# NEWTONGEN: PHYSICS-CONSISTENT AND CONTROLLABLE TEXT-TO-VIDEO GENERATION VIA NEURAL NEWTONIAN DYNAMICS

**Yu Yuan**
Purdue University
yuan418@purdue.edu

**Xijun Wang**
Purdue University
wang6661@purdue.edu

**Tharindu Wickremasinghe**
Purdue University
lwickrem@purdue.edu

**Zeeshan Nadir**
Samsung Research America
zeeshan.nadir@samsung.com

**Bole Ma**
Purdue University
ma929@purdue.edu

**Stanley H. Chan**
Purdue University
stanchan@purdue.edu

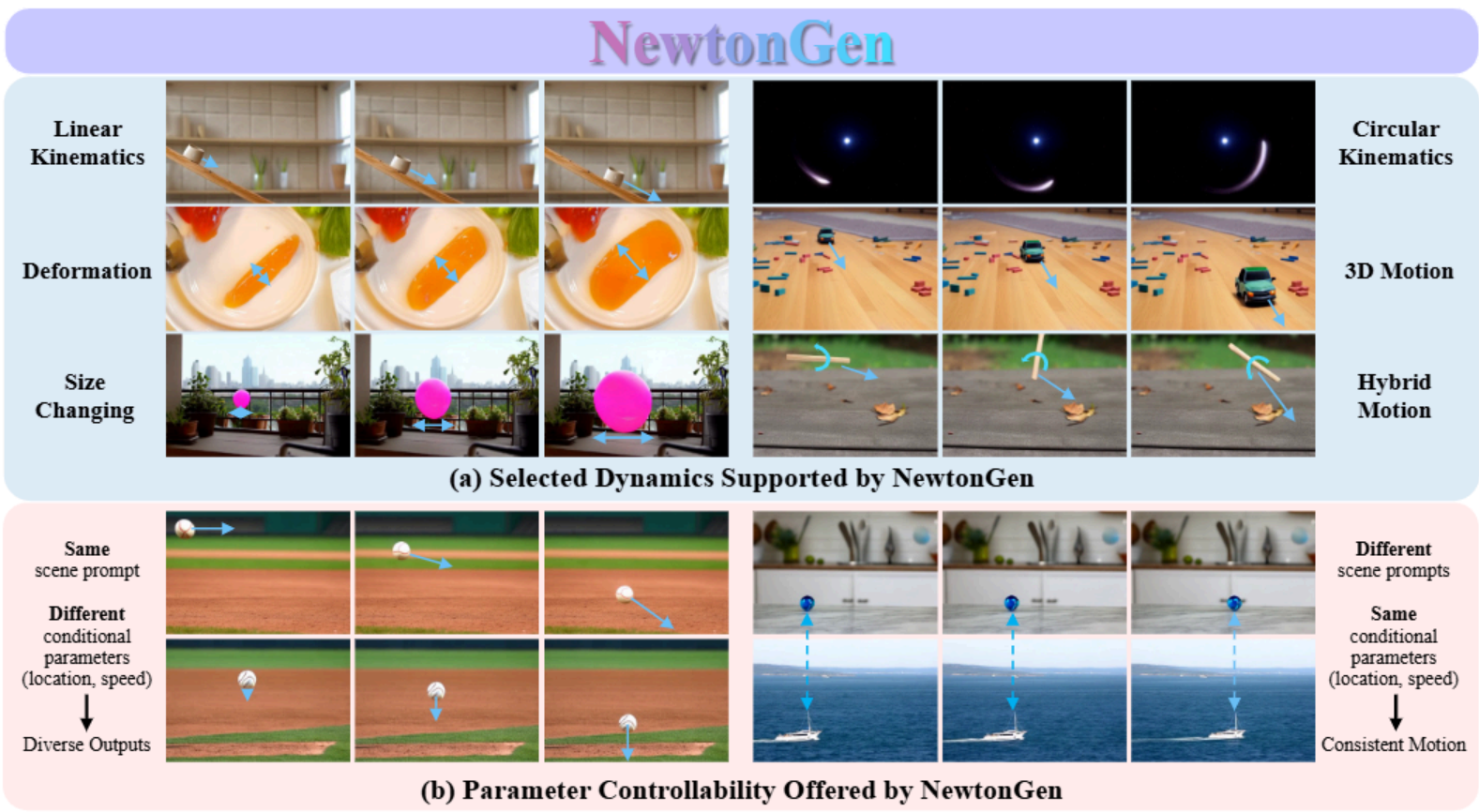



Figure 1: NewtonGen generates physically-consistent videos from text prompts, with diverse dynamic perception (a), and precise parameter control (b).

## ABSTRACT

A primary bottleneck in large-scale text-to-video generation today is physical consistency and controllability. Despite recent advances, state-of-the-art models often produce unrealistic motions, such as objects falling upward, or abrupt changes in velocity and direction. Moreover, these models lack precise parameter control, struggling to generate physically consistent dynamics under different initial conditions. We argue that this fundamental limitation stems from current models learning motion distributions solely from appearance, while lacking an understanding of the underlying dynamics. In this work, we propose NewtonGen, a framework that integrates data-driven synthesis with learnable physical principles. At its core lies trainable Neural Newtonian Dynamics (NND), which can model and predict a variety of Newtonian motions, thereby injecting latent dynamical constraints into the video generation process. By jointly leveraging data priors and dynamical guidance, NewtonGen enables physically consistent video synthesis with precise parameter control. All data and code will be public at here.

# NEWTONGEN：基于神经牛顿动力学的物理一致且可控的文本到视频生成

Yu Yuan
普渡大学
yuan418@purdue.edu

Xijun Wang
普渡大学
wang6661@purdue.edu

Tharindu Wickremasinghe
普渡大学
lwickrem@purdue.edu

Zeeshan Nadir
三星美国研究院
zeeshan.nadir@samsung.com

Bole Ma
普渡大学
ma929@purdue.edu

斯坦利·H·陈
普渡大学
stanchan@purdue.edu

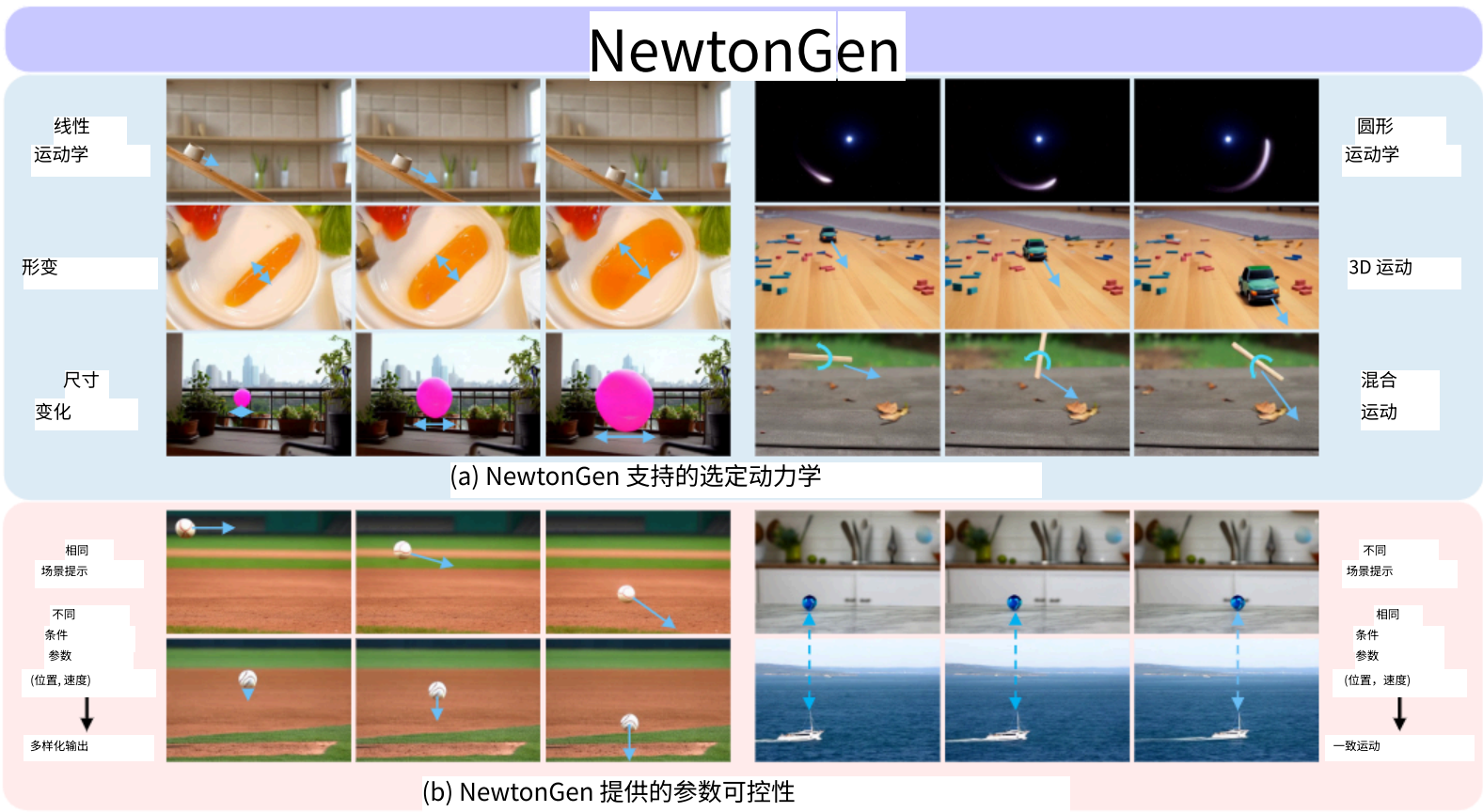



图 1：NewtonGen 根据文本提示生成符合物理规律的视频，具有多样的动态感知（a），以及精确的参数控制（b）。

## 摘要

当前大规模文本到视频生成的主要瓶颈在于物理一致性和可控性。尽管近年来取得了进展，最先进的模型仍然常常产生不真实的运动，例如物体向上坠落，或速度和方向发生突变。此外，这些模型缺乏精确的参数控制，难以在不同初始条件下生成物理一致的动力学。我们认为这一根本性限制源于当前模型仅从外观学习运动分布，而缺乏对底层动力学的理解。在这项工作中，我们提出了 NewtonGen，一个将数据驱动合成与可学习物理原理相结合的框架。其核心是可训练的神经牛顿动力学（NND），能够模拟和预测各种牛顿运动，从而将潜在的动力学约束注入视频生成过程。通过联合利用数据先验和动力学指导，NewtonGen 实现了物理一致的视频合成和精确的参数控制。所有数据和代码将在这里公开。

## 1 INTRODUCTION

Since the breakthrough of probabilistic diffusion models in the early 2020's (See, e.g., Ho et al. (2020); Song et al. (2021); Ramesh et al. (2021); Rombach et al. (2022)), foundational vision models have created unprecedented opportunities for digital content generation. While contemporary video generators can synthesize visually appealing frames (Ho et al., 2022; OpenAI, 2024b; Hong et al., 2023; Peebles & Xie, 2023; Kong et al., 2024; Yang et al., 2025b), they struggle to produce dynamic sequences that adhere to physically plausible motion. For instance, many videos generated by these methods violate basic physical laws such as objects falling upward, or abruptly changing velocity and direction (Bansal et al., 2024; 2025; Zhang et al., 2025a; Li et al., 2025b; Duan et al., 2025; Motamed et al., 2025; Gu et al., 2025). The goal of this paper is to provide a solution to these undesirable outcomes.

The failures in the above situations, according to some literature, can potentially be remedied by scaling laws (Kaplan et al., 2020). However, recent researches such as Kang et al. (2025); Li et al. (2025a); Chefer et al. (2025); Lin et al. (2025); Bansal et al. (2024; 2025) consistently point to a deeper reason that current models only learn the distribution of visual appearances. They lack an understanding of the underlying physical laws. Existing frameworks typically treat videos as spatio-temporal tokens and optimize the likelihood at the pixel level. During inference, the models mainly rely on **memorization and imitation**, making it difficult to generalize to out-of-distribution scenarios (Kang et al., 2025). To bridge this gap, we argue that we need to explicitly incorporate physical laws into the learning process. This is not only a crucial step for video generation, but also essential for connecting generative AI with the physical world.

In this paper, we introduce **NewtonGen**, a novel framework that integrates a data-driven, pre-trained video generator with physics-informed, Neural Newtonian Dynamics (NND). In NND, we introduce a neural ordinary differential equation (neural ODE) model to learn and predict the Newtonian motion from physics-clean data. By learning the dynamics of motion and manipulating its initial physical states, we can predict physics-consistent trajectories, orientations, and shapes. Subsequently, a motion-controlled video generator produces diverse and realistic videos by conditioning on both the predicted states and scene prompts. In summary, the contribution of this paper is twofold:

1. We propose NewtonGen, a **physics-consistent and controllable** text-to-video framework that explicitly incorporates dynamics into the generation process, allowing for interpretable, white-box control over generated motion.
2. We introduce **Neural Newtonian Dynamics (NND)**, which models different dynamics via unified neural ordinary differential equations (ODEs). NND can efficiently learn the latent dynamics from a small amount of physics-clean data.

We conducted extensive experiments, showing that NewtonGen achieves physical consistency and controllability across various dynamics, as illustrated in Figure 1, and outperforms other baselines.

## 2 RELATED WORK

### 2.1 VIDEO GENERATION MODELS

The emergence of diffusion models (Ho et al., 2020; Song et al., 2021) has greatly enhanced the ability of generative models to produce visually realistic images (Ramesh et al., 2021; Rombach et al., 2022; SD2). In video generation, models learn the distribution of real-world motion from large-scale datasets (OpenAI, 2024b; Blattmann et al., 2023; Hong et al., 2023; Yang et al., 2025b; Kong et al., 2024; NVIDIA, 2025; Wan et al., 2025). The method DiT proposed by Peebles & Xie (2023) introduces transformer architectures into diffusion models, further enhancing their scalability (Kaplan et al., 2020) for video generation tasks. Following this trend, representative video generation models (e.g., Sora OpenAI (2024b)) aim to leverage extensive video data to evolve toward general-purpose world simulators.

However, current video generation models still lack an understanding of real-world physics. Studies show that increasing data or model size does not help them learn the physical rules behind video content (Kang et al., 2025; Liu et al., 2025; Motamed et al., 2025; Lin et al., 2025). As a result,

## 1 引言

自 2020 年代初概率扩散模型取得突破以来（例如，Ho 等人（2020）；Song 等人（2021）；Ramesh 等人（2021）；Rombach 等人（2022）），基础视觉模型为数字内容生成创造了前所未有的机遇。尽管现代视频生成器能够合成视觉上吸引人的帧（Ho 等人，2022；OpenAI，2024b；Hong 等人，2023；Peebles 和 Xie，2023；Kong 等人，2024；Yang 等人，2025b），但它们难以生成符合物理上合理运动的动态序列。例如，这些方法生成的许多视频违反了基本的物理定律，如物体向上坠落，或突然改变速度和方向（Bansal 等人，2024；2025；Zhang 等人，2025a；Li 等人，2025b；Duan 等人，2025；Motamed 等人，2025；Gu 等人，2025）。本文的目标是为这些不希望的结果提供一个解决方案。

根据一些文献，上述情况中的失败有可能可以通过规模定律（Kaplan 等，2020）来弥补。然而，近期的研究如 Kang 等（2025）；Li 等（2025a）；Chefer 等（2025）；Lin 等（2025）；Bansal 等（2024；2025）一致指出，一个更深层次的原因是当前模型仅学习了视觉外观的分布，缺乏对底层物理规律的理解。现有框架通常将视频视为时空标记，并在像素级别优化似然性。在推理过程中，模型主要依赖记忆和模仿，难以泛化到分布外场景（Kang 等，2025）。为了弥补这一差距，我们认为需要将物理规律明确地融入学习过程。这不仅对于视频生成是一个关键步骤，也是将生成式 AI 与物理世界连接起来的必要条件。

在本文中，我们介绍了 NewtonGen，这是一个将数据驱动、预训练的视频生成器与物理信息、神经牛顿动力学（NND）相结合的新型框架。在 NND 中，我们引入了一个神经常微分方程（神经 ODE）模型，用于从物理干净的 数据中学习和预测牛顿运动。通过学习运动动力学并操纵其初始物理状态，我们可以预测物理一致的轨迹、方向和形状。随后，一个受运动控制的视频生成器通过结合预测状态和场景提示来生成多样化和逼真的视频。总之，本文的贡献有两个方面：

1. 我们提出了 NewtonGen，这是一个物理一致且可控的文本到视频框架，它明确地将动力学纳入生成过程，允许对生成的运动进行可解释的、白盒控制。

2. 我们引入了神经牛顿动力学（NND），它通过统一的神经常微分方程（ODEs）对不同的动力学进行建模。NND 能够从少量物理干净的数据中高效地学习潜在动力学。

我们进行了广泛的实验，表明 NewtonGen 在各种动力学中实现了物理一致性和可控性，如图 1 所示，并且优于其他基线。

## 2 相关工作

### 2.1 VGM

扩散模型（Ho 等人，2020；Song 等人，2021）的出现极大地增强了生成模型生成视觉真实图像的能力（Ramesh 等人，2021；Rombach 等人，2022；SD2）。在视频生成中，模型从大规模数据集中学习现实世界运动的分布（OpenAI，2024b；Blattmann 等人，2023；Hong 等人，2023；Yang 等人，2025b；Kong 等人，2024；NVIDIA，2025；Wan 等人，2025）。Peebles 和 Xie（2023）提出的 DiT 方法将 Transformer 架构引入扩散模型，进一步增强了其在视频生成任务中的可扩展性（Kaplan 等人，2020）。遵循这一趋势，具有代表性的视频生成模型（例如，Sora OpenAI（2024b））旨在利用大量视频数据进化为通用世界模拟器。

然而，当前的视频生成模型仍然缺乏对现实世界物理的理解。研究表明，增加数据或模型大小并不能帮助它们学习视频内容背后的物理规则（Kang 等人，2025；Liu 等人，2025；Motamed 等人，2025；Lin 等人，2025）。因此，

these models often produce videos that look realistic but contain physically incorrect dynamics when applied to out-of-distribution cases (Kang et al., 2025; Bansal et al., 2025; 2024; Meng et al., 2025; Zhang et al., 2025a; Li et al., 2025a; Gu et al., 2025; Chefer et al., 2025). The main reason is that they focus on appearance-level **motion** rather than the underlying **dynamics**.

### 2.2 Physics-aware Generation

To address the challenge of physical plausibility in video generation, recent research efforts have started incorporating explicit physical priors into generative pipelines. Based on the stage and approaches of injecting physical knowledge, these methods can be broadly categorized into three types:

**Generation then Physical Simulation.** These methods first generate static 3D models or images conditioned on textual or visual inputs using generative models. Subsequently, physical simulation techniques such as Material Point Method (MPM) (Stomakhin et al., 2013) is applied to animate these static outputs into dynamic 3D scenes or videos (Lin et al., 2024; Xie et al., 2024; Tan et al., 2024; Zhang et al., 2024; Hsu et al., 2024). Although the post hoc physics-based rendering process is explicit and controllable, it demands significantly more manual effort. These methods can be summarized in the following Equation:

$$\widehat{\mathbf{V}} = \underbrace{P}_{\text{Physical Simulation}} \Big( \overbrace{G_\psi(\mathbf{I})}^{\text{Video Generation}} \Big) \tag{1}$$

where $P$ denotes the physical simulation, $G$ denotes the video generator parameterized by network weight $\psi$. $\mathbf{I}$ is the input conditional prompt or image, $\widehat{\mathbf{V}}$ is the video we want.

**Physical Simulation then Generation.** Approaches in this category (Yuan et al., 2023; Liu et al., 2024; Savant Aira et al., 2024; Chen et al., 2025; Xie et al., 2025; Li et al., 2025d) first apply physical simulation to conditionally specified images to generate plausible dynamic behaviors. The simulated dynamics are then utilized as conditional inputs for video generation models. For instance, PhysGen (Liu et al., 2024) segments dynamic objects from input images, simulates their motion according to Newtonian mechanics, and then refines the rendering by conditioning a video generation model on both simulated object positions and static backgrounds. However, generative models themselves lack any inherent physical reasoning or simulation capability: users must predefine the physical simulation parameters and rules for each scenario, and these settings cannot readily generalize to other contexts or different physical laws. These approaches can be summarized as:

$$\widehat{\mathbf{V}} = G_\psi\big(P(\mathbf{I})\big) \tag{2}$$

**Generation with Learned Physics Priors.** As illustrated in Equation 3, these methods leverage physical priors extracted from large-scale pretrained models to guide the generative process directly (Li et al., 2024; Lv et al., 2024; Xu et al., 2024; Yang et al., 2025a; Pandey et al., 2025; Xue et al., 2025; Cao et al., 2024; Wang et al., 2025; Yuan et al., 2025; Zhang et al., 2025b; Chefer et al., 2025; Zhang et al., 2025c; Feng et al., 2025; Yang et al., 2025a). For example, PhysT2V (Xue et al., 2025) employs a large language model (LLM) ChatGPT (OpenAI, 2024a) and a vision–language model (VLM) (Wang et al., 2024a) as physics-consistency evaluators, performing multiple rounds of self-refinement to generate videos with improved physical plausibility. The main limitation of this line of work lies in the implicit assumption that existing models are capable of physical reasoning. In practice, however, these large-scale models, much like conventional video generation models, derive their so-called "physical understanding" purely from data fitting, and thus struggle when faced with physically challenging out-of-distribution scenarios. Our method broadly fits within this paradigm; however, our physical prior is driven by both explicit physics models and physics-clean data, which gives it stronger conditional controllability and better out-of-distribution generalization.

$$\widehat{\mathbf{V}} = G_\psi\big(P_\phi(\mathbf{I})\big) \tag{3}$$

where $\phi$ is the learned physical parameters.

### 2.3 Learn Physics from Videos

Leveraging the spatiotemporal information in videos, methods such as Wu et al. (2015); Watters et al. (2017); Wu et al. (2017); Belbute-Peres et al. (2018); Raissi et al. (2019); Chari et al. (2019);

这些模型在应用于分布外情况时，往往生成看起来真实但包含物理上不正确的动态的视频（Kang 等人，2025；Bansal 等人，2025；2024；Meng 等人，2025；Zhang 等人，2025a；Li 等人，2025a；Gu 等人，2025；Chefer 等人，2025）。主要原因是它们专注于外观级别的运动，而不是底层动态。

## 2.2 P-G

为了解决视频生成中的物理合理性挑战，近期研究工作开始将显式的物理先验知识融入生成流程。根据注入物理知识的阶段和方法，这些方法可以大致分为三种类型：

生成然后物理模拟。这些方法首先使用生成模型根据文本或视觉输入生成静态 3D 模型或图像。随后，应用物理模拟技术（如材料点法（MPM）（Stomakhin 等人，2013 年））将这些静态输出动画化为动态 3D 场景或视频（Lin 等人，2024 年；Xie 等人，2024 年；Tan 等人，2024 年；Zhang 等人，2024 年；Hsu 等人，2024 年）。尽管基于物理的后期渲染过程是显式且可控的，但它需要显著更多的手动工作。这些方法可以总结为以下公式：

$$\hat{V} = P(\underbrace{\underbrace{G(I)}_{\text{视频生成}}}_{\text{物理模拟}}) \tag{1}$$

其中 P 表示物理模拟，G 表示由网络权重ψ参数化的视频生成器。I 是输入的条件提示或图像，bV 是我们想要的视频。

物理模拟然后生成。这类方法（Yuan 等人，2023；Liu 等人，2024；Savant Aira 等人，2024；Chen 等人，2025；Xie 等人，2025；Li 等人，2025d）首先将物理模拟应用于条件指定的图像，以生成合理的动态行为。然后，模拟的动态被用作视频生成模型的条件输入。例如，PhysGen（Liu 等人，2024）从输入图像中分割动态对象，根据牛顿力学模拟它们的运动，然后通过在模拟对象位置和静态背景上对视频生成模型进行条件化来细化渲染。然而，生成模型本身缺乏任何固有的物理推理或模拟能力：用户必须预先定义每个场景的物理模拟参数和规则，这些设置不能轻易地推广到其他环境或不同的物理定律。这些方法可以总结为：

(2) 基于学习物理先验的生成。如方程 3 所示，这些方法利用从大规模预训练模型中提取的物理先验来直接指导生成过程（Li 等人，2024；Lv 等人，2024；Xu 等人，2024；Yang 等人，2025a；Pandey 等人，2025；Xue 等人，2025；Cao 等人，2024；Wang 等人，2025；Yuan 等人，2025；Zhang 等人，2025b；Chefer 等人，2025；Zhang 等人，2025c；Feng 等人，2025；Yang 等人，2025a）。例如，PhysT2V（Xue 等人，2025）采用大型语言模型（LLM）ChatGPT（OpenAI，2024a）和视觉-语言模型（VLM）（Wang 等人，2024a）作为物理一致性评估器，通过多轮自我优化生成具有更高物理合理性的视频。这项工作的主要局限性在于隐含地假设现有模型能够进行物理推理。然而在实践中，这些大规模模型与传统视频生成模型类似，其所谓的“物理理解”完全源于数据拟合，因此在面对物理挑战性分布外场景时会遇到困难。 我们的方法大体上符合这一范式；然而，我们的物理先验由显式物理模型和物理干净的训练数据共同驱动，这使得它具有更强的条件可控性和更好的分布外泛化能力。

$$\hat{\phi} = G(P(I)) \tag{3}$$

其中 ϕ 是学习到的物理参数。

## 2.3 LPV

利用视频中的时空信息，Wu 等人（2015）；Watters 等人（2017）；Wu 等人（2017）；Belbute-Peres 等人（2018）；Raissi 等人（2019）；Chari 等人（2019）；等方法

Greydanus et al. (2019); Lutter et al. (2019); Zhong & Leonard (2020); Jaques et al. (2020); Le Guen & Thome (2020); Hofherr et al. (2023); Garrido et al. (2025); Garcia et al. (2025); Deng et al. (2025); Li et al. (2025c) estimate the parameters of known governing equation, This enables tasks such as future-frame prediction and physical reasoning. These methods usually adopt an encoder-decoder structure. Each frame is encoded into a latent physical state using models like a variational autoencoder (VAE) (Kingma & Welling, 2013; 2019). The latent state is then processed by a physics engine and decoded back to reconstruct the frame for training. Most of these approaches are designed for a single type of simple dynamical system. They are difficult to generalize to different systems within a single framework.

