PhysCtrl: Generative Physics for Controllable and Physics-Grounded Video Generation

Chen Wang[1*], **Chuhao Chen**[1*], **Yiming Huang**[1], **Zhiyang Dou**[1,2]
Yuan Liu[3], **Jiatao Gu**[1], **Lingjie Liu**[1]
[1]University of Pennsylvania, [2]HKU, [3]HKUST [*] equal contribution
{chenw30,chuhaoc,ymhuang9,zydou,jgu32,lingjie.liu}@seas.upenn.edu
yuanly@ust.hk
https://cwchenwang.github.io/physctrl

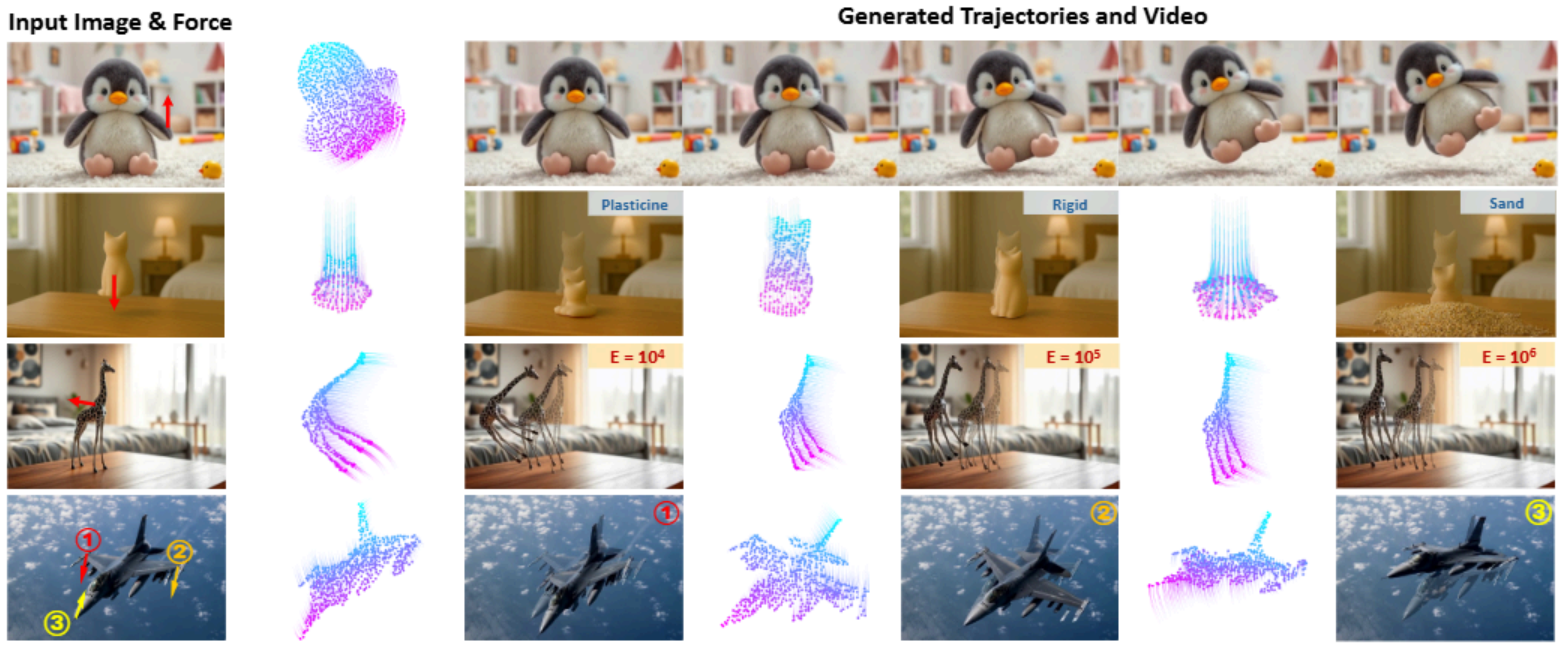

Figure 1: We propose PhysCtrl, a novel framework for physics-grounded image-to-video generation with physical material and force control. PhysCtrl supports generating physics-plausible motion trajectories across multiple materials as control signals (second row), and allows controls over physics parameters (e.g., **Young's Modulus** E of elastic material (third row)) and **force** (last row). Note that in the bottom three rows, overlaid trajectories and frames use lighter hues for earlier time steps and darker hues for later ones.

Abstract

Existing video generation models excel at producing photo-realistic videos from text or images, but often lack physical plausibility and 3D controllability. To overcome these limitations, we introduce PhysCtrl, a novel framework for physics-grounded image-to-video generation with physical parameters and force control. At its core is a generative physics network that learns the distribution of physical dynamics across four materials (elastic, sand, plasticine, and rigid) via a diffusion model conditioned on physics parameters and applied forces. We represent physical dynamics as 3D point trajectories and train on a large-scale synthetic dataset of 550K animations generated by physics simulators. We enhance the diffusion model with a novel spatiotemporal attention block that emulates particle interactions and incorporates physics-based constraints during training to enforce physical plausibility. Experiments show that PhysCtrl generates realistic, physics-grounded motion trajectories which, when used to drive image-to-video models, yield high-fidelity, controllable videos that outperform existing methods in both visual quality and physical plausibility.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

PhysCtrl: 可控与物理基础视频生成的生成式物理方法

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou Yuan Liu,
Jiatao Gu, Lingjie Liu
[1]宾夕法尼亚大学、香港大学、香港科技大学共同贡献
{chenw30,chuhaoc,ymhuang9,zydou,jgu32,lingjie.liu}@seas.upenn.edu
yuanly@ust.hk https://cwchenwang.github.io/physctrl

图 1：我们提出了 PhysCtrl，一个用于物理基础图像到视频生成的全新框架，支持物理材质和力的控制。PhysCtrl 能够生成跨越多种材质的物理合理运动轨迹作为控制信号（第二行），并允许控制物理参数（例如弹性材料的杨氏模量 E（第三行））和力（最后一行）。请注意，在底部三行中，覆盖的轨迹和帧使用较浅的色调表示较早的时间步，较深的色调表示较晚的时间步。

摘要

现有的视频生成模型擅长从文本或图像生成逼真的视频，但往往缺乏物理合理性和三维可控性。为了克服这些局限性，我们引入了 PhysCtrl，这是一个具有物理参数和力控制的物理基础图像到视频生成的新框架。其核心是一个生成物理网络，通过一个基于物理参数和施加力的扩散模型学习四种材料（弹性、沙子、黏土和刚性）的物理动力学分布。我们将物理动力学表示为三维点轨迹，并在由物理模拟器生成的 550K 动画的大规模合成数据集上进行训练。我们通过一个新颖的时空注意力模块增强了扩散模型，该模块模拟粒子相互作用，并在训练过程中结合基于物理的约束来确保物理合理性。实验表明，PhysCtrl 生成的逼真、基于物理的运动轨迹，当用于驱动图像到视频模型时，能够产生高保真、可控的视频，在视觉质量和物理合理性方面均优于现有方法。

第 39 届神经信息处理系统大会（NeurIPS 2025）。

1 Introduction

Video generation has emerged as a transformative technology, powering applications in gaming [7, 80, 14], animation [10, 92, 26], autonomous driving [84, 86], digital avatars [30, 101], robotics [49]. Modern video generative models [62, 92, 94, 4] can produce photo-realistic videos from text or single images. However, they often lack physical plausibility, controllability over dynamic physical behaviors and high fidelity, because they are trained on massive 2D videos in a pure data-driven manner [2, 3].

To achieve physics-grounded video generation, incorporating inductive biases of physical dynamics is crucial. Driven by this, recent works have combined physics simulators [37, 1, 56] with neural representations (e.g., Gaussian splats) to simulate rigid or non-rigid dynamics and render them into videos [91, 38, 51, 8, 76] under scene-specific settings. While physics simulators based on Newtonian mechanics can model the dynamics of diverse real-world systems—including soft/rigid bodies, fluids, and gases [37, 59, 56], they suffer from high computational cost, sensitivity to hyperparameters (e.g., simulation substeps, grid size), numerical instabilities, and trade-offs between generality and accuracy. As a result, when directly using a physics simulator for video generation, people have to tune several hyperparameters and might need to switch simulators with regard to object material (*e.g*, MPM for elastic and rigid body simulators for rigid). It might also lack robustness and suffer from slow speed (especially for inverse problems).

To address these issues, we propose PhysCtrl, a framework for physics-grounded image-to-video generation with explicit control over physical parameters and external forces. A key component of our framework is a generative physics network, a diffusion-based model that learns the distribution of physical dynamics. It works on various material types, requires minimal user input and supports fast forward and backward. Conditioned on physical parameters and applied forces, it predicts physical dynamics that serve as control signals for pretrained video generative models [24]. In our design, we address two fundamental questions to achieve robust, efficient, and generalizable physics priors for controllable video generation:

1. What is an appropriate representation of physical dynamics for providing control in video models? We seek a representation that enables efficient control of video models while generalizing across a wide range of materials. Very recent work on controllable video generation [21, 24] has shown that video models can synthesize rich and coherent content from only sparse and explicit point controls. Meanwhile, point clouds offer greater flexibility and generalization for modeling different materials than other explicit representations, such as meshes or voxel grids, making them more suitable for learning-based generative physics networks. Considering these two aspects, we propose to represent physical dynamics as 3D point trajectories, enabling compact motion encoding and seamless integration with video generative models while supporting diverse material types.

2. How to embed generative physics priors across various materials into a network? High-quality and diverse data are essential for learning the distribution of physical dynamics (i.e., *generative physics*). We therefore collect a large-scale synthetic dataset of 550K object animations across four material types (elastic, sand, plasticine, and rigid), capturing complex, physics-grounded dynamics via physics simulators. Using this dataset, we design a diffusion model to generate physics-plausible 3D motion trajectories conditioned on physical conditions. Inspired by particle dynamics [37], where particles interact with neighbors to determine their next state, we introduce a novel spatiotemporal attention block in the diffusion model to emulate these interactions: it first aggregates spatial influences from neighboring points and then predicts each point's trajectory over time. Finally, to embed explicit physical knowledge directly into the network, we incorporate physics-based constraints during training, ensuring that the generated motions are physics-plausible.

We conduct comprehensive evaluations of our method, demonstrating our model can produce physics-plausible motion trajectories. We further show that the generated trajectories can be used as the input for a trajectory-conditioned video model for image-to-video generation, outperforming existing video generative models in both visual fidelity and physics plausibility. Our key contributions are:

- We introduce PhysCtrl, a novel and scalable framework that represents physics dynamics as 3D point trajectories over time, enabling physics-grounded image-to-video generation with explicit control over physical parameters and external forces.

1 引言

视频生成已成为一项变革性技术，推动了游戏[7, 80 , 14]、动画[10 , 92 , 26]、自动驾驶[84 , 86]、数字化身[30 , 101]、机器人[49]等领域的应用。现代视频生成模型[62 , 92 , 94 , 4]能够从文本或单张图像生成照片级真实感的视频。然而，它们往往缺乏物理合理性，难以控制动态物理行为，且保真度不高，因为它们是以纯粹数据驱动的方式在大量二维视频上进行训练的[2, 3]。

为了实现基于物理的视频生成，结合物理动力学的归纳偏置至关重要。受此驱动，近期研究将物理模拟器[37, 1, 56]与神经表示（例如高斯 splat）相结合，模拟刚体或非刚体动力学，并在特定场景设置下将其渲染成视频[91, 38, 51, 8, 76]。虽然基于牛顿力学的物理模拟器可以模拟多种现实世界系统的动力学——包括软/刚体、流体和气体[37, 59, 56]，但它们存在计算成本高、对超参数（例如模拟子步长、网格大小）敏感、数值不稳定性以及通用性与准确性之间的权衡等问题。因此，当直接使用物理模拟器进行视频生成时，人们必须调整多个超参数，并且可能需要根据物体材质（例如，MPM 用于弹性体模拟，刚体模拟器用于刚体）切换模拟器。这也可能导致缺乏鲁棒性，并且速度较慢（尤其是对于逆问题）。

为解决这些问题，我们提出了 PhysCtrl，一个具有物理参数和外部力显式控制的物理基础图像到视频生成框架。我们框架的关键组件是一个生成物理网络，这是一个基于扩散的模型，学习物理动力学的分布。它适用于各种材料类型，需要极少的用户输入，并支持快速正向和反向。在物理参数和应用力的条件下，它预测物理动力学，这些动力学作为预训练视频生成模型的控制信号[24]。在我们的设计中，我们解决了两个基本问题，以实现可控视频生成的稳健、高效和泛化的物理先验：

1. 为视频模型提供控制，物理动力学的适当表示是什么？
我们寻求一种能够高效控制视频模型，同时在广泛材料范围内泛化的表示方法。最近关于可控视频生成的研究[21, 24]表明，视频模型可以从稀疏和显式的点控制中合成丰富且连贯的内容。同时，与网格或体素网格等其他显式表示相比，点云为建模不同材料提供了更大的灵活性和泛化能力，使其更适合用于基于学习的生成物理网络。考虑到这两个方面，我们提出将物理动力学表示为 3D 点轨迹，这能够实现紧凑的运动编码，并与视频生成模型无缝集成，同时支持多种材料类型。

2. 如何将各种材料中的生成式物理先验嵌入到网络中？高质量和多样化的数据对于学习物理动态的分布（即生成式物理）至关重要。因此，我们收集了一个包含 550K 个物体动画的大规模合成数据集，涵盖四种材料类型（弹性、沙子、油泥和刚性），通过物理模拟器捕捉复杂且基于物理的动态。利用这个数据集，我们设计了一个扩散模型，根据物理条件生成具有物理合理性的 3D 运动轨迹。受粒子动力学[37]的启发，其中粒子通过与邻居相互作用来确定其下一个状态，我们在扩散模型中引入了一个新的时空注意力模块来模拟这些相互作用：它首先从邻近点聚合空间影响，然后预测每个点随时间的轨迹。最后，为了将显式的物理知识直接嵌入到网络中，我们在训练过程中引入了基于物理的约束，确保生成的运动是physics-plausible.

我们对我们的方法进行了全面评估，证明我们的模型能够生成符合物理规律的运动轨迹。我们进一步展示，生成的轨迹可以作为轨迹条件视频模型的输入，用于图像到视频的生成，在视觉保真度和物理合理性方面均优于现有的视频生成模型。我们的主要贡献包括：

- 我们介绍了 PhysCtrl，这是一个新颖且可扩展的框架，它将物理动力学表示为随时间变化的 3D 点轨迹，实现了具有显式物理参数和外部力控制的物理基础图像到视频生成。

- We develop a diffusion-based point trajectory generative model equipped with a spatiotemporal attention mechanism and physics-based constraints, efficiently learning generative physical dynamics across four material types.
- We collect a large-scale synthetic dataset of 550K object animations, spanning elastic, sand, plasticine, and rigid materials, using physics simulators. We will release this dataset to support future research in physical dynamics learning.
- We demonstrate the effectiveness of PhysCtrl in generating realistic, physics-grounded dynamics and achieve high-quality image-to-video generation results given user-specified physics parameters and external forces.

2 Related Work

Neural Physical Dynamics Traditionally, physical dynamics are solved with numerical methods such as finite element method (FEM) [102], position-based dynamics (PBD) [60, 55], material point method (MPM) [37], smoothed-particle hydrodynamics (SPH) [17, 63, 43] and mass-spring systems [52]. Physical Informed Neural Networks (PINNs) [64] use neural networks to approximate the solution of partial differential equations and incorporate physics constraints in the loss functions. Combined with neural fields [58], PINNs achieve success in domains like fluids [13, 85] but are limited in per-scene optimization setting. Concurrent work, ElastoGen [19], replaces part of the physics simulation with neural networks for faster inference, but relies on a voxel representation, supports only elastic materials, and requires a full 3D model as input. Graph Neural Networks (GNNs) have emerged as an effective tool for modeling particle interactions with diverse material types [69, 93, 70, 95]. However, such approaches typically rely on next-step predictions for modeling dynamics, making them susceptible to drift and error accumulation over time. In contrast, our method represents objects as flexible point clouds and leverages a spatio-temporal trajectory diffusion model to robustly capture the dynamics of diverse materials in a unified framework.

Controllable Video Generative Models Video generative models are trained on massive text-video paired datasets and achieve high-quality video generation [29, 4, 41, 11, 94]. Existing works have shown that additional control signals can be injected into pretrained models for controllable video generation, such as camera movement [25, 20], human pose [30], and point movement [21, 24, 5]. However, these models lack an understanding of physical laws and thus generate outputs that are often not physically plausible. Furthermore, they cannot support explicit physics control. Our work focuses on generating physics-grounded dynamics that can be used as a physics control signal for video models.

Physics-Grounded Video Generation Existing methods leverage physics simulators to produce physics-grounded videos. One approach reconstructs neural representations from multi-view images, applies simulation on these representations, and then renders the results into video. For example, PhysGaussian [91], Spring-Gaus [99], and Vid2Sim [9] integrate MPM, spring–mass systems, and LBS-based simulation [59] into 3D Gaussians for simulation and rendering. VR-GS [38] applies physics-aware Gaussian Splatting in VR/MR for real-time, intuitive 3D interaction and physics-based editing. PhysDreamer [97] distills motions from video models to estimate physics parameters. These methods are scene-specific and require high-quality 3D reconstruction to achieve good results. Recently, researchers started to combine physics simulators with video generative models. PhysGen [51], PhysGen [8] and PhysMotion [76] generate videos of 2D rigid body dynamics or deformable dynamics. These methods rely on physics simulators to generate dynamics and coarse texture and only use video models for texture refinement. PhysAnimator [90] combines physical simulators and a sketch-guided video diffusion model for animations. Compared with methods that rely on physics simulators, our method embeds physics priors into a diffusion model, which avoids manual hyperparameter tuning and improves numerical stability for dynamics prediction. The predicted dynamics can be used as guidance for video generative models to synthesize physics-grounded and controllable videos. Concurrent works WonderPlay [48] and Force Prompting [22] also investigate using force as the condition signal for video generation.

4D Dynamics Parametric models have been widely used to represent category-specific deformable shapes, such as SMPL and SMAL [54, 103] for human and animal bodies, FLAME [47] for faces, MANO [67] for hands. Recent advances in 4D dynamics have been exploring to capture object dynamics of arbitrary topologies [61, 57, 77, 45, 77, 12] with Neural-ODE and coordinate-MLPs.

- 我们开发了一种基于扩散的点轨迹生成模型，配备了时空注意力机制和基于物理的约束，有效地学习了四种材料类型的生成物理动力学。

- 我们使用物理模拟器收集了一个包含 550K 物体动画的大规模合成数据集，涵盖了弹性、沙子、黏土和刚性材料，我们将发布这个数据集以支持未来物理动力学学习的研究。

- 我们展示了 PhysCtrl 在生成逼真、基于物理的动力学方面的有效性，并在给定用户指定的物理参数和外部力的情况下，实现了高质量的图像到视频生成结果。

2 相关工作

神经物理动力学 传统上，物理动力学通过有限元方法（FEM）[102]、基于位置的动力学（PBD）[60, 55]、材料点法（MPM）[37]、平滑粒子流体动力学（SPH）[17, 63, 43] 和质点弹簧系统 [52] 等数值方法求解。物理信息神经网络（PINNs）[64] 利用神经网络逼近偏微分方程的解，并在损失函数中融入物理约束。结合神经场 [58]，PINNs 在流体等领域 [13, 85] 取得成功，但在单场景优化设置中存在局限。同时进行的 ElastoGen [19] 研究，用神经网络替代部分物理模拟以实现更快推理，但依赖于体素表示，仅支持弹性材料，并需要完整的三维模型作为输入。图神经网络（GNNs）已成为模拟不同材料类型粒子交互的有效工具 [69, 93, 70, 95]。然而，这类方法通常依赖下一步预测来建模动力学，使其容易随时间产生漂移和误差累积。 相比之下，我们的方法将对象表示为灵活的点云，并利用时空轨迹扩散模型在一个统一的框架中稳健地捕捉不同材料的动态。

可控视频生成模型可控视频生成模型在大量的文本-视频配对数据集上进行训练，实现高质量的视频生成[29, 4 , 41, 11 , 94]。现有工作表明，可以通过预训练模型注入额外的控制信号来实现可控视频生成，例如摄像机运动[25 , 20]、人体姿态[30]和点运动[21 , 24 , 5]。然而，这些模型缺乏对物理定律的理解，因此生成的输出往往不具备物理合理性。此外，它们无法支持显式的物理控制。我们的工作专注于生成基于物理的动态，这些动态可以用作视频模型的物理控制信号。

基于物理的视频生成 现有方法利用物理模拟器来生成基于物理的视频。一种方法是从多视角图像中重建神经表示，对这些表示进行模拟，然后将结果渲染成视频。例如，PhysGaussian [91]、Spring-Gaus [99]和 Vid2Sim [9]将 MPM、弹簧-质量系统和基于 LBS 的模拟[59]集成到 3D 高斯中进行模拟和渲染。VR-GS [38]在 VR/MR 中应用物理感知的高斯喷溅，以实现实时、直观的 3D 交互和基于物理的编辑。PhysDreamer [97]从视频模型中提取运动来估计物理参数。这些方法是场景特定的，并且需要高质量的 3D 重建才能获得良好的结果。最近，研究人员开始将物理模拟器与视频生成模型相结合。PhysGen [51]、PhysGen [8]和 PhysMotion [76]生成 2D 刚体动力学或可变形动力学的视频。这些方法依赖于物理模拟器来生成动力学和粗纹理，并且仅使用视频模型进行纹理细化。PhysAnimator [90] 结合了物理模拟器和基于草图的视频扩散模型用于动画制作。与依赖物理模拟器的方法相比，我们的方法将物理先验嵌入到扩散模型中，避免了手动超参数调整，并提高了动力学预测的数值稳定性。预测的动力学可用于作为视频生成模型的指导，以合成物理基础可控的视频。同时工作的 WonderPlay [48] 和 Force Prompting [22] 也研究了使用力作为视频生成的条件信号。

4D 动力学参数模型已被广泛用于表示特定类别的可变形形状，例如用于人体和动物身体的 SMPL 和 SMAL [54 , 103]，用于面部的 FLAME [47]，用于手的 MANO [67]。4D 动力学的最新进展正在探索使用 Neural-ODE 和坐标-MLPs 捕获任意拓扑结构的物体动力学 [61, 57, 77, 45, 77, 12]。

With the success of diffusion models [28, 72, 73, 74] on high-quality generation on several modalities, including text [23], image [66, 68], audio [42, 44, 33], video [29, 27] and 3D [50, 71, 96, 53], researchers have started to learn the distribution of object dynamics with diffusion models [18, 6, 98, 88]. Motion2VecSets [6] introduced a 4D representation with latent vector sets, and trained a conditional diffusion model for dynamic reconstruction from sparse point cloud sequences. DNF [98] leverages a dictionary-based neural field to learn a compact motion space for unconditional 4D generation. However, these methods are only trained on datasets with a limited number of shapes that contain only human and animal motions, while our method focuses on learning physics-grounded dynamics, which contain a large variety of dynamic phenomena. We also use a more flexible point representation that is better suited for downstream tasks.

