MotionPhysics: Learnable Motion Distillation for Text-Guided Simulation

Miaowei Wang, Jakub Zadrożny, Oisin Mac Aodha, Amir Vaxman

School of Informatics, The University of Edinburgh, Edinburgh EH8 9AB, United Kingdom
m.wang-123@sms.ed.ac.uk, {jakub.zadrozny, oisin.macaodha, a.vaxman}@ed.ac.uk

Abstract

Accurately simulating existing 3D objects and a wide variety of materials often demands expert knowledge and time-consuming physical parameter tuning to achieve the desired dynamic behavior. We introduce *MotionPhysics*, an end-to-end differentiable framework that infers plausible physical parameters from a user-proved natural language prompt for a chosen 3D scene of interest, removing the need for guidance from ground-truth trajectories or annotated videos. Our approach first utilizes a multimodal large language model to estimate material parameter values, which are constrained to be within plausible ranges. We further propose a learnable motion distillation loss, which extracts robust motion priors from pretrained video diffusion models while minimizing appearance and geometry inductive biases to guide the simulation. We evaluate *MotionPhysics* across more than thirty scenarios, including real-world, human-designed, and AI-generated 3D objects, spanning a wide range of materials such as elastic solids, metals, foams, sand, and both Newtonian and non-Newtonian fluids. We demonstrate that it produces visually realistic dynamic simulations guided by natural language, surpassing the state of the art, with physically plausible parameters that are automatically determined.

Introduction

The specification of plausible physical parameters is essential for realistic simulation. For instance, Young's modulus controls a material's stiffness, while yield stress marks the onset of irreversible plastic deformation. Traditional methods for identifying such parameters often rely on expert intuition or laborious trial-and-error, making simulation pipelines time-consuming and inaccessible to non-experts.

This has motivated a wide body of work focused on automatic physical parameter estimation. Early approaches relied on direct observations (Asenov et al. 2019). More recent work in novel view synthesis, such as Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and Gaussian Splatting (GS) (Kerbl et al. 2023), offers alternative strategies. GS represents scene geometry explicitly via Gaussian kernels, which enables straightforward integration with existing simulators such as PBD for fluids (Feng et al. 2025), XPBD for elastic bodies (Jiang et al. 2024), and MPM for general materials (Xie et al. 2024). This integration has revived interest in system identification, which recovers physical parameters from multi-view videos of synthetic objects (Cai et al. 2024; Li et al. 2023). When ground-truth (GT) dynamics are unavailable, recent methods guide video diffusion models with text or image prompts to infer plausible parameter values (Zhang et al. 2024). While video diffusion models offer promising zero-shot parameter estimation capabilities, their accuracy remains limited. Recent work (Kang et al. 2025; Bansal et al. 2024) shows that both open-source (Yang et al. 2025; Wang et al. 2025b) and closed-source (Pika 2024; Bar-Tal et al. 2024) video generation systems struggle to produce videos that satisfy even basic physical common sense. As a result, for many cases—including AI-generated shapes (Fig. 1, Top), human-designed models (Fig. 1, Bottom), and real-world objects under novel viewpoints or varying forces (Fig. 7)—current video diffusion methods frequently infer incorrect simulation parameters.

Building fully-fledged video diffusion models capable of generating physically plausible outputs is impractical due to their massive computational demands and the need for diverse GT motion data across a wide range of novel objects. Instead, we leverage existing pretrained video diffusion models to assist with out-of-distribution scene and object simulations. We aim to distill plausible motion cues from pre-trained models guided by high-level, user-provided language instructions, while mitigating the models' inductive shape and appearance biases. To achieve this, we introduce a novel Learnable Motion Distillation (LMD) loss that extracts pure *motion* signals from a pretrained video diffusion model to steer our differentiable simulations. Concretely, LMD minimizes appearance and geometry discrepancies between the simulation and the diffusion model's predictions by combining a lightweight, trainable motion extractor with augmented perturbations in both geometry and appearance during training. Accurate initialization of simulation parameters is critical as poor starting values waste computation and hinder convergence. We extend PhysFlow's multimodal initialization with constraint-aware prompts that embed domain-specific parameter bounds (e.g., typical Young's modulus or density ranges for metal, foam, plasticine). By forcing the LLM to select values within these limits, we leverage its internal knowledge of real-world materials. This approach both anchors simulations in physically plausi-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ble ranges and suppresses LLM hallucinations and fabrications (Farquhar et al. 2024; Walters and Wilder 2023).

Our ultimate goal is to ensure that *"what you describe is exactly what you simulate"*. We validate our framework on over 30 simulation scenarios, spanning elastic materials, plasticine, metals, foams, sands, Newtonian, and non-Newtonian fluids. Our main contributions are: (i) the introduction of a learnable motion distillation loss to isolate and leverage true motion signals, with LLM initialization triggered by plausible material range values, and (ii) a fully automatic, text-guided system that achieves state-of-the-art (SOTA) simulation performance, surpassing existing methods seamlessly on human-designed, AI-generated, and real-world objects.

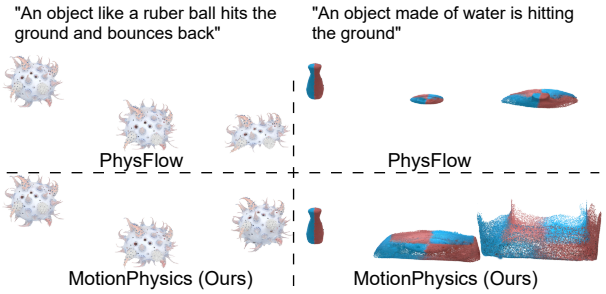

Figure 1: **MotionPhysics** automatically estimates plausible material parameters to support dynamic 3D simulation of diverse materials and object types. Compared to prior work (e.g., PhysFlow (Liu et al. 2025a)), it more accurately adheres to the user's input prompt (Top), particularly for AI-generated objects (Top: elastic simulation), human-designed objects (Bottom: water simulation), and real-world scans.