Our Neural Newtonian Dynamics (NND) is partly inspired by the aforementioned methods. We adopt an encoder-only architecture integrated with physics-informed general neural ordinary differential equations (ODEs) to explicitly capture diverse dynamics from videos.

## 3 Preliminary Concepts

### 3.1 Incorporating Physical Dynamics into Data-driven Video Generation

Existing video generation models (OpenAI, 2024b; Hong et al., 2023; Kong et al., 2024; Yang et al., 2025b; Blattmann et al., 2023) are mostly data-driven, relying on large-scale video datasets without physical annotations. While they achieve good performance within training domains, they often fail in out-of-distribution scenarios by violating basic physical laws (Chefer et al., 2025; Kang et al., 2025). In contrast, physics-driven dynamics methods explicitly incorporate governing constraints, yielding better physical plausibility and out-of-distribution generalization (Champion et al., 2019). To combine the strengths of both, we propose incorporating physical dynamics into data-driven video generation. This hybrid paradigm leverages the low-bias learning capacity of data-driven models while injecting lightweight dynamics priors to enforce consistency with fundamental laws, thereby achieving improved generalization and physically coherent video synthesis.

### 3.2 Modeling the Dynamics in a General Physics-Informed Neural ODE

To understand how NewtonGen works, we first ask: what is the best way to describe Newtonian motion? Physics textbooks tell us that if we are given the initial position, initial velocity, acceleration and mass, we can predict the trajectory of how the object moves in space and time. In mathematics, this is done through ordinary differential equations (ODEs). Based on this intuition, we consider a second-order system governed by autonomous ODEs with no explicit time-varying external forces. We constrain the ODEs to the second order, because most common physical motions in daily life (e.g., flying balls) can generally be described by second-order dynamics. Even in more complex motions and three-dimensional scenes, the dynamics can still be effectively characterized by second-order formulations over relatively short time intervals with sufficiently dense anchor points.

To handle a wide range of video generation tasks, we require the ODE framework capable of accommodating diverse dynamics. This raises the following question: how can we construct a universal ODE framework that can describe various types of motion? To this end, we introduce two key design principles:

1. **Latent Physical States.** We define a 9-dimensional latent physical state vector $\mathbf{Z} = [x, y, v_x, v_y, \theta, \omega, s, l, a]$. Here, $x, y$ represent the position, and $v_x, v_y$ represent velocity of the object's center of mass. $\theta, \omega$ encode the object's rotation or rotation about a pivot point. $s, l$ are the object's shortest and longest dimensions, and $a$ is its projected area. This formulation allows our physical states to capture translation, rotation, deformation, and other complex behaviors. 3D motion can also be equivalently realized through the combination of position and size control.
2. **Linear Physics-Informed Neural ODEs with a Residual MLP.** Different motions follow inherently different dynamical laws: for instance, free-fall can be described by a simple linear ODE, while a damped pendulum or other unknown motion cannot. To address this, we combine linear physics-informed neural ODEs with a residual multilayer perceptron (MLP) as illustrated in Equation 4 and Figure. 2(a). The linear ODEs capture the dominant linear

(2019); Greydanus 等人(2019); Lutter 等人(2019); Zhong 与 Leonard(2020); Jaques 等人(2020); Le Guen 与 Thome(2020); Hofherr 等人(2023); Garrido 等人(2025); Garcia 等人(2025); Deng 等人(2025); Li 等人(2025c)估计已知控制方程的参数，这使得能够执行未来帧预测和物理推理等任务。这些方法通常采用编码器-解码器结构。每帧都使用变分自编码器（VAE）（Kingma 与 Welling，2013 年；2019 年）等模型编码成潜在物理状态。然后通过物理引擎处理该潜在状态并解码回重建帧用于训练。这些方法大多数是为单一类型的简单动力系统设计的。它们难以在单个框架内泛化到不同的系统。

我们的神经牛顿动力学（NND）部分受到了上述方法的启发。我们采用了一种仅包含编码器的架构，该架构集成了物理信息通用神经常微分方程（ODEs），以显式地捕捉视频中的多样化动力学。

## 3 预备概念

### 3.1 IPDD-VG

现有的视频生成模型（OpenAI, 2024b; Hong 等人, 2023; Kong 等人, 2024; Yang 等人, 2025b; Blattmann 等人, 2023）大多是数据驱动的，依赖于大规模视频数据集而不需要物理标注。虽然它们在训练域内取得了良好的性能，但它们常常在分布外场景中失败，因为它们违反了基本的物理定律（Chefer 等人, 2025; Kang 等人, 2025）。相比之下，物理驱动的动力学方法明确地包含了控制约束，从而产生了更好的物理合理性和分布外泛化能力（Champion 等人, 2019）。为了结合两者的优势，我们提出将物理动力学整合到数据驱动的视频生成中。这种混合范式利用了数据驱动模型的低偏差学习能力，同时注入轻量级的动力学先验来确保与基本定律的一致性，从而实现改进的泛化和物理上连贯的视频合成。

### 3.2 MDGP-INODE

要理解 NewtonGen 的工作原理，我们首先会问：描述牛顿运动最好的方法是什么？物理学教科书告诉我们，如果我们知道初始位置、初始速度、加速度和质量，我们就可以预测物体在空间和时间中的运动轨迹。在数学上，这是通过常微分方程（ODEs）实现的。基于这种直觉，我们考虑一个由自主 ODEs 控制的二阶系统，其中没有显式的时间变化外部力。我们将 ODEs 限制为二阶，因为日常生活中最常见的物理运动（例如飞行的球）通常可以用二阶动力学来描述。即使在更复杂的运动和三维场景中，动力学仍然可以通过在具有足够密集锚点的相对短时间间隔内使用二阶公式来有效地表征。

为了处理各种视频生成任务，我们需要一个能够容纳不同动力学的 ODE 框架。这就引出了以下问题：我们如何构建一个通用的 ODE 框架来描述各种类型的运动？为此，我们引入了两个关键设计原则：

1. 隐式物理状态。我们定义一个 9 维的隐式物理状态向量 $Z = [x, y, v, v, \theta, \omega, s, l, a]$。其中，$x, y$ 表示位置，$v, v$ 表示物体质心的速度。$\theta, \omega$编码物体的旋转或绕枢轴点的旋转。$s, l$ 是物体的最短和最长尺寸，$a$ 是其投影面积。这种表述方式使我们的物理状态能够捕捉平移、旋转、形变以及其他复杂行为。三维运动也可以通过位置和尺寸控制的组合来等效实现。

2. 带有残差 MLP 的线性物理信息神经 ODE。不同的运动遵循本质上不同的动力学规律：例如，自由落体可以用简单的线性 ODE 描述，而阻尼摆或其他未知运动则不能。为此，我们将线性物理信息神经 ODE 与残差多层感知器（MLP）相结合，如图 4 和图 2(a)所示。线性 ODE 捕捉主导的线性

dynamics, while the residual MLP models nonlinear and unknown components, enabling the system to flexibly approximate a wide range of physical behaviors.

$$a_z\,\ddot{z} + b_z\,\dot{z} + c_z\,z + d_z + \mathrm{MLP}(\mathbf{Z}) = 0 \tag{4}$$

where $z$ is one element of the 9-dimensional latent physical state vector $\mathbf{Z}$, and $a_z, b_z, c_z, d_z$ are learnable parameters of the linear ODE. We can use multiple ODEs to predict future physical states in a compact autonomous form:

$$\mathbf{Z}_t = \mathbf{Z}_0 + \int_{t_0}^{t} \mathrm{Func}\big(\mathbf{Z}(\tau)\big)\,\mathrm{d}\tau, \tag{5}$$

where $\mathrm{Func}\big(\mathbf{Z}(\tau)\big)$ represents the collection of all individual $\mathrm{d}z/\mathrm{d}t$ ODEs, and $\mathbf{Z}_0 = \mathbf{Z}(t_0)$ is the known initial physical state at time $t_0$.

## 4 METHODOLOGY

**Overall Framework.** As shown in Figure. 2, NewtonGen consists of two main stages. As illustrated in Figure. 2(b), in the first stage, we train the proposed Neural Newtonian Dynamics (NND) on a small set of physics-clean data to learn the underlying motion dynamics and parameters. In the second stage shown in Figure. 2(c), we use the learned dynamics to predict future physical states from arbitrary initial conditions, and feed these predictions, together with the scene prompt, into a motion-controlled text-to-video generation model to produce the final video.

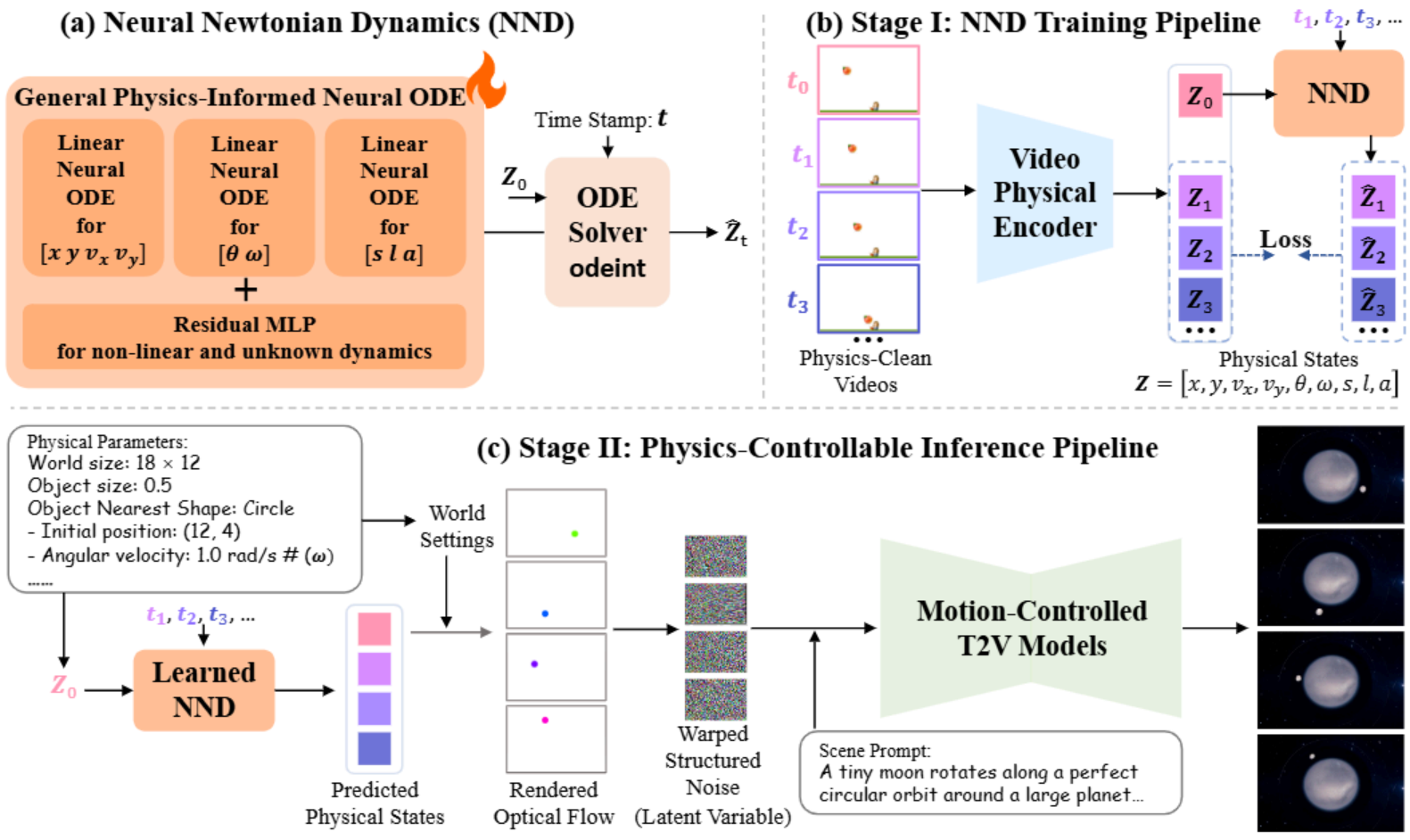



Figure 2: The overall framework of NewtonGen. a) Neural Newtonian Dynamics (NND) employs physics-informed linear neural ODEs combined with an MLP to build a general dynamics learning framework suitable for diverse motions. b) We train NND on a physics-clean dataset to capture the underlying dynamics. c) Using the learned NND together with a data-driven motion-controlled model, we generate physically plausible and controllable videos.

### 4.1 NEURAL NEWTONIAN DYNAMICS

As discussed in Subsection 3.2, our Neural Newtonian Dynamics (NND) aims to construct a unified model capable of capturing a wide range of dynamical behaviors. Its core is composed of physics-informed general neural ordinary differential equations (Neural ODEs) (Chen et al., 2018). As

动态特性，而残差 MLP 模型则处理非线性及未知部分，使系统能灵活地逼近广泛的物理行为。

(4) 线性常微分方程（ODE）捕捉了与导数的线性关系，其中 $Z$ 是 9 维潜在物理状态向量 $Z$ 的一个元素，而 a、b、c、d 是线性 ODE 的可学习参数。我们可以使用多个 ODE 以紧凑的自治形式预测未来的物理状态：

$$Z = Z + \int_{t}^{Z} \mathrm{Func}\left(Z(\tau)\right) \, d\tau, \tag{5}$$

其中 $\mathrm{Func}$ 表示所有单个 $dZ/dt$ 常微分方程的集合，$Z=Z(t)$ 是时间 t 时的已知初始物理状态。

# 详细框架

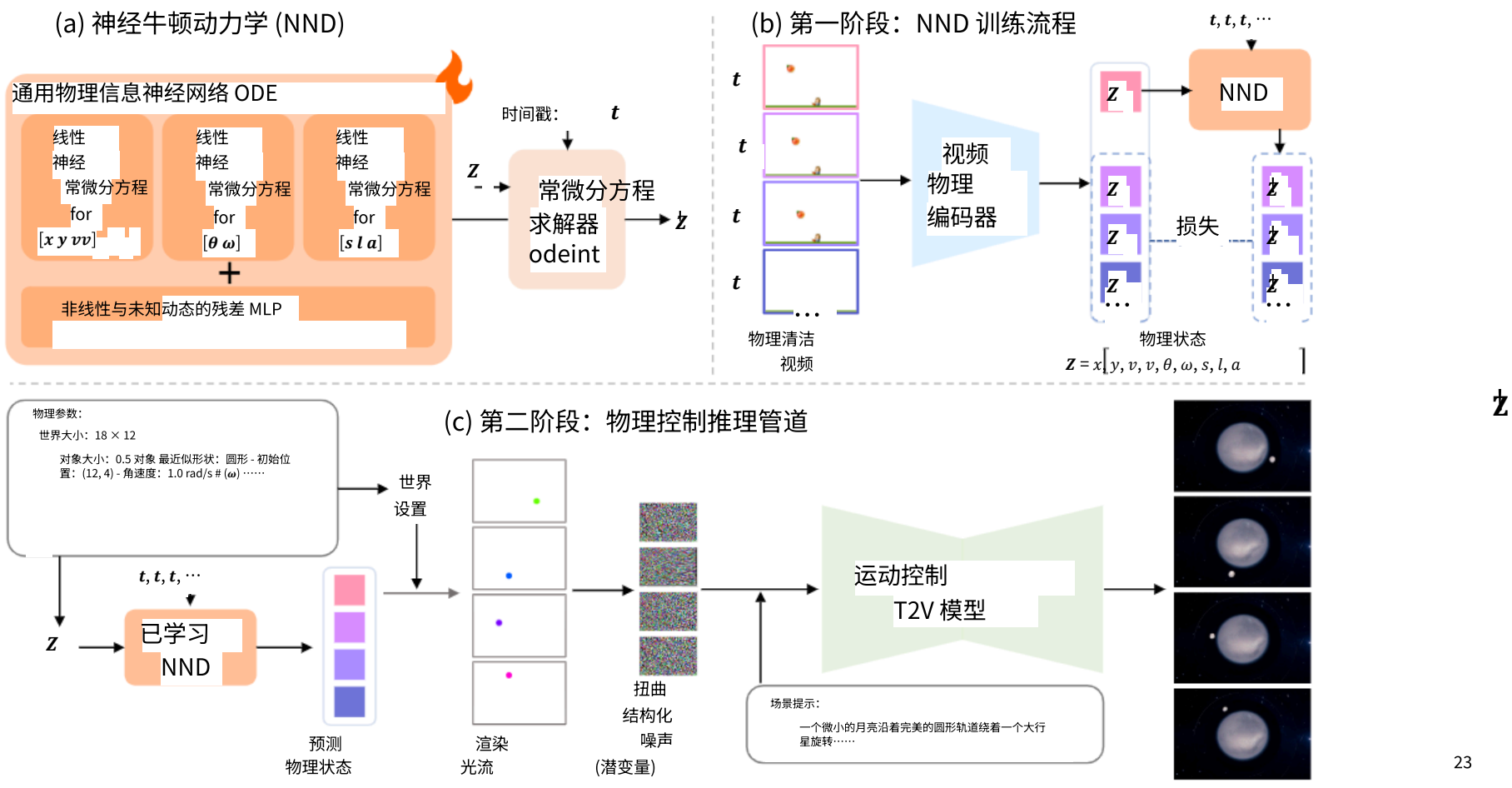



图 2：NewtonGen 的整体框架。a)神经牛顿动力学（NND）采用物理信息线性神经 ODE 结合 MLP 构建一个适用于各种运动的通用动力学学习框架。b)我们在物理干净的数据集上训练 NND 以捕捉底层动力学。c)使用学习到的 NND 与数据驱动的运动控制模型，我们生成物理上合理且可控的视频。

## 4.1 NND

如 3.2 小节所述，我们的神经牛顿动力学（NND）旨在构建一个能够捕捉广泛动力学行为的统一模型。其核心由物理信息通用神经常微分方程（Neural ODEs）（Chen 等人，2018）组成。

demonstrated in Figure. 2(a), physics-informed linear neural ODEs are employed to model the underlying linear dynamics, while a residual three-layer MLP captures nonlinear and unknown components of the dynamics. With this design, the learnable neural ODEs can represent more complex or real-world dynamics. Given an initial physical states $\mathbf{Z}_0$ and a time stamp $t$, the ODE solver `odeint` (Chen et al., 2018) can be used to predict the object's future physical states $\mathbf{Z}_t$.

### 4.2 Training for Neural Newtonian Dynamics

**Overall Training Pipeline.** Figure. 2(b) illustrates that, for training Neural Newtonian Dynamics (NND), we adopt an encoder-only architecture. This design does not require decoding back to images, and optimizes solely in the latent physical space, significantly reducing computational cost. Specifically, a Video Physical Encoder $\mathrm{E}_{\text{phys}}$ compresses each video frame into its corresponding physical state. The initial state $Z_0$ and the sequence of frame time stamps $(t_1, t_2, t_3, \dots)$ are fed into NND, which predicts $(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \widehat{\mathbf{Z}}_3, \dots)$. The loss is then computed between the predicted states and the states $(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \dots)$ extracted by the Encoder $\mathrm{E}_{\text{phys}}$ :

$$\text{Loss} = \frac{1}{T}\sum_{t=1}^{T} \| \underbrace{\mathrm{E}_{\text{phys}}(\mathbf{I}_t)}_{\mathbf{Z}_t} - \underbrace{\mathrm{NND}_{\kappa}(\mathrm{E}_{\text{phys}}(\mathbf{I}_0), t)}_{\widehat{\mathbf{Z}}_t} \|_2^2 \tag{6}$$

where $T$ denotes the number of sampled time stamps, $\mathbf{I}_t$ is the video frame at time $t$, and $\kappa$ represents the learnable parameters of the ODEs.

**Training Data.** To enable Neural Newtonian Dynamics to learn accurate and effective representations of physical dynamics, we require "physics-clean" video data. That is, the motion in the videos should be prominent and monotonic, with no motion blur or excessive noise in each frame, and minimal color, texture, or background distractions. However, to our knowledge, such high-quality datasets of physical dynamics are still lacking. To address this, we developed a Python-based physics data simulator that can render videos with precise timestamps for different world settings, initial conditions, and types of dynamics. Simulation code and some sample videos are shown in the Supplementary Material.

**Video Physical Encoder.** To extract physical state labels from videos, we first apply the visual segmentation foundation model SAM2 (Ravi et al., 2025) to obtain masks for the dynamic regions in each frame. From the extracted masks, we compute object attributes such as position, velocity, geometry, and area using morphological analysis and OpenCV-based tools. Finally, these attributes are uniformly quantized to form the physical states $\mathbf{Z}$.

### 4.3 Inference for Physical-Controllable Text-to-Video Generation

As illustrated in Figure. 2(c), during inference, we decouple physical dynamics reasoning from video generation. Physical dynamics reasoning focuses on modeling and predicting the motion of dynamic objects, while video generation leverages rich scene understanding and generation capabilities to render detailed and flexible visual content.

We adopt Go-with-the-Flow (Burgert et al., 2025) as our base video generation model, which achieves motion control through structured noise (Chang et al., 2024). By warping the independently initialized Gaussian noise of each frame according to the input optical flow, temporal correlations emerge between the initial noise of consecutive frames, leading to more effective motion control. Other motion-controlled video generation models (Yin et al., 2023; Wang et al., 2024b; Zhang et al., 2025d) typically encode trajectories or bounding boxes through ControlNet (Zhang et al., 2023) or additional encoders and inject the features into the base video generators. However, these approaches often struggle with handling deformations, rotations, or more complex motions. We choose Go-with-the-Flow for its generality and effectiveness.

To effectively transfer the physical knowledge from our NND to the video generation model, a multi-step procedure is required. First, based on the user's physical prompts, we parse the initial physical state $\mathbf{Z}_0$ and future time stamps. $\mathbf{Z}_0$ and the frame timestamps are fed into the trained NND to obtain the corresponding physical states for all future frames. Next, using the world setting information parsed from the physical prompts (e.g., scene dimensions, object size, and the closest simple geometric shape of the object), we compute an approximate pixel-level optical flow for

如图 2(a)所示，采用物理信息线性神经常微分方程来建模底层的线性动力学，而残差三层 MLP 则捕获动力学的非线性及未知成分。通过这种设计，可学习的神经常微分方程能够表示更复杂或现实世界的动力学。给定初始物理状态 Z 和时间戳 t，可以使用常微分方程求解器 odeint（Chen 等人，2018）来预测物体的未来物理状态 Z。

### 4.2 TNND

整体训练流程。图 2(b)说明，为训练神经牛顿动力学（NND），我们采用仅编码器架构。这种设计无需解码回图像，仅在潜在物理空间中优化，显著降低计算成本。具体而言，视频物理编码器 E 将每个视频帧压缩为其对应的物理状态。初始状态 Z 和帧时间戳序列（t, t, t, ...）被输入 NND，它预测（bZ, bZ, bZ, ...）。然后计算预测状态与编码器 E 提取的状态（Z, Z, Z, ...）之间的损失：

$$\text{Loss} = \frac{1}{T}\sum_{t=1}^{T} \left\| E(I) - \text{NND}(E(I), t) \right\| \quad (6)$$

其中 T 表示采样时间戳的数量，I 是时间 t 的视频帧，κ表示 ODEs 的可学习参数。

训练数据。为了使神经牛顿动力学能够学习物理动力学的准确和有效表示，我们需要“物理纯净”的视频数据。也就是说，视频中的运动应该突出且单调，每一帧都没有运动模糊或过度噪声，并且颜色、纹理或背景干扰最小。然而，据我们所知，目前仍然缺乏此类高质量的物理动力学数据集。为了解决这个问题，我们开发了一个基于 Python 的物理数据模拟器，该模拟器可以渲染具有精确时间戳的视频，这些时间戳对应于不同的世界设置、初始条件和动力学类型。模拟代码和一些示例视频显示在补充材料中。

视频物理编码器。为了从视频中提取物理状态标签，我们首先应用视觉分割基础模型 SAM2（Ravi 等人，2025）来获取每帧中动态区域的掩码。从提取的掩码中，我们使用形态学分析和基于 OpenCV 的工具计算对象属性，如位置、速度、几何形状和面积。最后，这些属性被统一量化形成物理状态 Z。

### 4.3 IP-CT--VG

如图 2(c)所示，在推理过程中，我们将物理动力学推理与视频生成解耦。物理动力学推理专注于建模和预测动态物体的运动，而视频生成则利用丰富的场景理解和生成能力来渲染详细且灵活的视觉内容。

我们采用 Go-with-the-Flow（Burgert 等人，2025）作为我们的基础视频生成模型，该模型通过结构化噪声（Chang 等人，2024）实现运动控制。通过根据输入的光流扭曲每帧独立初始化的高斯噪声，连续帧的初始噪声之间会产生时间相关性，从而实现更有效的运动控制。其他运动控制视频生成模型（Yin 等人，2023；Wang 等人，2024b；Zhang 等人，2025d）通常通过 ControlNet（Zhang 等人，2023）或额外的编码器编码轨迹或边界框，并将特征注入基础视频生成器。然而，这些方法通常难以处理变形、旋转或更复杂的运动。我们选择 Go-with-the-Flow 是因为其通用性和有效性。

为了有效地将我们的神经牛顿动力学（NND）中的物理知识转移到视频生成模型，需要多步流程。首先，根据用户的物理提示，我们解析初始物理状态 Z 和未来的时间戳。Z 和帧时间戳被输入到训练好的 NND 中，以获得所有未来帧的相应物理状态。接下来，使用从物理提示中解析的世界设置信息（例如，场景尺寸、物体大小以及物体最接近的简单几何形状），我们计算近似像素级的视差流。

each frame based on the predicted physical states. These flows are then temporally and spatially downsampled to match the resolution of the video generator's latent space, resulting in a structured optical flow sequence. Finally, combining the user's scene prompts, video frames are sampled to produce the final videos.