3 Preliminary

We generate ground-truth point trajectories for training our generative physics network (also referred to as "physics-grounded trajectory generative model") on data synthesized by physics simulators, including MPM and rigid body simulators. Here we review the basics of MPM, which form the basis for our physics-aware constraint in Section 4.

Material Point Method Material Point Method (MPM) [75, 65, 40, 37, 35, 31, 91] simulates the deformation of discrete material particles under the assumption of continuum mechanics, where the transformation of each particle from the material space to the world space is defined by a deformation mapping $\mathbf{x} = \phi(\mathbf{X}, t)$, and the associated deformation gradient $\mathbf{F} = \nabla_{\mathbf{X}}\phi(\mathbf{X}, t)$ measures the local deformation of the material such as rotation and stretch. The evolution of ϕ at time t is governed by the conservation of mass and momentum, which can be formulated as

$$\rho\frac{D\mathbf{v}}{Dt} = \nabla\cdot\boldsymbol{\sigma} + \mathbf{f}_{ext} \qquad \frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{v} = 0 \tag{1}$$

where ρ, $\mathbf{v}$ and $\mathbf{f}_{ext}$ denote the density, the velocity field and the per-unit volume external force respectively. The Cauchy stress $\boldsymbol{\sigma} = \frac{1}{\det(\mathbf{F})}\frac{\partial\Psi}{\partial\mathbf{F}}(\mathbf{F})\mathbf{F}^\top$ and the energy density function $\Psi(\mathbf{F})$ are derived from the deformation gradient $\mathbf{F}$ and physics parameters (e.g. Young's modulus E and Poisson's ratio ν) related to specific constitutive models. Based on Equation (1), MPM associates particles with background grids in the simulation, performing a particle-to-grid (P2G) and grid-to-particle (G2P) transfer loop. For stepping t to $t+1$, the P2G transfer can be formulated as

$$\frac{m_i}{\Delta t}(\mathbf{v}_i^{t+1} - \mathbf{v}_i^t) = -\sum_p V_p^0\frac{\partial\Psi}{\partial\mathbf{F}}(\mathbf{F}_p^t){\mathbf{F}_p^t}^\top\nabla N_i(\mathbf{x}_p^t) \tag{2}$$

where p and i represent attributes for particle and grid. V_p^0 is the initial particle volume and $N_i(\mathbf{x}_p^t)$ is the B-spline kernel defined on i-th grid evaluated at $\mathbf{x}_p^t$. Grid mass $m_i^t = \sum_p N_i(\mathbf{x}_p^t)m_p$ and grid momentum $m_i^t\mathbf{v}_i^t = \sum_p N_i(\mathbf{x}_p^t)m_p(\mathbf{v}_p^t + \mathbf{C}_p^t(\mathbf{x}_i - \mathbf{x}_p^t))$ are obtained according to the standard APIC [36], where $\mathbf{C}_p^t$ is the affine matrix. The G2P transfer can be formulated as:

$$\mathbf{C}_p^{t+1} = \frac{4}{(\Delta x)^2}\sum_i N_i(\mathbf{x}_p^t)\mathbf{v}_i^{t+1}(\mathbf{x}_i - \mathbf{x}_p^t)^\top \quad \mathbf{F}_p^{t+1} = (\mathbf{I} + \Delta t\sum_i \mathbf{v}_i^{t+1}\nabla N_i(\mathbf{x}_p^t)^\top)\mathbf{F}_p^t \tag{3}$$

Afterwards, $\mathbf{v}_p$ and $\mathbf{x}_p$ are updated as $\mathbf{v}_p^{t+1} = \sum_i N_i(\mathbf{x}_p^t)\mathbf{v}_i^{t+1}$ and $\mathbf{x}_p^{t+1} = \mathbf{x}_p^t + \Delta t\mathbf{v}_p^{t+1}$.

4 Method

Given a monocular image, our method generates physics-grounded videos with the control signals of physics parameters and external forces. The core part of our method is a conditional diffusion model to generate physics-grounded point cloud trajectories (Section 4.1) with physics parameters and external forces as conditioning. To enable that, as illustrated in Figure 2, we first lift the input image into 3D points (Section 4.2). Once we obtain the generated trajectories, we leverage them as the condition to pre-trained video models for image-to-video synthesis (Section 4.2).

4.1 Physics-Grounded Generative Dynamics

Our goal is to learn the distribution of physical dynamics across various materials — termed *generative dynamics* — using a diffusion-based model, thereby avoiding the high cost, hyperparameter sensitivity,

随着扩散模型[28, 72, 73, 74]在多种模态（包括文本[23]、图像[66, 68]、音频[42, 44, 33]、视频[29, 27]和 3D[50, 71, 96, 53]）的高质量生成上取得成功，研究人员开始利用扩散模型学习物体动态的分布[18, 6, 98, 88]。Motion2VecSets[6]引入了具有潜在向量集的 4D 表示，并训练了一个条件扩散模型用于从稀疏点云序列进行动态重建。DNF[98]利用基于字典的神经场学习紧凑的运动空间，用于无条件的 4D 生成。然而，这些方法仅在包含仅有人类和动物运动的有限形状数据集上进行训练，而我们的方法专注于学习物理基础的动态，其中包含各种动态现象。我们还使用了一种更灵活的点表示，更适合下游任务。

3 初步

我们为训练生成物理网络（也称为“物理基础轨迹生成模型”）生成真实点轨迹，这些数据由物理模拟器合成，包括多粒子方法（MPM）和刚体模拟器。在这里，我们回顾 MPM 的基础知识，这些知识构成了我们第 4 节中物理感知约束的基础。

多粒子方法多粒子方法（MPM）[75, 65, 40, 37, 35, 31, 91] 在连续介质力学的假设下模拟离散材料粒子的变形，其中每个粒子从材料空间到世界空间的变换由变形映射 $x = \phi(X, t)$定义，相关的变形梯度 $F = \nabla\phi(X, t)$测量材料的局部变形，如旋转和拉伸。在时间 t 时，ϕ的演化受质量和动量守恒的支配，这可以表示为

$$\rho\frac{Dv}{Dt} = \nabla \cdot \sigma + f \qquad \frac{D\rho}{Dt} + \rho\nabla \cdot v = 0 \tag{1}$$

其中 ρ、v 和 f 分别表示密度、速度场和单位体积外部力。Cauchy 应力 $\sigma = \det(\partial(F)F)$ 和能量密度函数 Ψ(F) 由变形梯度 F 和与特定本构模型相关的物理参数（例如杨氏模量 E 和泊松比 ν）推导而来。基于公式（1），MPM 在模拟中将粒子与背景网格关联，执行粒子到网格（P2G）和网格到粒子（G2P）的传输循环。为了将时间步长从 t 跃迁到 t + 1，P2G 传输可以表示为

$$\frac{m}{\Delta t}(v_i - v) = -\sum_p V \frac{\partial\Psi}{\partial F}(F)F\nabla N(x) \tag{2}$$

其中 p 和 i 分别代表粒子和网格的属性。V 是初始粒子体积，N(x)是在第 i 个网格上定义并在 x 处评估的 B 样条核。网格质量 $m = \sum_p N(x)m$and 网格动量 $v = C(x - x))$ 是根据标准 APIC [36] 获得的，其中 C 是仿射矩阵。

G2P 转换可以表示为：

$$C_p = \frac{4}{(\Delta x)}\sum_i N(x)v_i(x - x)F_p = (I + \Delta t\sum_i v_i \nabla N(x))F \tag{3}$$

随后，v 和 x 被更新为 $v_p = \sum_i N(x)v_i$ 和 $x_p = x + \Delta t v_p$。

4 方法

给定单目图像，我们的方法通过物理参数控制信号和外部力的控制生成基于物理的视频。我们方法的核心部分是一个条件扩散模型，用于生成基于物理的点云轨迹（第 4.1 节），其中物理参数和外部力作为条件。为了实现这一点，如图 2 所示，我们首先将输入图像提升为 3D 点（第 4.2 节）。一旦我们获得生成的轨迹，我们就利用它们作为条件来预训练视频模型进行图像到视频的合成（第 4.2 节）。

4.1 基于物理的生成动态

我们的目标是使用基于扩散的模型学习不同材料中物理动态的分布——称为生成动态——从而避免高成本、超参数敏感性

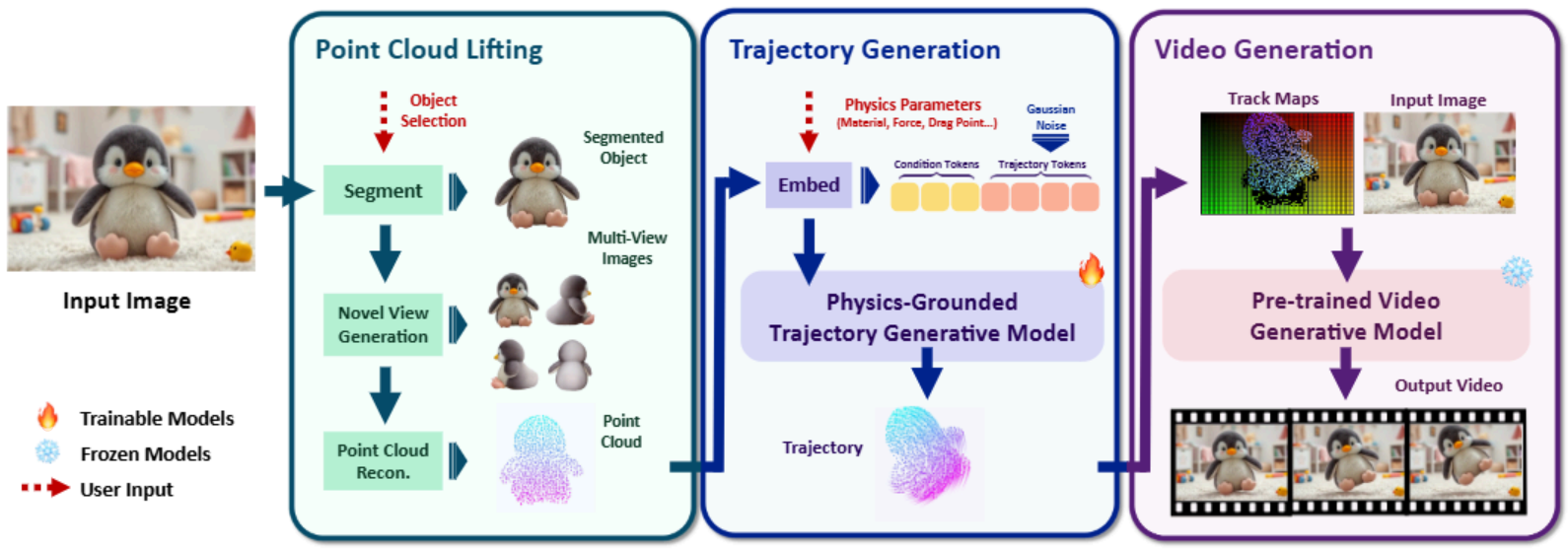

Figure 2: **An overview of PhysCtrl**. Given a single image, we first lift the object in that image into 3D points. We then generate physics-grounded motion trajectories conditioned on physics parameters and external force with a diffusion model, which are then used as strong physics-grounded guidance for image-to-video generation.

numerical instabilities, and generality–accuracy trade-offs of classical simulators. We select point clouds as our representation because they flexibly model diverse materials and suffice to control pretrained video models. Specifically, each object is represented by 2048 points in practice; we predict their trajectories over time and use them as control signals for video synthesis. We use 2048 points for guiding video model because prior work [24] show that it can achieve similar results with more points. Also, works on 4D reconstruction and generation [34, 46, 97] demonstrated that real-world motion can be represented with a sparse number of basis or control points.

4.1.1 Problem Setting

Given an object, represented as a 3D point cloud with N points $\mathbf{P}_0 = \{\mathbf{x}_i^0 \in \mathbb{R}^3\}_{i=1}^N$, and its physics parameters $\{E, \nu\}$, our trajectory generative model generates its dynamics given an initial force. Specifically, the dynamics of the object is represented by the position of each point in future F timesteps $\mathcal{P} = \mathcal{P}^{1:F} = \{\mathbf{P}^f\}_{f=1}^F = \{\{\mathbf{x}_p^f\}_{p=1}^N\}_{f=1}^F$. Denote the force, drag point and boundary condition (floor height) as $\mathbf{f} \in \mathbb{R}^3$, $\mathbf{D} \in \mathbb{R}^3$, and $h \in \mathbb{R}^1$. Thus, the goal of PhysCtrl is to predict $\mathcal{P}$ under the condition $c = \{\mathbf{P}_0, \mathbf{f}, \mathbf{D}, \{E, \nu\}, h, [\text{mat}]\}$. Here, we use an additional [mat] token to denote different materials. In this paper, we cover four different materials: elastic, plasticine, sand, and rigid. Notably, because of our flexible point cloud representation, the model is not limited to these four categories and can be readily extended to other materials, such as fluids, given sufficient computational resources.

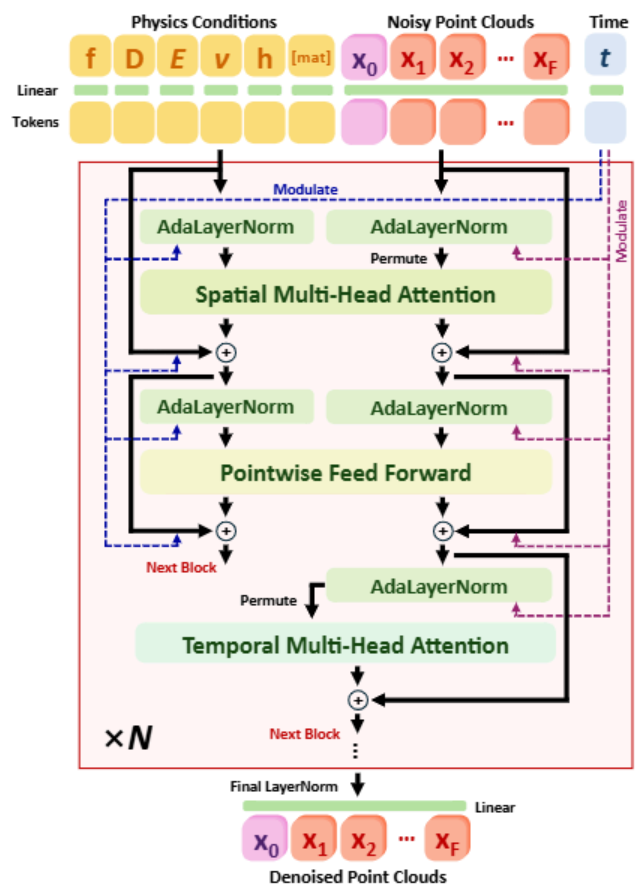

Figure 3: Our trajectory generation architecture which consists of spatial attention and temporal attention in each block.

We train our trajectory generative model on data from physics simulators—MPM [37] and a rigid-body solver. Simulator hyperparameters (e.g., substeps, grid size) introduce variability that our model, conditioning only on core physics parameters, does not capture directly. To account for this uncertainty, we employ a diffusion model to learn the conditional distribution $p(\mathcal{P}|c)$. Our method can also be extended to learning physics from more simulation methods since it requires only sampled points.

4.1.2 Physics-grounded Trajectory Generative model

Prior trajectory generative models for human motion synthesis [79, 100] typically project all point positions into a single latent space, applying attention to only temporal correlations. This approach is inadequate for our setting (see Figure 1), as it overlooks crucial spatial relationships. While naive 4D attention across both space and time can model spatio-temporal correlations in physics simulation

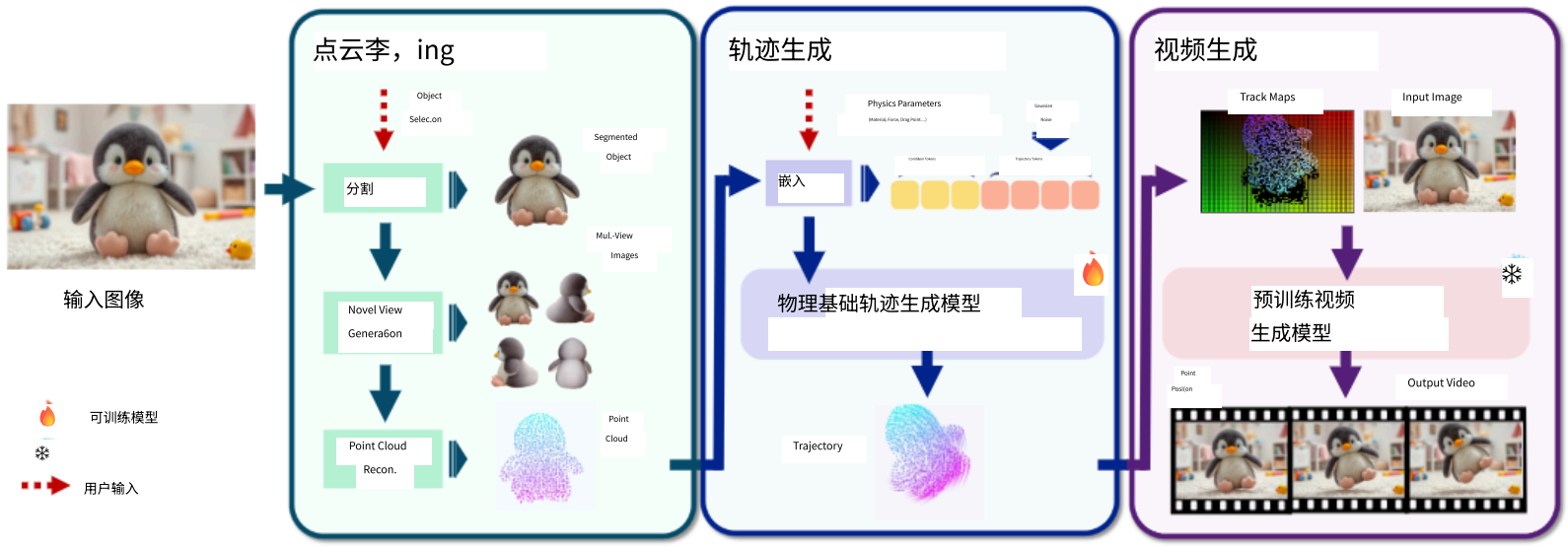

图 2：PhysCtrl 的概述。给定一张图像，我们首先将图像中的物体提升为 3D 点。然后，我们使用扩散模型生成基于物理参数和外部力的物理化运动轨迹，这些轨迹随后被用作图像到视频生成的强物理化指导。