Related Work

3D Dynamics Generation. Diffusion (Liu et al. 2022) and flow-matching (Jin et al. 2024) models have enabled medium duration, high-quality video synthesis from text or image prompts, exemplified by SORA (Liu et al. 2024b) and GOKU (Chen et al. 2025). To extend these dynamic priors into 4D (3D + time), several works fuse diffusion with dual shape–texture representations, using either NeRF-style encoders or Gaussian Splatting (GS) to generate view-consistent, dynamic scenes without explicit 4D supervision (Singer et al. 2023; Yuan et al. 2024; Ling et al. 2024). While implicit methods such as D-NeRF (Pumarola et al. 2021) and HyperNeRF (Park et al. 2021) often suffer from slow rendering and limited user control, explicit splatting pipelines (e.g., 4DGS (Wu et al. 2024), HYBRID3D–4DGS (Oh et al. 2025). These splatting frameworks have been further generalized across input modalities: multi-view static reconstruction (Chen et al. 2024; Huang et al. 2024a), dynamic captures (Xu et al. 2024), single uncalibrated images (Yi et al. 2024; Smart et al. 2024), and mesh-to-Gaussian-field conversion (Waczyńska et al. 2024). Despite these advances, most methods lack physics grounding, as they do not model deformations or motion driven by forces and material properties, limiting realism.

Physics-grounded Dynamic Generation. Embedding physical laws into generative models (Zhong et al. 2024) yields more realistic interactions and using differentiable simulators enables gradient-based motion synthesis while simultaneously tuning material parameters. Simple spring–mass systems, such as SPRINGGAUS (Zhong et al. 2024) and PHYSTWIN (Jiang et al. 2025), effectively capture elastic deformation, while the differentiable Material Point Method (MPM) (Jiang et al. 2016) excels at modeling diverse material behaviors. For example, PHYSMOTION (Tan et al. 2024) leverages MPM for single-image dynamics, whereas PAC-NERF (Li et al. 2023) fuses NeRF and MPM via particles, trading off efficiency and fidelity in complex scenes. PHYS-GAUSSIAN (Xie et al. 2024) further boosts visual quality by combining 3DGS with MPM, yet still requires hand-tuned material settings. System identification can recover these settings but requires ground-truth (GT) videos (Cai et al. 2024) or markers (Ma et al. 2023), limiting scalability. Thus, automatic estimation of material properties without any GT dynamics is an open challenge for real-world applications.

Physical Parameter Estimation. To inject physical realism, recent method (Zhang et al. 2024; Liu et al. 2024a; Huang et al. 2025; Lin et al. 2025; Liu et al. 2025a) leverage video-diffusion priors (Blattmann et al. 2023; Meng et al. 2024) to infer material properties such as elasticity and plasticity. PHYSDREAMER (Zhang et al. 2024) models elastic behavior in real scenes, and PHYS3D (Liu et al. 2024a) extends this approach to plastic deformations. DREAMPHYSICS and PHYSFLOW handle a broader range of materials, while OMNIPHYSGS integrates constitutive models into each particle to support heterogeneous interactions. These methods optimize material parameters by backpropagating through differentiable simulators using either (i) direct perceptual objectives, such as image similarity (Zhang et al. 2024) or optical-flow divergence (Liu et al. 2025a) between simulated and generated frames, or (ii) score-distillation losses derived from a diffusion model (Huang et al. 2025; Liu et al. 2024a). However, these methods rely on real-world footage, where objects are often anchored, occluded, or subject to noise (Wang et al. 2025a), limiting their applicability compared to the increasingly prevalent human-designed and AI-generated assets. Human-designed objects often lack consistent textures (Fig.1, Bottom), while AI-generated meshes (Zhao et al. 2025; Tochilkin et al. 2024) can exhibit atypical geometry or appearance (Fig.1, Top), confusing appearance-driven supervision. To address these limitations, we introduce a *learnable motion distillation* loss that extracts motion cues from pretrained diffusion models guided by text prompts, while suppressing appearance and geometry biases. While some works (Jeong, Park, and Ye 2024) extract inter-frame motion for video-based transfer, our approach focuses on text-guided estimation of physical parameters for physics-based simulation. Text prompts often implicitly convey physical properties by indicating material types and object categories. We enable more effective automated estimation of these material properties by using pretrained LLMs with plausible material value ranges, a crucial aspect overlooked by prior works (Huang et al. 2025; Liu et al. 2025a; Lin et al. 2024).

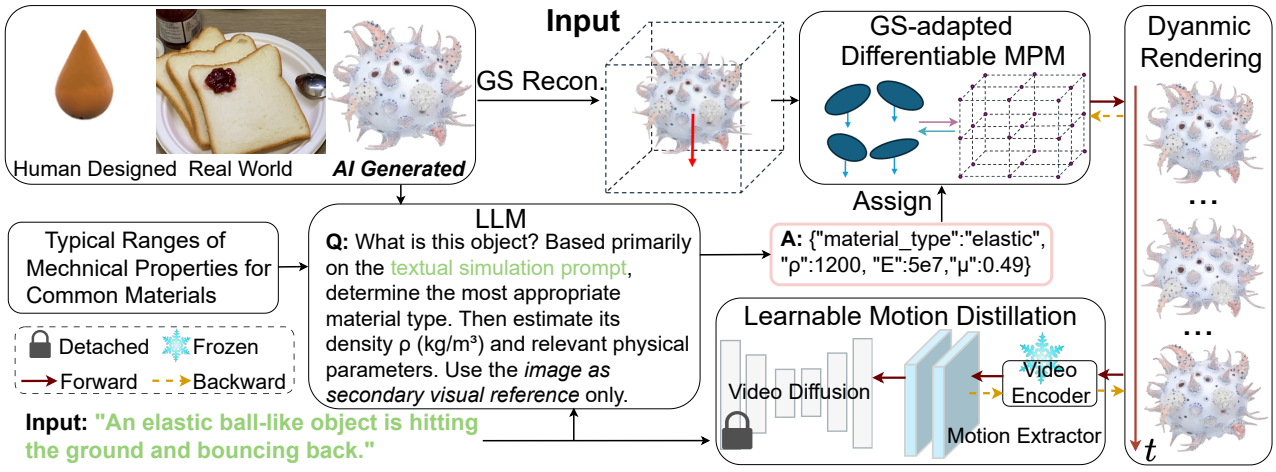

Figure 2: **Overview.** MotionPhysics simulates physically consistent dynamics from text-guided input prompts by automatically estimating physical parameters for diverse input scenes, including AI-generated, real-world, and human-designed assets.