# 5 EXPERIMENTS

In this section, we evaluate the applications of our framework for physically-consistent and controllable video generation. Subsection 5.1 presents implementation details, Subsection 5.2 compares NewtonGen with other baselines, and Subsection 5.3 discusses the results of ablation study.

## 5.1 IMPLEMENTATION DETAILS

**Supported Motion Types.** In NewtonGen, we evaluate 12 distinct types of motion: **uniform velocity, uniform acceleration, deceleration, parabolic motion, 3D motion, slope sliding, circular motion, rotation around an axis, parabolic motion with rotation, damped oscillation, size changing, and deformation**. These categories cover the most common fundamental motion patterns encountered in everyday scenarios. The tested velocity magnitudes are mostly within the range of 0–15 m/s, while the duration of the generated motions is typically concentrated within 1–2 seconds.

**Training Details for NND.** We optimize the Neural Newtonian Dynamics (NND) with the AdamW optimizer (initial learning rate $1 \times 10^{-4}$) and a CosineAnnealingLR scheduler (Loshchilov & Hutter, 2017). For each type of motion, we collect 100 physical videos with different initial conditions from the physics simulator mentioned in Subsection 4.2 as training data. The model is trained with a batch size of 64 for a total of 20,000 epochs, which requires about 2 hours on a single NVIDIA A100 80 GB GPU.

**Metrics.** Assessing the physical consistency of different video generation models is hampered by the absence of a shared ground truth and by the fact that each synthetic sequence is defined in its own coordinate frame and scale and time. Consequently, a single, unified physical evaluation metric cannot be directly applied. In this work, we are inspired by the Physical Invariance Score (PIS) (Zhang et al., 2025a), which evaluates physical plausibility by checking whether a motion preserves its expected invariants $C$. For example, in parabolic motion and the horizontal velocity $v_x$ should be constant. We use SAM2 (Ravi et al., 2025) to segment the object in each frame and obtain its centroid and shape features. Velocities are estimated from frame-to-frame centroid differences. The Physical Invariance Score for a quantity $C$ is defined as the relative standard deviation of $C$ over time:

$$\text{PIS} = (1 + C_\sigma / (|C_\mu| + \epsilon))^{-1} \tag{7}$$

where $C$ denotes one of the quantities introduced above (i.e., the horizontal velocity, the vertical acceleration, or the angular speed). $\epsilon = 1 \times 10^{-5}$ is added to the denominator to prevent division by zero. The PIS score is bounded in $[0, 1]$, with a value of 1 indicating that the evaluated physical quantity remains perfectly invariant.

## 5.2 COMPARISONS WITH OTHER METHODS

**General Comparisons.** We compare our method with five baselines: SORA (OpenAI, 2024b), Veo3 (Google, 2025), CogVideoX-5B (Yang et al., 2025b), Wan2.2 (Wan et al., 2025) and PhysT2V (Xue et al., 2025). These baselines represent the current state-of-the-art in both closed-source and open-source video generation models, as well as physics-based generation methods. We standardize the video generation settings across all methods to ensure maximum fairness in comparison. For evaluation, we collected 24 prompts for each motion type to assess the physical generation capabilities of all methods.

In Figure. 3, the video sequences generated by NewtonGen exhibit the highest degree of physical consistency across all 12 motion types. The motions display smooth and realistic trajectories without abrupt changes in direction or speed, realistic 3D movement effects (with object scale gradually increasing as distance decreases), physically plausible self-rotation (objects preserve shape with uniform angular velocity), smooth deformations (edges stretch or shrink progressively), and natural size variations (e.g., balloon diameter increases over time but at a decelerating rate). Table 1 further

我们基于预测的物理状态，为每一帧计算近似像素级光流。然后，这些光流在时间和空间上进行下采样，以匹配视频生成器的潜在空间分辨率，从而得到结构化的光流序列。最后，结合用户的场景提示，对视频帧进行采样，生成最终视频。

## 5 实验

在本节中，我们评估了我们的框架在物理一致且可控的视频生成中的应用。第 5.1 节介绍了实现细节，第 5.2 节将 NewtonGen 与其他基线进行了比较，第 5.3 节讨论了消融研究的结果。

### 5.1 ID

支持的运动物理类型。在 NewtonGen 中，我们评估了 12 种不同的运动类型：匀速运动、匀加速运动、减速运动、抛物线运动、三维运动、斜坡滑动、圆周运动、绕轴旋转、带旋转的抛物线运动、阻尼振荡、尺寸变化和变形。这些类别涵盖了日常生活中最常见的最基本的运动模式。测试的速度大小大多在 0–15 m/s 的范围内，而生成的运动的持续时间通常集中在 1–2 秒内。

NND 训练细节。我们使用 AdamW 优化器（初始学习率 1 × 10^-4）和 CosineAnnealingLR 调度器（Loshchilov & Hutter, 2017）优化神经牛顿动力学（NND）。对于每种类型的运动，我们从第 4.2 节提到的物理模拟器中收集 100 个具有不同初始条件的物理视频作为训练数据。模型以 64 的批处理大小进行训练，总共进行 20,000 个 epoch，这需要大约 2 小时的单个 NVIDIA A100 80 GB GPU。

指标。评估不同视频生成模型的物理一致性受到缺乏共享真实值以及每个合成序列定义在其自己的坐标框架和尺度及时间的影响。因此，无法直接应用单一、统一的物理评估指标。在本工作中，我们受到物理不变性分数（PIS）（Zhang 等人，2025a）的启发，该分数通过检查运动是否保持其预期的不变量 C 来评估物理合理性。例如，在抛物线运动中，水平速度 v 应保持恒定。我们使用 SAM2（Ravi 等人，2025）对每帧中的对象进行分割，并获取其质心和形状特征。速度通过帧间质心差异来估计。量 C 的物理不变性分数定义为 C 随时间的相对标准偏差：

$$\mathrm{PIS} = (1 + C/(|C| + \epsilon)) \tag{7}$$

其中 C 表示上述引入的量之一（即水平速度、垂直加速度或角速度）。在分母中添加了ε = 1 × 10^-5 以防止除以零。PIS 分数被限制在[0, 1]范围内，值为 1 表示评估的物理量保持完美不变。

### 5.2 与 OM 的比较

总体比较。我们将我们的方法与五个基线进行比较：SORA（OpenAI，2024b）、Veo3（Google，2025）、CogVideoX-5B（Yang 等人，2025b）、Wan2.2（Wan 等人，2025）和 PhysT2V（Xue 等人，2025）。这些基线代表了闭源和开源视频生成模型以及基于物理的生成方法的当前最先进水平。我们标准化了所有方法的视频生成设置，以确保比较的最大公平性。在评估方面，我们为每种运动类型收集了 24 个提示，以评估所有方法的物理生成能力。

在图 3 中，NewtonGen 生成的视频序列在所有 12 种运动类型中表现出最高的物理一致性。这些运动显示出平滑且逼真的轨迹，方向或速度没有突然变化，具有逼真的 3D 运动效果（随着距离减小，物体尺寸逐渐增大），物理上合理的自旋（物体保持形状且角速度均匀），平滑的变形（边缘逐渐拉伸或收缩），以及自然的大小变化（例如，气球直径随时间增加但增速逐渐减慢）。表 1 进一步

shows that our model achieves significantly higher physical consistency scores than competing methods across different motion categories. Notably, some motion types do not admit perfect physical invariants; in these cases, we still compute quantities such as angular velocity and compare them against reference simulation videos under the same conditions.

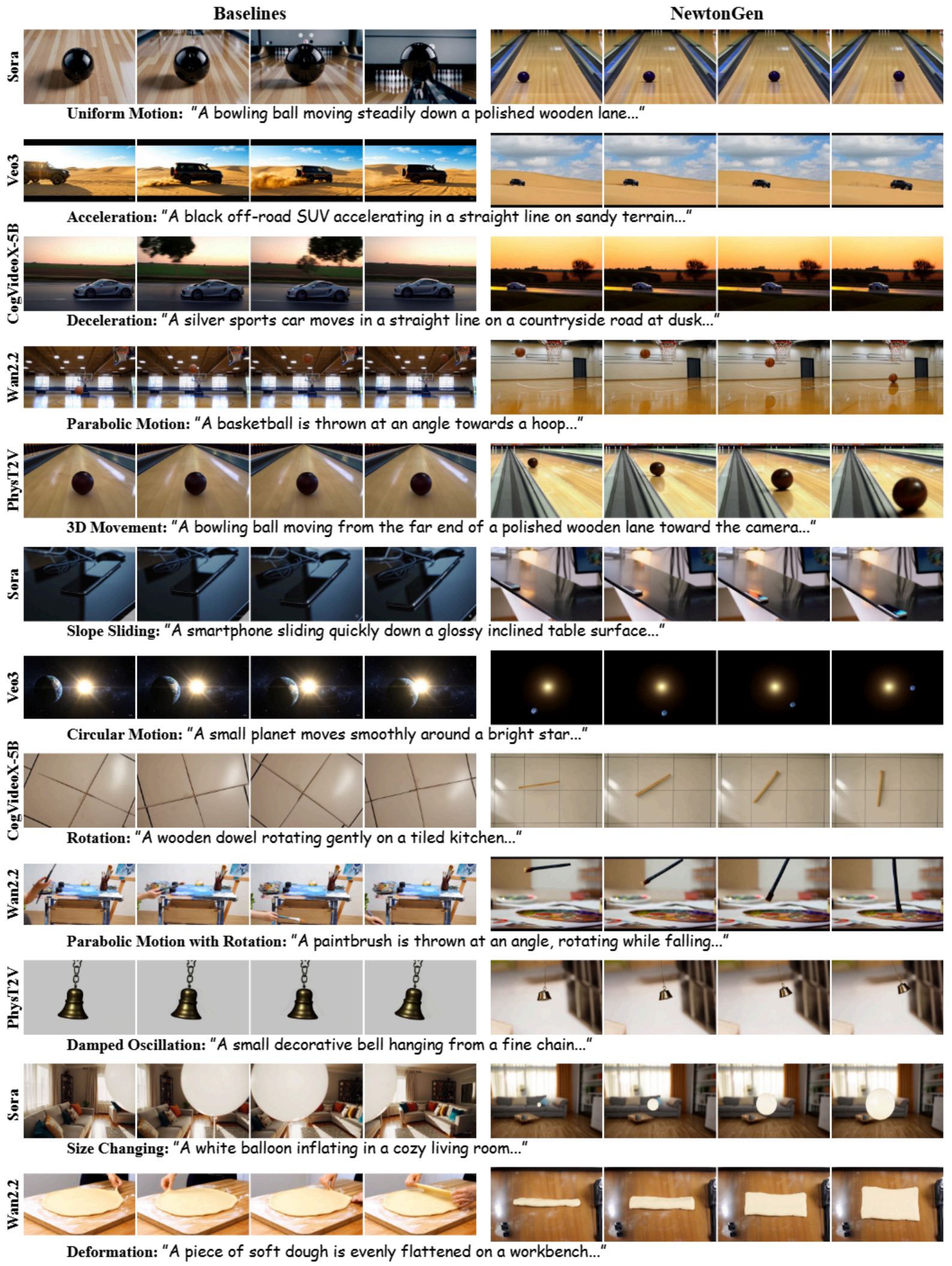



Figure 3: Visual comparisons of different text-to-video generation methods across diverse physical dynamics, where our method consistently shows strong physical consistency.

**Parameter Controllability Comparisons.** In Figure. 4, we demonstrate NewtonGen's ability to perceive physical parameters. Unlike other models, our method faithfully reflects world settings,

表 1 进一步表明，我们的模型在不同运动类别中均显著高于竞争方法的物理一致性得分。值得注意的是，某些运动类型并不允许完美的物理不变量；在这些情况下，我们仍然计算角速度等量，并在相同条件下将其与参考模拟视频进行比较。

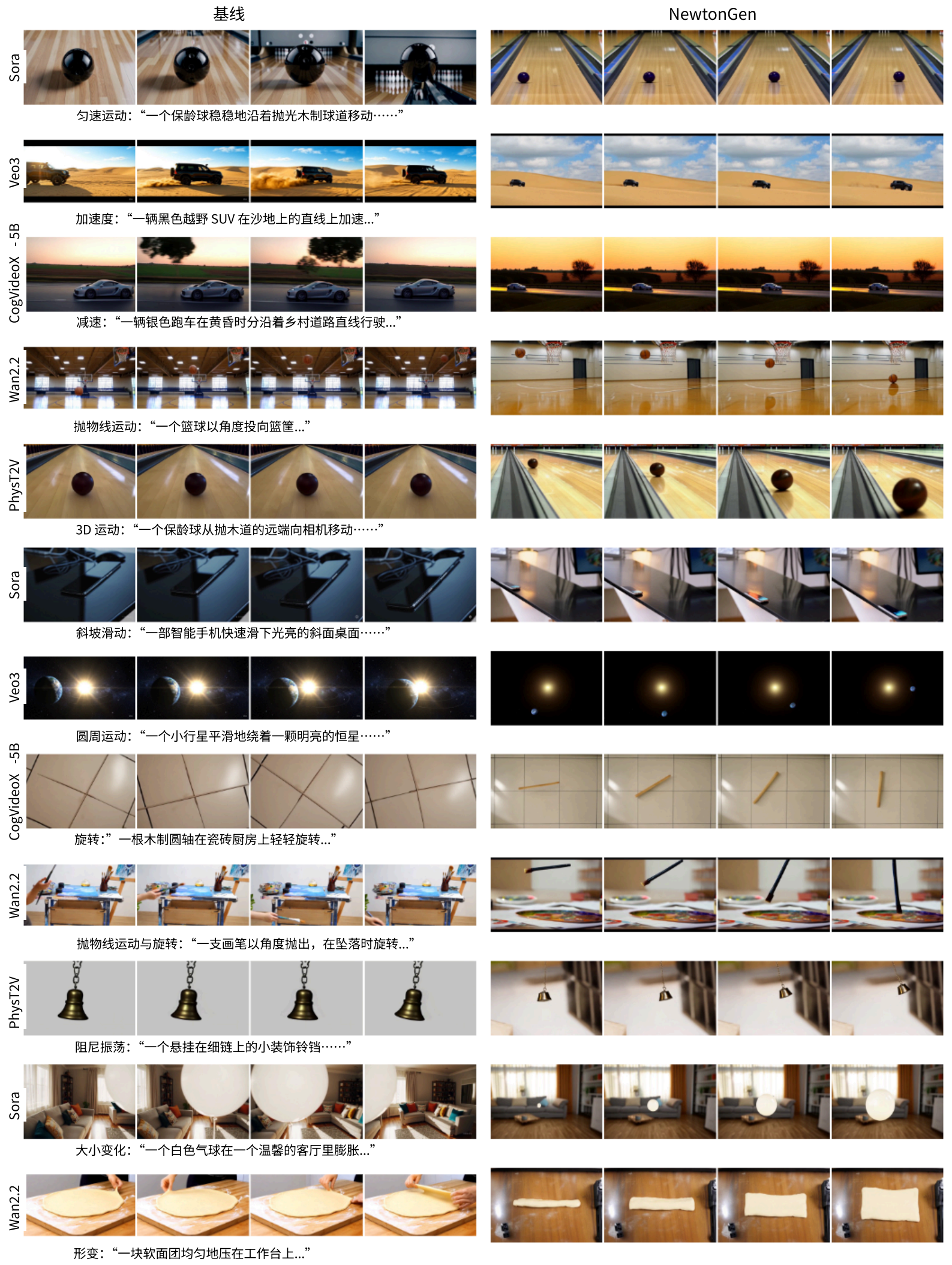



图 3：不同文本到视频生成方法在多种物理动态下的视觉比较，其中我们的方法始终表现出强大的物理一致性。

参数可控性比较。在图 4 中，我们展示了 NewtonGen 感知物理参数的能力。与其他模型不同，我们的方法忠实地反映了世界设置，

Table 1: Quantitative comparison with different methods. Reference is calculated on the simulated videos. The detailed definition of PIS for each type of motion is provided in the Appendix. We highlight the best and second-best values for each metric.

| Motion Types | PIS↑ | Methods | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Reference | Sora | Veo3 | CogVideoX-5B | Wan2.2 | PhysT2V | Ours |
| Uniform Motion | $v$ | 0.9972 | 0.6548 | 0.9784 | 0.5392 | 0.6395 | 0.5349 | 0.9830 |
| Acceleration (Uniform) | $a_x$ | 0.8489 | 0.3437 | 0.6187 | 0.5458 | 0.3077 | 0.5033 | 0.6568 |
| Deceleration (Uniform) | $a_x$ | 0.8872 | 0.6162 | 0.6173 | 0.4988 | 0.4705 | 0.5167 | 0.6891 |
| Parabolic Motion | $v_x$ | 0.9988 | 0.9095 | 0.9042 | 0.7392 | 0.7747 | 0.6370 | 0.9803 |
| | $a_y$ | 0.9487 | 0.5723 | 0.7662 | 0.4230 | 0.5571 | 0.3567 | 0.8189 |
| 3D Motion | $\Delta_l$ | 0.7388 | 0.5013 | 0.5932 | 0.3026 | 0.4583 | 0.2911 | 0.6472 |
| | $v_y$ | 0.9986 | 0.8481 | 0.8913 | 0.6690 | 0.8384 | 0.6510 | 0.9371 |
| Slope Sliding | $a_x$ | 0.8741 | 0.4931 | 0.6081 | 0.3533 | 0.3108 | 0.3570 | 0.6312 |
| | $a_y$ | 0.9148 | 0.4616 | 0.3815 | 0.4731 | 0.3967 | 0.4297 | 0.5840 |
| Circular Motion (Orbit) | $\omega$ | 0.9933 | 0.8393 | 0.8932 | 0.7726 | 0.4677 | 0.6391 | 0.9788 |
| Rotation (Uniform) | $\omega$ | 0.9836 | 0.4267 | 0.5285 | 0.6596 | 0.3425 | 0.7842 | 0.8838 |
| Parabolic Motion with Rotation | $v_x$ | 0.9990 | 0.5797 | 0.7029 | 0.6488 | 0.6558 | 0.7689 | 0.9446 |
| | $a_y$ | 0.9657 | 0.4903 | 0.5603 | 0.2614 | 0.4331 | 0.2879 | 0.5614 |
| | $\omega$ | 0.9829 | 0.6522 | 0.9019 | 0.3380 | 0.3474 | 0.4119 | 0.9289 |
| Damped Oscillation | $a_y$ | 0.9402 | 0.4418 | 0.3516 | 0.3083 | 0.3494 | 0.2841 | 0.5240 |
| Size Changing | $\Delta_r$ | 0.8501 | 0.2840 | 0.4167 | 0.5774 | 0.1972 | 0.4010 | 0.6362 |
| Deformation | $\Delta_l$ | 0.9247 | 0.3626 | 0.3466 | 0.3550 | 0.3515 | 0.3601 | 0.5492 |

Table 2: Quantitative results of ablation study. We compute the normalized absolute error between the predicted and ground-truth physical states across all time steps within the test batch.

| Motions | Uni | Acc | Dec | Para | 3DMot | Slope | Circ | Rota | ParaRota | Osci | Size | Def |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ablations | Normalized Absolute Error ↓ | | | | | | | | | | | |
| W/o MLP | 0.0174 | 0.0069 | 0.0104 | 0.0193 | 0.0937 | 0.0831 | 0.5388 | 0.0382 | 0.7451 | 0.2275 | 0.1239 | 0.0854 |
| Our-data10 | 0.0632 | 0.0260 | 0.0184 | 0.0284 | 0.1079 | 0.0935 | 0.1246 | 0.0739 | 0.0273 | 0.1045 | 0.2327 | 0.0555 |
| Our-data100 | 0.0142 | 0.0034 | 0.0078 | 0.0042 | 0.0182 | 0.0324 | 0.0255 | 0.0058 | 0.0064 | 0.0425 | 0.1193 | 0.0357 |
| Our-data500 | 0.0195 | 0.0051 | 0.0072 | 0.0040 | 0.0192 | 0.0307 | 0.0196 | 0.0049 | 0.0063 | 0.0694 | 0.1379 | 0.0290 |

object properties, and initial conditions, with trajectories and velocities that better follow physical laws (third row). More cases are provided in the Appendix.

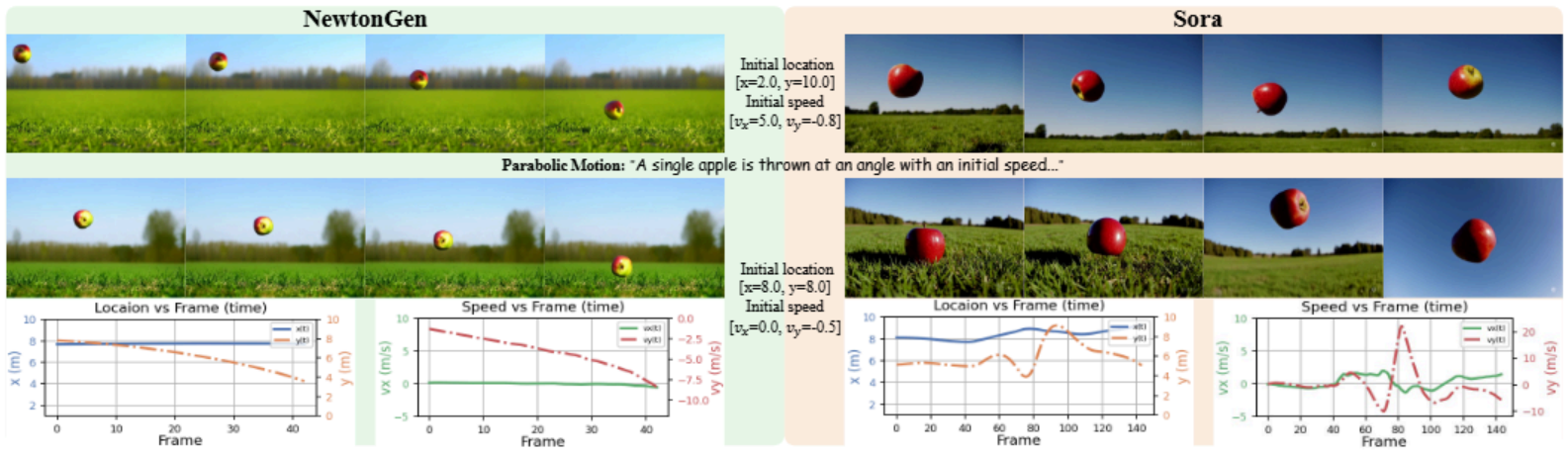



Figure 4: NewtonGen generates videos that can accurately reflect user-specified initial physical parameters, including object position, velocity, angle, shape and size.

### 5.3 Ablation Study

Our ablation study focuses on the effects of the MLP in Neural Newtonian Dynamics (NND) and the training data scale. As shown in Table 2, adding the MLP significantly improves NND's performance on nonlinear dynamics and noisy data. Increasing the training dataset size does not lead to notable gains, indicating that NND can accurately infer the underlying system dynamics from a relatively small number of physically clean samples.