数值不稳定性，以及经典模拟器的通用性-精度权衡。我们选择点云作为我们的表示形式，因为它们能灵活地模拟多种材料，并且足以控制预训练的视频模型。具体来说，在实践中每个物体由 2048 个点表示；我们预测它们随时间的轨迹，并使用这些轨迹作为视频合成的控制信号。我们使用 2048 个点来指导视频模型，因为先前的工作[24]表明它可以用比更多点更少的结果实现相似效果。此外，关于 4D 重建和生成的工作[34, 46, 97]表明，现实世界的运动可以用稀疏数量的基点或控制点来表示。

4.1.1 问题设定

给定一个物体，表示为包含 N 个点的 3D 点云 P={x∈R}，以及其物理参数{E, ν}，我们的轨迹生成模型根据初始力生成其动力学。具体而言，物体的动力学由未来 F 个时间步长内每个点的位置表示，即 P=P={{P}}={{x}}。将力、阻力点和边界条件（地板高度）分别记为 f∈R、D∈R 和 h∈R。因此，PhysCtrl 的目标是在条件 c={P, f, D, {E, ν}, h, [mat]}下预测 P。这里，我们使用额外的[mat]标记来表示不同的材料。在本文中，我们涵盖了四种不同的材料：弹性材料、油泥、沙子和刚性材料。值得注意的是，由于我们灵活的点云表示方法，模型不仅限于这四种类别，在拥有足够计算资源的情况下，可以方便地扩展到其他材料，例如流体。

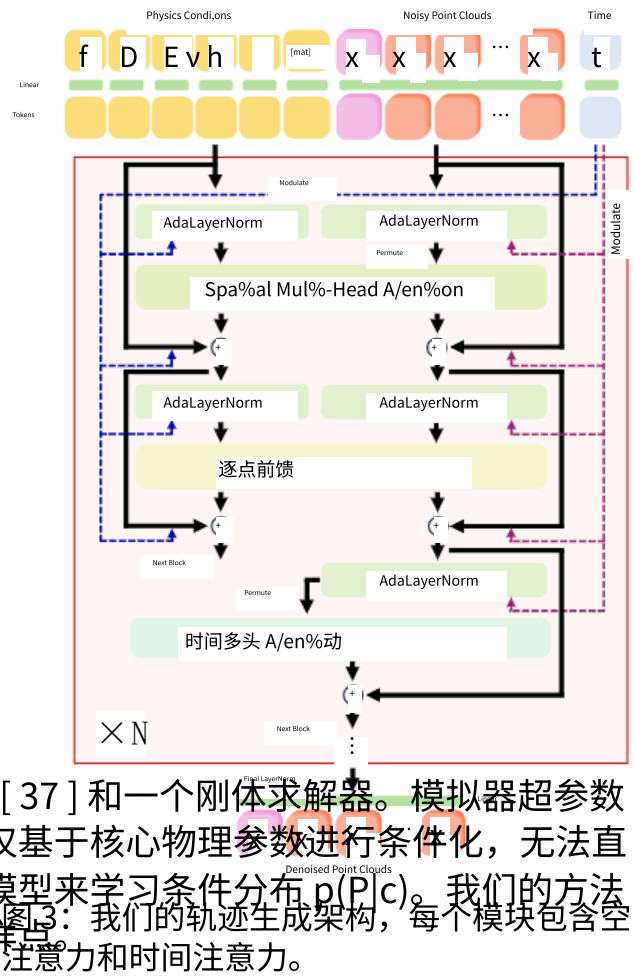

图 3：我们的轨迹生成架构，每个模块包含空间注意力和时间注意力。

我们在物理模拟器数据上训练我们的轨迹生成模型——MPM [37] 和一个刚体求解器。模拟器超参数（例如，子步长、网格大小）引入了变化性，而我们的模型仅基于核心物理参数进行条件化，无法直接捕捉这种变化性。为了解释这种不确定性，我们采用扩散模型来学习条件分布 p(P|c)。我们的方法也可以扩展到从更多模拟方法中学习物理，因为它只需要采样点。

4.1.2 基于物理的轨迹生成模型

先前用于人类运动合成的轨迹生成模型 [79 , 100] 通常将所有点位置投影到一个单一潜在空间中，仅对时间相关性应用注意力。这种方法不适用于我们的设置（见图 1），因为它忽略了关键的空间关系。虽然跨空间和时间进行简单的 4D 注意力可以模拟物理模拟中的时空相关性

data, it is suboptimal in terms of quality and efficiency due to the combinatorial explosion of spatial points across time steps. Instead, since we aim to model point cloud trajectories with a one-to-one point correspondence across frames, we introduce an efficient attention mechanism tailored for physics simulation data, which first applies spatial attention followed by temporal attention. This design reduces the computational complexity and, more importantly, reflects the underlying process of physics simulation: first integrating information from neighboring points, then propagating forward in time dimension.

Specifically, given noisy point cloud sequences, we apply point embedding and project it to latent dimensions, add sinusoidal positional embeddings in both space and time and predict its trajectory offset with our denoising network $\mathcal{D}$. The core of network $\mathcal{D}$ is a diffusion transformer consisting of a set of spatial-temporal attention blocks as shown in Figure 3. Each block contains two attention layers: spatial attention and temporal attention.

Spatial attention learns the correlation of each point with other points in the same frame with self-attention. To inject physical conditioning c into the attention layer, we first map them into additional tokens using MLPs: $\mathbf{cond} = \text{MLP}_{\text{phys}}([\mathbf{f}; \mathbf{D}; \{E, \nu\}, h, [\text{mat}]]) \in \mathbb{R}^{d_c}$. Then, we concatenate them with point positions along the sequence dimension. Motivated by CogVideoX [94], we apply the adaptive layer norm to positional tokens and physical tokens separately to facilitate the alignment between the two spaces:

$$\hat{\mathbf{P}}^f = \text{SelfAttn}\left(\text{AdaLN}([\mathbf{P}^f; \mathbf{cond}])\right), \quad \forall f \in [1, F] \tag{4}$$

Temporal attention mainly aggregates information of the same point across all timesteps for temporal consistency. We also apply attention to the input point cloud $\mathbf{P}_0$ for better trajectory learning.

$$\hat{\mathbf{T}}_p = \text{SelfAttn}\left(\text{AdaLN}([\mathbf{T}_p])\right), \quad \forall p \in [1, N] \tag{5}$$

where $\mathbf{T}_p = [\mathbf{x}_p^0, \mathbf{x}_p^1, \mathbf{x}_p^2, \ldots, \mathbf{x}_p^F] \in \mathbb{R}^{(F+1)\times d}$.

4.1.3 Training Losses

We train a standard diffusion model in which we add Gaussian noise $\boldsymbol{\epsilon}$ of different levels t to the entire point cloud sequence: $\mathcal{P}_t = \alpha_t \mathcal{P} + \sigma_t \boldsymbol{\epsilon}$ and then feed the noisy point cloud sequence into the denoising network $\mathcal{D}$. We use the signal-prediction formulation of diffusion models: $\hat{\mathcal{P}} = \mathcal{D}(\mathcal{P}_t, t, c)$.

Diffusion Loss We use MSE loss between the predicted and ground truth signal given noise samples:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathcal{P}\sim q(\mathcal{P}|c), t\sim[1,T]} \|\mathcal{D}(\mathcal{P}_t; t, c) - \mathcal{P}\|_2^2 \tag{6}$$

Velocity Loss We regulate the velocity across two frames, similar to that used in MDM [79]:

$$\mathcal{L}_{\text{vel}} = \frac{1}{F-1} \sum_{f=1}^{F-1} \|(\mathcal{P}^{f+1} - \mathcal{P}^f) - (\hat{\mathcal{P}}^{f+1} - \hat{\mathcal{P}}^f)\|_2^2 \tag{7}$$

Physics Loss To enable the model to learn physics-plausible motion trajectories, we introduce a physics-based supervision as regularization to enforce physical plausibility for the elastic, plasticine and sand material from MPM. Specifically, we constrain the position and velocity of the predicted points to adhere to the deformation gradient update (Equation (3)) across frames:

$$\mathcal{L}_{\text{phys}} = \frac{1}{N(F-2)} \sum_{f=1}^{F-2} \sum_{p=1}^{N} \|\mathbf{F}_p^{f+1} - g(\hat{\mathbf{x}}_p^f)\mathbf{F}_p^f\|_2 \quad g(\hat{\mathbf{x}}_p^f) = \mathbf{I} + \Delta T \sum_i \hat{\mathbf{v}}_i^{f+1} \nabla N(\mathbf{x}_i - \hat{\mathbf{x}}_p^f)^\top \tag{8}$$

where $\mathbf{F}_p^{f+1}$ and $\mathbf{F}_p^f$ are the ground-truth deformation gradient between adjacent frames and $\hat{\mathbf{x}}_p^f \in \hat{\mathcal{P}}^f$ is the predicted position. To obtain an approximation of grid velocity $\hat{\mathbf{v}}_i^{f+1}$ in Equation (8), we perform one P2G and G2P step (Equation (2)) at each frame in training. This can be formulated as

$$\hat{\mathbf{v}}_i^{f+1} = \frac{\sum_p N_i(\hat{\mathbf{x}}_p^f) m_p (\hat{\mathbf{v}}_p^{f+1} + \mathbf{C}_p^f(\mathbf{x}_i - \hat{\mathbf{x}}_p^f))}{\sum_p N_i(\hat{\mathbf{x}}_p^f) m_p} \tag{9}$$

虽然跨越时空的简单 4D 注意力机制能够对物理模拟数据中的时空相关性进行建模，但由于空间点在时间步长上的组合爆炸，它在质量和效率方面表现不佳。相反，由于我们旨在对帧间一一对应的点云轨迹进行建模，我们引入了一种针对物理模拟数据的高效注意力机制，该机制首先应用空间注意力，然后应用时间注意力。这种设计降低了计算复杂度，更重要的是，它反映了物理模拟的底层过程：首先整合来自邻近点的信息，然后在时间维度上向前传播。

具体来说，给定含噪声的点云序列，我们应用点嵌入并将其投影到潜在维度，在空间和时间中添加正弦位置嵌入，并使用我们的去噪网络 D 预测其轨迹偏移。网络 D 的核心是一个扩散 Transformer，由一组时空注意力块组成，如图 3 所示。每个块包含两个注意力层：空间注意力和时间注意力。

空间注意力通过自注意力学习每个点与同一帧中其他点的相关性。为了将物理条件 c 注入注意力层，我们首先使用 MLP 将它们映射到额外的 token：cond = MLP([f ; D; {E, v}, h, [mat]]) ∈ R。然后，我们将它们与沿序列维度的点位置连接起来。受 CogVideoX [94]的启发，我们对位置 token 和物理 token 分别应用自适应层归一化，以促进两个空间之间的对齐：

$$\hat{P} = \text{自我注意力} \quad \text{AdaLN}([P; \text{cond}]) \quad \boxtimes, \forall f \in [1, F] \tag{4}$$

时间注意力主要聚合所有时间步长中相同点的信息，以保持时间一致性。我们还对输入点云 P 应用注意力，以更好地学习轨迹。

$$\hat{T} = \text{SelfAttn}\left(\text{AdaLN}([T])\right), \forall p \in [1, N] \tag{5}$$

其中 T= [x, x, x, . . . , x] ∈ R.

4.1.3 训练损失

我们在标准扩散模型中训练，向整个点云序列添加不同级别的高斯噪声 ϵ：$P = \alpha P + \sigma\epsilon$，然后将带噪声的点云序列输入去噪网络 D。我们使用扩散模型的信号预测公式：$\hat{P} = D(P, t, c)$。

扩散损失 我们使用预测信号和真实信号之间的均方误差损失，给定噪声样本：

$$L = E \, /\!/ \, D(P; t, c) - P \, /\!/ \tag{6}$$

速度损失 我们调节两帧之间的速度，类似于 MDM [79]中使用的：

$$L = \frac{1}{F-1} \sum_{f=1}^{F-1} /\!/ (P - P) - (\hat{P} - \hat{P}) /\!/ \tag{7}$$

物理损失 为了使模型能够学习符合物理规律的运动轨迹，我们引入基于物理的监督作为正则化手段，以强制 MPM 弹性、塑性泥和沙材料符合物理合理性。具体来说，我们约束预测点的位置和速度在帧之间遵循变形梯度更新（公式(3)）：

$$L = \frac{1}{N(F-2)} \sum_{f=1}^{F-2} \sum_{p=1}^{N} /\!/ F_p - g(\hat{x})F /\!/ g(\hat{x}) = I + \Delta T \sum_{i} \hat{v}_i \nabla N(x - \hat{x}) \tag{8}$$

其中 Fp 和 Fare 相邻帧之间的真实变形梯度，$\hat{x} \in \hat{P}$ 是预测位置。为了获得方程(8)中网格速度 $\hat{v}_i$ 的近似值，我们在训练中的每一帧执行一次 P2G 和 G2P 步骤（方程(2)）。这可以表示为

$$\hat{v}_i = \frac{\sum_p N(\hat{x})m(\hat{v}_p + C(x - \hat{x}))}{\sum_p N(\hat{x})m} \tag{9}$$

where $\mathbf{C}_p^f$ is also from ground-truth and $\hat{\mathbf{v}}_p^{f+1} = (\hat{\mathbf{x}}_p^{f+2} - \hat{\mathbf{x}}_p^f)/(2\Delta T)$. Note that we ignore the stress term and use next-frame point velocity $\hat{\mathbf{v}}_p^{f+1}$ because it yields a more accurate approximation when the frame interval ΔT is much larger than the substep interval Δt for MPM simulation.

Boundary Loss To enforce the boundary condition of the ground, we add a penetration loss, preventing the points from passing through the surface:

$$\mathcal{L}_{\text{floor}} = \frac{1}{N} \sum_{f=1}^{F} \sum_{p=1}^{N} \left(\max(h - \hat{\mathbf{x}}_p^f, 0) \right)^2 \tag{10}$$

Overall, our training loss is: $\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}} + \lambda_{\text{floor}}\mathcal{L}_{\text{floor}}$.

4.2 Physics-grounded Image-to-Video Generation

Starting with a single image of 3D scene with objects, we first segment out [39] the objects and generate novel view images for each object. We then feed both the novel views and the segmented image into a multiview Gaussian reconstruction model [78] and extract a point cloud for the input objects. For input with floor conditions, we support user input to select the floor region and use VGGT [83] to reconstruct the 3D scene. Then we align the coordinate system of VGGT and the 3D points of the object and obtain the height of the floor using principal component analysis. We then use our trajectory generative model to generate the dynamics of object points. The generated 3D point trajectories are then projected to the image space of the input camera viewpoint to obtain the motion trajectories of each pixel. The projected pixel trajectories can be directly used as conditioning signals for a pre-trained video generative model to produce the final video. Specifically, we use DaS [24] as the video model. It takes a "tracking video" as condition, which is the projected 3D point trajectories of 2D grid anchor points at the first frame. For each anchor point, we associate it with the nearest 3D object point. Then, we project the 3D point trajectories into 2D and get the final tracking video.

Figure 4: Qualitative comparison between our method and existing video generation methods.

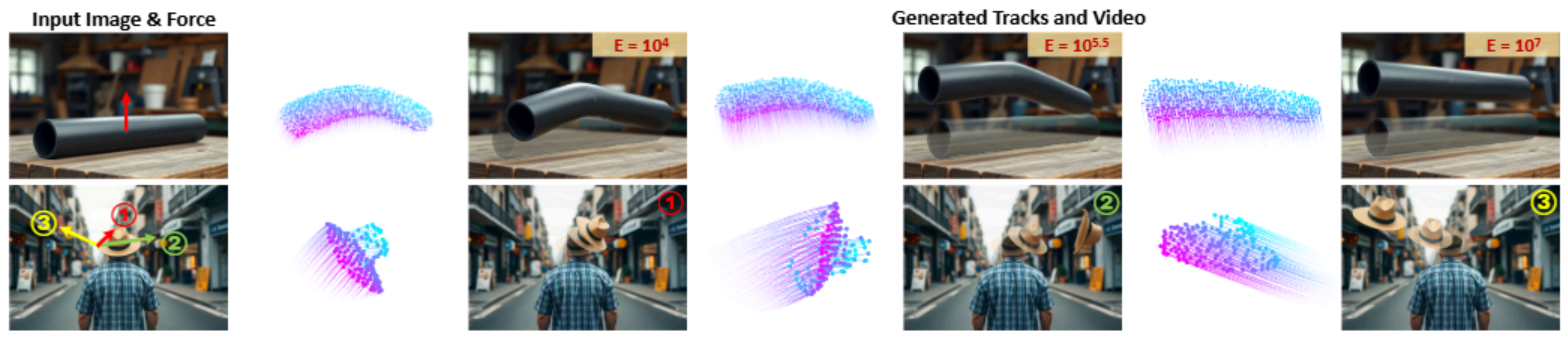

Figure 5: PhysCtrl generates videos of the same object under different physics parameters and forces.

其中 C 也来自真实值，ˆvp = (ˆxp − ˆx)/(2ΔT)。需要注意的是，我们忽略了应力项，并使用下一帧点的速度ˆvp，因为在 MPM 模拟中，当帧间隔ΔT 远大于子步间隔Δt 时，这能提供更精确的近似。

边界损失 为了强制执行地面的边界条件，我们添加了穿透损失，以防止点穿过表面：

$$L = \frac{1}{N} \sum_{f=1}^{F} \sum_{p=1}^{N} \max(h - \hat{x}, 0) \quad (10)$$

总体而言，我们的训练损失为：L = L+ λL+ λL+ λL。

4.2 物理基础图像到视频生成

从一张包含物体的 3D 场景图像开始，我们首先分割出[39]这些物体，并为每个物体生成新的视角图像。接着，我们将这些新视角图像和分割后的图像输入到一个多视角高斯重建模型[78]中，并提取输入物体的点云。对于具有地面条件的输入，我们支持用户选择地面区域，并使用 VGGT[83]来重建 3D 场景。然后我们使 VGGT 的坐标系与物体的 3D 点对齐，并使用主成分分析获得地面的高度。随后我们使用我们的轨迹生成模型来生成物体点的动力学。生成的 3D 点轨迹被投影到输入相机视点的图像空间中，以获得每个像素的运动轨迹。这些投影的像素轨迹可以直接用作预训练视频生成模型的条件信号，以生成最终视频。具体来说，我们使用 DaS[24]作为视频模型。它以“跟踪视频”作为条件，即第一帧中 2D 网格锚点的投影 3D 点轨迹。 对于每个锚点，我们将其与最近的 3D 物体点关联。然后，我们将 3D 点轨迹投影到 2D 空间，得到最终的跟踪视频。

图 4：我们方法与现有视频生成方法之间的定性比较。

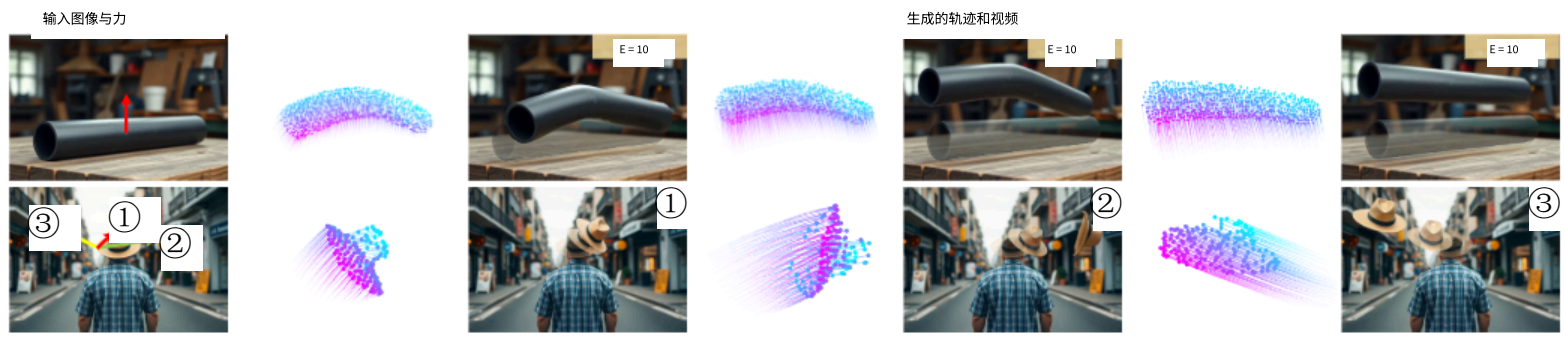

图 5：PhysCtrl 在不同物理参数和力的作用下生成同一物体的视频。

5 Experiments

5.1 Evaluation on Image-to-Video Generation

Baselines We compare PhysCtrl with state-of-the-art controllable video generative models, including Wan2.1-I2V-14B [82], CogVideoX [94], DragAnything [89], ObjCtrl-2.5D [87]. The first two methods support image-to-video generation with text prompts. We use ChatGPT-4o to generate text prompts based on the direction of the object movement. The last two achieve controllable video generation with user-specified single-point trajectories. We use the trajectories of the drag point generated by our model to prompt the model.