Methodology

Problem Statement

We consider objects and scenes that are real, synthetic, human-designed (often lacking high-quality textures), or AI-generated (often with uncommon geometry or appearance). These inputs may be provided as multi-view static images, dynamic videos, single images, or meshes. Compared to traditional mesh representations, 3DGS is more suitable for reproducing real-world scenes and supports high-quality, real-time rendering. Methods such as PGSR, GIC (Cai et al. 2024), Splat3R, and GaMes can convert inputs into collections of 3D Gaussians, $\mathcal{G} := \{\mathbf{x}_g, \sigma_g, \mathbf{\Sigma}_g, \mathcal{S}_g\}$, where each splat g is defined by its center $\mathbf{x}_g \in \mathbb{R}^3$, opacity $\sigma_g \in [0,1]$, covariance $\mathbf{\Sigma}_g \in \mathbb{R}^{3\times3}$, and color coefficients $\mathcal{S}_g$.

To simulate dynamic behavior over discrete time t, we denote time-varying splats as $\mathcal{G}^t := \{\mathbf{x}_g^t, \sigma_g^t, \mathbf{\Sigma}_g^t, \mathcal{S}_g^t\}$. We adopt a differentiable GS-adapted MLS-MPM simulator (Xie et al. 2024), which evolves splat states using a Markovian update $\mathcal{T}$, consisting of particle-to-grid and grid-to-particle mappings (velocity omitted for clarity):

$$(\mathbf{x}_g^{t+1},\ \mathbf{\Sigma}_g^{t+1},\ \mathcal{S}_g^{t+1}) = \mathcal{T}(\mathbf{x}_g^t,\ \mathbf{\Sigma}_g^t,\ \mathcal{S}_g^t \mid \boldsymbol{\theta},\ \mathbf{f}^{\text{ext}}), \quad (1)$$

where $\mathbf{\Sigma}_g^t$ and $\mathcal{S}_g^t$ evolve through the deformation gradient F^{t+1}, which captures stretching, rotation, and shear (Xie et al. 2024), and $\mathbf{f}^{\text{ext}}$ denotes external forces. Following PhysFlow (Liu et al. 2025a), all splats share the same material parameters defined as $\boldsymbol{\theta} := \{\rho, c, \boldsymbol{\theta}_c\}$, where ρ is density, c is a material class (e.g., elastic, plasticine, metal, foam, sand, Newtonian or non-Newtonian fluid) corresponding to different material constitutive models, and $\boldsymbol{\theta}_c$ contains associated class-specific coefficients.

Objective: Our goal is to *automatically infer* the full material parameter set $\boldsymbol{\theta}$ from a natural language prompt $\mathcal{P}_{\text{text}}$, without supervision from ground-truth dynamics, motion capture markers, or videos. Once obtained, the parameters enable high-fidelity and diverse 3D dynamic simulations under various force fields and physical conditions.

Preliminaries

Score Distillation Sampling (SDS) (Poole et al. 2022) was initially proposed to distill 3D priors from large-scale 2D diffusion models for text-to-3D generation (Lin et al. 2023). Recent extensions (Lin et al. 2025; Huang et al. 2025; Liu et al. 2024a) adapt SDS to optimize physical parameters $\boldsymbol{\theta}_c$ using a diffusion model ϕ. Let $z_0 = \mathcal{E}_\phi(\{\mathcal{I}_l\})$ be the latent encoding of frames $\{\mathcal{I}_l\}$ via the video encoder $\mathcal{E}_\phi$. At video diffusion step k, z_0 is perturbed by noise $\epsilon \sim \mathcal{N}(0, I)$:

$$z_k = \sqrt{\bar{\alpha}_k}\, z_0 \ + \ \sqrt{1-\bar{\alpha}_k}\, \epsilon, \quad \bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i, \quad \alpha_i \in [0,1].$$

The SDS update is computed as:

$$s_{\text{SDS}}(k,\epsilon) \ = \ w_k \left(\epsilon_\phi(z_k, k \mid \mathcal{P}_{\text{text}}) \ - \ \epsilon\right), \quad (2)$$

where ϵ_ϕ is the predicted noise from ϕ and w_k is a step-dependent weight. The gradient with respect to $\boldsymbol{\theta}_c$ is:

$$\nabla_{\boldsymbol{\theta}_c} L_{\text{SDS}} \ = \ \mathbb{E}_{k,\epsilon}\Big[\, s_{\text{SDS}}(k,\epsilon)\, \frac{\partial\{\mathcal{I}_l\}}{\partial\{\mathcal{G}^t\}}\, \frac{\partial\{\mathcal{G}^t\}}{\partial \boldsymbol{\theta}_c}\Big]. \quad (3)$$

Here, $\{\mathcal{I}_l\}$ is differentiably rendered from $\{\mathcal{G}^t\}$ by 3DGS, and $\{\mathcal{G}^t\}$ depends on $\boldsymbol{\theta}_c$ via GS-adapted MLS-MPM simulation. This objective aligns simulated videos with the data distribution of the diffusion model, thereby distilling physical priors from the latter to optimize the parameters $\boldsymbol{\theta}_c$.

Method Overview

As shown in Fig. 2, we first reconstruct the static object or scene into an initial GS representation $\mathcal{G}^0$. Simultaneously, we prompt a multimodal LLM (e.g., GPT-4 (Achiam et al. 2023)) with prescribed parameter ranges defined for each possible material type, conditioning primarily on the user's text prompt $\mathcal{P}_{\text{text}}$ and secondarily on a reference rendered image, to obtain an initial material parameters estimate $\boldsymbol{\theta}^{\text{ini}}$. Using $\boldsymbol{\theta}^{\text{ini}}$ and external forces $\mathbf{f}^{\text{ext}}$, our differentiable GS-adapted MLS-MPM solver simulates splat dynamics over time, yielding a sequence $\{\mathcal{G}^t\}$. We alpha-blend

a sparse subset of these into frames $\{\mathcal{I}_l\}$. To refine $\boldsymbol{\theta}^{\text{ini}}$, we introduce a learnable motion distillation loss (Eq. 5), which extracts dynamic priors from a pretrained video diffusion model ϕ conditioned on the same $\mathcal{P}_{\text{text}}$. By iterating simulation, rendering, and gradient-based optimization, we converge to a physically plausible parameter set $\boldsymbol{\theta}_c$ that faithfully reproduces the motion specified by the user.

