[89] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21551–21561, 2024. 2

[90] Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. Aniclipart: Clipart animation with text-to-video priors. *arXiv:2404.12347*, 2024. 2

[91] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2025. 2, 6, 7

[92] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv:2311.12198*, 2023. 2, 3, 8

[93] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2, 6, 8

[94] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360 {\deg} sparse view synthesis using gaussian splatting. *arXiv:2312.00206*, 2023. 2

[95] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8254–8263, 2023. 2

[96] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2, 3

[97] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[98] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv:2408.06072*, 2024. 2, 6, 8

[99] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8466–8476, 2024. 2

[100] Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, and Saining Xie. Image sculpting: Precise object editing with 3d geometry control. *arXiv:2401.01702*, 2024. 2, 5

[101] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv:2308.08089*, 2023. 1

[102] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 2

[103] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv:2406.09394*, 2024. 2

[104] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023. 2

[105] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024. 2

[106] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 5, 6

[107] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video syn-

thesis via cascaded diffusion models. *arXiv:2311.04145*, 2023. 2, 6, 8

[108] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406. Springer, 2025. 2, 3, 8

[109] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv:2406.19680*, 2024. 1

[110] Zhenghao Zhang, Junchao Liao, Menghao Li, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv:2407.21705*, 2024. 2

[111] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv:2310.08465*, 2023. 2

[112] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, pages 407–423. Springer, 2025. 2

[113] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European Conference on Computer Vision*, pages 145–163. Springer, 2025. 2

[114] Zeshun Zong, Xuan Li, Minchen Li, Maurizio M Chiaramonte, Wojciech Matusik, Eitan Grinspun, Kevin Carlberg, Chenfanfu Jiang, and Peter Yichen Chen. Neural stress fields for reduced-order elastoplasticity and fracture. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3, 4

[115] Zeshun Zong, Chenfanfu Jiang, and Xuchen Han. A convex formulation of frictional contact for the material point method and rigid bodies. *arXiv:2403.13783*, 2024. 3