Quantitative Evaluation Since we are the first method to inject physics prior into a video model, we utilize GPT-4o to evaluate three aspects of 12 generated videos in a 5-Likert score inspired by VideoPhy [2]: (1) Semantic Adherence (SA): how well the content and motion in the video match the description in the text prompt, especially the alignment with the force direction and position; (2) Physical commonsense (PC): whether the object's motion follows intuitive, physically plausible dynamics given the applied force direction and position; (3) Video Quality (VQ): overall visual and temporal quality of the video. Results in Table 1 show that our method achieve the best results across all baselines. Results of user study can be found in the supplemental.

Qualitative Evaluation The qualitative results between our method and baselines can be found in Figure 4. CogVideoX-5B [94] and Wan2.1 [82] have strong generation ability and partly follow the text prompts. However, they only use text prompts as conditions and lack precise control, thus, they cannot produce motions that fully reflect physics conditions. For example, the *chair* in Figure 4 doesn't move according to the force direction. DragAnything [89] uses purely 2D trajectories and cannot distinguish between camera motion and object motion, thus sometimes generating camera motions while objects remain static. More importantly, both DragAnything [89] and ObjCtrl2.5D [87] only use coarse trajectory as a condition and struggle to generate more complex motions, *e.g*, the UFO case in Figure 4 that contains both rotations and depth change. In comparison, PhysCtrl produces physics-plausible videos that follow the given forces by generating physics-grounded 3D trajectories as a strong conditional signal to guide the superior generation capability of pretrained video generative models for video synthesis.

Table 1: Results of video evaluation.

	SA↑	PC↑	VQ↑
DragAnything [89]	2.9	2.8	2.8
ObjCtrl [87]	1.5	1.3	1.4
Wan2.1 [82]	3.8	3.7	3.6
CogVideoX [94]	3.2	3.2	3.1
Ours	**4.5**	**4.5**	**4.3**

Table 2: Quantitative comparison on trajectory generation.

Method	vIoU↑	CD↓	Corr↓
M2V [6]	24.92%	0.2160	0.1064
MDM [79]	53.78%	0.0159	0.0240
Ours	**77.59%**	**0.0028**	**0.0015**

Results on Varying Physical Conditions Since our trajectory generative model is conditioned on external forces and physics parameters, we can generate videos of the same object under varying conditions. As shown in Figure 5, we can change the Young's modulus in elastic material to produce results with different deformations given the same force. The direction and amplitude of the force can also be adjusted to match the user's desired motion. We found that Poisson's ratio ν has negligible influence on the generated trajectories, similar to the findings in PhysDreamer [97].

5.2 Evaluation on Generative Dynamics

Baselines We compare our approach with existing methods that focus on generative dynamics, including Motion2VecSets [6] and MDM [79]. Motion2VecSets is a method for reconstructing sparse point cloud sequences; we eliminate the sparse point cloud condition and introduce physics conditions instead. MDM is primarily aimed at human motion generation, so we substitute human joints with point clouds and incorporate physics conditions as additional tokens. For computation efficiency, we trained all baselines and ablations on our elastic subset of 160K objects that contains complex deformations for metrics comparison.

Evaluation Metrics Following [45, 6], we adopt volume Intersection over Union (vIoU), Chamfer Distance (CD) and L_2-distance error for evaluation. vIoU measures the overlap between predicted and ground truth point clouds, CD measures the averaged per-point pairwise nearest neighbor distance

5 实验

5.1 对图像到视频生成的评估

基线 我们将 PhysCtrl 与最先进的可控视频生成模型进行比较，包括 Wan2.1-I2V-14B [82]、CogVideoX [94]、DragAnything [89] 和 ObjCtrl-2.5D [87]。前两种方法支持使用文本提示进行图像到视频的生成。我们使用 ChatGPT-4o 根据物体运动的方向生成文本提示。后两种方法通过用户指定的单点轨迹实现可控视频生成。我们使用我们模型生成的拖动点轨迹来提示模型。

定量评估 由于我们是首个将物理先验注入视频模型的方法，我们利用 GPT-4o 对 12 个生成的视频在受 VideoPhy [2]启发的 5 分李克特量表上评估了三个方面：(1) 语义一致性（SA）：视频中的内容和运动与文本提示中的描述匹配程度，特别是与力的方向和位置的符合程度；(2) 物理常识（PC）：物体的运动是否遵循在给定力的方向和位置下的直观、物理上合理的动力学；(3) 视频质量（VQ）：视频的整体视觉和时序质量。表 1 中的结果表明，我们的方法在所有基线中取得了最佳结果。用户研究的结果可以在补充材料中找到。

定性评估我们的方法与基线方法的定性结果如图 4 所示。CogVideoX-5B [94] 和 Wan2.1 [82] 具有较强的生成能力，并在一定程度上遵循文本提示。然而，它们仅使用文本提示作为条件，缺乏精确控制，因此无法生成完全反映物理条件的运动。例如，图 4 中的椅子没有按照力的方向移动。DragAnything [89] 使用纯二维轨迹，无法区分相机运动和物体运动，因此有时会生成相机运动而物体保持静止。更重要的是，DragAnything [89] 和 ObjCtrl2.5D [87] 都仅使用粗略轨迹作为条件，难以生成更复杂的运动，例如图 4 中的 UFO 案例，它包含旋转和深度变化。相比之下，PhysCtrl 通过生成基于物理的 3D 轨迹作为强条件信号来指导预训练视频生成模型的优越生成能力，从而产生遵循给定力的物理合理的视频。

表 1：视频评估结果。

	SA ↑	PC ↑	VQ ↑
拖动任何 [89]	2.9	2.8	2.8
对象控制 [87]	1.5	1.3	1.4
万	2.1		
[82]	3.8	3.7	3.6
CogVideoX [94]	3.2	3.2	3.1
Ours	4.5	4.5	4.3

表 2：轨迹生成的定量比较。

方法	vIoU ↑	CD ↓	Corr ↓
M2V [6]	24.92%	0.2160	0.1064
MDM [79]	53.78%	0.0159	0.0240
Ours	77.59%	0.0028	0.0015

不同物理条件下的结果由于我们的轨迹生成模型基于外部力和物理参数进行条件化，因此我们可以生成在不同条件下同一物体的视频。如图 5 所示，我们可以改变弹性材料中的杨氏模量，在相同力的作用下产生具有不同变形的结果。力的方向和幅度也可以调整以匹配用户的期望运动。我们发现泊松比v对生成的轨迹影响可以忽略不计，这与 PhysDreamer [97]中的发现相似。

5.2 在生成动力学上的评估

基线 我们将我们的方法与现有的关注生成动力学的现有方法进行比较，包括 Motion2VecSets [6] 和 MDM [79]。Motion2VecSets 是一种用于重建稀疏点云序列的方法；我们消除了稀疏点云条件，并引入了物理条件。MDM 主要针对人体运动生成，因此我们将人体关节替换为点云，并将物理条件作为附加标记纳入。为了计算效率，我们在包含复杂变形的 160K 对象的弹性子集上训练了所有基线和消融实验，用于指标比较。

评估指标遵循[45, 6]，我们采用体积交并比（vIoU）、Chamfer 距离（CD）和 L 距离误差进行评估。vIoU 衡量预测点云与真实点云之间的重叠程度，CD 衡量每对最近邻点之间的平均距离

between two point clouds, L_2-distance is the Euclidean distance between two corresponding point clouds. Each metric is calculated at each timestep separately and averaged across all frames.

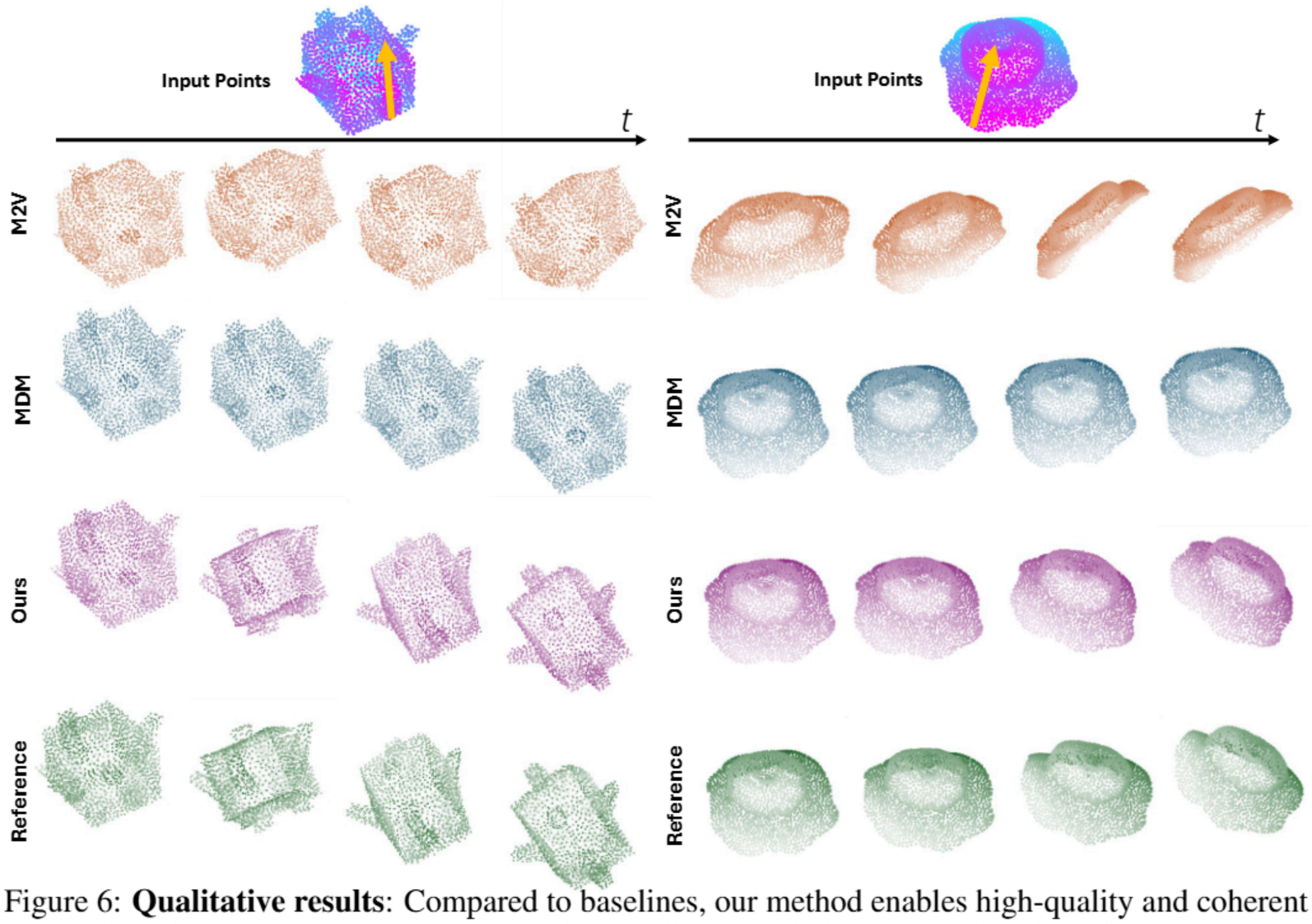

Figure 6: **Qualitative results**: Compared to baselines, our method enables high-quality and coherent generation of motion sequences from physics conditions and closely matches the reference.

Table 3: Ablation study on trajectory generative model.

Method	vIoU↑	CD↓	Corr↓
w/o spatial attention	33.76%	0.2348	0.1163
w/o temporal attention	53.63%	0.0480	0.0507
w/o physics loss	76.30%	0.0030	0.0016
Ours	**77.59%**	**0.0028**	**0.0015**

Table 4: Ablation study on using traditional simulator and our model for video generation.

Method	SA↑	PC↑	VQ↑
Traditional (2048 points)	4.3	4.3	**4.4**
Traditional (8192 points)	4.3	4.3	4.2
Ours (2048 points)	**4.5**	**4.5**	4.3

Results Table 2 shows the quantitative comparison of our method with other baselines. Our method demonstrates the best performance over all metrics on the testing set. The qualitative comparison can be found in Figure 1. Our model achieves physics-grounded and consistent generation of motion trajectories. Motion2vecsets struggles to generate time-coherent motions because in our experiments, there is no sparse point cloud condition in their original setting. M2V struggles to generate coherent motions in our experiments. There are two potential reasons for this. Firstly, their model is originally designed for point cloud completion, but in our setting, there is no sparse point cloud condition. Prior work [98] also found that M2V does not work well in this situation. Secondly, their deformation latent is encoded frame-by-frame without temporal interaction. MDM can generate consistent motion sequences, but fails to capture detailed deformations because all points in a frame are projected into a single latent. The superiority of our method is based on our spatial-temporal attention block, which leverages explicit per-point correspondence.

5.3 Ablation Study

The qualitative and quantitative results of the ablation study for trajectory generation can be found in Figure 7 and Table 3. Our physics loss improved all the metrics and makes the results of our trajectory generation close to the ground truth. The physics loss aligns the updated deformation gradient with the ground truth and constrains the predicted positions. Although without physics loss, our model can achieve good results, it can be further improved with physical guidance as regularization.

Table 4 presents the ablation study for video generation. Results show that using our trajectory generation model for video generation is on par with using a traditional simulator. Also, results also show that using more points didn't bring a performance gain for video generation.

CD 衡量两个点云之间的平均每点成对最近邻距离，L-distance 是两个对应点云之间的欧几里得距离。每个指标在每个时间步分别计算，并在所有帧上取平均值。

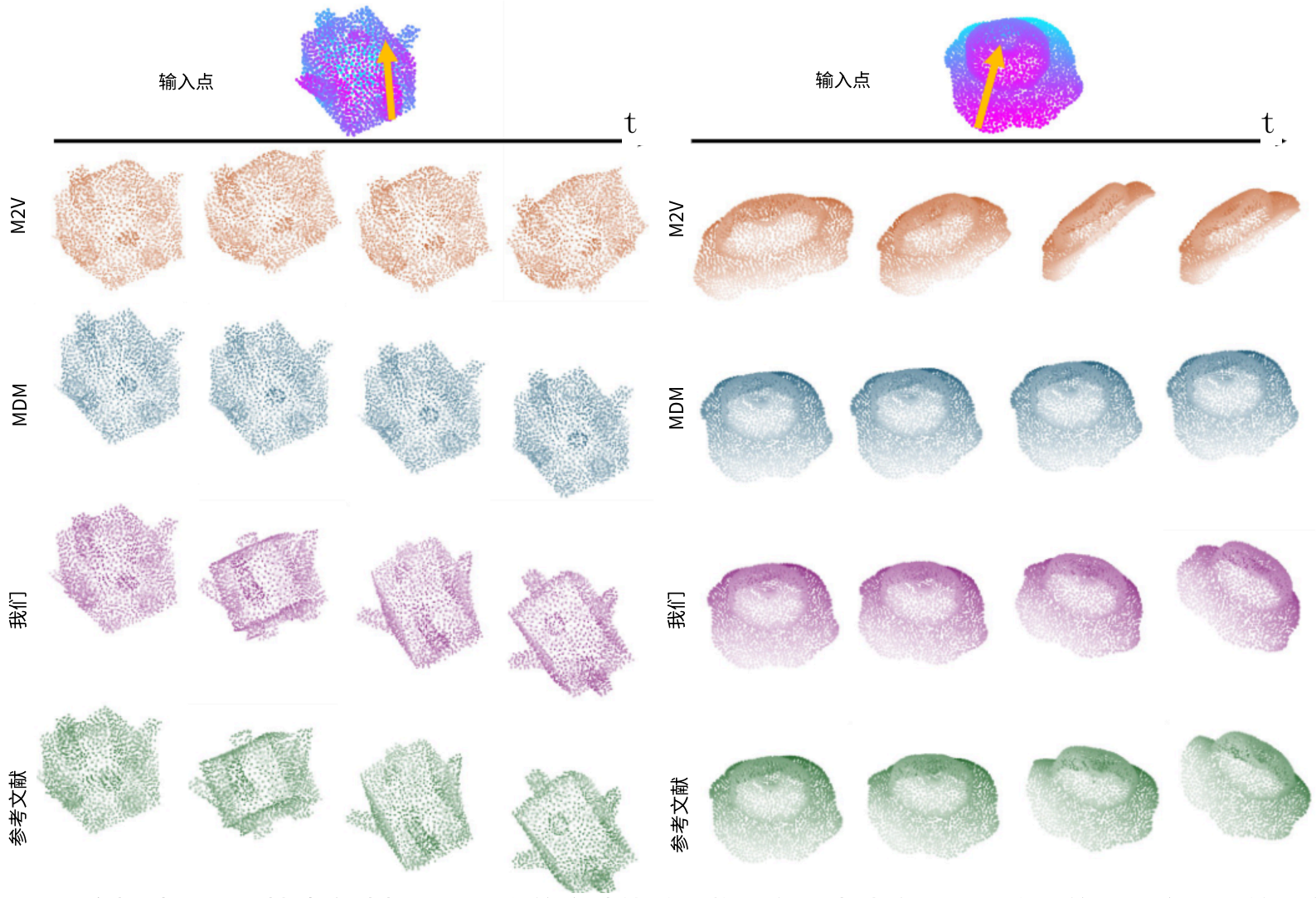

图 6：定性结果：与基线方法相比，我们的方法能够从物理条件生成高质量且连贯的运动序列，并与参考结果高度匹配。

表 3：轨迹生成模型的消融研究。

方法	vIoU ↑	CD ↓	Corr ↓
无空间注意力	33.76%	0.2348	0.1163
无时间注意力	53.63%	0.0480	0.0507
无物理损失	76.30%	0.0030	0.0016
我们	77.59%	0.0028	0.0015

表 4：使用传统模拟器和我们的模型进行视频生成的消融研究。

方法	SA ↑	PC ↑	VQ ↑
传统（2048 点）	4.3	4.3	4.4
传统（8192 点）	4.3	4.3	4.2
我们（2048 点）	4.5	4.5	4.3

结果表 2 展示了我们的方法与其他基线的定量比较。我们的方法在测试集上所有指标上均表现出最佳性能。定性比较可参见图 1。我们的模型实现了物理基础且一致的轨迹生成。Motion2vecsets 难以生成时间一致的运动，因为在我们的实验中，它们的原始设置中没有稀疏点云条件。M2V 在我们的实验中也难以生成一致的运动。这存在两个潜在原因。首先，它们的模型原本设计用于点云补全，但在我们的设置中不存在稀疏点云条件。先前工作[98]也发现 M2V 在这种情况下表现不佳。其次，它们的变形潜空间是逐帧编码的，没有时间交互。MDM 可以生成一致的动序列，但无法捕捉详细的变形，因为一帧中的所有点都被投影到一个单一潜空间中。我们方法的优势基于我们的时空注意力模块，该模块利用了显式的逐点对应关系。

5.3 消融实验

轨迹生成的消融研究的定性和定量结果可以在图 7 和表 3 中找到。我们的物理损失改进了所有指标，并使我们的轨迹生成结果接近真实值。物理损失使更新后的变形梯度与真实值对齐，并约束预测位置。尽管没有物理损失，我们的模型也能取得良好结果，但通过物理指导作为正则化，可以进一步改进。

表 4 展示了视频生成的消融研究。结果表明，使用我们的轨迹生成模型进行视频生成与传统模拟器相当。此外，结果还显示，使用更多点并未为视频生成带来性能提升。

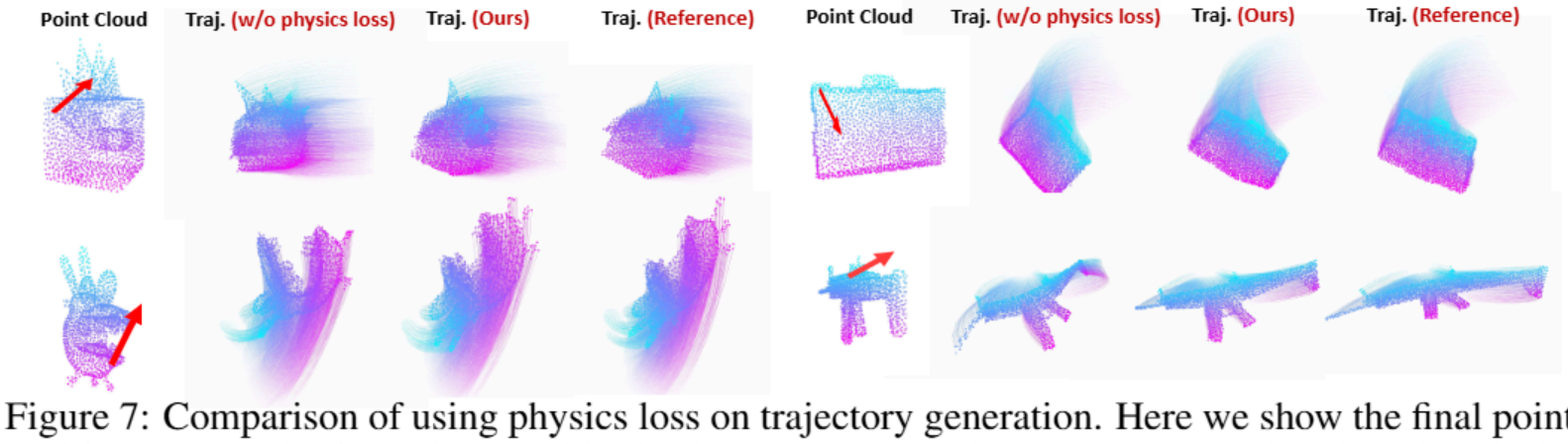

Figure 7: Comparison of using physics loss on trajectory generation. Here we show the final point position and tracks for points. With physics loss, the results are more closely aligned with the reference (simulated by MPM).