Initialization via Multimodal LLMs

Accurately identifying the material type c is crucial for selecting the appropriate constitutive model in MLS-MPM. Given a user prompt, e.g., *"A rubber ball-like object hits the ground and bounces back."*, we aim to infer the material type c, density ρ, and a corresponding set of material-specific parameters $\boldsymbol{\theta}_c$. If the prompt changes, the inferred material and parameters should adapt accordingly, enabling flexible, user-driven simulation.

To achieve this, we leverage GPT-4 (Achiam et al. 2023) to estimate initial material parameters $\boldsymbol{\theta}^{\text{ini}}$ primarily from text and secondarily from a reference image. However, naïvely querying GPT-4 as done by PhysFlow can lead to hallucinated (Farquhar et al. 2024) or fabricated (Walters and Wilder 2023) numerical values (see Fig. 6). LLMs inherently encode extensive real-world material knowledge, so to reduce hallucinations, we provide GPT-4 with prompt templates that contain value-range constraints grounded in standard material-property handbooks (Callister and Rethwisch 2020). This grounding steers the LLM toward realistic values, preventing implausible predictions and simulation failures in certain cases (Fig. 6). The Supplement includes the full prompt template and typical parameter ranges. For instance, Young's modulus spans 10^7–4×10^{11} Pa for elastic materials and $10^6 - 5 \times 10^6$ Pa for plasticine.

Learnable Motion Distillation

Density ρ and material class c can be reliably inferred using our LLM approach above. However, the class-specific coefficients $\boldsymbol{\theta}_c$ are coarse approximations that are insufficient for precise simulation and require additional supervision, since LLMs lack dynamic modelling and simulation capabilities. Following prior work (Liu et al. 2024a; Huang et al. 2025), we employ a video diffusion model ϕ to supervise motion optimization. However, diffusion-based predictions inherently entangle motion with appearance and geometric biases from their training data, making it challenging to extract pure motion signals, especially for human-designed or AI-generated objects whose appearance or structure falls outside the training distribution (see Fig. 5).

A key observation is that, despite changes in appearance or shape, identical simulated motions under the same initial physical parameters yield globally consistent latent codes from the video encoder $\mathcal{E}_\phi$. As Fig. 3 (Left) shows, whether the bird's color shifts from red to blue (Top) or its shape loses tail and beak (Bottom), the latents (visualized by PCA projection) share the same overall structure and vary only in local details. This holds across different diffusion models and motivates extracting motion signals directly from "clean" latents rather than noise via a learnable motion extractor M, by dynamically smoothing local latent-space disparities, i.e, the appearance and geometry gaps between the model's pre-learned distribution and the target scene.

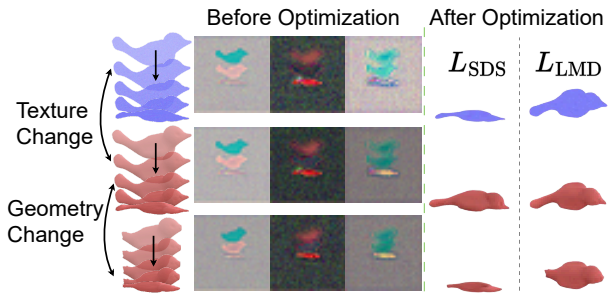

Figure 3: **Structure Similarity.** Left: the same motion pattern (with PCA-visualized latent codes in the middle) under varying textures and geometries, before optimization. Right: after applying our L_{LMD}, the dynamics become consistent and remain largely unaffected by both texture and geometry.

Concretely, to enforce motion learning, we augment the initial representation $\mathcal{G}^0$ (both kernel centers and color coefficients) with additive Gaussian noise ϵ' (see Suppl.), yielding $\tilde{\mathcal{G}}^0$. This produces augmented renderings $\{\tilde{\mathcal{I}}_l\}$ and latents $\tilde{z}_0 = \mathcal{E}_\phi(\{\tilde{\mathcal{I}}_l\})$. Then we compute one-step denoised latents $\tilde{z}_k^0 = \frac{1}{\sqrt{\bar{\alpha}_k}}\Big(\tilde{z}_k - \sqrt{1-\bar{\alpha}_k}\,\epsilon_\phi(\tilde{z}_k, k \mid \mathcal{P}_{\text{text}})\Big)$, geometrically projecting $\tilde{z}_k$ back onto the clean-latent manifold.

We denote the distilled motion targets and predictions as:

$$y_{\text{target}} = M(\tilde{z}_0), \quad y_{\text{pred}} = M(\tilde{z}_k^0), \tag{4}$$

where the learnable motion extractor M is a lightweight two-layer convolutional network initialized to the identity mapping and trained with a small learning rate of 2×10^{-5}. Our resulting learnable motion distillation (LMD) loss uses a Charbonnier variant for numerical stability:

$$L_{\text{LMD}} = w_k \sqrt{\|y_{\text{pred}} - y_{\text{target}}\|^2 + \beta^2}, \tag{5}$$

with small constant $\beta = 10^{-3}$, and M is kept synchronized via exponential moving averaging. Finally, gradients to the coefficients $\boldsymbol{\theta}_c$ are estimated:

$$\nabla_{\boldsymbol{\theta}_c} L_{\text{LMD}} = \mathbb{E}_{k,\epsilon,\epsilon'}\Big[\frac{\partial L_{\text{LMD}}}{\partial\{\tilde{\mathcal{I}}_l\}}\,\frac{\partial\{\tilde{\mathcal{I}}_l\}}{\partial\{\tilde{\mathcal{G}}^t\}}\,\frac{\partial\{\tilde{\mathcal{G}}^t\}}{\partial\boldsymbol{\theta}_c}\Big]. \tag{6}$$

As shown in Fig. 3 (Right), our L_{LMD} loss captures consistent motion patterns across appearance and geometry, while L_{SDS} (Eq. 3) leads to inconsistent results after optimization.

Implementation Details

We build our differentiable simulator on NVIDIA's WARP implementation (Macklin 2022). Following (Xie et al. 2024), we mitigate skinny artifacts via anisotropic regularization and fill solid objects' internal volumes to enhance simulation realism. To stabilize gradient propagation and improve training speed over long MPM rollouts, we leverage a frame boosting scheme (Huang et al. 2025; Zhang et al. 2024). Given $M \times L$ total frames (with $M =$

Figure 4: **Qualitative Evaluation.** Results from (a) PhysDreamer, (b) DreamPhysics, (c) OmniPhysGS, (d) PhysFlow, and (e) Ours across diverse simulation scenarios, including human-designed objects (Toothpaste, Left), real-world scenes (Jam, Middle), and AI-generated shapes (Alien, Right). Red arrows indicate applied forces, and red circles mark key differences.