表 1：与不同方法的定量比较。参考值是在模拟视频上计算的。每种运动类型的 PIS 详细定义在附录中。我们突出了每个指标的最佳和次佳值。

| 方法 | 运动类型 | PIS↑ | 参考 | Sora | Veo3 | CogVideoX-5B | Wan2.2 | PhysT2V | 我们 |
|---|---|---|---|---|---|---|---|---|---|
| 匀速运动 | v | 0.9972 | 0.6548 | 0.9784 | 0.5392 | 0.6395 | 0.5349 | 0.9830 | |
| 加速度（匀速） | a | 0.8489 | 0.3437 | 0.6187 | 0.5458 | 0.3077 | 0.5033 | 0.6568 | |
| 减速度（匀速） | a | 0.8872 | 0.6162 | 0.6173 | 0.4988 | 0.4705 | 0.5167 | 0.6891 | |
| 抛物线运动 | v | 0.9988 | 0.9095 | 0.9042 | 0.7392 | 0.7747 | 0.6370 | 0.9803 | |
| | a | 0.9487 | 0.5723 | 0.7662 | 0.4230 | 0.5571 | 0.3567 | 0.8189 | |
| 3D 运动 | Δ | 0.7388 | 0.5013 | 0.5932 | 0.3026 | 0.4583 | 0.2911 | 0.6472 | |
| | v | 0.9986 | 0.8481 | 0.8913 | 0.6690 | 0.8384 | 0.6510 | 0.9371 | |
| 斜坡滑动 | a | 0.8741 | 0.4931 | 0.6081 | 0.3533 | 0.3108 | 0.3570 | 0.6312 | |
| | a | 0.9148 | 0.4616 | 0.3815 | 0.4731 | 0.3967 | 0.4297 | 0.5840 | |
| 圆周运动（轨道） | ω | 0.9933 | 0.8393 | 0.8932 | 0.7726 | 0.4677 | 0.6391 | 0.9788 | |
| 匀速旋转 | ω | 0.9836 | 0.4267 | 0.5285 | 0.6596 | 0.3425 | 0.7842 | 0.8838 | |
| | v | 0.9990 | 0.5797 | 0.7029 | 0.6488 | 0.6558 | 0.7689 | 0.9446 | |
| | a | 0.9657 | 0.4903 | 0.5603 | 0.2614 | 0.4331 | 0.2879 | 0.5614 | |
| | ω | 0.9829 | 0.6522 | 0.9019 | 0.4286 | 0.3474 | 0.4119 | 0.9289 | |
| 旋转抛物线运动 | 物线运动 | | | | | | | | |
| 阻尼振荡 | a | 0.9402 | 0.4418 | 0.3516 | 0.3083 | 0.3494 | 0.2841 | 0.5240 | |
| 尺寸变化 | Δ | 0.8501 | 0.2840 | 0.4167 | 0.5774 | 0.1972 | 0.4010 | 0.6362 | |
| 变形 | Δ | 0.9247 | 0.3626 | 0.3466 | 0.3550 | 0.3515 | 0.3601 | 0.5492 | |

表 2：消融实验的定量结果。我们在测试批次中所有时间步长内计算预测物理状态与真实物理状态之间的归一化绝对误差。

| 运动 | 单一加速度 | 减速度 | 参数 | 3DMot 斜率 | 圆周 | 参数旋转 | 旋转参数 | 振动 | 大小 | 定义 | 摘除 | 归一化绝对误差 ↓ | 无 | MLP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0174 | 0.0069 | 0.0104 | 0.0193 | 0.0937 | 0.0831 | 0.5388 | 0.0382 | 0.7451 | 0.2275 | 0.1239 | 0.0854 | Our-data10 | 0.0632 | 0.0260 |
| 0.0184 | 0.0284 | 0.1079 | 0.0935 | 0.1246 | 0.0739 | 0.0273 | 0.1045 | 0.2327 | 0.0555 | Our-data100 | 0.0142 | 0.0034 | 0.0078 | 0.0042 |
| 0.0182 | 0.0324 | 0.0255 | 0.0058 | 0.0064 | 0.0425 | 0.1193 | 0.0357 | Our-data500 | 0.0195 | 0.0051 | 0.0072 | 0.0040 | 0.0192 | 0.0307 |
| 0.0196 | 0.0049 | 0.0063 | 0.0694 | 0.1379 | 0.0290 | | | | | | | | | |

物体属性，以及初始条件，轨迹和速度更好地遵循物理定律（第三行）。附录中提供了更多案例。

## 结果

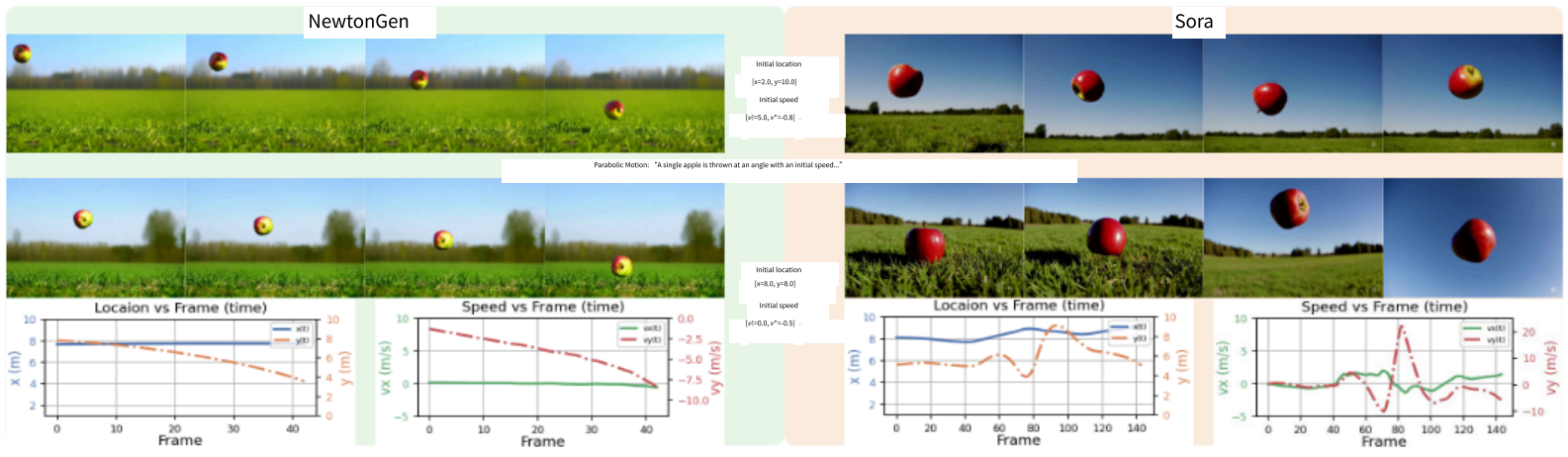



图 4：NewtonGen 生成的视频能够准确反映用户指定的初始物理参数，包括物体位置、速度、角度、形状和大小。

5.3 A

我们的消融研究主要关注神经牛顿动力学（NND）中的 MLP 的影响以及训练数据规模。如表 2 所示[38]，添加 MLP 显著提升了 NND 在非线性动力学和噪声数据上的性能。增加训练数据集规模并未带来显著提升，表明 NND 能够从相对较少的物理干净的样本中准确推断出潜在的系统动力学。

## 6 Conclusion

In this paper, we introduce NewtonGen, a physics-consistent and controllable text-to-video generation framework. NewtonGen integrates a Neural Newtonian Dynamics (NND) module, which learns latent dynamics for diverse motions from a small set of physically accurate examples and predicts future physical states. We validate NewtonGen on over twelve different dynamic video generation tasks, demonstrating its physical consistency and parameter controllability. NewtonGen holds the potential to narrow the gap between current generative models and the real physical world.

## 7 Limitations, Ethics Statement and Reproducibility Statement

**Limitations.** While our framework effectively models and predicts the dynamics of most common motions, it is based on continuous dynamics. This means that NewtonGen can be less effective for handling multi-object interactions (e.g., collisions or coalescence). We expect that future work incorporating event-based or discrete neural architectures will address these limitations.

**Ethics Statement.** This model is designed to generate high-quality content and educational videos; however, when misused without labels and watermarks, it can produce fake videos and lead to the spread of misinformation.

**Reproducibility Statement.** All data, code and model weights will be made publicly available. In our model evaluation, we fix the random seed and provide the test prompts and generated videos in the supplementary materials and appendix. In addition, we include more detailed explanations of the evaluation metrics in the Appendix.

### Acknowledgments

This work is supported, in part, by the United States National Science Foundation under the grants 2133032, 2431505, and a research award from Samsung Research America.

## References

Stable Diffusion. https://github.com/Stability-AI/StableDiffusion.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, and Josh Tenenbaum. End-to-end differentiable physics for learning and control. In *Advances in Neural Information Processing Systems*, 2018.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Qinglong Cao, Ding Wang, Xirui Li, Yuntian Chen, Chao Ma, and Xiaokang Yang. Teaching video diffusion model with latent physical phenomenon knowledge. *arXiv preprint arXiv:2411.11343*, 2024.

## 6 结论

在本文中，我们介绍了 NewtonGen，一个物理一致且可控的文本到视频生成框架。NewtonGen 集成了一个神经牛顿动力学（NND）模块，该模块从少量物理准确的示例中学习各种运动的潜在动力学，并预测未来的物理状态。我们在超过十二种不同的动态视频生成任务上验证了 NewtonGen，展示了其物理一致性和参数可控性。NewtonGen 有潜力缩小当前生成模型与真实物理世界之间的差距。

## 7 局限性、伦理声明和可复现性声明

局限性。尽管我们的框架能有效模拟和预测大多数常见运动的动力学，但它基于连续动力学。这意味着 NewtonGen 在处理多物体交互（例如碰撞或合并）时可能效果较差。我们预计未来结合事件驱动或离散神经架构的工作将解决这些局限性。

伦理声明。该模型旨在生成高质量内容和教育视频；然而，在没有标签和水印的情况下被误用时，它可能产生虚假视频并导致错误信息的传播。

可复现性声明。所有数据、代码和模型权重都将公开提供。在我们的模型评估中，我们固定了随机种子，并在补充材料和附录中提供了测试提示和生成的视频。此外，我们在附录中提供了更详细的评估指标说明。

## A

这项工作部分由美国国家科学基金会通过项目 2133032、2431505 以及三星美国研究奖提供支持。

## 参考文献

Stable Diffusion. https://github.com/Stability-AI/StableDiffusion.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520, 2024.

Hritik Bansal、Clark Peng、Yonatan Bitton、Roman Goldenberg、Aditya Grover 和 Kai-Wei Chang。VideoPhy-2: 一个在视频生成中进行挑战性动作中心物理常识评估的系统。arXiv 预印本 arXiv:2503.06800, 2025。

Filipe de Avila Belbute-Peres、Kevin Smith、Kelsey Allen 和 Josh Tenenbaum。用于学习和控制的端到端可微物理。在《神经信息处理系统进展》，2018 年。

安德烈亚斯・布莱特曼，蒂姆・道克霍恩，苏米特・库拉尔，丹尼尔・门德列维奇，马塞伊・基利亚恩，多米尼克・洛伦茨，亚姆・利维，齐翁・英格利什，维克拉姆・沃莱蒂，亚当・莱茨，瓦鲁纳・贾马帕尼，和罗宾・罗姆巴赫。稳定视频扩散：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。

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, 和 Ning Yu. Go-with-the-flow: 使用实时扭曲噪声的运动可控视频扩散模型。在 IEEE/CVF 计算机视觉与模式识别会议，2025 年。

Qinglong Cao, Ding Wang, Xirui Li, Yuntian Chen, Chao Ma, and Xiaokang Yang. Teaching video diffusion model with latent physical phenomenon knowledge. arXiv preprint arXiv:2411.11343, 2024.

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116 (45):22445–22451, 2019.

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

Pradyumna Chari, Chinmay Talegaonkar, Yunhao Ba, and Achuta Kadambi. Visual physics: Discovering physical laws from videos. *arXiv preprint arXiv:1911.11893*, 2019.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. In *International Conference on Machine Learning*, 2025.

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6572–6583, 2018.

Congyue Deng, Brandon Y. Feng, Cecilia Garraffo, Alan Garbarz, Robin Walters, William T. Freeman, Leonidas Guibas, and Kaiming He. Denoising hamiltonian network for physical reasoning. *arXiv preprint arXiv:2503.0759*, 2025.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

Tao Feng, Xianbing Zhao, Zhenhua Chen, Tien Tsin Wong, Hamid Rezatofighi, Gholamreza Haffari, and Lizhen Qu. Physics-grounded motion forecasting via equation discovery for trajectory-guided image-to-video generation. *arXiv preprint arXiv:2507.06830*, 2025.

Alejandro Castañeda Garcia, Jan van Gemert, Daan Brinks, and Nergis Tömen. Learning physics from video: Unsupervised physical parameter estimation for continuous dynamical systems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.08987*, 2025.

Google. Veo 3: Our state-of-the-art video generation model. https://aistudio.google.com/models/veo-3/, 2025.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangriu Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. Phyworldbench: A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022.

Florian Hofherr, Lukas Koestler, Florian Bernard, and Daniel Cremers. Neural implicit representations for physical parameter inference from a single video. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

凯瑟琳·钱普顿、贝萨尼·卢什、J·内森·库茨和史蒂文·L·布伦顿。数据驱动的发现坐标和支配方程。美国国家科学院院报，116 (45):22445–22451, 2019。

Pascal Chang、Jingwei Tang、Markus Gross 和 Vinicius C. Azevedo。我是如何扭曲你的噪声：用于扩散模型的时序相关噪声先验。在 2024 年国际学习表征会议。

Pradyumna Chari、Chinmay Talegaonkar、Yunhao Ba 和 Achuta Kadambi。视觉物理：从视频中发现物理定律。arXiv 预印本 arXiv:1911.11893，2019。

Hila Chefer、Uriel Singer、Amit Zohar、Yuval Kirstain、Adam Polyak、Yaniv Taigman、Lior Wolf 和 Shelly Sheynin。Videojam：用于视频模型中增强运动生成的联合外观-运动表示。在 2025 年机器学习国际会议。

Boyuan Chen、Hanxiao Jiang、Shaowei Liu、Saurabh Gupta、Yunzhu Li、Hao Zhao 和 Shenlong Wang。Physgen3d：从单张图像构建微型交互世界。在 2025 年 IEEE/CVF 计算机视觉与模式识别会议。

Ricky T. Q. Chen、Yulia Rubanova、Jesse Bettencourt 和 David Duvenaud。神经常微分方程微分方程。在《神经信息处理系统进展》中，第 6572-6583 页，2018 年。

Congyue Deng, Brandon Y. Feng, Cecilia Garraffo, Alan Garbarz, Robin Walters, William T. Freeman, Leonidas Guibas, 和 Kaiming He。去噪哈密顿网络用于物理推理。arXiv 预印本 arXiv:2503.0759，2025 年。

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, 和 Jiajun Wu。WorldScore：一个用于世界生成的统一评估基准。arXiv 预印本 arXiv:2504.00983，2025 年。

Tao Feng, Xianbing Zhao, Zhenhua Chen, Tien Tsin Wong, Hamid Rezatofighi, Gholamreza Haffari, 和 Lizhen Qu。基于物理的轨迹引导图像到视频生成运动预测，通过方程发现。arXiv 预印本 arXiv:2507.06830，2025 年。

亚历杭德罗·卡斯坦耶达·加西亚、扬·范·盖尔特、达恩·布林克斯和内尔吉斯·托门。从视频中学习物理：连续动力系统的无监督物理参数估计。在 IEEE/CVF 计算机视觉与模式识别会议，2025。

Quentin Garrido，Nicolas Ballas，Mahmoud Assran，Adrien Bardes，Laurent Najman，Michael Rabbat，Emmanuel Dupoux 和 Yann LeCun。直观的物理理解来自于在自然视频上的自监督预训练。arXiv 预印本 arXiv:2502.08987，2025。

Google. Veo 3: 我们最先进的视频生成模型。https://aistudio.google.com/models/veo-3/, 2025.

Samuel Greydanus、Misko Dzamba 和 Jason Yosinski。哈密顿神经网络。在《神经信息处理系统进展》，2019 年。

Jing Gu、Xian Liu、Yu Zeng、Ashwin Nagarajan、Fangriu Zhu、Daniel Hong、Yue Fan、Qianqi Yan、Kaiwen Zhou、Ming-Yu Liu 和 Xin Eric Wang。Phyworldbench：对文本到视频模型物理真实性的综合评估。arXiv 预印本 arXiv:2507.13428，2025 年。

Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。在《神经信息处理系统进展》，第 6840-6851 页，2020 年。

乔纳森·何，蒂姆·萨拉曼斯，阿列克谢·格里特森科，威廉·陈，穆罕默德·诺鲁齐和戴维·J舰队。视频扩散模型。在《神经信息处理系统进展》第 35 卷，第 8633-8646 页，2022 年。

弗洛里安·霍费尔、卢卡斯·科斯特勒、弗洛里安·贝尔纳德和丹尼尔·克雷默斯。从单个视频中推断物理参数的神经隐式表示。在 IEEE/CVF 计算机视觉应用冬季会议，2023 年。

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

Hao-Yu Hsu, Zhi-Hao Lin, Albert Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically realistic video editing from natural language instructions. *arXiv preprint arXiv:2411.02394*, 2024.

Miguel Jaques, Michael Burke, and Timothy Hospedales. Physics-as-inverse-graphics: Unsupervised physical parameter estimation from video. In *International Conference on Learning Representations*, 2020.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv: 2001.08361*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11471–11481, 2020.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning*, 2025a.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025b.

Shiqian Li, Ruihong Shen, Chi Zhang, and Yixin Zhu. Neural force field: Learning generalized physical representation from a few examples. *arXiv preprint arXiv:2502.08987*, 2025c.

Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. In *International Conference on Computer Vision*, 2025d.

Jiajing Lin, Zhenzhong Wang, Shu Jiang, Yongjie Hou, and Min Jiang. Phys4dgen: A physics-driven framework for controllable and efficient 4d content generation from a single image. *arXiv preprint arXiv:2411.16800*, 2024.

Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, and Donglin Wang. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025.

Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, and Chang Xu. Generative physical ai in vision: A survey, 2025.

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, 和 Jie Tang. Cogvideo: 基于 transformer 的文本到视频生成的大规模预训练。在 2023 年国际学习表示会议。

Hao-Yu Hsu, Zhi-Hao Lin, Albert Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically

从自然语言指令进行逼真的视频编辑。arXiv 预印本 arXiv:2411.02394, 2024。

米格尔・雅克斯、迈克尔・伯克和蒂莫西・霍斯佩德斯。物理作为逆图形：无监督
从视频中估计物理参数。在 2020 年学习表示国际会议。

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng.
视频生成与世界模型有多远：从物理定律的角度看。在 2025 年机器学习国际会议。

Jared Kaplan、Sam McCandlish、Tom Henighan、Tom B. Brown、Benjamin Chess、Rewon Child、Scott Gray、Alec Radford、Jeffrey Wu 和 Dario Amodei。神经语言模型的规模定律。
arXiv 预印本 arXiv: 2001.08361, 2020.

Diederik P. Kingma 和 Max Welling. 自动编码变分贝叶斯. arXiv 预印本 arXiv:1312.6114, 2013.

Diederik P. Kingma 和 Max Welling. 变分自编码器导论. arXiv 预印本 arXiv:1906.02691, 2019.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

Vincent Le Guen 和 Nicolas Thome. 将物理动力学与未知因素分离

用于无监督视频预测。在 IEEE/CVF 计算机视觉与模式识别会议，第 11471–11481 页，2020 年。

李晨宇，奥斯卡・米歇尔，潘希晨，刘山南，罗伯茨，谢山宁。比萨实验：通过观察物体下落探索视频扩散模型的物理后训练。在机器学习国际会议，2025a。

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong

王宏旭，约瑟夫・E・冈萨雷斯，伊奥恩・斯托伊卡，韩松，卢瑶。世界模型基准：将视频生成模型作为世界模型进行评判。arXiv 预印本 arXiv:2502.20694，2025b。

李石倩，沈瑞红，张驰，朱奕欣。神经力场：从少量示例中学习广义物理表示。arXiv 预印本 arXiv:2502.08987，2025c。

郑琦、理查德・塔克、诺亚・斯奈弗利和亚历山大・霍林斯基。生成图像动力学。
在 IEEE/CVF 计算机视觉与模式识别会议，2024。

李子张、余宏兴、刘伟、杨寅、查尔斯・赫尔曼、戈登・韦茨斯坦和吴佳俊。Wonderplay：从单张图像和动作生成动态 3D 场景。在计算机视觉国际会议，2025d。

林佳静、王振中、蒋舒、侯永杰和蒋敏。Phys4dgen：一种基于物理的可控高效 4D 内容生成框架，从单张图像。arXiv 预印本 arXiv:2411.16800，2024。

林明辉、王翔、王奕山、王舒、戴峰奇、丁鹏翔、王存祥、左正荣、桑农、黄思腾和王东林。探索视频生成中物理认知的演化：一项调查。arXiv 预印本 arXiv:2503.21765，2025。

Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, and Chang Xu.

视觉中的生成式物理 AI：一项调查，2025。

刘少伟、任钟正、Saurabh Gupta 和王深龙。Physgen：基于刚体物理的图像到视频生成。在 2024 年欧洲计算机视觉会议。

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Michael Lutter, Christian Ritter, and Jan Peters. Deep lagrangian networks: Using physics as model prior for deep learning. In *International Conference on Learning Representations*, 2019.

Jiaxi Lv, Yi Huang Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Cheng Yu, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025.

NVIDIA. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

OpenAI. Introducing openai o1-preview. ttps://openai.com/index/introducing-openai-o1-preview/, 2024a.

OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024b.

Karran Pandey, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, Niloy J. Mitra, and Paul Guerrero. Motion modes: What could happen next? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685, 2022.

Luca Savant Aira, Antonio Montanaro, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motion-craft: Physics-based zero-shot video generation. In *Advances in Neural Information Processing Systems*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics*, 2013.

Ilya Loshchilov 和 Frank Hutter。Sgdr：带热重启动的随机梯度下降。在 2017 年国际学习表征会议。

Michael Lutter、Christian Ritter 和 Jan Peters。深度拉格朗日网络：将物理作为深度学习模型先验。在 2019 年国际学习表征会议。

Jiaxi Lv, Yi Huang Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: 通过 blender 导向的 GPT 规划在文本到视频生成中脚本化物理运动。在 IEEE/CVF 计算机视觉与模式识别会议，2024。

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Cheng Yu, Dianqi Li, Yu Qiao, and Ping Luo. 迈向世界模拟器：为视频生成打造基于物理常识的基准。在国际机器学习会议，2025。

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini Jaini, and Robert Geirhos. 生成式视频模型理解物理原理吗？，2025。

NVIDIA. 物理人工智能的宇宙世界基础模型平台。arXiv 预印本 arXiv:2501.03575，

2025.

OpenAI. 发布 openai o1-preview。https://openai.com/index/ introducing-openai-o1-preview/, 2024a.

OpenAI. 视频生成模型作为世界模拟器。https://openai.com/index/ video-generation-models-as-world-simulators/, 2024b.

Karran Pandey, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, Niloy J. Mitra 和 Paul Guerrero. 运动模式：接下来可能发生什么？在 IEEE/CVF 计算机视觉与模式识别会议，2025。

William Peebles 和 Saining Xie. 基于 transformer 的可扩展扩散模型。在计算机视觉国际会议，2023。

Maziar Raissi、Paris Perdikaris 和 George E Karniadakis。物理信息神经网络：用于解决涉及非线性偏微分方程的正向和反向问题的深度学习框架。计算物理杂志，378:686–707，2019。

Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本到图像生成。在机器学习国际会议，2021。

Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman Rädle、Chloe Rolland、Laura Gustafson、Eric Mintun、Junting Pan、Kalyan Vasudev

Alwala、Nicolas Carion、Chao-Yuan Wu、Ross Girshick、Piotr Dollár 和 Christoph Feichtenhofer。Sam 2：图像和视频中的任何分割。在学习表示国际会议，2025。

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser 和 Björn Ommer. 基于潜在扩散模型的超分辨率图像合成. 在 IEEE/CVF 计算机视觉与模式识别会议, 第 10674–10685 页, 2022.

Luca Savant Aira, Antonio Montanaro, Emanuele Aiello, Diego Valsesia, 和 Enrico Magli. Motioncraft: 基于物理的零样本视频生成. 在《神经信息处理系统进展》, 2024.

杨松，贾沙・索尔-迪克斯坦，迪德里克・P・金玛，阿比什克・库马尔，斯特凡诺・埃尔蒙，本・普尔。基于分数的生成模型通过随机微分方程。在 2021 年学习表示国际会议。URL https://openreview.net/forum?

id=PxTIG12RRHS.

亚历克谢・斯托马金、克雷格・施罗德、劳伦斯・蔡、约瑟夫・特兰和安德鲁・塞尔。雪的模拟材料点方法。ACM 图形学汇刊，2013。

Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. Physmotion: Physics-grounded dynamics from a single image. *arXiv preprint arXiv:2411.17189*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024a.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. WISA: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2502.08153*, 2025.

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

Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in Neural Information Processing Systems*, 2017.

Jiajun Wu, Ilker Yildirim, Joseph J. Lim, William T. Freeman, , and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*, 2015.

Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems*, 2017.

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

Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative cartoon animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tianshuo Xu, Zhifei Chen, Leyi Wu, Hao Lu, Yuying Chen, Lihui Jiang, Bingbing Liu, and Yingcong Chen. Motion dreamer: Realizing physically coherent video generation through scene-aware motion reasoning. *arXiv preprint arXiv:2412.00547*, 2024.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. PhyT2V: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In *International Conference on Computer Vision*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025b.

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

Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. 物理动力的运动：从单张图像生成的物理动力。arXiv 预印本 arXiv:2411.17189，2024。

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. arXiv 预印本 arXiv:2503.20314, 2025.

王嘉伟、袁丽萍、张宇辰和孙浩淼。Tarsier：用于训练和评估大型视频描述模型的方案。arXiv 预印本 arXiv:2407.00634，2024a。

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. WISA: World simulator assistant for physics-aware text-to-video generation. arXiv preprint arXiv:2502.08153, 2025.

周夏王，袁子阳，王新涛，李耀伟，陈天水，夏梦涵，罗平，山英。Motionctrl: 一种用于视频生成的统一且灵活的运动控制器。在 ACM SIGGRAPH 2024 会议论文集中，第 1-11 页，2024b。

尼古拉斯·沃特斯、丹尼尔·佐兰、西奥万·韦伯、彼得·巴塔利亚、拉兹万·帕斯卡努和阿德里亚娜·塔切蒂。视觉交互网络：从视频中学习物理模拟器。发表于《神经信息处理系统进展》，2017。

贾俊武、伊尔克·伊尔迪里姆、约瑟夫·J·林、威廉·T·弗里曼、和约书亚·B·坦南鲍姆。伽利略：通过将物理引擎与深度学习相结合来感知物理对象属性。发表于《神经信息处理系统进展》，2015。

贾俊武、艾丽卡·卢、普什梅特·科利、比尔·弗里曼和乔什·坦南鲍姆。通过视觉去动画学习物理。发表于《神经信息处理系统进展》，2017。

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian：用于生成动力学的物理集成 3D 高斯。发表于 IEEE/CVF 计算机视觉与模式识别会议，2024。

Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative 卡通动画。在 IEEE/CVF 计算机视觉与模式识别会议，2025。

徐天舒，陈志飞，吴磊，卢浩，陈雨莹，蒋丽辉，刘冰冰，陈英丛。运动梦想家：通过场景感知运动推理实现物理一致的生成视频。arXiv 预印本 arXiv:2412.00547，2024。

薛奇瑶，尹祥宇，杨博远，高伟。PhyT2V：基于 LLM 指导的迭代自细化物理基础文本到视频生成。在 IEEE/CVF 计算机视觉与模式识别会议，2025。

杨新迪，李宝路，张一鸣，尹振飞，白雷，马丽倩，王志勇，蔡建飞，王天霖，卢胡川，贾旭。Vlipp：基于视觉和语言物理先验的物理合理生成视频。在计算机视觉国际会议，2025a。

朱耀义，滕嘉言，郑文迪，丁明，黄诗宇，徐嘉政，杨元明，洪文怡，张晓寒，冯冠宇等. Cogvideox: 基于专家变压器的文本到视频扩散模型. 国际学习表征会议，2025b.