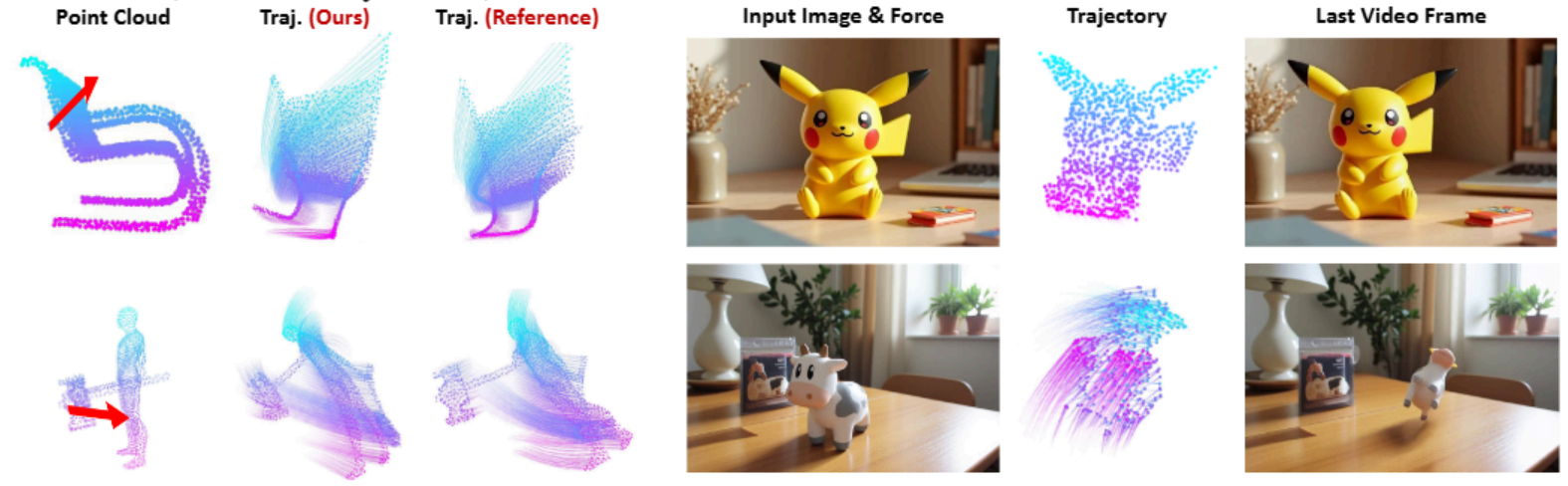

Figure 8: Failure cases.

6 Discussion

Failure Cases As shown in Figure 2, our model contains three components: point cloud lifting, trajectory generation and video generation. (1) **Failure due to point cloud lifting is extremely rare**. Single image-to-3D produces reasonable geometry overall and is unlikely to yield severely distorted or implausible shapes. While minor artifacts (e.g. geometry not perfectly smooth or noisy points on the surface) may occur, they have minimal impact on our results. We achieve such robustness to geometric variations because our trajectory generation model on the diverse Objaverse dataset and applying data augmentation of surface noise. (2) **Failure cases of our trajectory generation model**: our model cannot handle thin structures well and sometimes might fail to accurately capture complex internal deformations. (3) **Failure cases due to video generation**: The video model cannot fully follow the trajectory when it conflicts with the prior ofthe video model. For example, when the user input for the object material is very stiff, but it appears soft according to RGB information. The video model might also hallucinate unexpected content in the occluded region, *e.g*, for an animal, it might generate five legs. See Figure 8 for example failure cases.

Extension to Multiple Objects MPM achieves multi-object interaction by representing scene dynamics as point movement. Our model also predicts point trajectories, so it is also inherently capable of multi-object interaction. We did a preliminary experiment on multi-objects on a simplied setting: we create a dataset that has an object dragged towards a cube and colliding with the cube from different angles and distances. We trained our model on it and achieved 93.70% vIOU on the held out testing set.

7 Conclusion and Limitations

In this paper, we introduce PhysCtrl, a novel framework for physics-grounded video generation with physics parameters and force control. We design a diffusion model with spatial-temporal attention blocks and physics-based supervision to effectively and efficiently learn complex physical deformations directly on point cloud sequences. The generated motion trajectories can be used as a strong conditional signal for pre-trained video generative models. Our experiments demonstrate that PhysCtrl can generate physics-grounded dynamics and enable high-quality image-to-video generation results conditioned on external forces and physics parameters.

Our approach mostly focuses on single-object dynamics for four material types and does not cover all possible materials. We only do initial study on multiple objects and more complex phenomena should be investigated, such as intricate boundary conditions. Future work includes addressing these limitations and extending PhysCtrl to more diverse and complex physics phenomena in the real world.

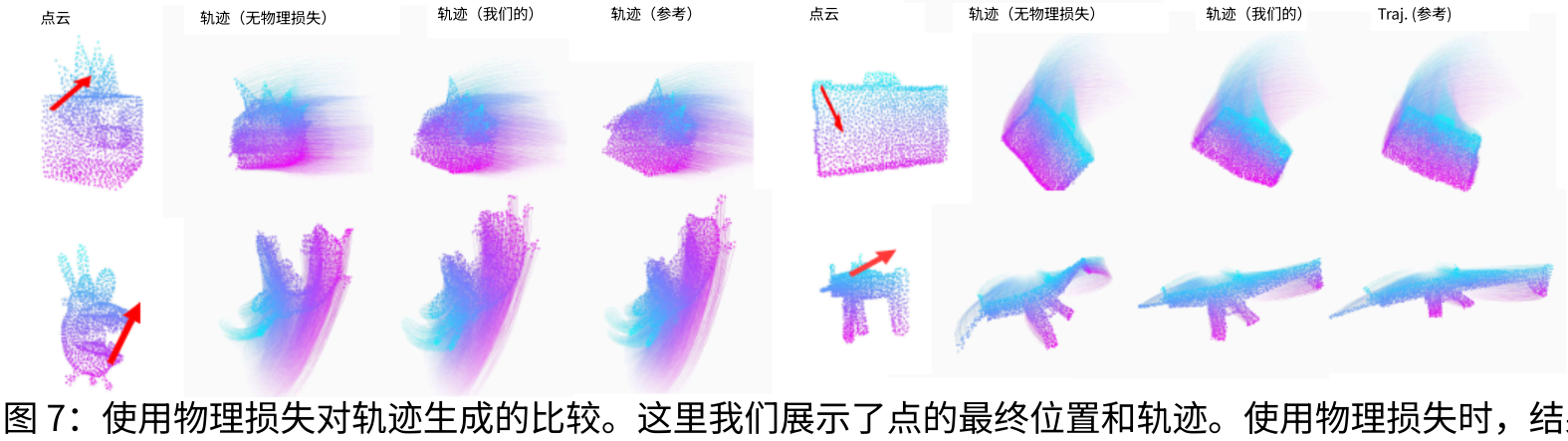

图 7：使用物理损失对轨迹生成的比较。这里我们展示了点的最终位置和轨迹。使用物理损失时，结果更紧密地与参考（由 MPM 模拟）对齐。

图 8：失败案例。

6 讨论

失败案例 如图 2 所示，我们的模型包含三个组件：点云提升、轨迹生成和视频生成。(1) 由于点云提升导致的失败案例极为罕见。单图像到 3D 生成的整体几何形状合理，不太可能产生严重扭曲或不合理的形状。虽然可能会出现轻微的伪影（例如几何形状不完全平滑或表面上的噪声点），但它们对我们结果的影响最小。我们之所以能实现这种对几何变化的鲁棒性，是因为我们的轨迹生成模型在多样化的 Objaverse 数据集上训练，并应用了表面噪声的数据增强。(2) 我们轨迹生成模型的失败案例：我们的模型处理细长结构的效果不佳，有时可能无法准确捕捉复杂的内部变形。(3) 由于视频生成导致的失败案例：当轨迹与视频模型的先验知识冲突时，视频模型无法完全遵循轨迹。例如，当用户输入的物体材质非常坚硬，但根据 RGB 信息却显得柔软时。视频模型还可能在遮挡区域产生意想不到的内容，e.g，对于动物来说，它可能会生成五条腿。例如，图 8 展示了失败的案例。

扩展到多个物体 MPM 通过将场景动态表示为点的运动来实现多物体交互。我们的模型也预测点的轨迹，因此它本质上也具有多物体交互的能力。我们在简化的设置上进行了初步的多物体实验：我们创建了一个数据集，其中包含一个物体被拖向一个立方体并与立方体从不同的角度和距离发生碰撞。我们在该数据集上训练了我们的模型，并在保留的测试集上达到了 93.70%的 vIOU。

7　结论和局限性

在本文中，我们介绍了 PhysCtrl，一个具有物理参数和力控制的物理基础视频生成新框架。我们设计了一个具有时空注意力模块和基于物理的监督的扩散模型，以有效且高效地直接在点云序列上学习复杂的物理变形。生成的运动轨迹可以作为预训练视频生成模型的强条件信号。我们的实验表明，PhysCtrl 可以生成物理基础的动态，并能在外部力和物理参数的条件下实现高质量图像到视频的生成结果。

我们的方法主要关注四种材料类型的单物体动力学，并未涵盖所有可能的材料。我们仅对多物体进行了初步研究，更复杂的现象（如复杂的边界条件）应进一步研究。未来的工作包括解决这些局限性，并将 PhysCtrl 扩展到更多样化和复杂的现实世界物理现象。

References

[1] Pymunk (2023), https://pymunk.org

[2] Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)

[3] Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025)

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

[5] Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., et al.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13–23 (2025)

[6] Cao, W., Luo, C., Zhang, B., Nießner, M., Tang, J.: Motion2vecsets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking. In: CVPR. pp. 20496–20506 (2024)

[7] Che, H., He, X., Liu, Q., Jin, C., Chen, H.: Gamegen-x: Interactive open-world game video generation. arXiv preprint arXiv:2411.00769 (2024)

[8] Chen, B., Jiang, H., Liu, S., Gupta, S., Li, Y., Zhao, H., Wang, S.: Physgen3d: Crafting a miniature interactive world from a single image. arXiv preprint arXiv:2503.20746 (2025)

[9] Chen, C., Dou, Z., Wang, C., Huang, Y., Chen, A., Feng, Q., Gu, J., Liu, L.: Vid2sim: Generalizable, video-based reconstruction of appearance, geometry and physics for mesh-free simulation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)

[10] Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)

[11] Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR. pp. 7310–7320 (2024)

[12] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. NeurIPS **31** (2018)

[13] Chu, M., Liu, L., Zheng, Q., Franz, E., Seidel, H.P., Theobalt, C., Zayer, R.: Physics informed neural fields for smoke reconstruction with sparse data. ACM TOG **41**(4), 1–14 (2022)

[14] Decart, E., Campbell, S., McIntyre, Q., Chen, X., Quevedo, J.: Oasis: A universe in a transformer (2024)

[15] Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36**, 35799–35813 (2023)

[16] Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR. pp. 13142–13153 (2023)

[17] Desbrun, M.: Smoothed particles: A new paradigm for animating highly deformable bodies. Computer Animation and Simulation/Springer Vienna (1996)

[18] Erkoç, Z., Ma, F., Shan, Q., Nießner, M., Dai, A.: Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In: ICCV. pp. 14300–14310 (2023)

[19] Feng, Y., Shang, Y., Feng, X., Lan, L., Zhe, S., Shao, T., Wu, H., Zhou, K., Su, H., Jiang, C., et al.: Elastogen: 4d generative elastodynamics. arXiv preprint arXiv:2405.15056 (2024)

[20] Fu, X., Liu, X., Wang, X., Peng, S., Xia, M., Shi, X., Yuan, Z., Wan, P., Zhang, D., Lin, D.: 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. arXiv preprint arXiv:2412.07759 (2024)

参考文献

[1] Pymunk (2023), https://pymunk.org [2] Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: 评估视频生成中的物理常识。arXiv 预印本 arXiv:2406.03520 (2024) [3] Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: 视频生成中具有挑战性的动作中心物理常识评估。arXiv 预印本 arXiv:2503.06800 (2025) [4] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., 等.: Stable video diffusion: 将潜在视频扩散模型扩展到大型数据集。arXiv 预印本 arXiv:2311.15127 (2023) [5] Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., 等.: Go-with-the-flow: 使用实时扭曲噪声的运动可控视频扩散模型。在《计算机视觉与模式识别会议论文集》中。第 13-23 页 (2025) [6] Cao, W., Luo, C., Zhang, B., Nießner, M., Tang, J.: Motion2vecsets: 用于非刚性形状重建和跟踪的四维潜在向量集扩散。在 CVPR 中。第 20496 - 20506 (2024) [7] Che, H., He, X., Liu, Q., Jin, C., Chen, H.: Gamegen-x: 交互式开放世界游戏视频生成。arXiv 预印本 arXiv:2411.00769 (2024) [8] Chen, B., Jiang, H., Liu, S., Gupta, S., Li, Y., Zhao, H., Wang, S.: Physgen3d: 从单张图像构建微型交互世界。arXiv 预印本 arXiv:2503.20746 (2025) [9] Chen, C., Dou, Z., Wang, C., Huang, Y., Chen, A., Feng, Q., Gu, J., Liu, L.: Vid2sim: 基于视频的通用外观、几何和物理重建，用于无网格模拟。IEEE 计算机视觉与模式识别会议 (CVPR) (2025) [10] Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X.,

等：Videocrafter1：用于高质量视频生成的开放扩散模型。arXiv 预印本 arXiv:2310.19512 (2023) [11] Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y：Videocrafter2：克服高质量视频扩散模型的数据限制。发表于 CVPR。第 7310–7320 页（2024）[12] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential

方程。发表于 NeurIPS 31 (2018) [13] Chu, M., Liu, L., Zheng, Q., Franz, E., Seidel, H.P., Theobalt, C., Zayer, R.: 物理信息神经场用于稀疏数据下的烟雾重建。发表于 ACM TOG 41(4), 1–14 (2022) [14] Decart, E., Campbell, S., McIntyre, Q., Chen, X., Quevedo, J.: Oasis: 一个在

Transformer 中的宇宙（2024） [15] Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte,

C., Voleti, V., Gadre, S.Y., 等：Objaverse-xl：一个包含 10m+ 3d 对象的宇宙。神经信息处理系统进展 36, 35799–35813 (2023) [16] Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L.,

Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: 一个标注的 3D 物体宇宙。在：CVPR。第 13142–13153 页（2023）[17] Desbrun, M.: 光滑粒子：一种用于动画化高度可变形物体的新范式。

计算机动画与模拟/Springer 维也纳（1996）[18] Erkoç, Z., Ma, F., Shan, Q., Nießner, M., Dai, A.: Hyperdiffusion：使用权重空间扩散生成隐式神经

场。在：ICCV。第 14300–14310 页（2023）[19] Feng, Y., Shang, Y., Feng, X., Lan, L., Zhe, S., Shao, T., Wu, H., Zhou, K., Su, H., Jiang, C，

等：Elastogen：4D 生成弹性动力学。arXiv 预印本 arXiv:2405.15056（2024）[20] Fu, X., Liu, X., Wang, X., Peng, S., Xia, M., Shi, X., Yuan, Z., Wan, P., Zhang, D., Lin,

D.: 3dtrajmaster: 掌握视频生成中多实体运动的 3D 轨迹。arXiv 预印本 arXiv:2412.07759 (2024)

[21] Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Doersch, C., Aytar, Y., Rubinstein, M., et al.: Motion prompting: Controlling video generation with motion trajectories. arXiv preprint arXiv:2412.02700 (2024)

[22] Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: Force prompting: Video generation models can learn and generalize physics-based control signals. arXiv preprint arXiv:2505.19386 (2025)

[23] Gong, S., Li, M., Feng, J., Wu, Z., Kong, L.: Diffuseq: Sequence to sequence text generation with diffusion models. ICLR (2023)

[24] Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. arXiv preprint arXiv:2501.03847 (2025)

[25] He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)

[26] He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)

[27] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

[28] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)

[29] Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: ICLR Workshop on Deep Generative Models for Highly Structured Data (2022), https://openreview.net/forum?id=BBelR2NdDZ5

[30] Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: CVPR. pp. 8153–8163 (2024)

[31] Hu, Y., Fang, Y., Ge, Z., Qu, Z., Zhu, Y., Pradhana, A., Jiang, C.: A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018)

[32] Hu, Y., Li, T.M., Anderson, L., Ragan-Kelley, J., Durand, F.: Taichi: a language for high-performance computation on spatially sparse data structures. ACM Transactions on Graphics (TOG) **38**(6), 1–16 (2019)

[33] Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In: International Conference on Machine Learning. pp. 13916–13932. PMLR (2023)

[34] Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4220–4230 (2024)

[35] Jiang, C., Gast, T., Teran, J.: Anisotropic elastoplasticity for cloth, knit and hair frictional contact. ACM Transactions on Graphics (TOG) **36**(4), 1–14 (2017)

[36] Jiang, C., Schroeder, C., Selle, A., Teran, J., Stomakhin, A.: The affine particle-in-cell method. ACM Transactions on Graphics (TOG) **34**(4), 1–10 (2015)

[37] Jiang, C., Schroeder, C., Teran, J., Stomakhin, A., Selle, A.: The material point method for simulating continuum materials. In: Acm siggraph 2016 courses, pp. 1–52 (2016)

[38] Jiang, Y., Yu, C., Xie, T., Li, X., Feng, Y., Wang, H., Li, M., Lau, H., Gao, F., Yang, Y., et al.: Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–1 (2024)

[39] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)

[40] Klár, G., Gast, T., Pradhana, A., Fu, C., Schroeder, C., Jiang, C., Teran, J.: Drucker-prager elastoplasticity for sand animation. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)

[41] Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)

07759 (2024) [21] Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Doersch, C., Aytar, Y., Rubinstein, M., 等：运动提示：通过运动控制视频生成
轨迹。arXiv 预印本 arXiv:2412.02700 (2024) [22] Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: 力
提示：视频生成模型可以学习和泛化基于物理的控制信号。arXiv 预印本 arXiv:2505.19386 (2025) [23] Gong, S., Li, M., Feng, J., Wu, Z., Kong, L.: Diffuseq: 序列到序列文本生成
使用扩散模型。ICLR (2023) [24] Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., 等：
扩散作为着色器：用于多功能视频生成控制的 3D 感知视频扩散。arXiv 预印本 arXiv:2501.03847 (2025) [25] He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera
用于文本到视频生成的控制。arXiv 预印本 arXiv:2404.02101 (2024) [26] He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: 用于高保真度的潜在视频扩散模型
长视频生成。arXiv 预印本 arXiv:2211.13221 (2022) [27] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B.,
Norouzi, M., Fleet, D.J., 等：Imagen video：使用扩散模型的超高清视频生成。arXiv 预印本 arXiv:2210.02303 (2022) [28] Ho, J., Jain, A., Abbeel, P：去噪扩散概率模型。NeurIPS 33, 6840–6851

(2020)

[29] Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., Fleet, D.J.: 视频扩散模型。在：2022 年高度结构化数据的深度生成模型研讨会，https://openreview.net/forum?id=BBelR2NdDZ5

[30] Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character
动画。在：CVPR。第 8153-8163 页（2024 年）[31] Hu, Y., Fang, Y., Ge, Z., Qu, Z., Zhu, Y., 基于位移不连续性和双向刚体耦合的质点法。ACM Transactions on Graphics (TOG) 37(4), 1–14 (2018)