8), we split them into M interleaved subsequences $V_i = \{I_i, I_{i+M}, \ldots, I_{i+M(L-1)}\}$ for $i = 1, \ldots, M$, and alternate supervision across these M groups. Each simulation spans 5 seconds, generating 150 rendered frames in total, with 256 internal substeps per frame. With frame boosting, for each subsequence, we perform $256 \times M$ intermediate updates between adjacent frames, computing gradients only at the final step. Besides, learnable motion distillation is distilled using the CogVideoX model (Yang et al. 2025) with classifier-free guidance (CFG=100), following PhysFlow. Training converges in approximately 40 iterations, with each forward–backward pass taking about 28 seconds on an NVIDIA A100 80 GB GPU. Our framework supports diverse manual specifications of boundary conditions (e.g., see Fig. 1) and force applications (see Fig. 7), enabling precise spatiotemporal control of material response.

Results and Discussion

Evaluation Settings

Datasets. We conduct experiments on three dataset types. 1) **Human Designed:** We evaluate eight PAC-NeRF models (Torus, Bird, Playdoh, Cat, Trophy, Droplet, Letter Cream, and Toothpaste), which exhibit uniform colors rather than detailed textures. We use the static 3DGS reconstructions from GIC (Cai et al. 2024). Since our focus is text-guided physical simulation using various material prompts (see Suppl.), rather than system identification, we do not use their rendered dynamic frames as ground-truth labels. Those frames rely on manually specified parameters and cannot capture the full diversity of realistic distributions.
2) **Real World:** We include four PhysDreamer scenes (Alocasia, Carnation, Hat, and Telephone) (Zhang et al. 2024) and four additional scenes: Fox from InstantNGP (Müller et al. 2022), Plane from NeRFStudio (Tancik et al. 2023), Kitchen from Mip-NeRF 360 (Barron et al. 2022), and Jam and Sandcastle from PhysFlow. We use the text prompts provided by PhysFlow for all real-world scenes.
3) **AI Generated:** We use the meshes Urchin, Alien, Gentleman, and Axe from Hunyuan3D (Zhao et al. 2025), which are further processed with GaMes (Waczyńska et al. 2024) to obtain their corresponding 3DGS representations.
Baselines. We compare with: 1) *PhysDreamer* (Zhang et al. 2024), which estimates material properties and initial velocity via an image appearance loss between simulated renderings and generated videos, supporting only elastic materials. 2) *DreamPhysics* (Huang et al. 2025), which applies Score Distillation Sampling (SDS). 3) *PhysFlow* (Liu et al. 2025a), which minimizes an optical flow loss between simulated and generated videos. 4) *OmniPhysGS* (Lin et al. 2025), which models heterogeneous objects with constitutive 3D Gaussians guided by video diffusion. We use open-source implementations, where the first three are based on NVIDIA WARP, and OmniPhysGS relies on a slower, less stable PyTorch simulator (see Tab. 2), so we shorten its video lengths for consistent frame rates. All baselines share identical simulation settings (boundary conditions, external forces, and text prompts). For material initialization, we follow PhysFlow using GPT-4 predictions, except OmniPhysGS, which directly optimizes constitutive models without specific material parameters. All initializations and optimizations, including our method, are run once to ensure fairness.
Metrics. We conduct a two-alternative forced-choice (2AFC) user study (Zhang et al. 2024), in which 79 participants compare 15 randomly selected side-by-side video pairs, ours versus a baseline under the same scene, prompt, and applied force, and answer: Q1) **Physical Realism**: *Which video demonstrates a more realistic physical response to the applied force?* and Q2) **Prompt Adherence**: *Which video demonstrates better adherence to the user prompt?* (see Suppl. for details). We separately quantify prompt adherence objectively via Overall Consistency (OC) from VBench (Huang et al. 2024b), measured with ViCLIP (Wang et al. 2023) to capture global semantic alignment between prompt and video, CLIPSIM (Wu et al. 2021) (as in OmniPhysGS), averaging the cosine similarity between CLIP (Radford et al. 2021) embeddings of the prompt and each video frame, and we assess motion realism using the Energy-Constrained Motion Score (ECMS) from PhysFlow, grounded in the energy-minimization principle.

Results

To assess generalization, we varied input forces across diverse scenes and settings, resulting in notable deformations and dynamic responses (see Suppl.). Qualitative ex-

Comparison	Physical Realism			Prompt Adherence		
	Human Designed ↑	Real World ↑	AI Generated ↑	Human Designed ↑	Real World ↑	AI Generated ↑
Ours vs. PhysDreamer	**96.77%**	77.63%	**93.80%**	**95.88%**	**86.08%**	**96.45%**
Ours vs. DreamPhysics	**80.27%**	69.77%	**94.65%**	**82.35%**	75.05%	**93.80%**
Ours vs. OmniPhysGS	**97.62%**	**92.73%**	**84.60%**	**96.48%**	**91.43%**	**80.75%**
Ours vs. PhysFlow	**90.30%**	66.13%	**84.20%**	**91.30%**	71.53%	**85.60%**

Table 1: **User Study Results.** Mean preference percentages. Values above 80% (bold) indicate strong user preference.

amples are shown in Fig. 4, and average user preferences over each baseline are reported in Tab. 1 (detailed votes in Suppl.). Our method outperforms all baselines in physical realism and prompt adherence, with preferences exceeding 50% across all datasets and over 80% on human-designed and AI-generated scenes, demonstrating strong generalization to novel geometries and textures. For example, in the Jam (Fig. 4, Middle), our method captures Newtonian viscosity with a smooth, cavity-free spread, unlike the baselines (red circle). Similarly, in Toothpaste (Fig. 4, Left), it reproduces non-Newtonian behavior, initially flowing and spreading before settling into a stable mound, while others fail.