尹胜明，吴晨飞，梁健，石杰，李厚强，明公，段楠. Dragnuwa: 通过整合文本、图像和轨迹实现视频生成的细粒度控制. arXiv 预印本 arXiv:2308.08089, 2023.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *International Conference on Computer Vision*, 2023.

Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chenyu Zhang, Daniil Cherniavskii, Andrii Zadaianchuk, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025a.

Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M. Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision*, 2023.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, 2024.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025c.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025d.

Yaofeng Desmond Zhong and Naomi Ehrich Leonard. Unsupervised learning of lagrangian dynamics from images for prediction and control. In *Advances in Neural Information Processing Systems*, 2020.

叶元，宋佳明，乌马尔·伊克巴尔，阿拉斯·瓦哈特和简·考茨。Physdiff：物理引导的人类运动扩散模型。在 2023 年国际视觉会议。

Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. 生成式摄影：场景一致的相机控制，用于逼真的文本到图像合成。IEEE/CVF 计算机视觉与模式识别会议，2025 年。

张晨宇，丹尼尔·切尔尼亚夫斯基，安德烈·扎达亚诺丘克，安东尼奥斯·特拉戈达拉斯，安东尼奥斯·沃齐基斯，蒂姆·尼贾姆，德克·W·E·普林霍恩，马克·博德拉斯卡，尼库·塞贝，以及埃夫斯特拉托斯·加夫斯。Mimeuq：使用真实物理实验对视频生成模型的物理推理进行基准测试。arXiv 预印本 arXiv:2504.02918，2025a。

张科，萧晨，梅艺群，徐嘉聪，以及 Vishal M. Patel。扩散前思考：LLMs 指导的物理感知视频生成。arXiv 预印本 arXiv:2505.21653，2025b。

张吕敏、饶安怡和 Agrawala Maneesh。为文本到图像扩散模型添加条件控制。在 2023 年 IEEE 国际计算机视觉会议。

张天元、于宏兴、吴润迪、冯 Brandon Y.、郑长西、Snavely Noah、吴嘉俊和 William T. Freeman。PhysDreamer：通过视频生成实现基于物理的 3D 物体交互。在 2024 年欧洲计算机视觉会议。

张祥东、廖嘉琪、张少峰、孟芳庆、万祥鹏、闫俊驰和程宇。Videorepa：通过基础模型的关系对齐学习物理以生成视频。arXiv 预印本 arXiv:2505.23656，2025c。

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025d.

钟瑶峰 Desmond 和 Leonard Naomi Ehrich。拉格朗日动力学的不监督学习

用于预测和控制图像。在《神经信息处理系统进展》中，2020.

# Appendix

## A Appendix Introduction

This appendix provides additional discussions and details on the physics-clean video simulator (Section B), Neural Newtonian Dynamics network design and prediction accuracy analysis (Section C), evaluation details (Section D), more visual results (Section E), and a Q & A section (Section F). To illustrate the continuity and effects of physical coherence and controllability, **we recommend that readers view the Videos** included in the Supplementary Materials.

## B More Details of the Data Simulator

For each type of motion, we construct a physics-clean dataset for training the NND model. Our simulator is built upon physical principles and renders videos with time stamps. The simulator supports multi-parameter control, including initial position, velocity, orientation, angular velocity, world settings (world size, friction coefficient, acceleration/deceleration coefficient, damping coefficient, pivot point), and object properties (size, shape). Representative samples are shown in Figure. 5, while the complete simulator code and additional video examples are provided in the Supplementary Materials.

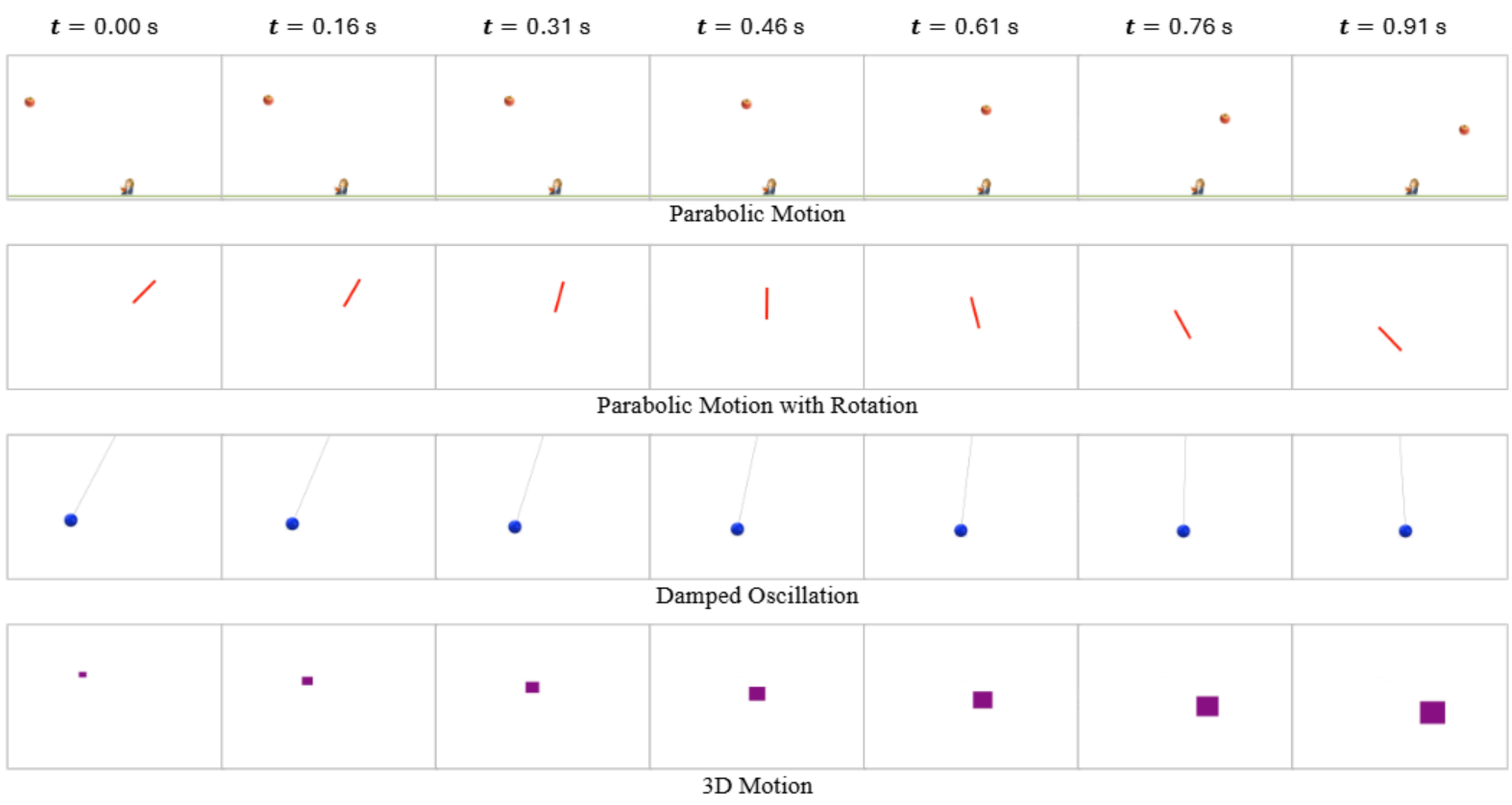



Figure 5: Sample physics-clean videos generated by our simulator.

## C More Details of Neural Newtonian Dynamics

### C.1 Neural Newtonian Dynamics Network

In Algorithm 1 we present the detailed architecture of the Neural Newtonian Dynamics network. We model the most salient dynamics using a physics-driven linear Neural ODEs, and augment it with a learnable MLP to capture nonlinear and unknown dynamics. The full network implementation is provided in the Supplementary Materials.

# 附录

附录介绍 A

本附录提供了关于物理清洁视频模拟器（第 B 节）、神经牛顿动力学网络设计和预测精度分析（第 C 节）、评估细节（第 D 节）、更多视觉结果（第 E 节）以及问答部分（第 F 节）的额外讨论和细节。为了说明物理一致性和可控性的连续性和效果，我们建议读者观看补充材料中包含的视频。

B 数据模拟器的更多细节

对于每种运动类型，我们构建了一个物理干净的数据库来训练 NND 模型。我们的模拟器基于物理原理构建，并渲染带有时间戳的视频。模拟器支持多参数控制，包括初始位置、速度、方向、角速度、世界设置（世界大小、摩擦系数、加速度/减速度系数、阻尼系数、枢轴点）以及物体属性（大小、形状）。代表性样本显示在图 5 中，完整的模拟器代码和额外的视频示例在补充材料中提供。

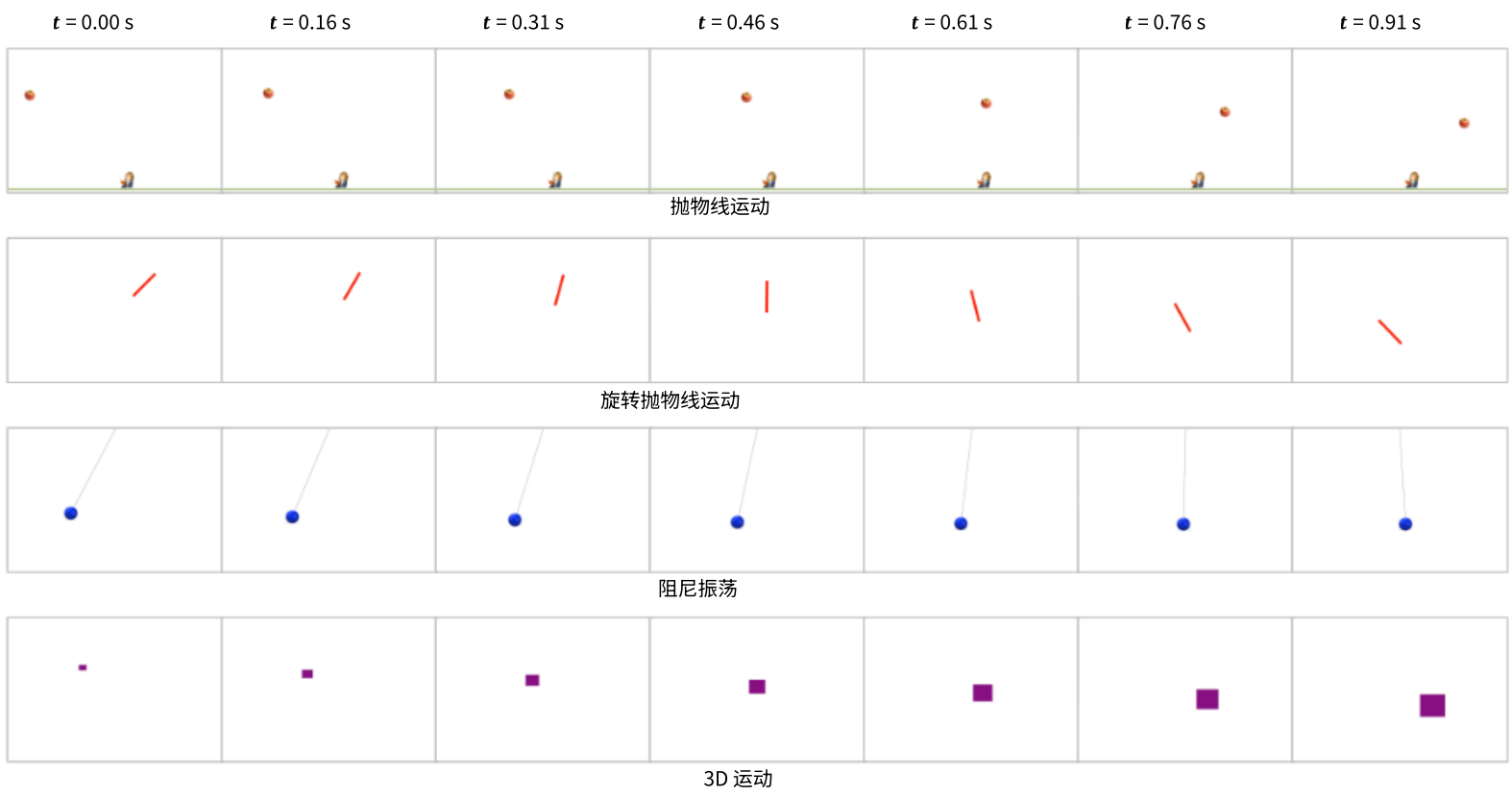



图 5：我们模拟器生成的样本物理清洁视频。

C 更详细的神经牛顿动力学

C.1 NNDN

在算法 1 中，我们展示了神经牛顿动力学网络（NNDN）的详细架构。我们使用物理驱动的线性神经常微分方程来建模最显著的动力学，并使用可学习的多层感知机（MLP）来增强它，以捕获非线性动力学和未知动力学。完整的网络实现提供在补充材料中。

**Algorithm 1** Neural Newtonian Dynamics Network Architecture

**Require:** Initial physical state $\mathbf{Z}_0 = [x, y, v_x, v_y, \theta, \omega, s, l, a]$, time stamps $t_0, \ldots, t_T$
**Ensure:** Future Latent physical states $\mathbf{Z}(t)$

1: Define learnable parameters:
- $(a_x, b_x, c_x), (a_y, b_y, c_y)$ for linear 2nd-order dynamics of $(x, y)$
- $(g/L, \gamma)$ for linearized pendulum or circular motion $(\theta, \omega)$
- $(\alpha_s, \beta_s), (\alpha_l, \beta_l), (\alpha_a, \beta_a)$ for 1st-order dynamics of $(s, l, a)$
- Residual scale $\epsilon$

2: Define residual MLP: $\text{ResMLP} : \mathbb{R}^9 \to \mathbb{R}^6$ (initialized to 0)
3: **for** each time $t$ **do**
4: Split $\mathbf{Z} = [x, y, v_x, v_y, \theta, \omega, s, l, a]$
5: Compute linear dynamics:

$$
\begin{aligned}
a_x^{\text{lin}} &= a_x x + b_x v_x + c_x \\
a_y^{\text{lin}} &= a_y y + b_y v_y + c_y \\
d\theta/dt &= \omega \\
d\omega^{\text{lin}}/dt &= -(g/L)\theta - \gamma\omega \\
ds^{\text{lin}}/dt &= \alpha_s s + \beta_s, \quad dl^{\text{lin}}/dt = \alpha_l l + \beta_l, \quad da^{\text{lin}}/dt = \alpha_a a + \beta_a
\end{aligned}
$$

6: Compute residual correction:

$$
[a_x^{\text{res}}, a_y^{\text{res}}, d\omega^{\text{res}}, ds^{\text{res}}, dl^{\text{res}}, da^{\text{res}}] = \epsilon \cdot \tanh(\text{ResMLP}(\mathbf{Z}))
$$

7: Update derivatives:

$$
\frac{d\mathbf{Z}}{dt} = [v_x, v_y, a_x^{\text{lin}} + a_x^{\text{res}}, a_y^{\text{lin}} + a_y^{\text{res}}, d\theta/dt, d\omega^{\text{lin}} + d\omega^{\text{res}}, ds^{\text{lin}} + ds^{\text{res}}, dl^{\text{lin}} + dl^{\text{res}}, da^{\text{lin}} + da^{\text{res}}]
$$

8: **end for**
9: Integrate ODE by odeint to obtain $\mathbf{Z}(t)$ over $t_0, \ldots, t_T$

### C.2 Accuracy of Neural Newtonian Dynamics Predictions

Figure. 6 to Figure. 17 show the predictions of the trained Neural Newtonian Dynamics (NND) model for each type of motion. Given the initial physical state $\mathbf{Z}_0$, the model's predicted physical states closely follow the ground truth over time.

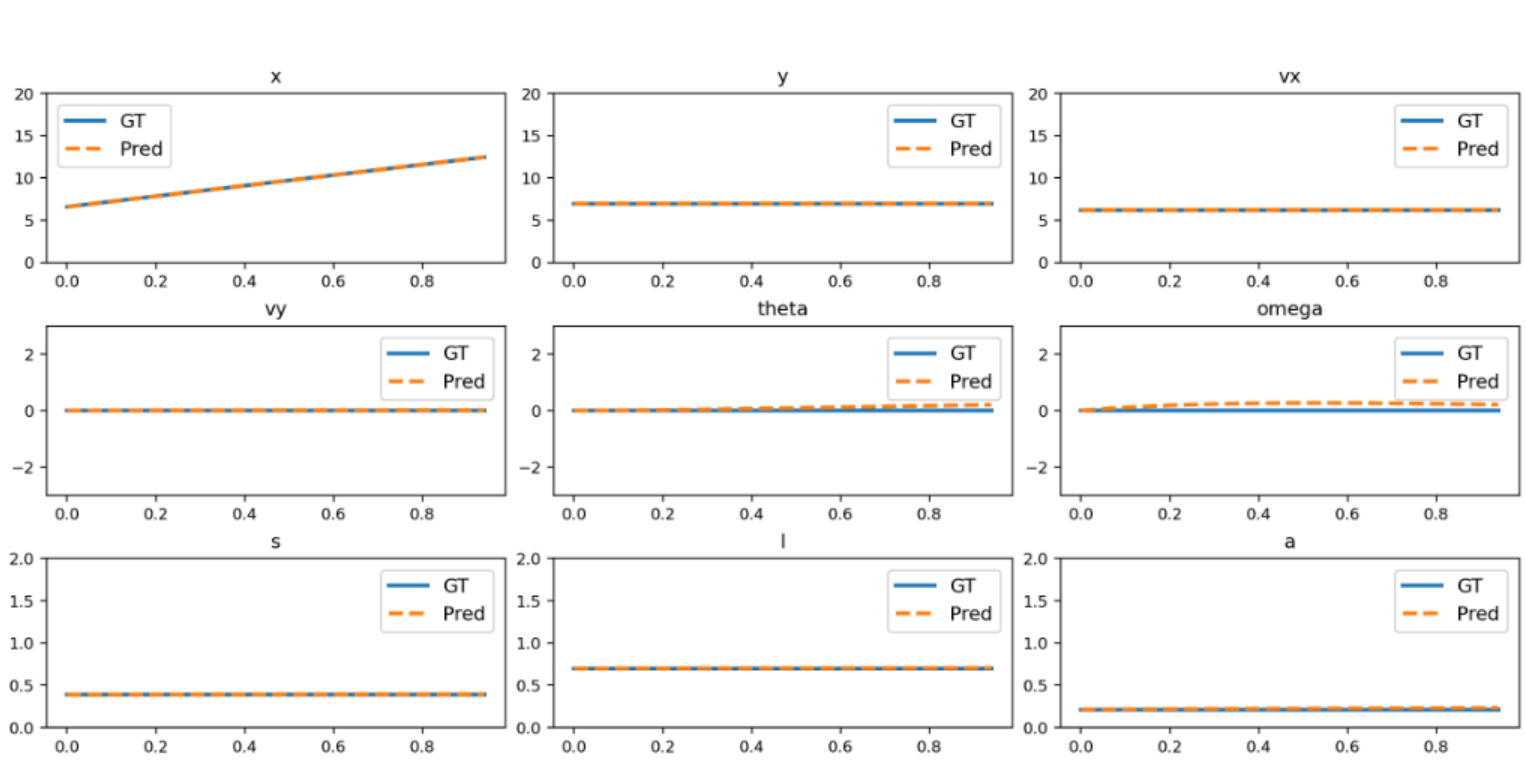



Figure 6: Comparison of NND predictions and ground truth for uniform motion.

算法 1 神经牛顿动力学网络架构

要求：初始物理状态 Z=[x, y, v, v, θ, ω, s, l, a]，时间戳 t, ..., t 确保未来潜在物理状态
Z(t) 1: 定义可学习参数：

- (a, b, c), (a, b, c) for linear 2nd-order dynamics of (x, y)
- (g/L, γ) 对于线性化摆动或圆周运动 (θ, ω)

• (α, β), (α, β), (α, β) 用于 (s, l, a) 的一阶动力学 • 残差尺度 $\epsilon^2$：定义
残差 MLP：ResMLP : R→ R（初始化为 0） 3: 对每个时间 t 进行 4: 分割
Z = [x, y, v, v, θ, ω, s, l, a] 5: 计算线性动力学：

$$a = ax + bv + c$$

$$a = ay + bv + c$$

$$d\theta/dt = \omega$$

$$d\omega/dt = -(g/L)\theta - \gamma\omega \quad ds/dt = \alpha s + \beta,\ dl/dt = \alpha l + \beta,\ da/dt = \alpha a + \beta$$

6: 计算残差校正：

$$[a, a, d\omega, ds, dl, da] = \epsilon \cdot \tanh(\text{ResMLP}(Z))$$

7: 更新导数：

$$\frac{dZ}{dt} = [v, v, a + a, a + a, d\theta/dt, d\omega + d\omega, ds + ds, dl + dl, da + da]$$

8: end for
9: 使用 odeint 积分 ODE 以在 t，...，t 上获得 Z(t)

C.2 ANNDP

图 6 至图 17 展示了训练好的神经牛顿动力学（NND）模型对每种运动类型的预测结果。给定初始物理状态 Z，模型预测的物理状态随时间紧密跟随真实值。

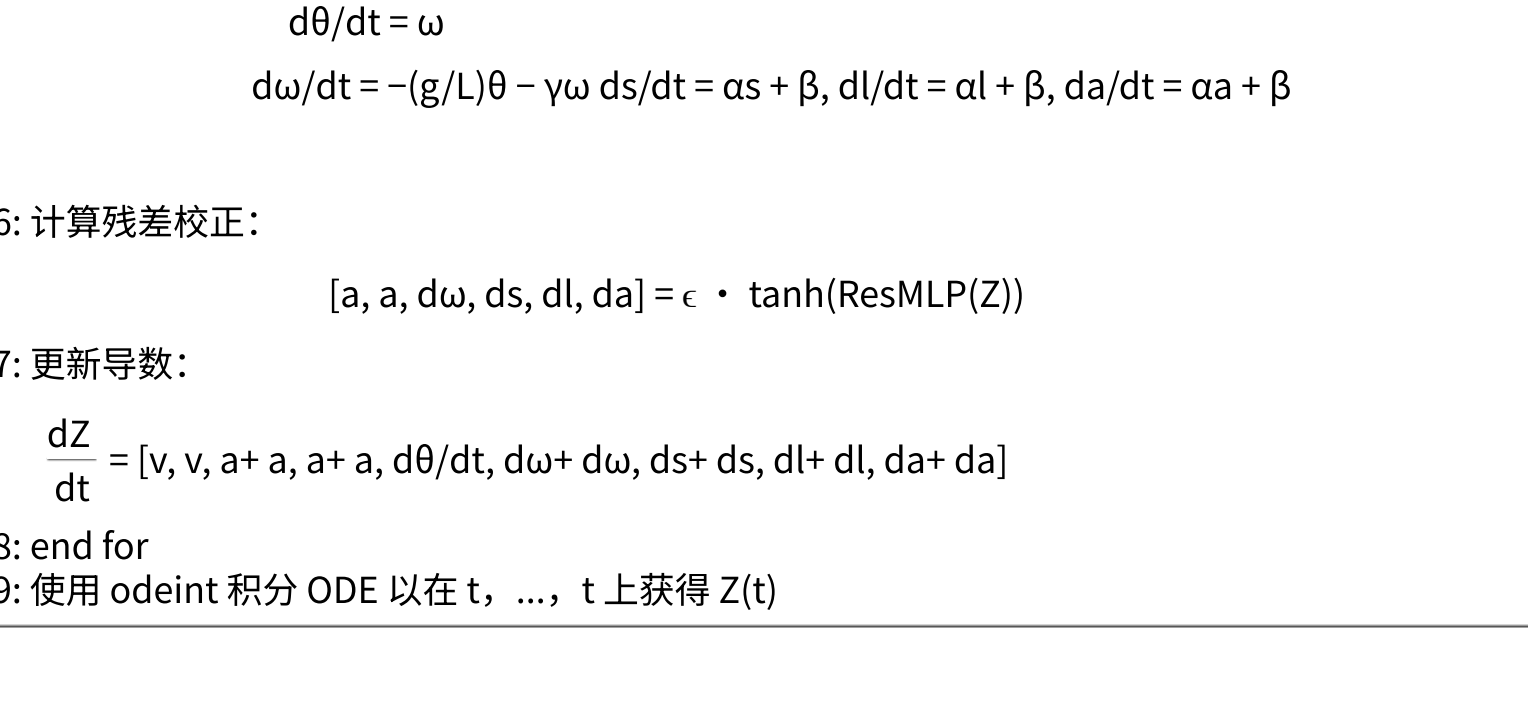



图 6：匀速运动中 NND 预测与真实值的比较。

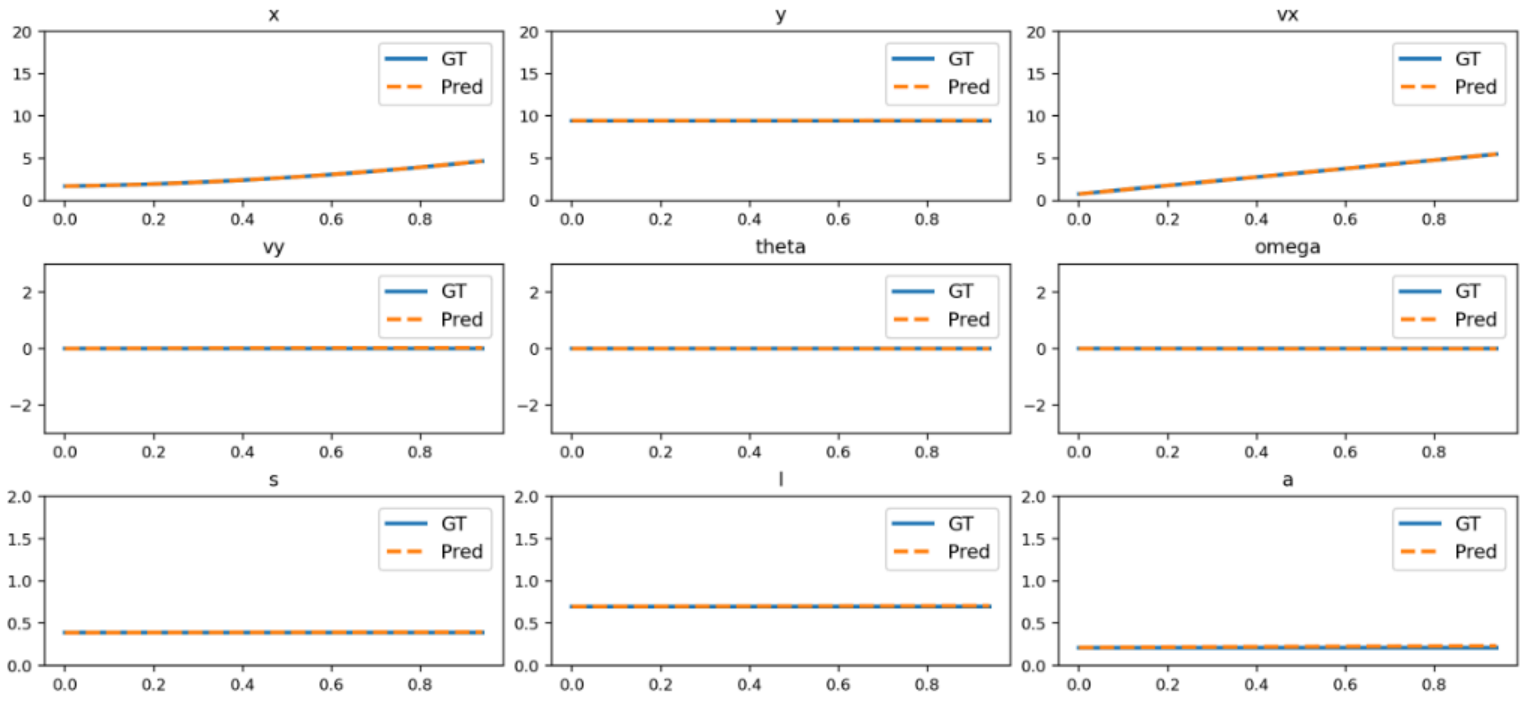



Figure 7: Comparison of NND predictions and ground truth for acceleration.