[32] Hu, Y., Li, T.M., Anderson, L., Ragan-Kelley, J., Durand, F.: Taichi: 一种用于高性能的编程语言
在空间稀疏数据结构上的性能计算。ACM Transactions on Graphics (TOG) 38(6), 1–16 (2019) [33] 黄荣，黄俊，杨东，任宇，刘亮，李明，叶子，刘杰，尹翔，赵
Z.: Make-an-audio: 文本到音频生成，使用提示增强扩散模型。发表于：机器学习国际会议。第 13916–13932 页。PMLR (2023) [34] 黄奕宏，孙宇彤，杨哲，吕翔，曹宇鹏，齐欣：Sc-gs：稀疏控制

高斯喷溅技术用于可编辑动态场景。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集。第 4220-4230 页（2024 年）[35] Jiang, C., Gast, T., Teran, J.: 各向异性弹塑性模型用于布料、针织品和毛发摩擦
联系。ACM Transactions on Graphics (TOG) 36(4), 1–14 (2017) [36] 江 C., 施罗德 C., 塞勒 A., 泰兰 J., 斯托马金 A.: 仿射粒子单元法。
ACM Transactions on Graphics (TOG) 34(4), 1–10 (2015) [37] 江晨，施罗德，特兰，斯托马金，塞尔：材料点法
模拟连续介质材料。在：ACM SIGGRAPH 2016 课程，第 1–52 页 (2016) [38] 江宇，余晨，谢涛，李翔，冯毅，王浩，李明，劳海，高飞，杨阳，等：
VR-GS：虚拟现实中的物理动力学感知交互高斯点系统。在：ACM SIGGRAPH 2024 会议论文集，第 1–1 页 (2024) [39] 基里洛夫，明图恩，拉维，毛华，罗兰，古斯塔夫森，肖天，怀特黑德，
伯格，洛维，等：Segment Anything。在：ICCV，第 4015–4026 页 (2023) [40] 克拉尔，加斯特，普拉达纳，付超，施罗德，江晨，特兰：Drucker-Prager
沙动画的弹塑性。ACM Transactions on Graphics (TOG) 35(4), 1–12 (2016) [41] Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang,

J., 等：Hunyuanvideo: 一种用于大型视频生成模型的系统框架。arXiv preprint arXiv:2412.03603 (2024)

[42] Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)

[43] Kugelstadt, T., Bender, J., Fernández-Fernández, J.A., Jeske, S.R., Löschner, F., Longva, A.: Fast corotated elastic sph solids with implicit zero-energy mode control. Proceedings of the ACM on Computer Graphics and Interactive Techniques **4**(3), 1–21 (2021)

[44] Lan, Z., Hao, Y., Zhao, M.: Guiding audio editing with audio language model. arXiv preprint arXiv:2509.21625 (2025)

[45] Lei, J., Daniilidis, K.: Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In: CVPR. pp. 6624–6634 (2022)

[46] Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6165–6177 (2025)

[47] Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM TOG **36**(6), 194–1 (2017)

[48] Li, Z., Yu, H.X., Liu, W., Yang, Y., Herrmann, C., Wetzstein, G., Wu, J.: Wonderplay: Dynamic 3d scene generation from a single image and actions. arXiv preprint arXiv:2505.18151 (2025)

[49] Liang, J., Liu, R., Ozguroglu, E., Sudhakar, S., Dave, A., Tokmakov, P., Song, S., Vondrick, C.: Dreamitate: Real-world visuomotor policy learning via video generation. arXiv preprint arXiv:2406.16862 (2024)

[50] Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV. pp. 9298–9309 (2023)

[51] Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: ECCV. pp. 360–378. Springer (2024)

[52] Liu, T., Bargteil, A.W., O'Brien, J.F., Kavan, L.: Fast simulation of mass-spring systems. ACM TOG **32**(6), 1–7 (2013)

[53] Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: CVPR. pp. 9970–9980 (2024)

[54] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)

[55] Macklin, M., Müller, M.: Position based fluids. ACM Transactions on Graphics (TOG) **32**(4), 1–12 (2013)

[56] Macklin, M., Müller, M., Chentanez, N.: Xpbd: position-based simulation of compliant constrained dynamics. In: Proceedings of the 9th International Conference on Motion in Games. pp. 49–54 (2016)

[57] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: CVPR. pp. 4460–4470 (2019)

[58] Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

[59] Modi, V., Sharp, N., Perel, O., Sueda, S., Levin, D.I.: Simplicits: Mesh-free, geometry-agnostic elastic simulation. ACM TOG **43**(4), 1–11 (2024)

[60] Müller, M., Heidelberger, B., Hennix, M., Ratcliff, J.: Position based dynamics. Journal of Visual Communication and Image Representation **18**(2), 109–118 (2007)

[61] Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Occupancy flow: 4d reconstruction by learning particle dynamics. In: ICCV. pp. 5379–5389 (2019)

[62] OpenAI: (2024), https://openai.com/index/sora

[63] Peer, A., Gissler, C., Band, S., Teschner, M.: An implicit sph formulation for incompressible linearly elastic solids. In: Computer Graphics Forum. vol. 37, pp. 135–148. Wiley Online Library (2018)

03603 (2024) [42] 孔子, 王, 黄, 赵, 卡塔扎罗: Diffwave: 一种通用的音频合成扩散模型. arXiv 预印本 arXiv:2009.09761 (2020)

[43] Kugelstadt, T., Bender, J., Fernández-Fernández, J.A., Jeske, S.R., Löschner, F., Longva, A.: 快速旋转弹性球壳固体与隐式零能量模式控制。ACM 计算机图形学与交互技术会议论文集 4(3), 1–21 (2021)

[44] Lan, Z., Hao, Y., Zhao, M.: 基于音频语言模型的音频编辑引导。arXiv 预印本 arXiv:2509.21625 (2025) [45] Lei, J., Daniilidis, K.: Cadex: 通过神经同胚学习规范变形坐标空间以实现动态表面表示. 在: CVPR. pp. 6624–6634 (2022) [46] Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: 动态高斯融合

通过 4D 运动支架从随兴视频中生成。收录于《计算机视觉与模式识别会议论文集》。第 6165-6177 页（2025 年）

[47] 李涛，博尔卡尔特，布莱克，李浩，罗梅罗：学习面部形状的模型和从 4D 扫描中获取的表达式。ACM TOG 36(6), 194–1 (2017)

[48] 李铮，余海星，刘伟，杨阳，赫尔曼，韦茨斯坦，吴俊：Wonderplay：动态从单张图像和动作生成 3D 场景。arXiv 预印本 arXiv:2505.18151 (2025) [49] Liang, J., Liu, R., Ozguroglu, E., Sudhakar, S., Dave, A., Tokmakov, P., Song, S., Vondrick

C.: Dreamitate：通过视频生成实现现实世界视运动策略学习。arXiv 预印本 arXiv:2406.16862 (2024)

[50] 刘瑞，吴瑞，Van Hoorick, B.，Tokmakov, P.，Zakharov, S.，Vondrick, C.：Zero-1-to-3：零样本单图到 3D 物体生成。In: ICCV. pp. 9298–9309 (2023)

[51] 刘思，任哲，古普塔，王思：Physgen：刚体物理基础图像到视频生成。收录于 ECCV 会议论文集，第 360-378 页。Springer 出版社（2024 年）[52] Liu, T., Bargteil, A.W., O’Brien, J.F., Kavan, L.: 快速模拟质点弹簧系统。ACM TOG 32(6), 1–7 (2013) [53] 龙晓，郭一超，林晨，刘宇，窦峥，刘磊，马岩，张书航，哈贝曼，特奥巴尔特，等：Wonder3d：单图像到 3D 的跨域扩散。在：CVPR.

pp. 9970–9980 (2024)

[54] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: 一个带皮肤的多人线性模型。在：经典图形论文：突破边界，第 2 卷，pp. 851–866 (2023)

[55] Macklin, M., Müller, M.: 基于位置的水体。ACM 图形学汇刊（TOG）32(4), 1–12 (2013) [56] 麦克林, M., 米勒, M., 申塔内兹, N.: Xpbd: 基于位置的柔顺体模拟

约束动力学。发表于《第九届游戏动画国际会议论文集》。第 49-54 页（2016 年）

[57] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: 占据网络：在函数空间中学习三维重建。发表于 CVPR，第 4460-4470 页（2019 年）

[58] Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., Ng, R.: Nerf: Repre-将场景呈现为神经辐射场用于视图合成。In: ECCV (2020) [59] Modi, V., Sharp, N., Perel, O., Sueda, S., Levin, D.I.: Simplicits: 无网格、几何无关

弹性模拟。ACM TOG 43(4), 1–11 (2024)

[60] Müller, M., Heidelberger, B., Hennix, M., Ratcliff, J.: 基于位置的动力系统。Journal of 视觉传达与图像表示 18(2), 109–118 (2007)

[61] Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Occupancy flow: 4d reconstruction by 学习粒子动力学。In: ICCV. pp. 5379–5389 (2019) [62] OpenAI: (2024), https://openai.com/index/sora

[63] Peer, A., Gissler, C., Band, S., Teschner, M.: 一种用于不可压缩流体的隐式 SPH 公式线性弹性固体。In: Computer Graphics Forum. vol. 37, pp. 135–148. Wiley Online Library (2018)

[64] Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational physics **378**, 686–707 (2019)

[65] Ram, D., Gast, T., Jiang, C., Schroeder, C., Stomakhin, A., Teran, J., Kavehpour, P.: A material point method for viscoelastic fluids, foams and sponges. In: Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 157–163 (2015)

[66] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)

[67] Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM TOG **36**(6) (2017)

[68] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS **35**, 36479–36494 (2022)

[69] Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., Battaglia, P.: Learning to simulate complex physics with graph networks. In: International conference on machine learning. pp. 8459–8468. PMLR (2020)

[70] Shi, H., Xu, H., Huang, Z., Li, Y., Wu, J.: Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks. The International Journal of Robotics Research **43**(4), 533–549 (2024)

[71] Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion. In: CVPR. pp. 20875–20886 (2023)

[72] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2020)

[73] Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. NeurIPS **32** (2019)

[74] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2020)

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

[76] Tan, X., Jiang, Y., Li, X., Zong, Z., Xie, T., Yang, Y., Jiang, C.: Physmotion: Physics-grounded dynamics from a single image. arXiv preprint arXiv:2411.17189 (2024)

[77] Tang, J., Xu, D., Jia, K., Zhang, L.: Learning parallel dense correspondence from spatio-temporal descriptors for efficient and robust 4d reconstruction. In: CVPR. pp. 6022–6031 (2021)

[78] Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: ECCV. pp. 1–18. Springer (2024)

[79] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICCV (2023)

[80] Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837 (2024)

[81] Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: ECCV. pp. 439–457. Springer (2024)

[82] Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

[83] Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)

Wiley Online Library (2018) [64] Raissi, M., Perdikaris, P., Karniadakis, G.E.: 物理信息神经网络：深度用于解决涉及非线性偏微分方程的正向和逆向问题的学习框架。计算物理杂志 378, 686–707 (2019)

[65] Ram, D., Gast, T., Jiang, C., Schroeder, C., Stomakhin, A., Teran, J., Kavehpour, P.: 一种材料点方法用于粘弹性流体、泡沫和海绵。收录于《第 14 届 ACM SIGGRAPH/Eurographics 计算机动画研讨会论文集》。第 157-163 页（2015 年）[66] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 高分辨率图像使用潜在扩散模型的合成。在：CVPR。第 10684–10695 页（2022 年）[67] Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and将物体结合在一起。ACM TOG 36(6) (2017) [68] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gon-

tijo Lopes, R., Karagol Ayan, B., Salimans, T., 等：具有深度语言理解的逼真文本到图像扩散模型。NeurIPS 35, 36479–36494 (2022)

[69] Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., Battaglia, P.: 学习使用图网络模拟复杂物理。国际机器学习会议。第 8459-8468 页。PMLR (2020)

[70] Shi, H., Xu, H., Huang, Z., Li, Y., Wu, J.: Robocraft: 学习观察、模拟和塑造具有图网络的 3D 弹塑性物体。国际机器人研究杂志 43(4), 533–549 (2024) [71] Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3D 神经场生成

使用三平面扩散。发表于 CVPR，第 20875-20886 页（2023 年）[72] 宋杰，孟超，Ermon：去噪扩散隐式模型（2020）[73] 宋阳，Ermon：通过估计数据分布的梯度进行生成建模。

NeurIPS 32 (2019) [74] 宋杨，索尔-迪克斯坦，金玛，库马尔，埃尔蒙，普尔：基于分数的通过随机微分方程的生成模型。In: ICLR (2020) [75] Stomakhin, A., Schroeder, C., Chai, L., Teran, J., Selle, A.: 一种用于雪的物质点方法模拟。ACM Transactions on Graphics (TOG) 32(4), 1–10 (2013) [76] Tan, X., Jiang, Y., Li, X., Zong, Z., Xie, T., Yang, Y., Jiang, C.: Physmotion: 物理基础

从单张图像中生成动力学。arXiv 预印本 arXiv:2411.17189 (2024) [77] Tang, J., Xu, D., Jia, K., Zhang, L.: 从时空信息中学习并行密集对应关系。用于高效和鲁棒的 4D 重建的时间描述符。In: CVPR. pp. 6022–6031 (2021) [78] Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: 大多视角高斯用于高分辨率 3D 内容创建的模型。In: ECCV. pp. 1–18. Springer (2024) [79] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion扩散模型。In: ICCV (2023) [80] Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: 扩散模型是实时游戏

引擎。arXiv 预印本 arXiv:2408.14837 (2024) [81] Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach,

R., Jampani, V.: Sv3d: 基于单图像的新型多视图合成与 3D 生成技术，使用潜在视频扩散。发表于 ECCV，第 439-457 页。Springer 出版社（2024 年）[82] Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J.,

王, J., 张, J., 周, J., 王, J., 陈, J., 朱, K., 赵, K., 严, K., 黄, L., 冯, M., 张, N., 李, P., 吴, P., 崔 R., 冯, R., 张, S., 孙, S., 方, T., 王, T., 谷, T., 温, T., 沈, T., 林, W., 王, W., 王, W., 周, W., 王, W., 沈, W., 鱼, W., 石, X., 黄, X., 徐, X., 口, Y., 吕, Y., 李, Y., 刘, Y., 王, Y., 张, Y., 黄, Y., 李, Y., 吴, Y., 刘, Y., 潘 Y., 郑, Y., 红, Y., 石, Y., 冯, Y., 姜, Z., 韩, Z., 吴, Z.F., 刘, Z.: Wan: 开放式和高级大规模视频生成模型. arXiv 预印本 arXiv:2503.20314 (2025)

[83] 王, J., 陈, M., 卡拉夫, N., 维达尔迪, A., 罗普雷希特, C., 纳沃特尼, D.: Vggt: 视觉几何尝试使用 grounded transformer。发表于《计算机视觉与模式识别会议论文集》，第 5294–5306 页（2025）

[84] Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV. pp. 55–72. Springer (2024)

[85] Wang, Y., Tang, S., Chu, M.: Physics-informed learning of characteristic trajectories for smoke reconstruction. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)

[86] Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14749–14759 (2024)

[87] Wang, Z., Lan, Y., Zhou, S., Loy, C.C.: Objctrl-2.5 d: Training-free object control with camera poses. arXiv preprint arXiv:2412.07721 (2024)

[88] Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. arXiv preprint arXiv:2411.18613 (2024)

[89] Wu, W., Li, Z., Gu, Y., Zhao, R., He, Y., Zhang, D.J., Shou, M.Z., Li, Y., Gao, T., Zhang, D.: Draganything: Motion control for anything using entity representation. In: ECCV. pp. 331–348. Springer (2024)

[90] Xie, T., Zhao, Y., Jiang, Y., Jiang, C.: Physanimator: Physics-guided generative cartoon animation. arXiv preprint arXiv:2501.16550 (2025)

[91] Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In: CVPR. pp. 4389–4398 (2024)

[92] Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: ECCV. pp. 399–417. Springer (2024)

[93] Yang, C., Gao, W., Wu, D., Wang, C.: Learning to simulate unseen physical systems with graph neural networks. arXiv preprint arXiv:2201.11976 (2022)

[94] Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)

[95] Zhang, K., Li, B., Hauser, K., Li, Y.: Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation. In: Proceedings of Robotics: Science and Systems (RSS) (2024)

[96] Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. ACM TOG **43**(4), 1–20 (2024)

[97] Zhang, T., Yu, H.X., Wu, R., Feng, B.Y., Zheng, C., Snavely, N., Wu, J., Freeman, W.T.: Physdreamer: Physics-based interaction with 3d objects via video generation. In: ECCV. pp. 388–406. Springer (2024)

[98] Zhang, X., Li, N., Dai, A.: Dnf: Unconditional 4d generation with dictionary-based neural fields. arXiv preprint arXiv:2412.05161 (2024)

[99] Zhong, L., Yu, H.X., Wu, J., Li, Y.: Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In: ECCV. pp. 407–423. Springer (2024)

[100] Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In: ECCV. pp. 18–38. Springer (2024)

[101] Zhu, S., Chen, J.L., Dai, Z., Dong, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In: ECCV. pp. 145–162. Springer (2024)

[102] Zienkiewicz, O.C., Taylor, R.L., Nithiarasu, P., Zhu, J.: The finite element method, vol. 3. Elsevier (1977)

[103] Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: CVPR. pp. 6365–6373 (2017)

[84] 王晓，朱志，黄刚，陈翔，朱军，陆进：Drivedreamer：迈向真实世界
用于自动驾驶的驱动世界模型。发表于 ECCV，第 55-72 页。Springer 出版社（2024 年）

[85] Wang, Y., Tang, S., Chu, M.: 基于物理信息的烟雾特征轨迹学习
重建。载于 ACM SIGGRAPH 2024 会议论文集。第 1-11 页（2024 年）

[86] Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual
使用世界模型进行自动驾驶的预测和规划。在：IEEE/CVF 计算机视觉与模式识别会议论文集。
第 14749-14759 页 (2024)

[87] Wang, Z., Lan, Y., Zhou, S., Loy, C.C.: Objctrl-2.5 d: 无需训练的对象控制与相机
姿势。arXiv 预印本 arXiv:2412.07721 (2024)

[88] Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d:
使用多视角视频扩散模型在 4D 中创造任何事物。arXiv 预印本 arXiv:2411.18613（2024）

[89] 吴伟，李铮，顾宇，赵锐，何垚，张东建，寿茂智，李阳，高天，张东：Draganything：使用实体表示的任意运动控制。在：ECCV。pp.
331–348. Springer (2024)

[90] 谢涛，赵宇，蒋宇，蒋晨：Physanimator：物理引导的生成卡通
动画。arXiv 预印本 arXiv:2501.16550 (2025)

[91] 谢涛，宗志，邱宇，李翔，冯毅，杨阳，蒋晨：Physgaussian：物理
集成 3D 高斯用于生成动态。发表于 CVPR，第 4389-4398 页（2024 年）

[92] Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: 用视频扩散先验动态生成开放域图像。在：ECCV.
第 399–417 页。Springer (2024)

[93] 杨晨，高伟，吴东，王晨：学习模拟未见过的物理系统
图神经网络。arXiv 预印本 arXiv:2201.11976 (2022)

[94] 杨哲，滕静，郑伟，丁明，黄思，徐杰，杨阳，洪伟，张翔，
冯, G., 等: Cogvideox: 基于专家变压器的文本到视频扩散模型. arXiv 预印本 arXiv:2408.06072 (2024)

[95] 张凯，李博，Hauser，李阳：Adaptigraph：材料自适应图神经网络
机器人操作动力学。载于《机器人：科学与系统》（RSS）（2024）

[96] Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A
用于创建高质量 3D 资产的可控大规模生成模型。ACM TOG 43(4), 1–20 (2024)

[97] 张涛，余海星，吴睿，冯博宇，郑超，斯奈弗利，吴俊，弗里曼：Physdreamer：通过视频生成实现基于物理的 3D 物体交互。在：ECCV。pp.
388–406. Springer (2024)

[98] 张翔，李娜，戴安：Dnf：基于词典的无条件 4D 生成
fields. arXiv 预印本 arXiv:2412.05161 (2024)

[99] Zhong, L., Yu, H.X., Wu, J., Li, Y.: 弹性物体的重建与模拟
弹簧-质量 3D 高斯。发表于 ECCV，第 407-423 页。Springer 出版社（2024 年）

5294–5306 (2025) [100] 周伟，窦志，曹铮，廖铮，王杰，王伟，刘岩，小仓隆，王伟，刘岩
L.: Emdm: 高效运动扩散模型，用于快速生成高质量运动。发表于：ECCV。第 18-38 页。
Springer 出版社（2024 年）

[101] Zhu, S., Chen, J.L., Dai, Z., Dong, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: 基于三维参数引导的可控且一致的人像动画。发表于：ECCV。
第 145-162 页。Springer 出版社（2024 年）

[102] Zienkiewicz, O.C., Taylor, R.L., Nithiarasu, P., Zhu, J.: 有限元法，第三卷。
Elsevier (1977)

[103] Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: 建立动物的 3d 形状和
姿态。在：CVPR。第 6365–6373 页 (2017)

The supplementary material covers the following sections: Implementation Details Section A, User study Section B, Physics Parameter Estimation Section C, Results Section D, Societal Impacts Section E, Data and Model Safeguards Section F. **We also encourage readers to refer to our supplementary videos for demonstrations of animatable results.**.