In Tab. 2, our method achieves the highest average scores across all objective quantitative metrics while maintaining competitive optimization speed (tested on Bird scene from Fig. 5, Bottom), thanks to its lightweight motion extractor. However, these metrics do not fully align with human perception. For example, the Droplet scene (see Suppl.) scores slightly lower in CLIPSIM and ECMS than some baselines, despite producing the expected water-splashing behavior. We attribute this gap to two factors: 1) current video-text consistency and motion metrics cannot distinguish dynamic differences across identical scenes, forces, and prompts, and 2) pretrained metric models focus on static appearance and lack the ability to capture diverse materials and motion patterns specified by text (see Suppl. for detailed analysis). One remedy is to compare physical parameters as system identification. However, obtaining GT distributions of physically plausible outcomes is challenging. Similar limitations have been noted in 3D-generation tasks (Yu et al. 2025; Tang et al. 2024). Therefore, we emphasize visual comparisons and present qualitative evaluations in subsequent experiments.

Ablation Study

Impact of LMD. To demonstrate the advantage of our learnable motion distillation (LMD) objective, we compare three losses: 1) the optical flow loss L_{Flow} from PhysFlow, which extracts flow from generated videos; 2) the SDS loss L_{SDS} (Eq. 3); and 3) our LMD loss L_{LMD} (Eq. 5). In Fig. 5, because the prompts specify the same materials as in the raw simulation data (Playdoh: Top, plasticine; Bird: Bottom, elastic), we use PAC-NeRF's manually tuned outputs (Li et al. 2023) as a coarse reference for material behavior, even though it is originally developed for system identification. The models optimized with L_{LMD} align most closely with this reference, confirming its superior material-parameter optimization (see quantitative results in Tab. 3 and additional ablations of L_{LMD} in Suppl.). Note, the results on PhysFlow's project page (Liu et al. 2025b) closely resemble the manual references since they perform system identification on human-designed objects trained on the paired GT videos.

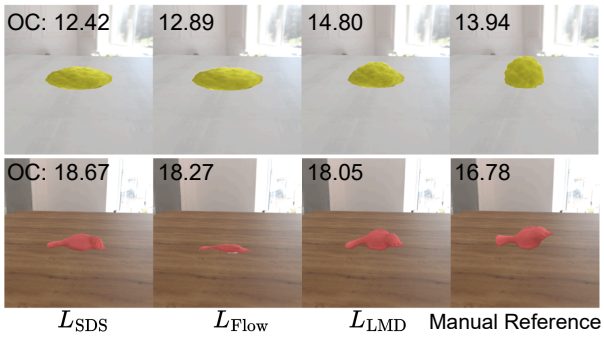

Figure 5: **Ablation of L_{LMD}.** The corresponding Overall Consistency (OC $\times 10^{-2}$ ↑) scores are also provided.

Impact of Initialization. Our method employs material-specific range prompts, in contrast to PhysFlow's naïve LLM reasoning. To validate this, in Fig. 6 we simulate the Hat scene under six conditions: PhysFlow initialization with (using LMD) and without optimization, our initialization with (using LMD) and without optimization, and median-value initialization with (using LMD) and without optimization (using the median of the constrained value ranges). PhysFlow's initialization results in exaggerated early deformations and, even after optimization, still produces unrealis-

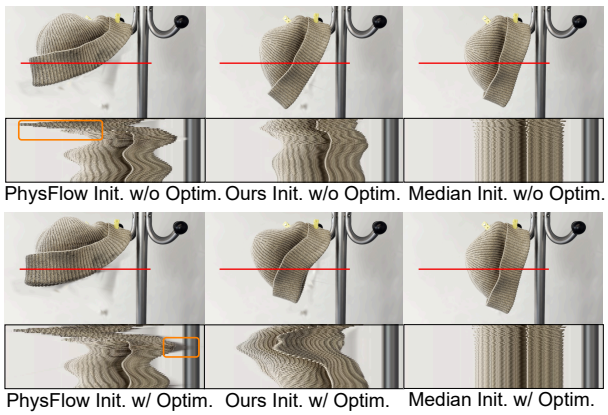

Figure 6: **Initialization Ablation.** We display oscillations with space–time slices, with time on the y-axis and the object's cross-section (red lines in the top "Object") on the x-axis, revealing both oscillations in amplitude and frequency.

Metrics	PhysDreamer	DreamPhysics	OmniPhysGS	PhysFlow	Ours
OC $\times 10^{-2}$ ↑	17.00	18.02	17.10	17.96	**18.18**
CLIPSIM $\times 10^{-2}$ ↑	21.62	21.64	21.27	21.32	**21.69**
ECMS ↓	27.48	15.76	13.70	13.07	**11.37**
Opt. Time ↓	16.48 min	18.20 min	∼ 9 h	18.14 min	18.39 min

Table 2: **Quantitative Results.** Best scores are shown in **bold**. "Opt. Time" denotes the post-initialization optimization time.

(a)	Ours Init. w/o Optim.	$L_{\text{LMD}} \rightarrow L_{\text{SDS}}$	**Ours**
18.04	18.05	18.13	**18.18**

Table 3: **Ablation Results.** Reported using OC ($\times 10^{-2}$ ↑). We denote (a): Ours Init. → PhysFlow Init.

tic artifacts, as highlighted in the orange rectangles. Median-value initialization is highly sensitive to the upper bound. In this case, a large Young's modulus of 2×10^{11} leads to overly rigid behavior with minimal deformation failing adhere to the provided force and textual description, even after optimization. In contrast, our approach provides a stable and plausible starting point (see Tab. 2). When combined with LMD, it yields material dynamics that are both accurate and visually convincing. These results demonstrate that our range prompts guide the LLM to select appropriate parameters values rather than hallucinating spurious guesses.

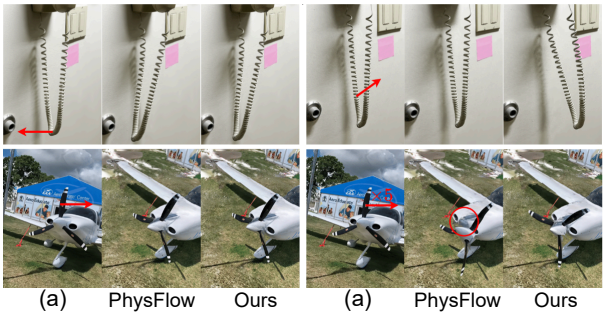

Figure 7: **Robustness to Varying Simulation Conditions (a).** (Top) Varying force directions and viewpoint direction. (Bottom) The propeller rotation speed is increased fivefold.