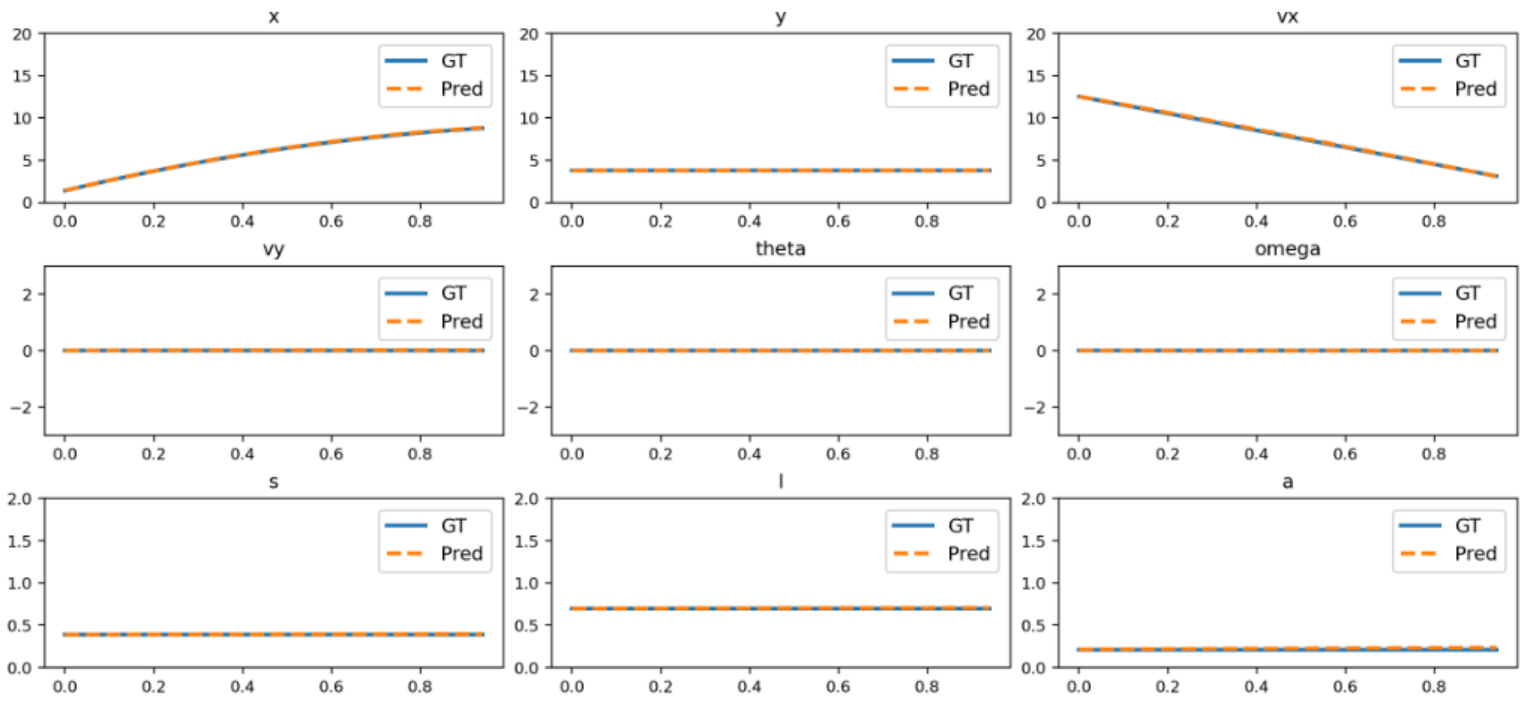



Figure 8: Comparison of NND predictions and ground truth for deceleration.

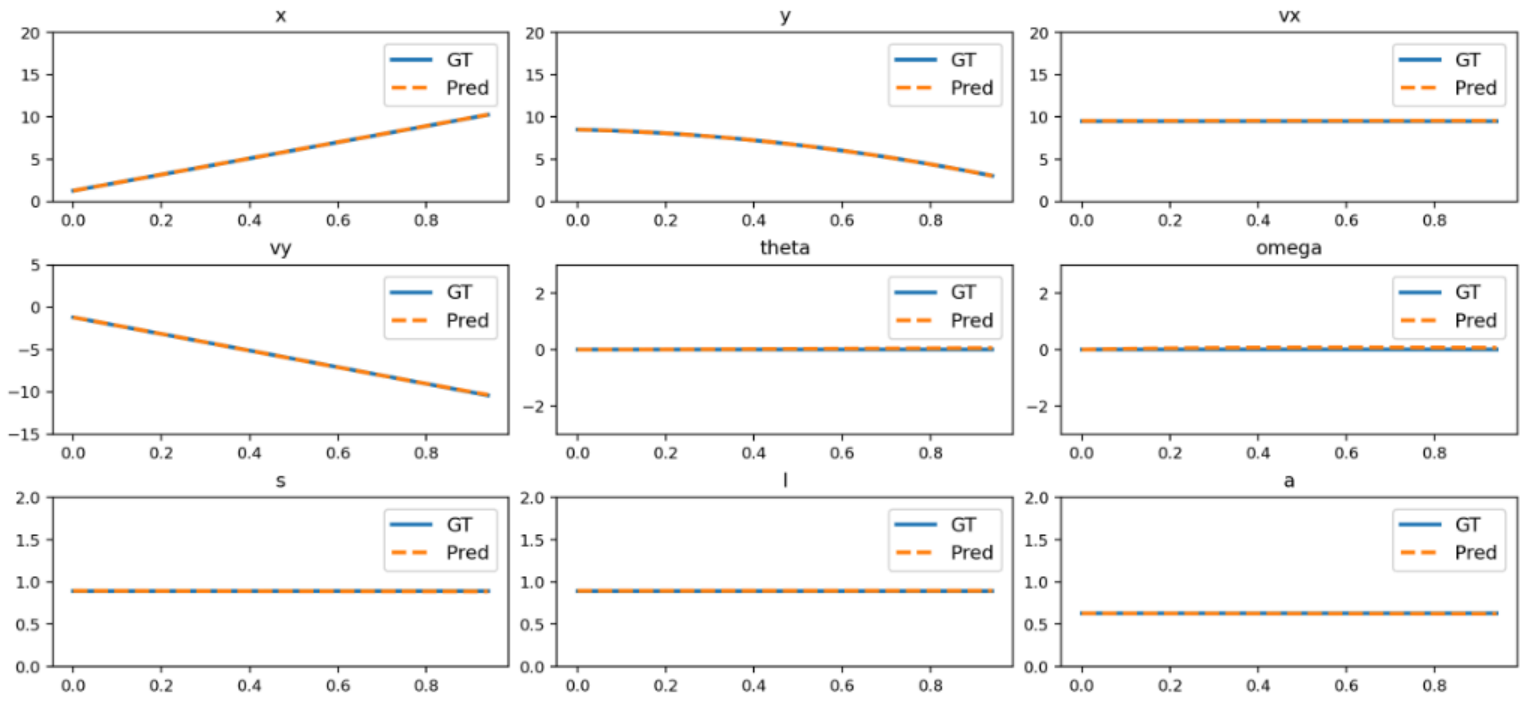



Figure 9: Comparison of NND predictions and ground truth for parabolic motion.

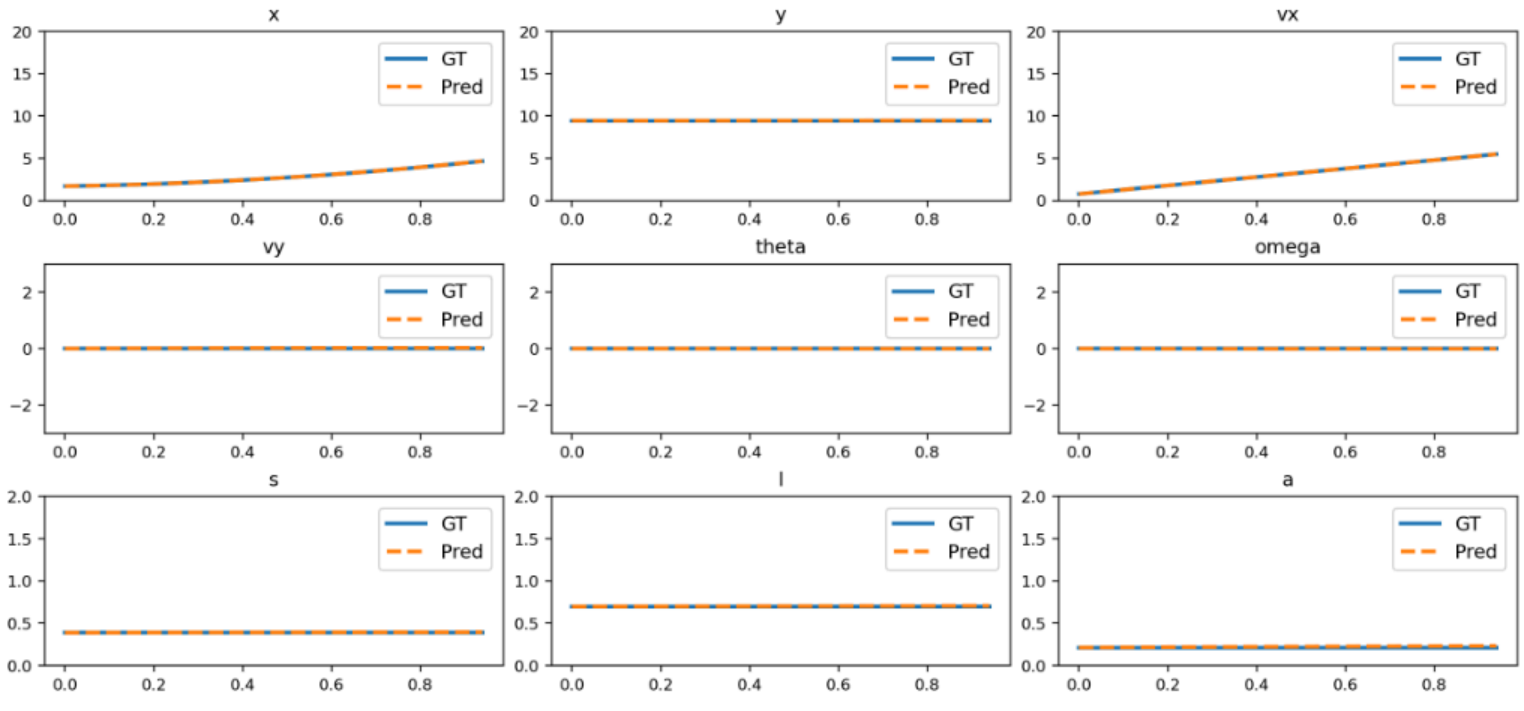



图 7：加速度的 NND 预测与真实值的比较。

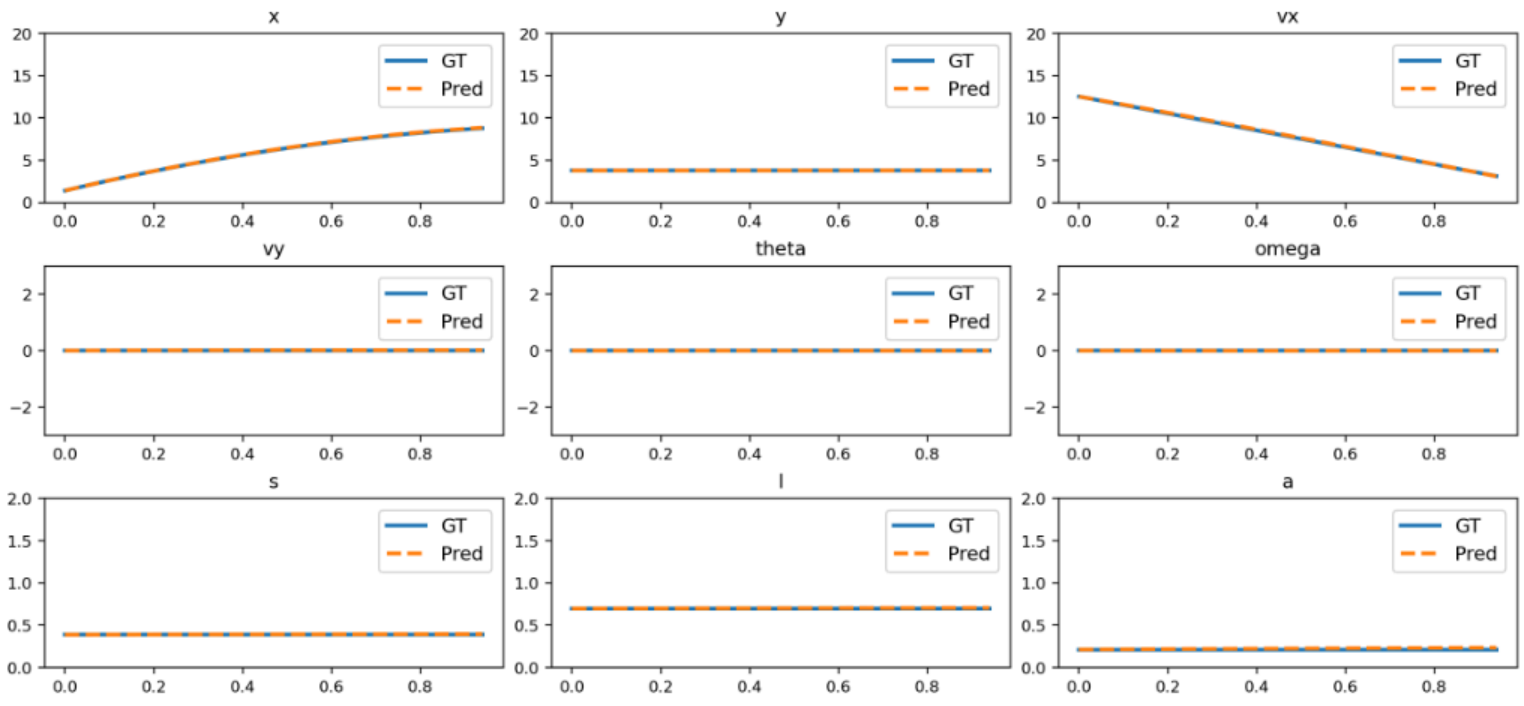



图 8：减速度的 NND 预测与真实值的比较。

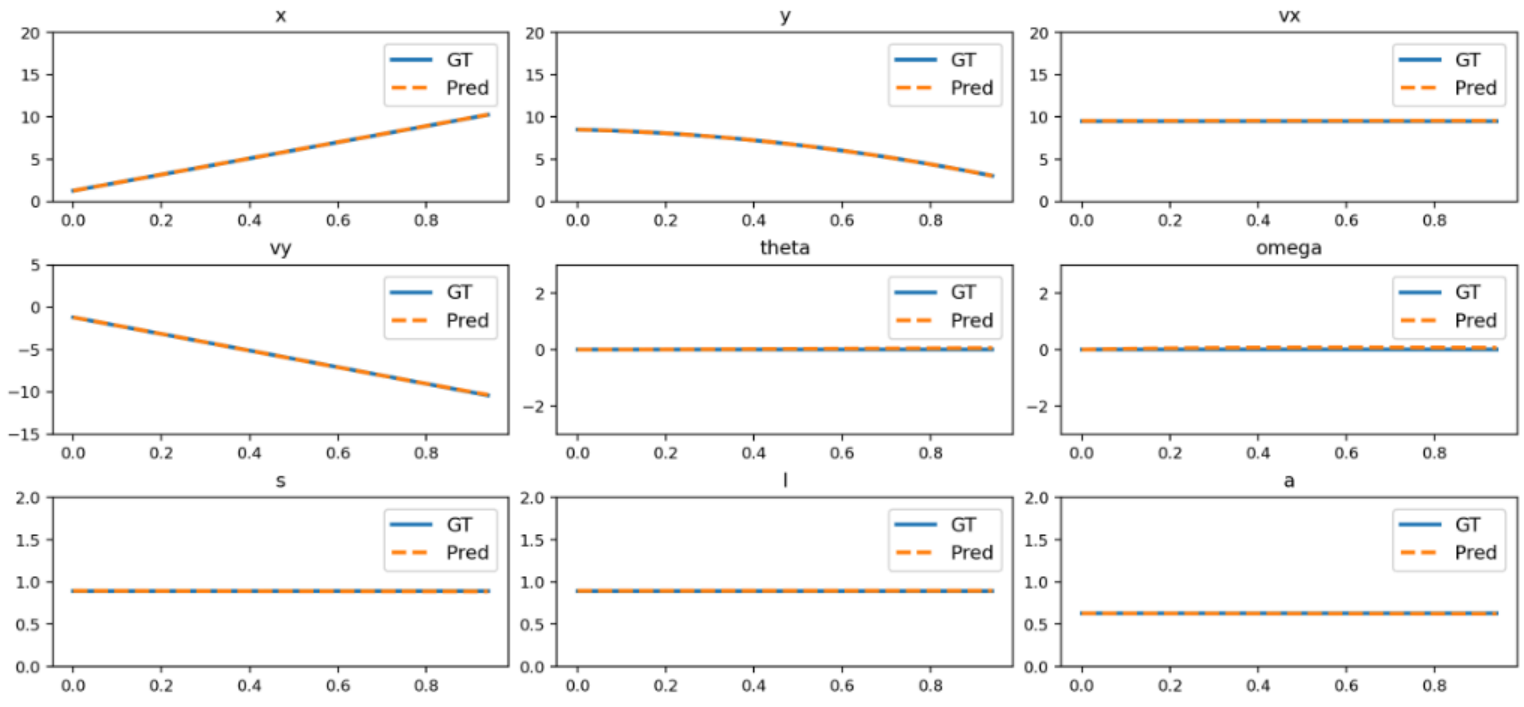



图 9：抛物线运动的 NND 预测与真实值的比较。

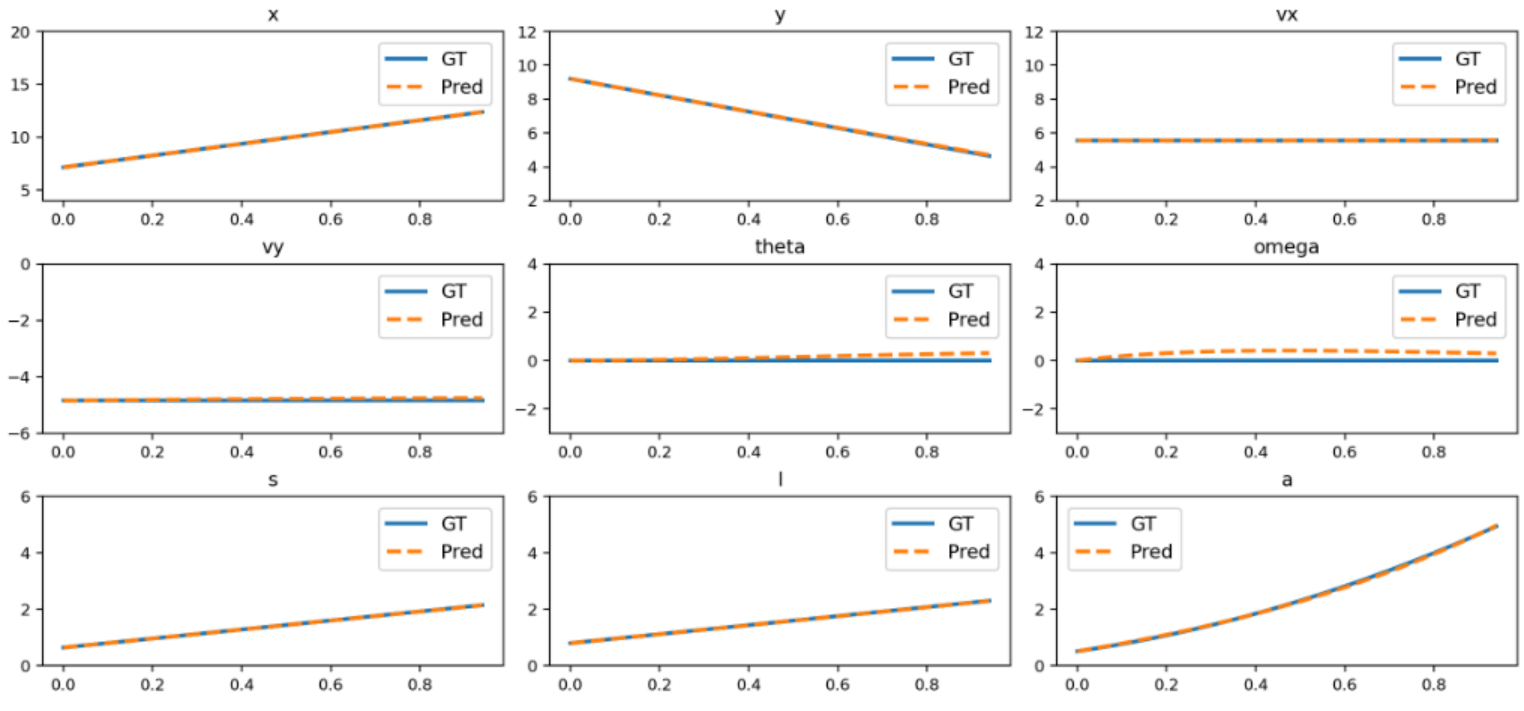



Figure 10: Comparison of NND predictions and ground truth for 3D motion.

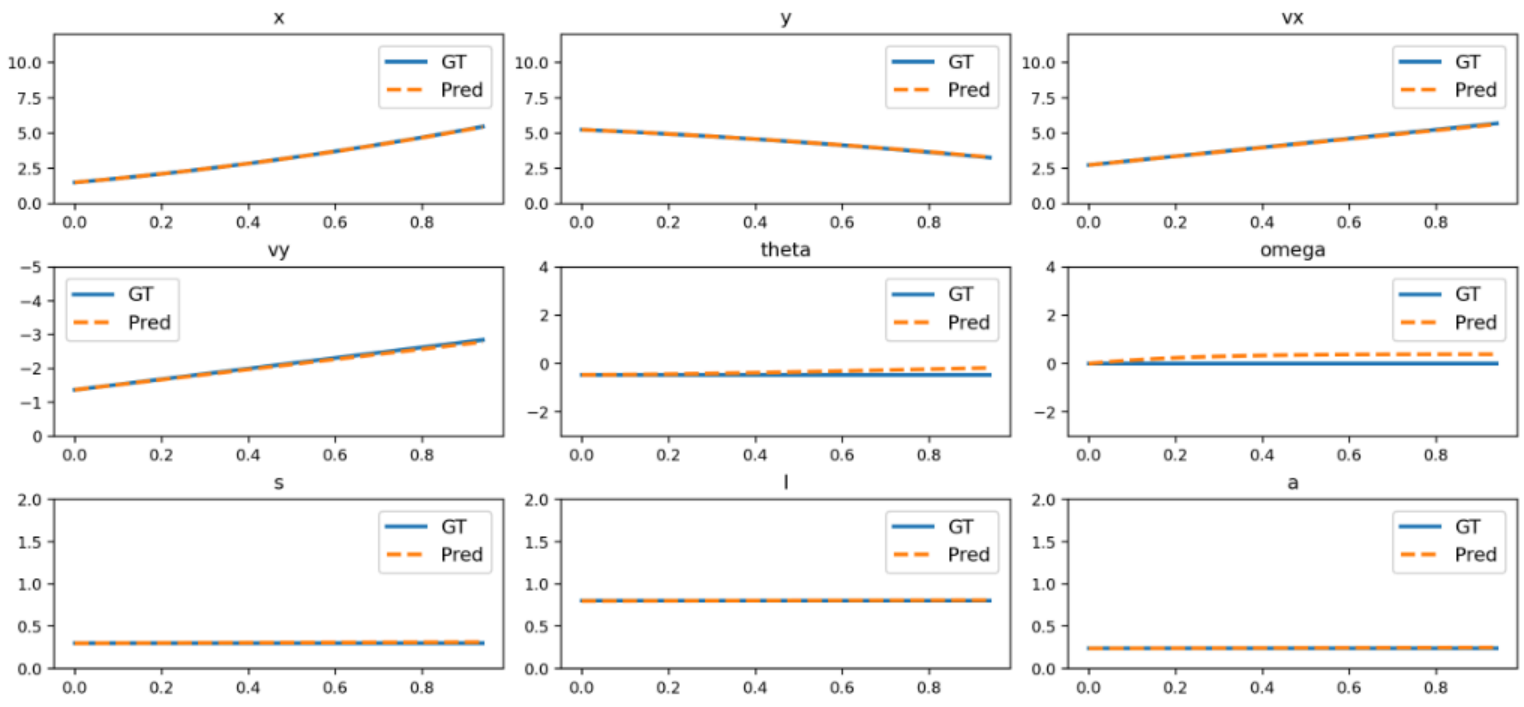



Figure 11: Comparison of NND predictions and ground truth for slope sliding.