A Implementation Details

Dataset. To make our model handle diverse objects and motion trajectories, we generate data using physics simulation using high-quality 3D objects selected from ObjaverseXL [16, 15]. We simulate animations for each object with the MPM simulator [37] as the ground-truth. We use a fixed number of simulated points $N = 2048$ (uniformly sampled on the faces of the mesh) and frames $F = 24$ to align with our model's input. For data augmentation, we randomly rotate the object around y-axis and add noise $\epsilon_p^{aug} \sim \mathcal{N}(0, 0.01^2)$ to each sampled initial point. Our whole dataset contains 550K objects, including 150K elastic objects of different drag force directions, 100K objects of gravity across elastic, sand, plasticine and rigid respectively. For the simulated animation of varying drag force, we randomly sample a constant force $\mathbf{f}$, a drag point $\mathbf{D} \in \mathbf{P}_0$ and physics parameters $E \in [10^4, 10^7]$, $\nu \in [0.05, 0.45]$. The force $\mathbf{f}$ has an outward direction of the object surface and a magnitude between $0.02G$ and $0.3G$ in total (G is the gravity of the whole object) and is only applied to points close to the drag point $\mathbf{D}$.

Training For metric comparison and ablation, we train our base model on the 150K elastic subset that contains different force and physical parameters with 6 layers and 256 latent size on 8 NVIDIA L40 GPUs with 48GB GPU memory for 60K iterations with a total batch size of 32, which takes about 30 hours. We randomly leave out 100 animations from this dataset as the test set and keep the remaining ones for training. We train a large model of different materials with 12 layers and a 512 latent size on all 550K data with the same iterations and batch size, which takes about 80 hours. We use AdamW optimizer with betas (0.9, 0.999) and a learning rate of 1e-4 with a cosine schedule and a warmup of 100 steps. We clip the gradient with the maximum norm of 1.0 and train with bfloat16 precision. We use a DDIM scheduler for sampling. For 25 diffusion steps, it takes 1s and 3s for the base and large model. For 4 diffusion steps, it takes 0.13s and 0.48s for the base and large model. We found that 4 steps can already achieve great results due to the low uncertainty of the model.

Image-to-3D Pipeline We use SAM [39] to segment the object in the input image and run SV3D [81] to generate 20 novel-view images of that object with orbit camera poses, from which we pick three images with azimuth (90°, 180°, 270°) relative to the input and send them together with the input into LGM [78] for 3D Gaussian reconstruction. We then convert the 3D Gaussians to a plain point cloud and sample N points using farthest point sampling (FPS) for trajectory generation.

GPT-4o Evaluation We prompt GPT-4o with the following prompt to use it for evaluation (Results might vary with GPT updates):

```
You are tasked with evaluating the quality of image-to-video generation
produced by a model.
For each test case, you will be given: 1. A text prompt describing
a single object and a force applied to it. The force's position
and direction are visualized as a red arrow in the input image. 2.
An input image of the object. 3. Five sets of 10 evenly spaced
frames-each set corresponds to a video generated by a different model
from the same input.
Please evaluate this video based on the following three criteria using a
5-point Likert scale (1 = poor, 5 = excellent):
- Semantic Adherence: How well the content and motion in the video
match the description in the text prompt, especially the alignment with
the force direction and position. Note that the video should starts
with the input image.
- Physical Commonsense: Whether the object's motion follows intuitive,
physically plausible dynamics given the applied force direction and
position.
```

6365–6373 (2017) 补充材料包括以下部分：实现细节部分 A、用户研究部分 B、物理参数估计部分 C、结果部分 D、社会影响部分 E、数据和模型保护部分 F。我们还鼓励读者参考我们的补充视频，以展示可动画化的结果演示。

A 实现细节

数据集。为了使我们的模型能够处理多样化的物体和运动轨迹，我们使用从 ObjaverseXL [16, 15]中选取的高质量 3D 物体，通过物理模拟生成数据。我们使用 MPM 模拟器[37]作为真实情况，对每个物体进行动画模拟。我们使用固定数量的模拟点 N = 2048（在网格面上均匀采样）和帧数 F = 24，以匹配我们模型的输入。对于数据增强，我们随机围绕 y 轴旋转物体，并向每个采样的初始点添加噪声ϵp ~ N (0, 0.01)。我们的整个数据集包含 550K 个物体，包括 150K 个具有不同阻力方向弹性物体，100K 个分别具有弹性、沙子、油泥和刚性的重力物体。对于模拟具有不同阻力的动画，我们随机采样一个恒定力 f，一个阻力点 D ∈ P，以及物理参数 E ∈ [10, 10], ν ∈ [0.05, 0.45]。力 f 的方向为物体表面的外向方向，大小在 0.02G 和 0.3G 之间（G 是整个物体的重力），并且只施加在靠近阻力点 D 的点上。

为了进行指标比较和消融实验，我们在包含不同力和物理参数的 150K 弹性子集上训练基础模型，使用 6 层和 256 个潜在大小的设置，在 8 块 NVIDIA L40 GPU（每块 48GB 显存）上进行 60K 次迭代，总批大小为 32，耗时约 30 小时。我们从该数据集中随机选出 100 个动画作为测试集，其余用于训练。我们使用相同的迭代次数和批大小，在全部 550K 数据上训练一个包含不同材料的 12 层 512 个潜在大小的模型，耗时约 80 小时。我们采用 AdamW 优化器，设置 beta 值为(0.9, 0.999)，学习率为 1e-4，使用余弦调度和 100 步的预热。梯度使用最大范数 1.0 进行裁剪，并以 bfloat16 精度进行训练。我们使用 DDIM 调度器进行采样。对于 25 步扩散，基础模型和大型模型分别耗时 1 秒和 3 秒。对于 4 步扩散，基础模型和大型模型分别耗时 0.13 秒和 0.48 秒。我们发现 4 步扩散已经能取得很好的效果，因为模型的方差较低。

图像到 3D 流程 我们使用 SAM [39] 对输入图像中的对象进行分割，并运行 SV3D [81] 生成该对象的三十个新视角图像，其中我们选择三个相对于输入的方位角（90, 180, 270）的图像，并将它们连同输入一起发送到 LGM [78] 进行 3D 高斯重建。然后我们将 3D 高斯转换为普通点云，并使用最远点采样（FPS）采样 N 个点用于轨迹生成。

GPT-4o 评估 我们使用以下提示来提示 GPT-4o 进行评估（结果可能因 GPT 更新而变化）：

您需要评估模型生成的图像到视频的质量。对于每个测试案例，您将获得：1. 描述单个物体及其所受力文本提示。力的位置和方向在输入图像中用红色箭头可视化。2. 物体的输入图像。3. 五组 10 个均匀分布的帧——每组对应由不同模型从相同输入生成的视频。请使用 5 分李克特量表（1=差，5=优秀）根据以下三个标准评估该视频：- 语义一致性：视频中的内容和运动与文本提示的描述匹配程度，特别是与力的方向和位置的符合程度。注意视频应从输入图像开始。- 物理常识：物体的运动是否遵循直观的、符合物理的动力学规律，考虑到所施加的力的方向和位置。

Table 5: Results of user study.

	Ours	CogVideoX	Wan	DragAnything	ObjCtrl2.5D
Physics Plausibility	**81.0%**	5.5%	10.2%	1.2%	2.1%
Video Quality	**66.0%**	6.2%	18.3%	4.5%	5.0%

```
- Video Quality:  The overall visual and temporal quality of the video
(note that static or nearly-static sequences are less preferred).
Provide your evaluation for each video strictly in the following
one-line format:
Video i, Semantic Adherence score, Physical Commonsense score, Video
Quality score
```

B User Study

We conducted a user study to evaluate the physics plausibility and overall quality of the videos generated by our model and other baselines. The study consisted of 12 questions, each including an input image with the force location and direction marked on the image, a text prompt describing the image and applied force, and generated video results produced by five different methods. The users are asked to carefully observe the videos and evaluate them from two aspects: (1) **Physics plausibility**: select the one that best matches the force direction (red arrow) and corresponding text prompt. The force and text prompt are assumed to match each other. (2) **Overall Video quality**: Select the one that has the best visual and temporal quality.

We received a total of 35 responses (35×12) and computed the percentage of times each method was selected as the best-performing video for each question. The results are summarized in Table 5, showing the preference rates for each method. The findings indicate that our model consistently outperforms baseline methods in terms of both physics plausibility and video quality. Although Wan received the second-best video quality, some of these high-quality videos suffer from low physics plausibility.

C Physics Parameter Estimation

Our trained trajectory generation model learns the conditional distribution of physically plausible motion trajectories, so it can also be used for inverse problems, *i.e*, to estimate the condition c given ground truth trajectories $\mathcal{P}$. The intuition is that **a c that is closer to the ground truth will introduce less discrepancy between the denoised trajectories and ground truth trajectories**. To this end, we define an energy function that measures how well the model can denoise a noisy version of $\mathcal{P}_t$ under that condition:

$$\mathcal{E}(c) = \mathbb{E}_{t\sim[1,T]} \|\mathcal{P}_t - \mathcal{D}(\mathcal{P}_t; t, c)\|^2, \tag{11}$$

During optimization, the denoiser $\mathcal{D}$ is frozen and only c is optimizable. We add random noise to the ground truth trajectory and feed it into the trained network to denoise. The gradient of the energy function will be backpropagated to optimize c.

We simulate 15 trajectories for elastic materials to test our physics parameter estimation pipeline. We compare our method with differentiable MPM [32], which needs to accumulate gradients over hundreds of substeps for one backward pass (costing more than 3min compared to 0.1s for ours). Table 6 shows that our method only takes about 2 minutes while achieving relatively good results, which also **demonstrates that our trained diffusion model captures physics-plausible motion trajectories**.

D More Results

More results of our method and baseline comparisons can be found in Figure 9. **We strongly encourage the readers to look at our video for better comparison, as isolated frames cannot fully represent the physical dynamics well.**

表 5：用户研究的结果。

	我们的	CogVideoX	Wan	DragAnything	ObjCtrl2.5D
物理合理性	81.0%	5.5%	10.2%	1.2%	2.1%
视频质量	66.0%	6.2%	18.3%	4.5%	5.0%

- 视频质量：视频的整体视觉和时序质量（注意：静态或接近静态的序列不太受欢迎）。请严格使用以下单行格式为每个视频提供您的评估：视频 i，语义一致性得分，物理常识得分，视频质量得分

B 用户研究

我们进行了一项用户研究，以评估我们模型和其他基线生成的视频的物理合理性和整体质量。该研究包括 12 个问题，每个问题都包含一个输入图像，图像上标明了力的位置和方向，一个描述图像和施加力的文本提示，以及由五种不同方法生成的视频结果。要求用户仔细观察视频，从两个方面进行评估：(1) 物理合理性：选择最符合力的方向（红色箭头）和相应文本提示的那个。假设力和文本提示是匹配的。(2) 整体视频质量：选择视觉和时序质量最好的那个。

我们共收到 35 份回复（35 × 12），并计算了每个方法被选为每个问题最佳视频的百分比。结果汇总在表 5 中，显示了每种方法的偏好率。研究结果表明，我们的模型在物理合理性和视频质量方面始终优于基线方法。尽管 Wan 获得了第二好的视频质量，但其中一些高质量视频存在物理合理性较低的问题。

C 物理参数估计

我们训练的轨迹生成模型学习物理上合理的运动轨迹的条件分布，因此它也可以用于逆问题，即根据真实轨迹 P 估计条件 c。其原理是，一个更接近真实轨迹的 c 会减少去噪轨迹与真实轨迹之间的差异。为此，我们定义了一个能量函数，用于衡量模型在给定条件下对 P 的噪声版本进行去噪的效果：

$$E(P, c) = \mathbb{E}_{k,\epsilon}\left[\left\| D(P_k, k, c) - P \right\|^2\right], \quad P_k = \sqrt{\bar{\alpha}_k} P + \sqrt{1-\bar{\alpha}_k}\,\epsilon, \tag{11}$$

在优化过程中，去噪器 D 被冻结，仅 c 可优化。我们对真实轨迹添加随机噪声，并将其输入训练好的网络进行去噪。能量函数的梯度将被反向传播以优化 c。

我们模拟了 15 条弹性材料的轨迹来测试我们的物理参数估计流程。我们将我们的方法与可微分的 MPM [32] 进行比较，后者需要在数百个子步中累积梯度以进行一次反向传递（相比之下，我们的方法只需 0.1 秒，而其耗时超过 3 分钟）。表 6 显示，我们的方法仅需约 2 分钟即可获得相对较好的结果，这也证明了我们训练的扩散模型能够捕捉物理合理的运动轨迹。

D 更多结果

我们方法及基线比较的更多结果可以在图 9 中找到。我们强烈
鼓励读者观看我们的视频进行更好的比较，因为单独的帧无法很好地代表物理动态。

Table 6: Mean Absolute Error (MAE) of Young's Modulus on physics parameter estimation.

Method	Runtime (min.)	MAE of $\log_{10}(E)$
Ours	2	0.506
Diff. MPM (5 iters)	20	0.439
Diff. MPM (15 iters)	60	0.394

E Societal Impacts

Positive Impacts Our method integrates physically grounded simulation signals into video generative models, offering new avenues for controllable and physically plausible video synthesis. These can support people from amateurs to filmmakers and designers in rapidly prototyping ideas with accurate physical behavior, democratizing access to high-fidelity visual tools.

Negative Impacts High-fidelity generative models, especially when conditioned on physical signals, may be misused for creating deceptive content such as realistic yet fabricated disaster footage or physically plausible fake videos. This poses risks for misinformation and erosion of public trust. Although our approach enhances physical plausibility, it is important to note that the generated outputs are not real-world occurrences.

F Data and Model Safeguards

Given the dual-use nature of video generation models, we recognize that our pretrained model could be misused to generate deceptive, physically plausible videos for misinformation. As such, we will implement appropriate safeguards to support controlled access when we release our model, including: (1) requiring users to agree to usage guidelines and restrictions, (2) distributing the model under a research-only license, (3) investigating automatic safety filters that can flag potentially harmful uses. These steps aim to reduce the risk of malicious or unintended applications while still supporting reproducible research.

Our training data consists exclusively of synthetic point cloud trajectories representing object motion under simulated physics. These datasets contain no images, videos, or human-related content, and thus should pose no risk of visual misinformation, privacy violations, or unsafe content. All point clouds are generated in simulation environments and contain only geometric and physical information about object movement.

表 6：杨氏模量在物理参数估计中的平均绝对误差（MAE）

方法	运行时间（分钟）	对数期望的均方误差
我们	2	0.506
差异 MPM（5 次迭代）	20	0.439
差异 MPM（15 次迭代）	60	0.394

E 社会影响

积极影响我们的方法将物理上合理的模拟信号集成到视频生成模型中，为可控且物理上合理的视频合成提供了新的途径。这些可以支持从业余爱好者到电影制作人及设计师的人们快速原型化具有精确物理行为的想法，使人们更容易获得高保真视觉工具。

负面影响高保真生成模型，特别是当它们基于物理信号进行条件化时，可能会被误用于创建欺骗性内容，例如逼真但虚构的灾难画面或物理上合理的假视频。这给虚假信息和公众信任的侵蚀带来了风险。尽管我们的方法增强了物理合理性，但重要的是要指出，生成的输出并非真实世界的现象。

F 数据和模型安全措施

鉴于视频生成模型的军民两用特性，我们认识到我们的预训练模型可能被误用于生成具有误导性的、物理上可信的视频。因此，在发布模型时，我们将实施适当的保护措施以支持受控访问，包括：（1）要求用户同意使用指南和限制，（2）以仅供研究许可证的形式分发模型，（3）研究可以标记潜在有害用途的自动安全过滤器。这些步骤旨在降低恶意或非预期应用的风险，同时仍然支持可复现的研究。

我们的训练数据完全由表示物体在模拟物理条件下运动的合成点云轨迹组成。这些数据集不包含图像、视频或与人类相关的内容，因此不应存在视觉误导、隐私侵犯或不安全内容的风险。所有点云都是在模拟环境中生成的，并且只包含关于物体运动的几何和物理信息。

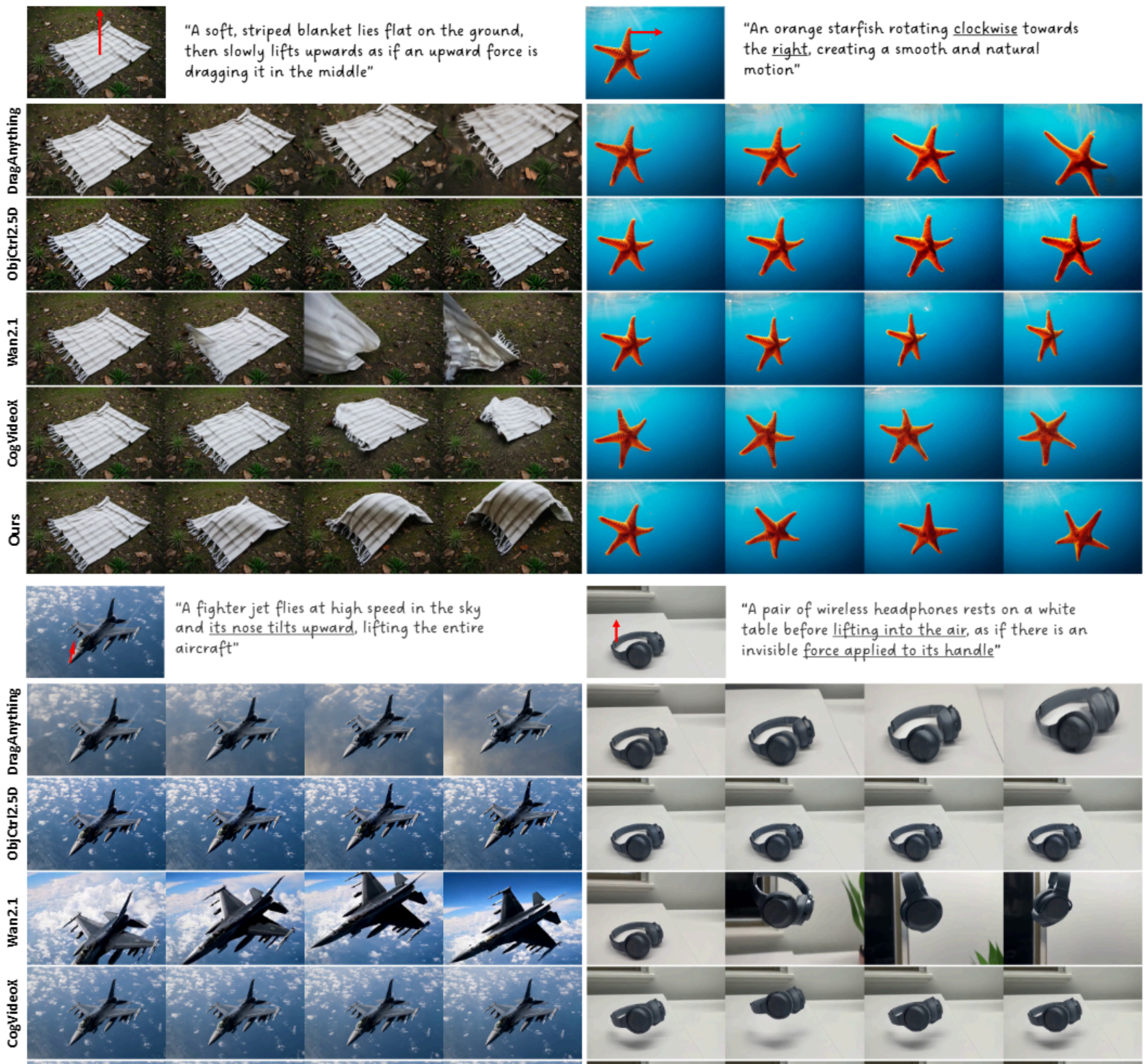

Figure 9: More qualitative comparison between our method and baselines.