Robustness to Varying Simulation Conditions. We evaluate each scene under varied external forces and slight camera perturbations. In Fig. 7 (Top), the telephone is subjected to a different force direction and a shifted viewpoint. Our method still produces the correct elastic deformation guided by the prompt: *"The telephone cord is gently vibrating"*, whereas PhysFlow remains nearly static as the generated videos themselves fail to provide the correct motion that serves as the GT labels for its loss (L_{Flow}) (see Figure A6 in Supplement). In Fig. 7 (Bottom), we quintuple the plane's propeller's rotational speed. Thanks to the high Young's modulus and yield strength of metals from our ranges, our simulation preserves structural integrity and yield stress. In contrast, PhysFlow detaches the propeller (red circle) even with the same prompt: *"The plane propeller is spinning"*.

Extension to Heterogeneous Materials. Our approach extends seamlessly to multi-object scenes with different materials. By combining GS segmentation (Wang et al. 2025a) with SAM2 (Ravi et al. 2024), a multimodal LLM cab infer each object's material properties from a single text prompt. As shown in Fig. 8, the axe is treated as metal and the toy as elastic rubber. In our MPM framework, every particle carries its own material parameters, yielding high-fidelity heterogeneous dynamics. For example, the rubber toy deforms under load while the axe remains rigid. Elastic artifacts in PhysDreamer are highlighted in red (see our attached video). Note, at the time of writing, multi-object simulation code for OmniPhysGS was not available. Adapting our pipeline to complex applications like cinematics or video games involves further engineering beyond this work's scope.

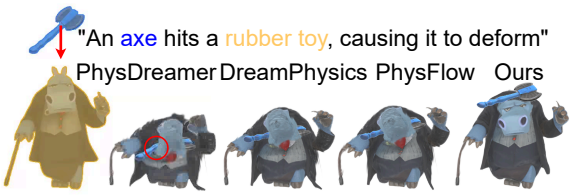

Figure 8: **Extension to Heterogeneous Materials.** Our method produces fewer visual artifacts compared to others.

Conclusion

MotionPhysics is a novel framework that uses video diffusion models and multimodal LLMs to drive 3D dynamic scene simulations guided by simple text prompts. A learnable motion distillation module extracts clean motion cues, while an LLM-based embedding initializes material-specific parameter priors, enabling high-fidelity, physically grounded animations. In future, we aim to support fully automatic configuration of fine-grained simulations from text and extend our motion distillation loss to other animation tasks, such as character rigging and deformation.

Limitations. Our method does not model shadow effects, which could improve visual realism. Additionally, while our estimated parameters enable plausible simulations, they are not intended for accurate real-world material measurement.

Acknowledgments

The authors wish to acknowledge the generous support of this research by Huawei Technologies Co. Ltd. We also thank the OpenAI Researcher Access Program for providing OpenAI Playground API credits used in this work.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Asenov, M.; Burke, M.; Angelov, D.; Davchev, T.; Subr, K.; and Ramamoorthy, S. 2019. Vid2param: Modeling of dynamics parameters from video. *Robotics and Automation Letters*.

Bansal, H.; Lin, Z.; Xie, T.; Zong, Z.; Yarom, M.; Bitton, Y.; Jiang, C.; Sun, Y.; Chang, K.-W.; and Grover, A. 2024. Videophy: Evaluating physical commonsense for video generation. *arXiv:2406.03520*.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Liu, G.; Raj, A.; et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 5470–5479.

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

Cai, J.; Yang, Y.; Yuan, W.; He, Y.; Dong, Z.; Bo, L.; Cheng, H.; and Chen, Q. 2024. Gaussian-Informed Continuum for Physical Property Identification and Simulation. *Advances in Neural Information Processing Systems*.

Callister, W. D.; and Rethwisch, D. G. 2020. *Materials Science and Engineering: An Introduction*. Wiley, 10th edition.

Chen, D.; Li, H.; Ye, W.; Wang, Y.; Xie, W.; Zhai, S.; Wang, N.; Liu, H.; Bao, H.; and Zhang, G. 2024. PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *arXiv:2406.06521*.

Chen, S.; Ge, C.; Zhang, Y.; Zhang, Y.; Zhu, F.; Yang, H.; Hao, H.; Wu, H.; Lai, Z.; Hu, Y.; et al. 2025. Goku: Flow Based Video Generative Foundation Models. *arXiv:2502.04896*.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Feng, Y.; Feng, X.; Shang, Y.; Jiang, Y.; Yu, C.; Zong, Z.; Shao, T.; Wu, H.; Zhou, K.; Jiang, C.; and Yang, Y. 2025. Gaussian Splashing: Unified Particles for Versatile Motion Synthesis and Rendering. In *Conference on Computer Vision and Pattern Recognition*.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024a. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery.

Huang, T.; Zhang, H.; Zeng, Y.; Zhang, Z.; Li, H.; Zuo, W.; and Lau, R. W. H. 2025. DreamPhysics: Learning Physics-Based 3D Dynamics with Video Diffusion Priors. In *AAAI*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024b. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Jeong, H.; Park, G. Y.; and Ye, J. C. 2024. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Conference on Computer Vision and Pattern Recognition*.

Jiang, C.; Schroeder, C.; Teran, J.; Stomakhin, A.; and Selle, A. 2016. The material point method for simulating continuum materials. In *Acm siggraph 2016 courses*, 1–52. Association for Computing Machinery.

Jiang, H.; Hsu, H.-Y.; Zhang, K.; Yu, H.-N.; Wang, S.; and Li, Y. 2025. PhysTwin: Physics-Informed Reconstruction and Simulation of Deformable Objects from Videos. *ICCV*.

Jiang, Y.; Yu, C.; Xie, T.; Li, X.; Feng, Y.; Wang, H.; Li, M.; Lau, H.; Gao, F.; Yang, Y.; et al. 2024. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 Conference Papers*.

Jin, Y.; Sun, Z.; Li, N.; Xu, K.; Jiang, H.; Zhuang, N.; Huang, Q.; Song, Y.; Mu, Y.; and Lin, Z. 2024. Pyramidal flow matching for efficient video generative modeling. *arXiv:2410.05954*.

Kang, B.; Yue, Y.; Lu, R.; Lin, Z.; Zhao, Y.; Wang, K.; Gao, H.; and Feng, J. 2025. How Far is Video Generation from World Model? A Physical Law Perspective. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR.