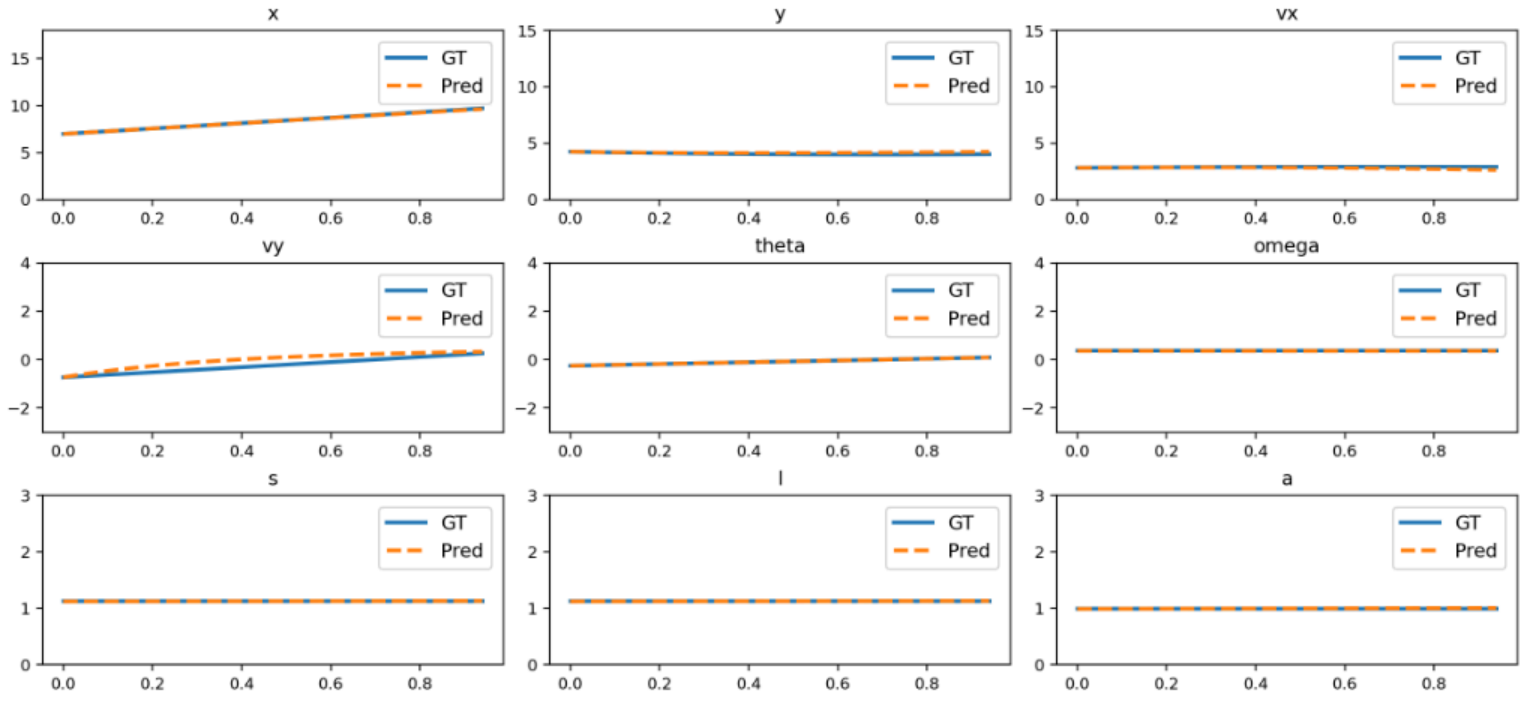



Figure 12: Comparison of NND predictions and ground truth for circular motion.

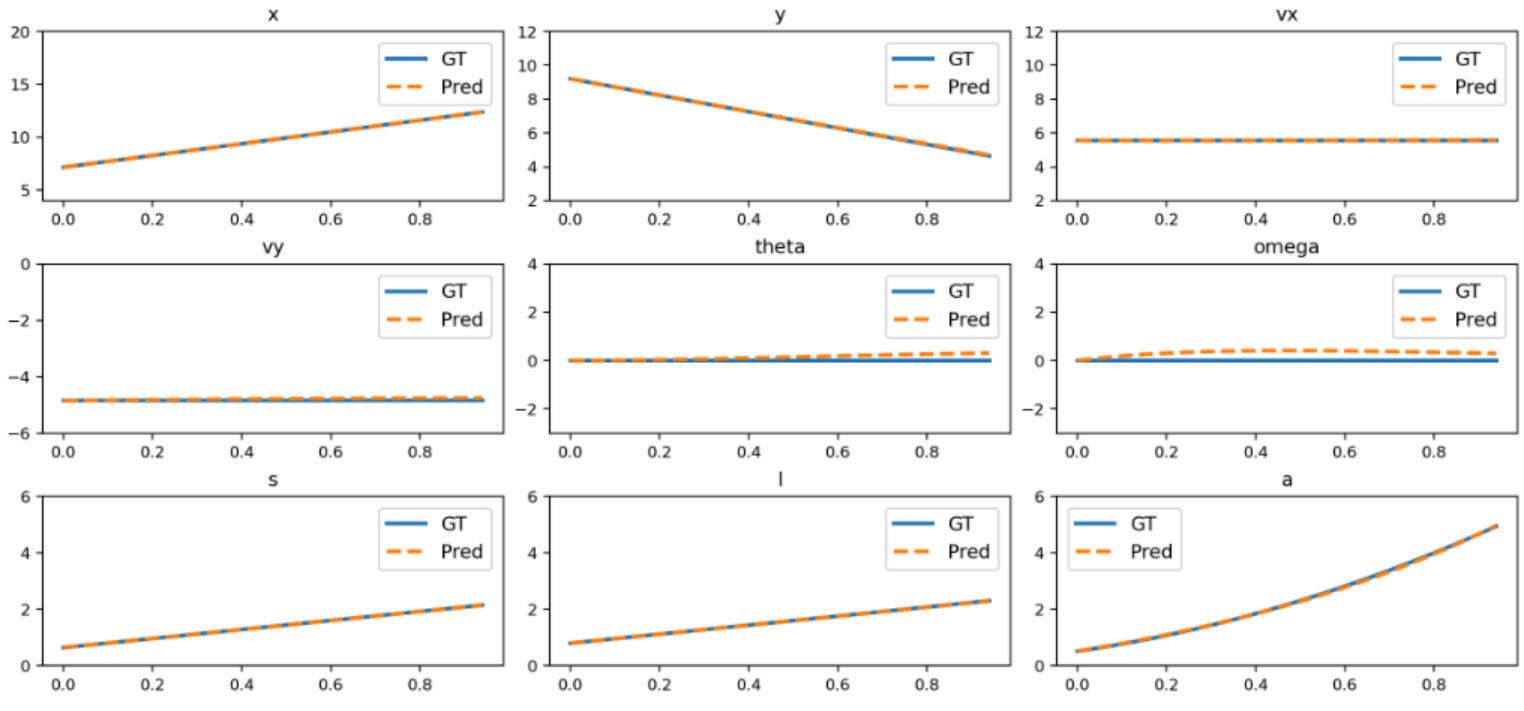



图 10：3D 运动中 NND 预测与真实值的比较

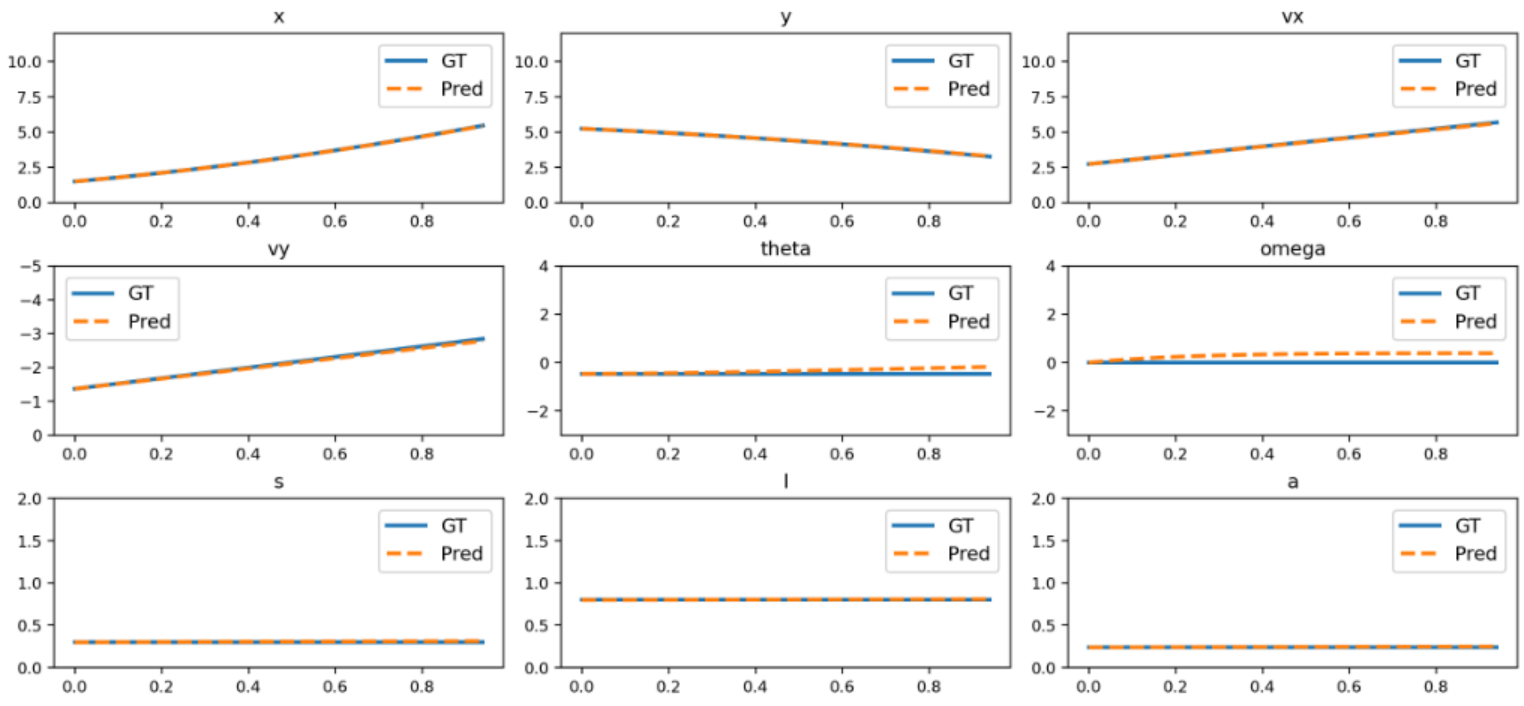



图 11：斜坡滑动中 NND 预测与真实值的比较。

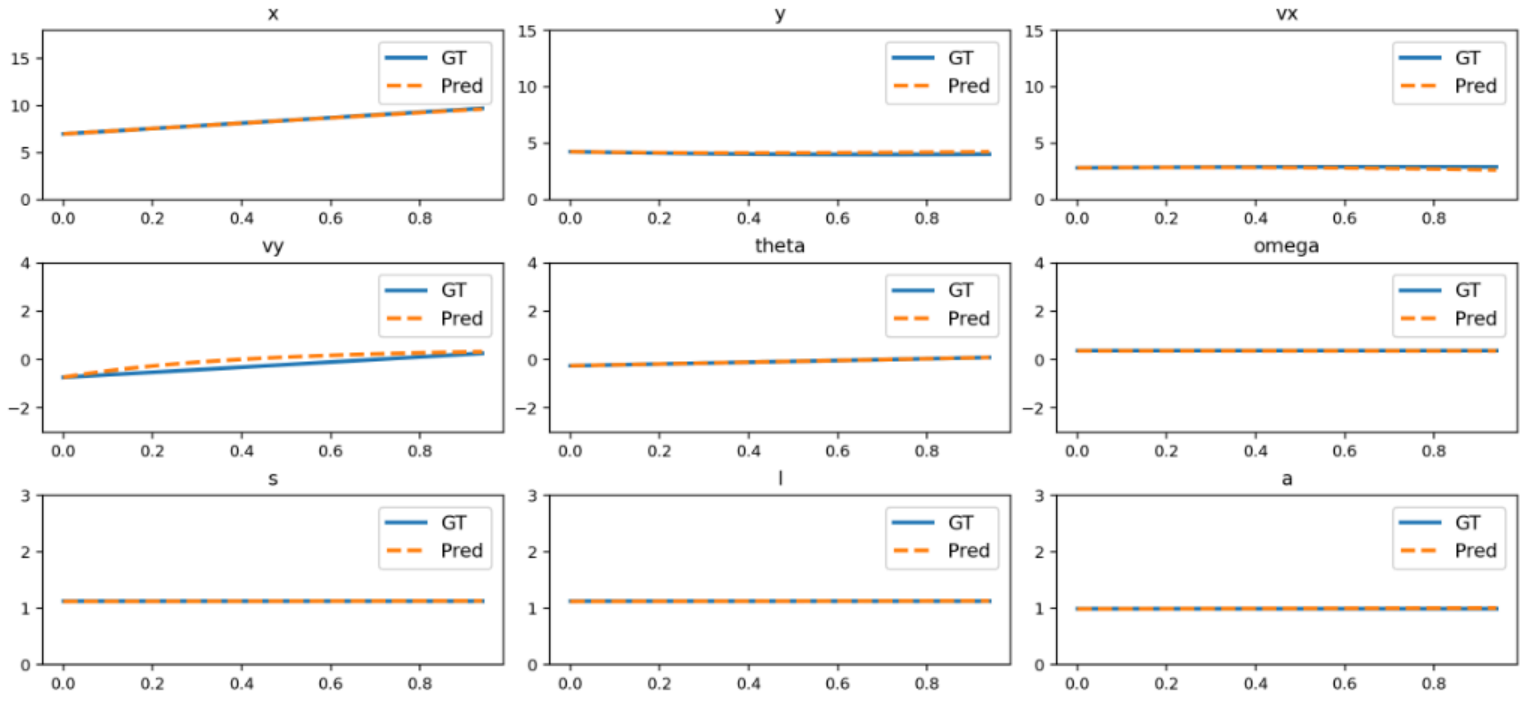



图 12：圆周运动中 NND 预测与真实值的比较。

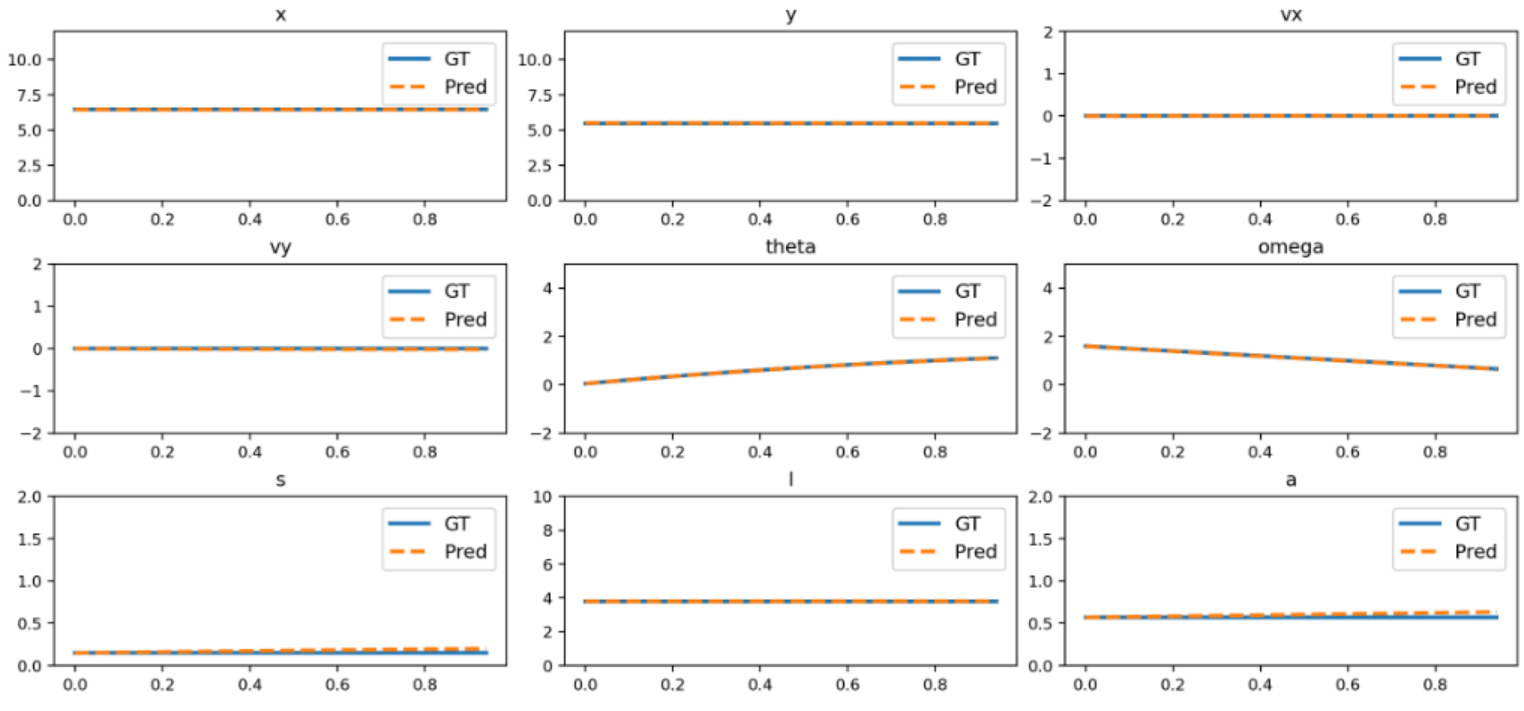



Figure 13: Comparison of NND predictions and ground truth for rotation.

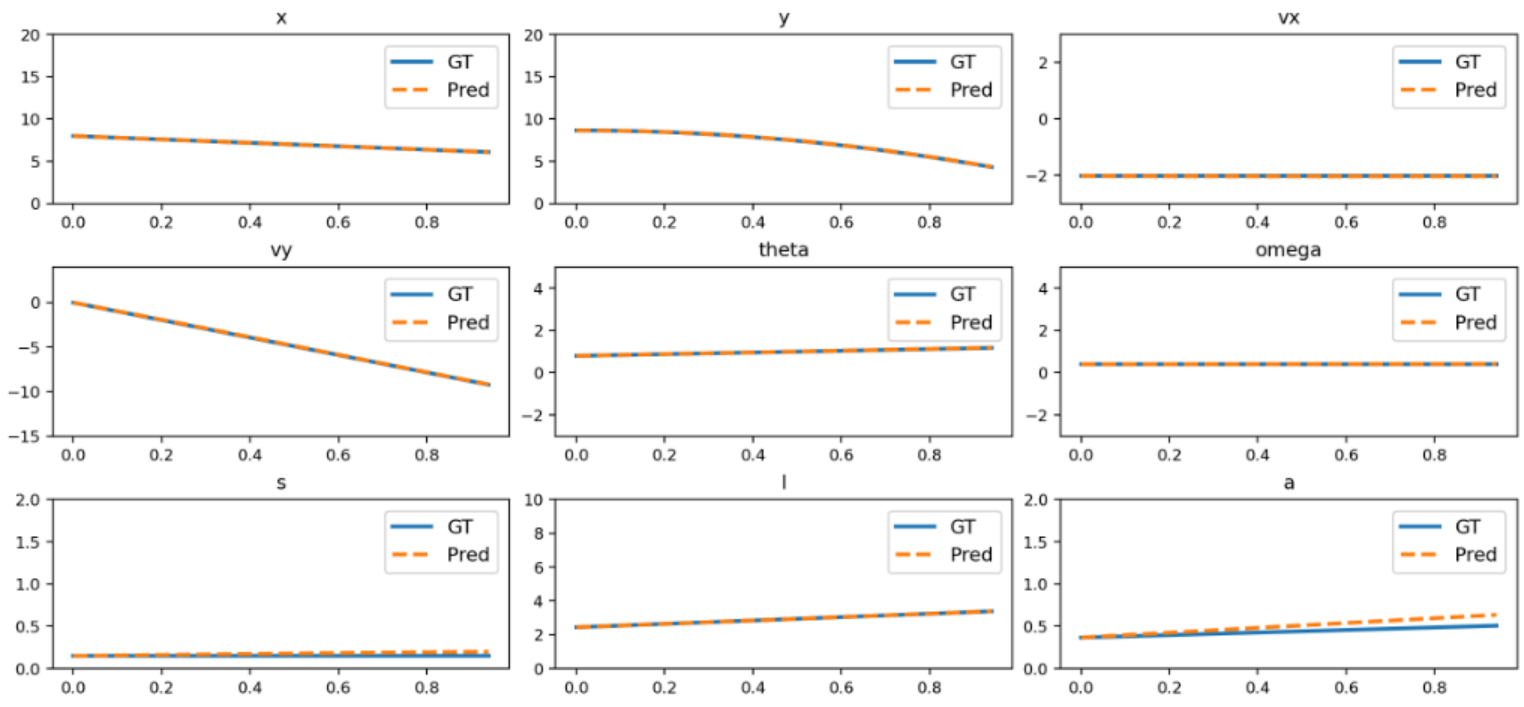



Figure 14: Comparison of NND predictions and ground truth for parabolic motion with rotation.

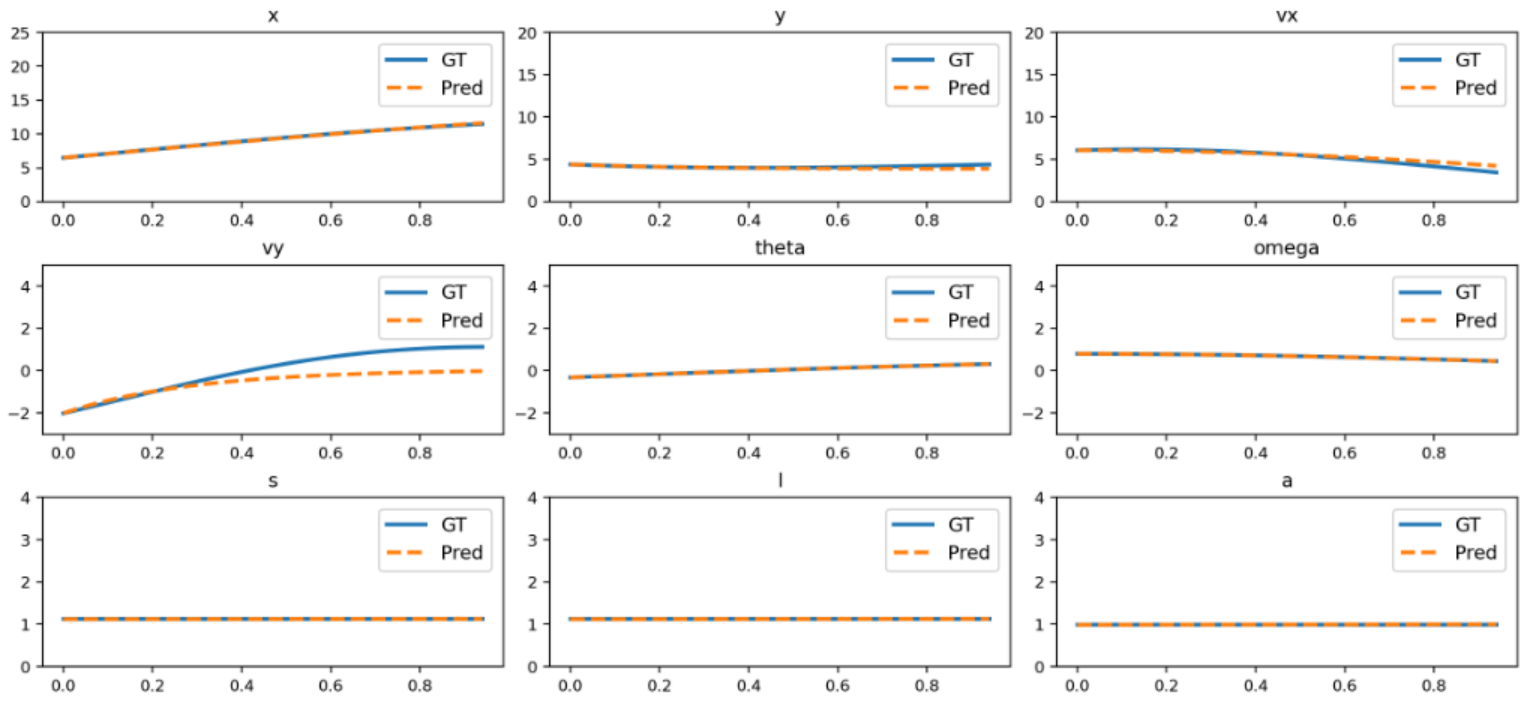



Figure 15: Comparison of NND predictions and ground truth for damped oscillation.