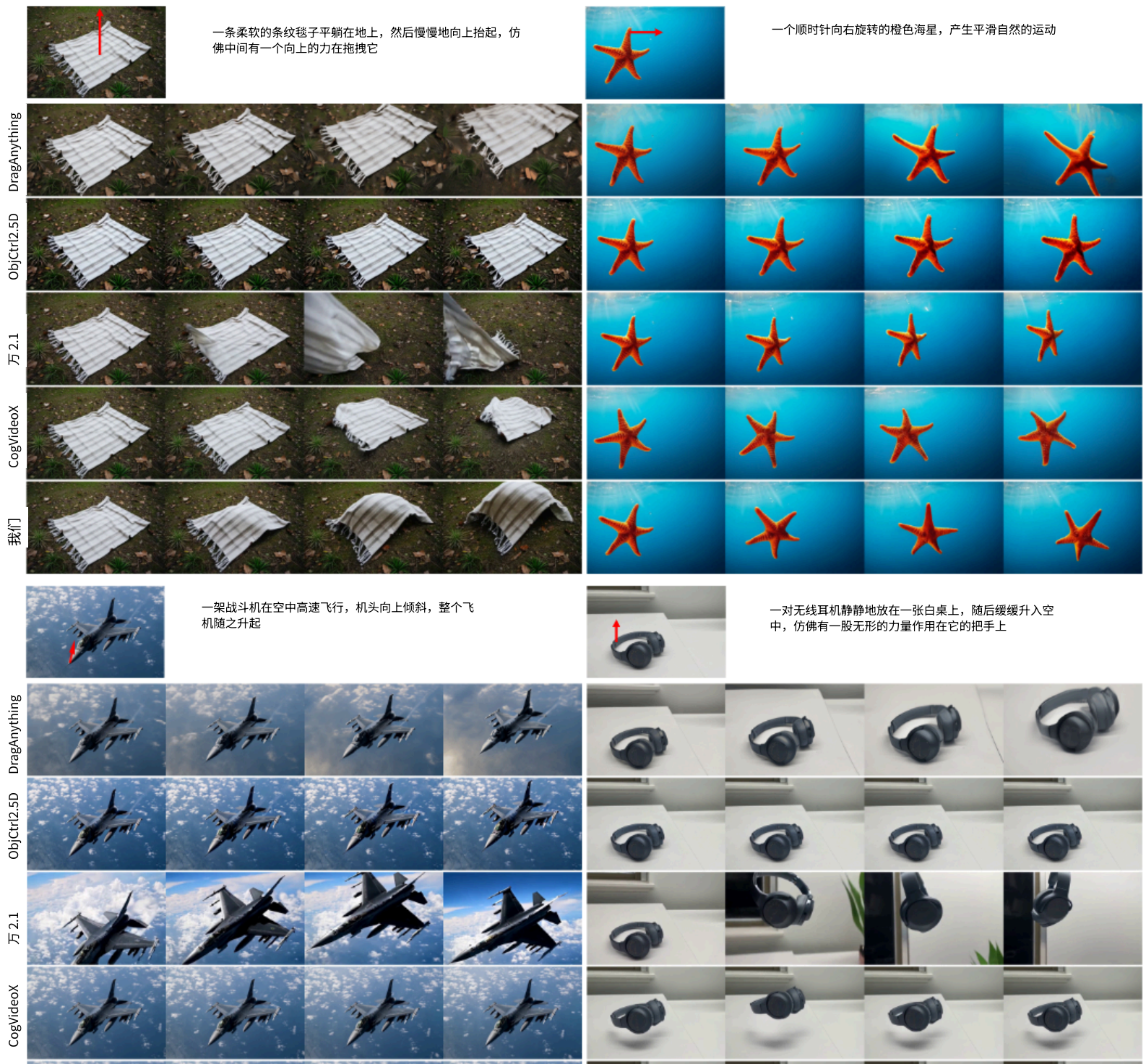

图 9：我们方法与基线方法的更多定性比较。

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims are validated by our quantitative and qualitative results in the experimental results section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

NeurIPS 论文检查清单

该清单旨在鼓励负责任的机器学习研究，解决可重复性、透明度、研究伦理和社会影响等问题。不要删除
清单：未包含清单的论文将被直接拒稿。清单应遵循参考文献，并遵循（可选的）补充材料。清单不计入页数限制。
请仔细阅读清单指南，了解如何回答这些问题。对于清单中的每个问题：

- 您应该回答 [是] ， [否] ，或 [不适用] 。
- [NA] 表示该问题对于特定论文不适用，或相关信息不可用。
- 请在答案后提供简短（1-2 句）的说明（即使对于“不适用”的情况）。

检查清单的答案是你论文提交的重要组成部分。它们对
审稿人、领域主席、高级领域主席和伦理审稿人。您将被要求在最终修改后将其（与论文的最终版本一起）提交，其最终版本将随论文一同发表。

审稿人将被要求使用检查清单作为他们评估中的一个因素。虽然“[是]”通常比“[否]”更可取，但如果给出适当的理由（例如，“由于计算成本过高，未报告误差线”或“我们无法找到所使用数据集的许可证”），回答“[否]”也是完全可以接受的。通常情况下，回答“[否]”或“[NA]”并不构成拒绝的理由。虽然问题以二元方式提出，但我们承认真正的答案往往更加复杂，因此请尽量使用自己的最佳判断并写明理由进行阐述。所有支持证据可以出现在正文中或补充材料中，以附录形式提供。如果你对某个问题回答[是]，请在理由中指明可以在哪些部分找到与该问题相关的内容。

重要提示：

- 删除这个指令块，但保留部分标题“NeurIPS 论文检查清单”
- 保留清单子部分标题、问题和答案以及下面的指南。
- 不要修改问题，仅使用提供的宏来回答。

1. 权利要求

   问题：摘要和引言中提出的主要权利要求是否准确反映了论文的贡献和范围？答案：[是] 论据：我们的权利要求通过实验结果部分中的定量和定性结果得到了验证。指南：

   - 答案为 NA 表示摘要和引言中未包含论文中提出的内容。
   - 摘要和/或引言应明确陈述提出的内容，包括论文中做出的贡献以及重要的假设和局限性。对此问题的答案为“否”或 NA 将不会受到审稿人的好评。
   - 所提出的论点应与理论和实验结果相符，并反映结果在其他设置中泛化的程度。
   - 只要明确这些目标并非由论文达成，将其作为动机是可以接受的。

2. 局限性

   问题：论文是否讨论了作者所做工作的局限性？答案：[是]

Justification: We provided our limitations in our last section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include the necessary details in the supplemental and will release the training code and checkpoints for reproducibility.

   Guidelines:

理由：我们在上一节中提供了我们的局限性。指南：

- 答案为 NA 表示该论文没有限制，而答案为 No 表示该论文存在限制，但这些限制在论文中并未讨论。
- 鼓励作者在论文中创建一个单独的"局限性"部分。
- 论文应当指出任何强假设，以及结果对违反这些假设的鲁棒性（例如，独立性假设、无噪声环境、模型良好规范、渐近近似仅在局部成立）。作者应当反思这些假设在实际中可能如何被违反，以及其可能带来的影响。

- 作者应反思所提出的声明范围，例如，如果该方法仅在少数数据集或少数运行中进行了测试。一般来说，实证结果往往依赖于隐含的假设，这些假设应该被明确说明。

- 作者应该反思影响方法性能的因素。例如，当图像分辨率低或图像在低光照条件下拍摄时，人脸识别算法可能会表现不佳。或者，语音转文本系统可能无法可靠地用于为在线讲座提供字幕，因为它无法处理专业术语。

- 作者应该讨论所提出算法的计算效率以及它们如何随数据集大小扩展。

- 如果适用，作者应讨论其方法在解决隐私和公平问题方面的可能局限性。

- 尽管作者可能担心完全诚实地提及局限性可能会被审稿人作为拒绝的理由，但更糟糕的结果可能是审稿人发现了论文中未提及的局限性。作者应运用自己的最佳判断，并认识到支持透明度的个人行为在建立维护社区诚信的规范中发挥着重要作用。审稿人将特别被指示不要因诚实提及局限性而进行惩罚。

3. 理论假设与证明

   问题：对于每项理论结果，论文是否提供了完整的假设集和完整（且正确）的证明？回答：[NA] 说明：我们的论文不包含理论结果。指南： • 回答 NA 表示论文不包含理论结果。

- 论文中的所有定理、公式和证明都应编号并相互引用。

- 所有假设都应在定理陈述中明确说明或引用。
- 证明可以出现在正文中或补充材料中，但如果出现在补充材料中，作者应提供简短的证明概要以提供直观理解。

- 相反，论文主体中提供的任何非正式证明都应通过附录或补充材料中的正式证明进行补充。
- 证明所依赖的定理和引理应正确引用。

4. 实验结果可复现性

   问题：论文是否充分披露了所有必要信息，以便在影响论文主要论点和/或结论的程度上复现论文的主要实验结果（无论代码和数据是否提供）？回答：[是] 说明：我们在补充材料中包含了必要细节，并将发布训练代码和检查点以确保可复现性。指南：

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We don't provide the code during submission. We will open-source the code, model and checkpoints after acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- 答案 NA 表示该论文不包含实验。
- 如果论文包含实验，对这个问题没有答案将不会给审稿人留下好印象：无论是否提供代码和数据，使论文可复现都很重要。

- 如果贡献是一个数据集和/或模型，作者应该描述他们为使结果可复现或可验证所采取的步骤。
- 根据贡献的不同，可重复性可以通过多种方式实现。例如，如果贡献是一个新架构，充分描述该架构可能就足够了；如果贡献是一个特定的模型和经验评估，可能需要让其他人使用相同的数据集复制该模型，或者提供模型访问权限。一般来说，发布代码和数据通常是实现这一目标的一种好方法，但可重复性也可以通过提供详细的复制结果说明、访问托管模型（例如，对于大型语言模型）、发布模型检查点或其他适合所进行研究的手段来提供。

- 虽然 NeurIPS 不要求公开代码，但该会议要求所有投稿提供某种合理的可复现性途径，这可能取决于贡献的性质。例如（a）如果贡献主要是新算法，论文应明确说明如何复现该算法。（b）如果贡献主要是新模型架构，论文应清晰完整地描述该架构。（c）如果贡献是新模型（例如大型语言模型），则应有途径访问该模型以复现结果，或提供复现模型的方法（例如使用开源数据集或构建数据集的说明）。（d）我们认识到在某些情况下复现性可能很复杂，此时作者欢迎描述他们提供的具体复现方法。

对于闭源模型，可能存在某种方式限制对模型的访问（例如，仅限注册用户），但其他研究人员应该有可能通过某种途径来复现或验证结果。

5. 开放数据和代码的访问

   问题：论文是否向公众开放数据和代码，并提供足够的说明来忠实地重现补充材料中描述的主要实验结果？

   回答：[否]

   理由：我们在提交时不会提供代码。我们将在论文被接受后开源代码、模型和检查点。

   指南：

   - 答案 NA 表示该论文不包含需要代码的实验。
   - 请参阅 NeurIPS 代码和数据提交指南（https://nips.cc/public/guides/CodeSubmissionPolicy）以获取更多详细信息。
   - 虽然我们鼓励发布代码和数据，但我们理解这可能并不可行，因此“否”是一个可接受的答案。除非这涉及到贡献的核心内容（例如，为新的开源基准），否则论文不能仅仅因为不包含代码而被拒绝。

   - 说明应包含运行以复现结果所需的精确命令和环境。有关详细信息，请参阅 NeurIPS 代码和数据提交指南（https: //nips.cc/public/guides/CodeSubmissionPolicy）。

   - 作者应提供数据访问和准备说明，包括如何访问原始数据、预处理数据、中间数据和生成数据等。
   - 作者应提供脚本以复现新提出的方法和基线的所有实验结果。如果只有部分实验可复现，他们应在脚本中说明哪些实验被省略以及原因。

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the creation of datasets, the training, and testing details in the paper and supplemental.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiments are computationally expensive (8x Nvidia L40 for nearly two days per experiment) to be run multiple times for error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the number of GPUs and running time for each experiment in the supplemental.

- 提交时，为保护匿名性，作者应发布匿名版本（如适用）。

- 建议在补充材料（附加在论文中）提供尽可能多的信息，但允许包含数据和代码的 URL 链接。

6. 实验设置/细节

   问题：论文是否详细说明了理解结果所需的全部训练和测试细节（例如，数据分割、超参数、选择方法、优化器类型等）？
   答案：[是]
   理由：我们在论文和补充材料中提供了数据集创建、训练和测试的详细情况。
   指南：• 答案 NA 表示论文不包含实验。

- 实验设置应在论文的核心部分以必要的详细程度呈现，以便读者能够理解结果并对其有所领会。

- 完整细节可以通过代码、附录或作为补充材料提供。

7. 实验统计显著性

   问题：论文是否适当且正确地报告了误差线或其他关于实验统计显著性的适当信息？
   答案：[否]
   理由：我们的实验计算成本很高（每个实验需要 8x Nvidia L40 近两天），无法多次运行以获取误差线。
   指南：• 答案 NA 表示论文不包含实验。

- 如果结果伴随着误差线、置信区间或统计显著性检验（至少对于支持论文主要论点的实验），作者应回答"是"。

- 误差线所捕捉的可变因素应明确说明（例如，训练/测试分割、初始化、某些参数的随机抽取，或给定实验条件下的整体运行）。

计算误差线的方法应进行说明（封闭形式公式、调用库函数、自助法等）• 所做的假设应给出（例如，误差正态分布）。

- 应明确误差线是标准差还是均值的标准误差。

- 报告 1-西格玛误差线是可以的，但应说明。如果误差正态性的假设未得到验证，作者最好报告 2-西格玛误差线，而不是说明他们有一个 96%的置信区间。

- 对于非对称分布，作者应谨慎避免在表格或图中展示对称误差线，因为这样会导致结果超出范围（例如负误差率）。

- 如果表格或图中报告了误差线，作者应在文中解释其计算方法，并在文中引用相应的图表或表格。

8. 实验计算资源

   问题：对于每个实验，论文是否提供了足够的信息来复现实验所需的计算机资源（计算工作类型、内存、执行时间）？ 答案：[是] 理由：我们在补充材料中报告了每个实验的 GPU 数量和运行时间。

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our paper conforms to the NeurIPS code of ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the impacts in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

指南：• 答案 NA 表示该论文不包含实验。

- 论文应标明计算工作者的类型，包括 CPU 或 GPU、内部集群或云服务提供商，以及相关的内存和存储。
- 论文应提供每个单独实验运行所需的计算量，并估计总计算量。

- 论文应披露整个研究项目是否需要比论文中报告的实验更多的计算量（例如，未纳入论文的初步或失败的实验）。

9. 伦理规范

   问题：论文中所进行的研究在所有方面是否符合 NeurIPS 伦理准则 https://neurips.cc/public/EthicsGuidelines？答案：[是] 理由：我们的论文在所有方面都符合 NeurIPS 伦理准则。指南：

- 回答为 NA 表示作者未审查 NeurIPS 伦理准则。
- 如果作者回答为否，他们应解释需要偏离伦理准则的特殊情况。

- 作者应确保保持匿名（例如，如果由于他们所在地区的法律或法规存在特殊考虑）。

10. 更广泛的影响

    问题：论文是否讨论了所进行工作的潜在积极社会影响和消极社会影响？答案：[是] 理由：我们在补充材料中讨论了这些影响。指南：• 答案为 NA 表示所进行的工作没有社会影响。

- 如果作者回答 NA 或否，他们应解释为什么他们的工作没有社会影响，或为什么论文没有涉及社会影响。
- 负面的社会影响示例包括潜在的恶意或非预期用途（例如，虚假信息、生成虚假资料、监控）、公平性考量（例如，部署可能对特定群体产生不公平影响的决策技术）、隐私考量以及安全考量。

- 会议预期许多论文将是基础性研究，与特定应用无关，更不用说部署了。然而，如果存在直接通往任何负面应用的路径，作者应当指出。例如，指出生成模型质量的提高可能被用于制造虚假信息以传播虚假信息是合理的。另一方面，指出一种通用的神经网络优化算法能够使人们更快地训练生成 Deepfakes 的模型则并非必要。

- 作者应当考虑技术按预期使用且功能正常时可能产生的危害、技术按预期使用但结果错误时可能产生的危害，以及技术（有意或无意）被滥用时可能产生的危害。

- 如果存在负面社会影响，作者也可以讨论可能的缓解策略（例如，模型的分阶段发布，提供防御措施而非攻击，监测滥用的机制，监测系统如何随时间从反馈中学习，提高机器学习（ML）的效率和可访问性）。

11. 安全措施

    问题：论文是否描述了为负责任地发布具有高风险被滥用的数据或模型（例如，预训练语言模型、图像生成器或抓取的数据集）而采取的安全措施？

Answer: [Yes]

Justification: We provide the safeguards in the supplemental.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original papers when using their code, data or models.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: We don't release the datasets in the submission. We will provide all the code for reproducing the dataset and the dataset itself after acceptance.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

答案：[是] 理由：我们在补充材料中提供了保护措施。指南：・
答案 NA 表示，论文不构成此类风险。

- 对于存在高风险被滥用或具有双重用途的已发布模型，应采取必要的保护措施以实现模型的受控使用，例如要求用户遵守使用指南或限制以访问模型，或实施安全过滤器。

- 从互联网上抓取的数据集可能存在安全风险。作者应该描述他们如何避免发布不安全的图像。
- 我们认识到提供有效的保护措施具有挑战性，许多论文并不要求这样做，但我们鼓励作者考虑这一点并尽最大努力。

12. 现有资产的许可证

    问题：论文中使用的资产（例如代码、数据、模型）的创建者或原始所有者是否得到了适当的致谢？是否明确提到了许可证和使用条款，并得到了适当的尊重？

    答案：[是] 理由：在使用他们的代码、数据或模型时，我们引用了原始论文。指南：・ 答案 NA 表示该论文未使用现有资源。

    - 作者应引用产生代码包或数据集的原始论文。
    - 作者应说明使用的资产版本，如果可能，应包含 URL。

    - 每个资产都应包含许可证名称（例如，CC-BY 4.0）。
    - 对于从特定来源（例如网站）抓取的数据，应提供该来源的版权和使用条款。

    - 如果资产被发布，应提供包中的许可证、版权信息和使用条款。对于流行的数据集，paperswithcode.com/datasets 为一些数据集整理了许可证。他们的许可证指南可以帮助确定数据集的许可证。

    - 对于重新打包的现有数据集，应同时提供原始许可证和派生资产的许可证（如果已更改）。
    - 如果这些信息无法在线获取，鼓励作者联系资产创建者。

13. 新资产

    问题：论文中引入的新资产是否得到了充分文档记录，并且文档是否随资产一同提供？答案：[否] 理由：我们未在提交中发布数据集。我们将在论文被接受后提供所有用于复现数据集的代码以及数据集本身。指南：・ 答案"NA"表示论文未发布新资产。

    - 研究人员应通过结构化模板在提交时详细说明数据集/代码/模型的相关信息。这包括训练细节、许可证、限制条件等。

    - 论文应讨论是否以及如何从资产使用者的个人那里获得同意。

    - 提交时，请记得对您的资产进行匿名化处理（如适用）。您可以创建一个匿名化的 URL，或包含一个匿名化的压缩文件。

14. 众包和涉及人类的研究

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use the reasoning ability of LLM to assess the quality of our video generation and described it in the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

问题：对于涉及众包实验和人类受试者的研究，论文是否包含提供给参与者的完整指令文本以及适用时的截图，以及关于补偿（如有）的详细信息？答案：[NA] 理由：我们的论文不涉及众包或人类受试者的研究。指南：

- 回答“不适用”表示该论文不涉及众包或涉及人类受试者的研究。
- 在补充材料中包含这些信息是可以的，但如果论文的主要贡献涉及人类受试者，那么应尽可能在正文中包含详细信息。
- 根据 NeurIPS 伦理准则，参与数据收集、数据整理或其他劳动的人员应至少获得数据收集所在国家的最低工资。

15. 机构审查委员会（IRB）批准或同等批准用于涉及人类受试者的研究

问题：论文是否描述了研究参与者可能承担的潜在风险，是否向受试者披露了这些风险，以及是否获得了机构审查委员会（IRB）的批准（或基于您国家或机构的相应批准/审查要求）？
答案：[NA] 理由：我们的论文不涉及众包或涉及人类受试者的研究。指南：

- 回答“不适用”表示该论文不涉及众包或涉及人类受试者的研究。
- 根据研究进行的国家，任何涉及人类受试者的研究可能都需要 IRB 批准（或同等批准）。如果您获得了 IRB 批准，您应在论文中明确说明。
- 我们认识到，此流程在不同机构及地点可能存在显著差异，并期望作者遵守 NeurIPS 道德准则及其所在机构的指南。
- 对于初次提交，不要包含任何可能破坏匿名性的信息（如适用），例如进行评审的机构信息。

16. LLM 使用声明

问题：如果 LLM 是本研究核心方法中一个重要、原创或非标准的组成部分，论文是否描述了其使用情况？请注意，如果 LLM 仅用于写作、编辑或格式化目的，且不影响研究的核心方法、科学严谨性或原创性，则无需声明。回答：[是] 理由：我们使用 LLM 的推理能力来评估视频生成的质量，并在论文中进行了描述。指南：

- 答案 NA 表示，本研究中的核心方法开发不涉及 LLMs 作为任何重要、原创或非标准组件。
- 请参考我们的 LLM 政策（https://neurips.cc/Conferences/2025/LLM），了解应该或不应该描述的内容。