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

Li, X.; Qiao, Y.-L.; Chen, P. Y.; Jatavallabhula, K. M.; Lin, M.; Jiang, C.; and Gan, C. 2023. PAC-NeRF: Physics Augmented Continuum Neural Radiance Fields for Geometry-Agnostic System Identification. In *ICLR*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Conference on computer vision and pattern recognition*.

Lin, J.; Wang, Z.; Xu, D.; Jiang, S.; Gong, Y.; and Jiang, M. 2024. Phys4DGen: Physics-Compliant 4D Generation with Multi-Material Composition Perception. *arXiv preprint arXiv:2411.16800*.

Lin, Y.; Lin, C.; Xu, J.; and MU, Y. 2025. OmniPhysGS: 3D Constitutive Gaussians for General Physics-Based Dynamics Generation. In *ICLR*.

Ling, H.; Kim, S. W.; Torralba, A.; Fidler, S.; and Kreis, K. 2024. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *conference on computer vision and pattern recognition*, 8576–8588.

Liu, F.; Wang, H.; Yao, S.; Zhang, S.; Zhou, J.; and Duan, Y. 2024a. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv:2406.04338*.

Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J.; et al. 2024b. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv:2402.17177*.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *conference on computer vision and pattern recognition*, 3202–3211.

Liu, Z.; Ye, W.; Luximon, Y.; Wan, P.; and Zhang, D. 2025a. Unleashing the Potential of Multi-modal Foundation Models and Video Diffusion for 4D Dynamic Physical Scene Simulation. *CVPR*.

Liu, Z.; Ye, W.; Luximon, Y.; Wan, P.; and Zhang, D. 2025b. Unleashing the Potential of Multi-modal Foundation Models and Video Diffusion for 4D Dynamic Physical Scene Simulation. *CVPR*. Available at https://zhuomanliu.github.io/PhysFlow/.

Ma, P.; Chen, P. Y.; Deng, B.; Tenenbaum, J. B.; Du, T.; Gan, C.; and Matusik, W. 2023. Learning neural constitutive laws from motion observations for generalizable pde dynamics. In *ICML*.

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

Meng, F.; Liao, J.; Tan, X.; Shao, W.; Lu, Q.; Zhang, K.; Cheng, Y.; Li, D.; Qiao, Y.; and Luo, P. 2024. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv:2410.05363*.

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

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.

Oh, S.; Lee, Y.; Jeon, H.; and Park, E. 2025. Hybrid 3D-4D Gaussian Splatting for Fast Dynamic Scene Representation. *arXiv:2505.13215*.

Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics (TOG)*.

Pika. 2024. Pika. https://pika.art/. Accessed: 2025-05-04.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. SAM 2: Segment Anything in Images and Videos. In *The International Conference on Learning Representations*.

Singer, U.; Sheynin, S.; Polyak, A.; Ashual, O.; Makarov, I.; Kokkinos, F.; Goyal, N.; Vedaldi, A.; Parikh, D.; Johnson, J.; and Taigman, Y. 2023. Text-To-4D Dynamic Scene Generation. In *ICML*.

Smart, B.; Zheng, C.; Laina, I.; and Prisacariu, V. A. 2024. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv:2408.13912*.

Tan, X.; Jiang, Y.; Li, X.; Zong, Z.; Xie, T.; Yang, Y.; and Jiang, C. 2024. PhysMotion: Physics-Grounded Dynamics From a Single Image. *arXiv preprint arXiv:2411.17189*.

Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Kerr, J.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; Ahuja, A.; McAllister, D.; and Kanazawa, A. 2023. Nerfstudio: A Modular Framework for Neural Radiance Field Development. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*.

Tochilkin, D.; Pankratz, D.; Liu, Z.; Huang, Z.; Letts, A.; Li, Y.; Liang, D.; Laforte, C.; Jampani, V.; and Cao, Y.-P. 2024. Triposr: Fast 3d object reconstruction from a single image. *arXiv:2403.02151*.

Waczyńska, J.; Borycki, P.; Tadeja, S.; Tabor, J.; and Spurek, P. 2024. Games: Mesh-based adapting and modification of gaussian splatting. *arXiv:2402.01459*.

Walters, W. H.; and Wilder, E. I. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1): 14045.

Wang, M.; Zhang, Y.; Ma, R.; Xu, W.; Zou, C.; and Morris, D. 2025a. DecoupledGaussian: Object-Scene Decoupling for Physics-Based Interaction. In *Conference on Computer Vision and Pattern Recognition*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2025b. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; et al. 2023. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. In *The International Conference on Learning Representations*.

Wu, C.; Huang, L.; Zhang, Q.; Li, B.; Ji, L.; Yang, F.; Sapiro, G.; and Duan, N. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv:2104.14806*.

Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Conference on Computer Vision and Pattern Recognition*.

Xie, T.; Zong, Z.; Qiu, Y.; Li, X.; Feng, Y.; Yang, Y.; and Jiang, C. 2024. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Conference on Computer Vision and Pattern Recognition*, 4389–4398.

Xu, J.; Fan, Z.; Yang, J.; and Xie, J. 2024. Grid4D: 4D Decomposed Hash Encoding for High-Fidelity Dynamic Gaussian Splatting. *arXiv:2410.20815*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2025. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*.

Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024. GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. In *CVPR*.

Yu, Q.; Li, X.; Tang, Y.; Han, X.; Hu, L.; Hao, Y.; and Chen, M. 2025. Fancy123: One Image to High-Quality 3D Mesh Generation via Plug-and-Play Deformation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 595–604.

Yuan, Y.-J.; Kobbelt, L.; Liu, J.; Zhang, Y.; Wan, P.; Lai, Y.-K.; and Gao, L. 2024. 4dynamic: Text-to-4d generation with hybrid priors. *arXiv:2407.12684*.

Zhang, T.; Yu, H.-X.; Wu, R.; Feng, B. Y.; Zheng, C.; Snavely, N.; Wu, J.; and Freeman, W. T. 2024. Physdreamer: Physics-based interaction with 3d objects via video generation. In *ECCV*.

Zhao, Z.; Lai, Z.; Lin, Q.; Zhao, Y.; Liu, H.; Yang, S.; Feng, Y.; Yang, M.; Zhang, S.; Yang, X.; et al. 2025. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv:2501.12202*.

Zhong, L.; Yu, H.-X.; Wu, J.; and Li, Y. 2024. Reconstruction and Simulation of Elastic Objects with Spring-Mass 3D Gaussians. *European Conference on Computer Vision*.