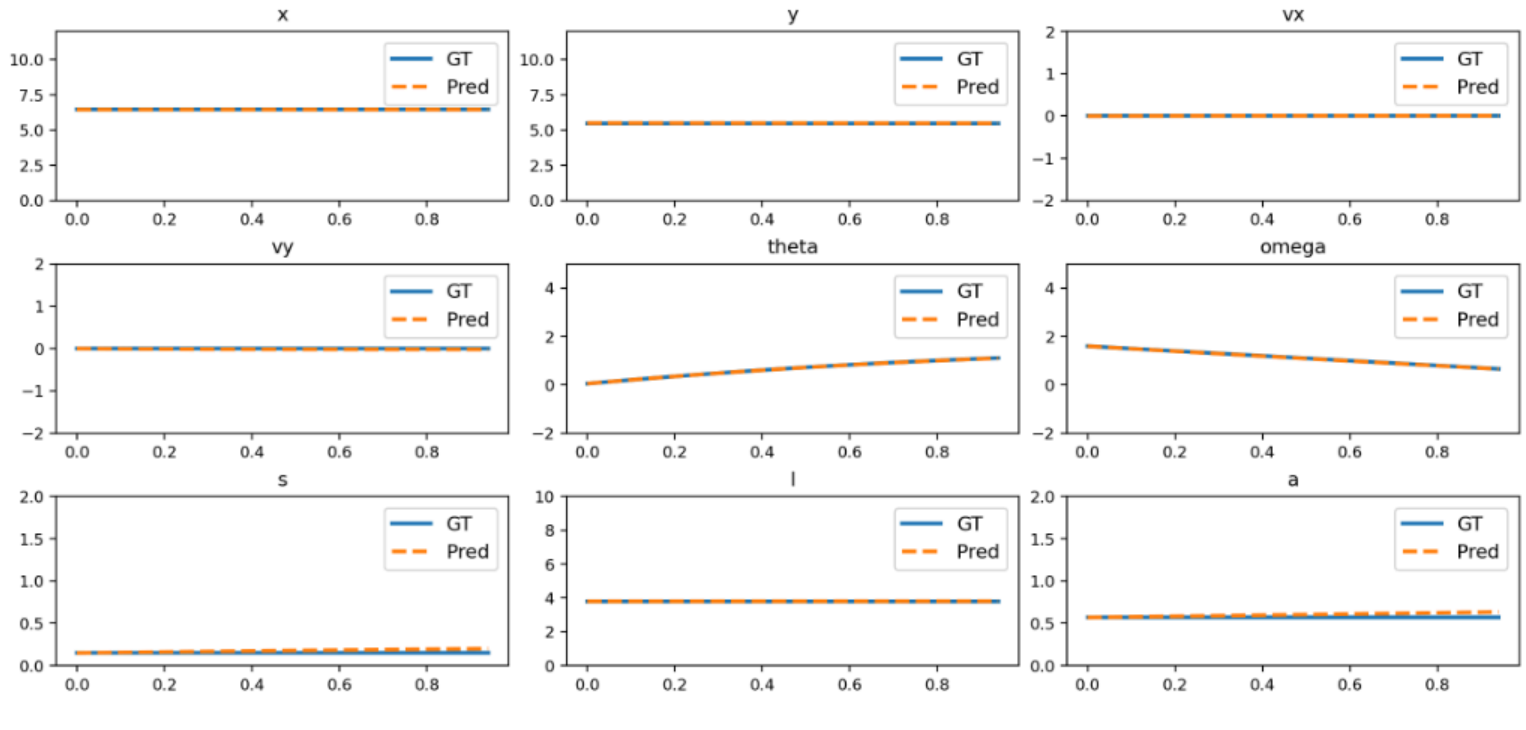



图 13：旋转的 NND 预测与真实值的比较

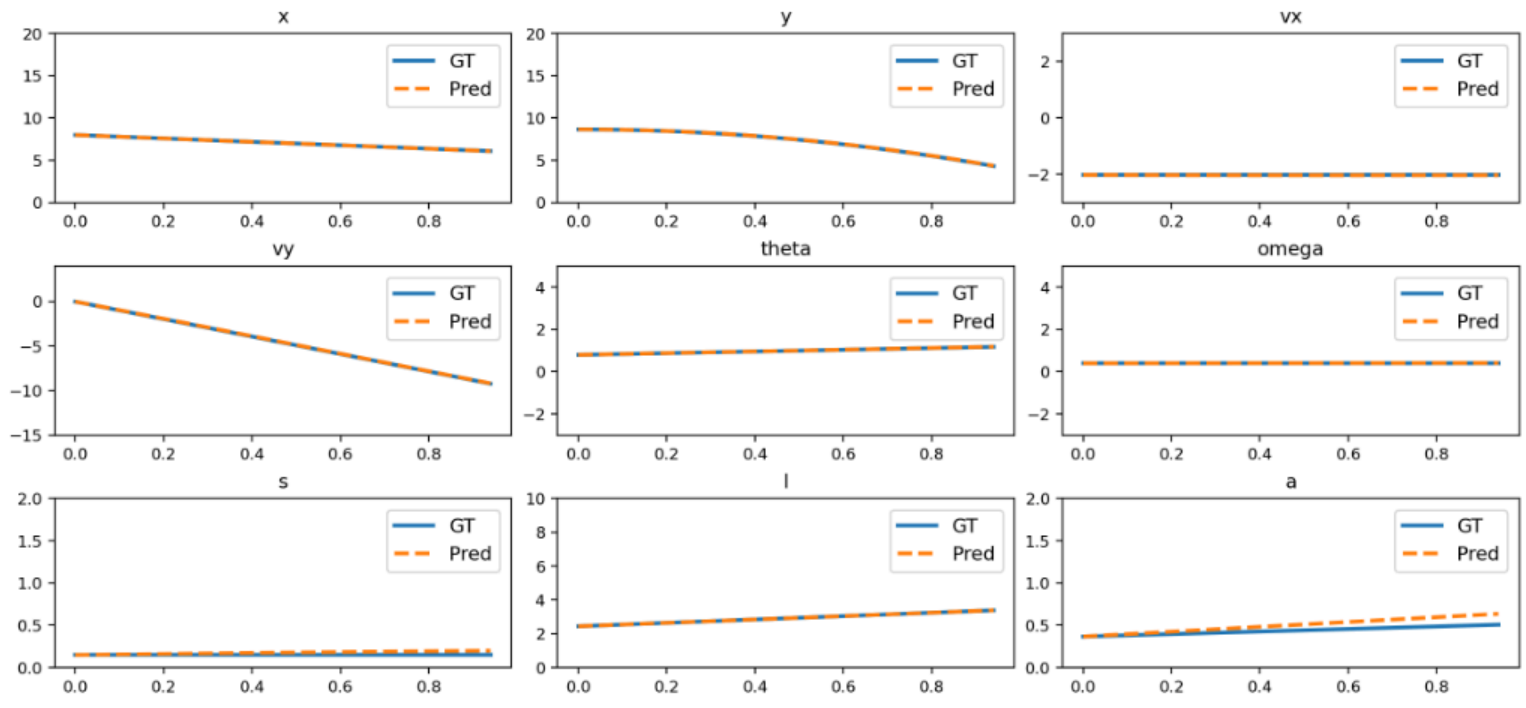



图 14：旋转抛物运动时 NND 预测与真实值的比较。

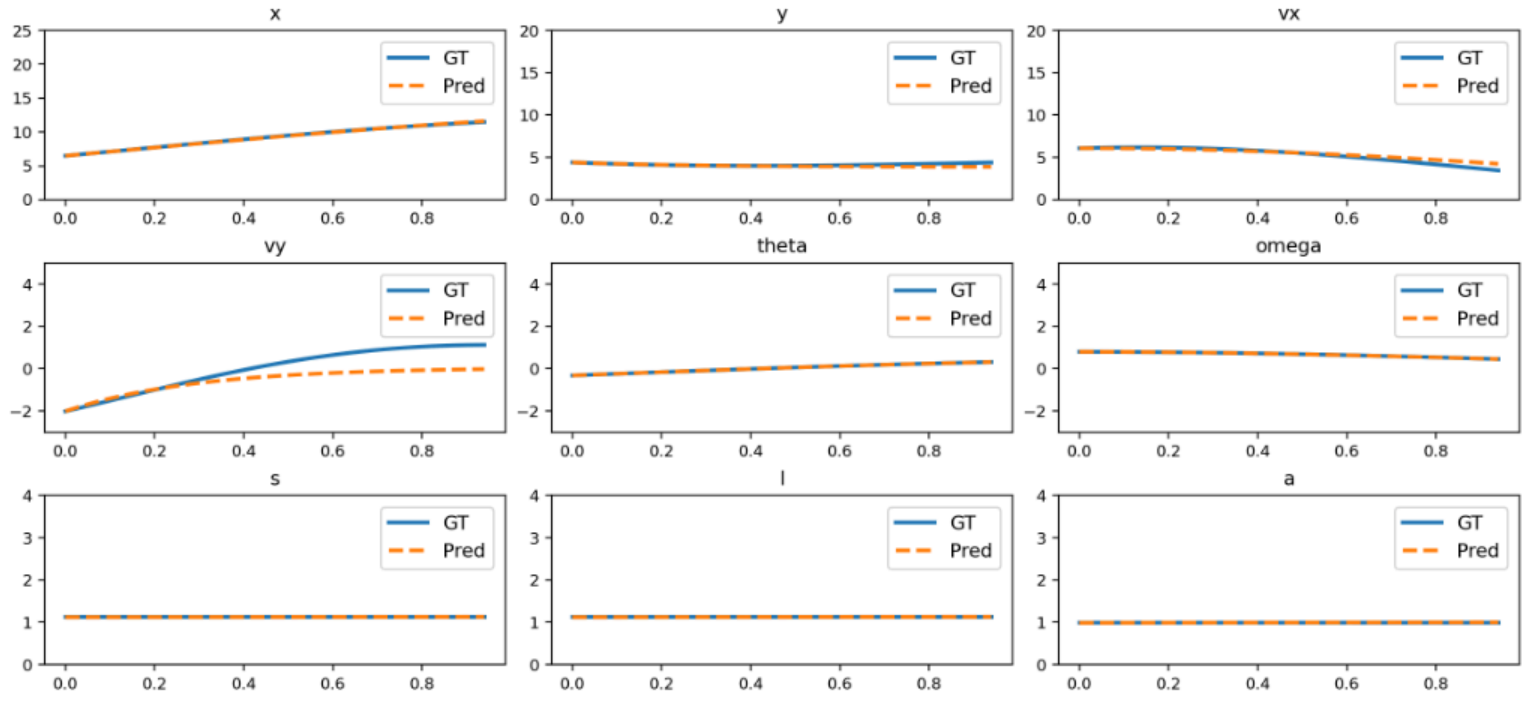



图 15：阻尼振荡的 NND 预测与真实值的比较

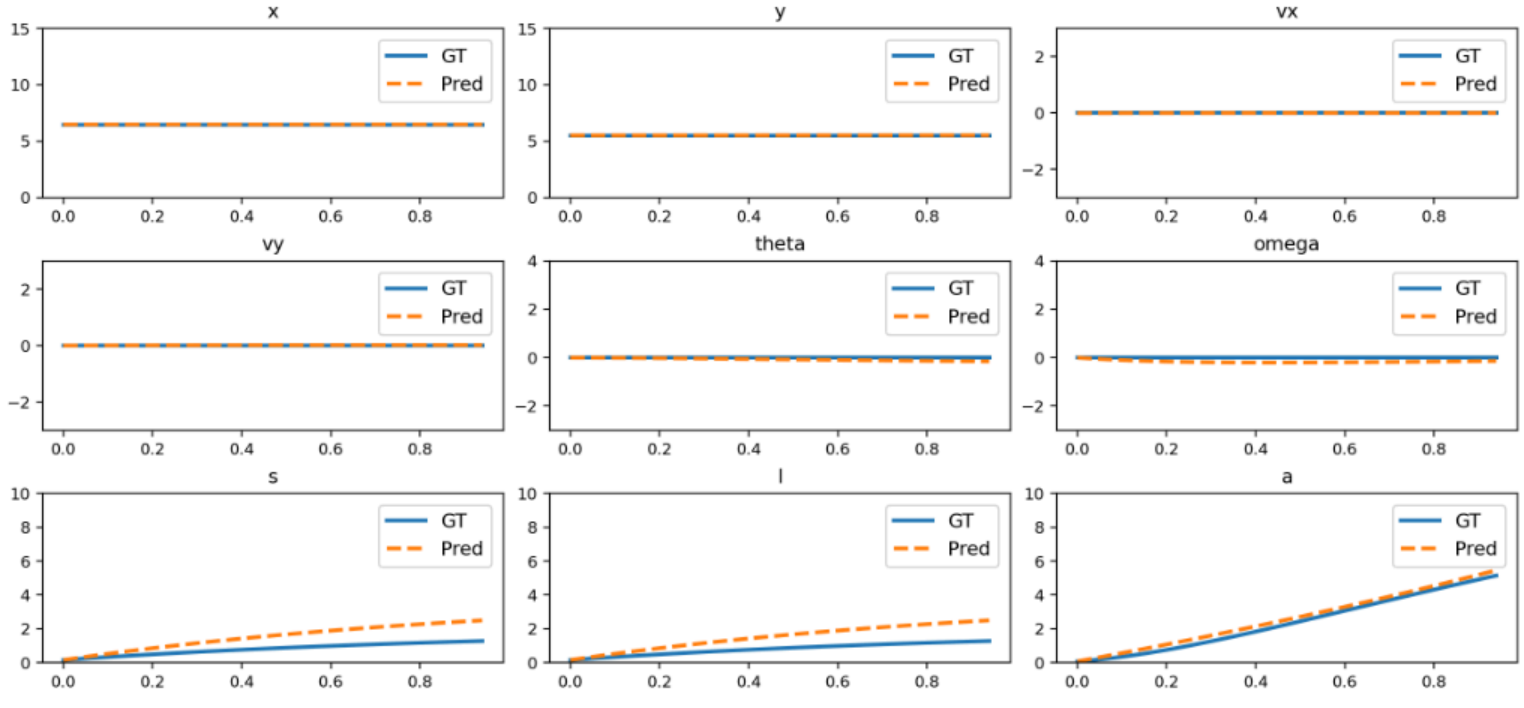



Figure 16: Comparison of NND predictions and ground truth for size changing.

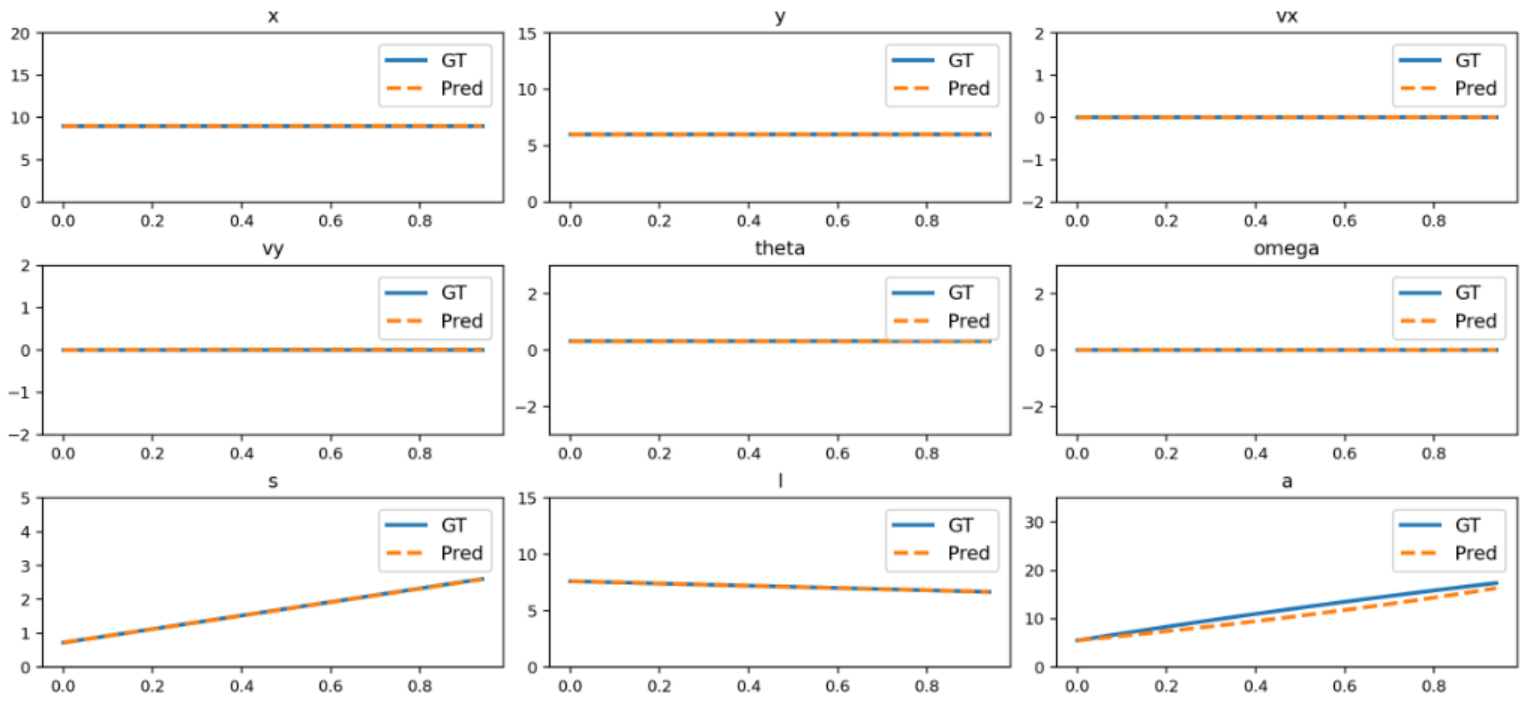



Figure 17: Comparison of NND predictions and ground truth for deformation.

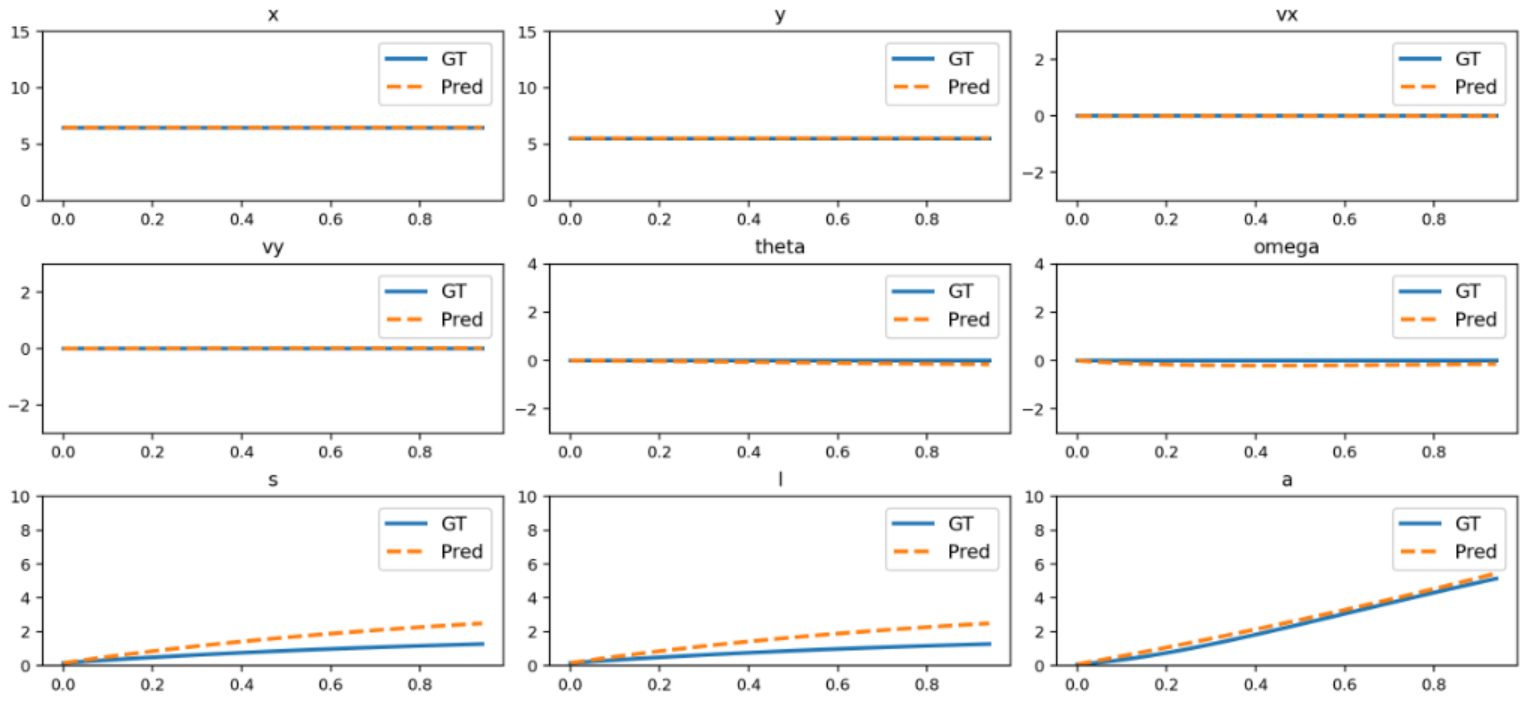



图 16：大小变化时 NND 预测与真实值的比较。

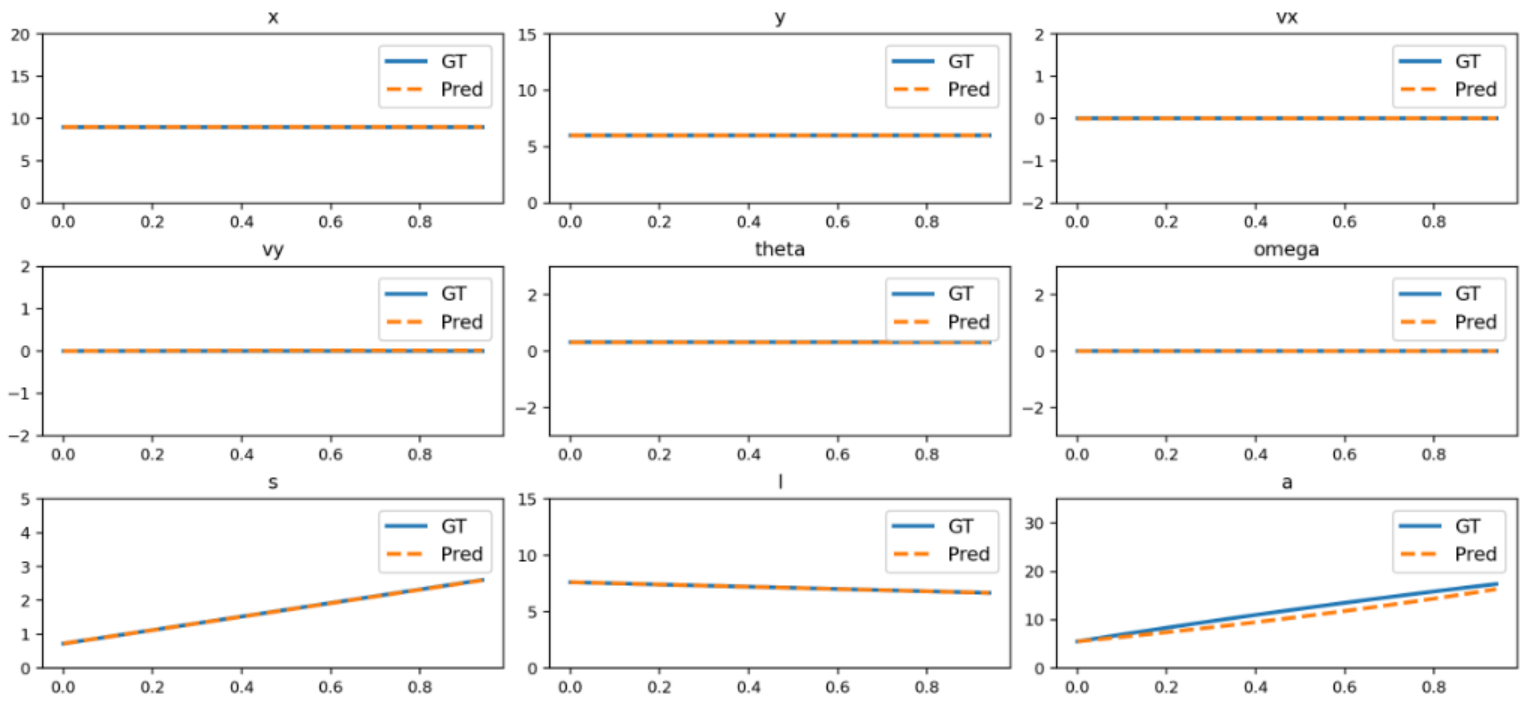



图 17：形变时 NND 预测与真实值的比较。

# D Evaluation Details

## D.1 Physical Invariance Score

The Physical Invariance Score (PIS) as described in equation 7 indicates whether a certain quantity $C$ remains invariant over time. If the laws of physics are replicated perfectly, $C$ remains constant, and $C_\sigma \to 0 \implies \text{PIS} \to 1$. For each type of motion, a suitable $C$ should be selected.

**Uniform Motion**: An object is prompted to travel horizontally in uniform velocity in each scene. Therefore, we select the horizontal velocity $v_x$ as the invariant feature.

**Uniform Acceleration and Deceleration**: Under these motions, we check if the object obeys the law of accelerating (or decelerating) at a constant rate. The guidance parameters and prompts specify horizontal motion. Therefore, we set $C = a_x$ for acceleration, and $C = -a_x$ for deceleration.

**Parabolic Motion**: Under this motion, there is no horizontal acceleration. Therefore $v_x$ is expected to be constant. Additionally, the vertical acceleration $a_y$ due to gravity should be constant.

**3D Motion**: The prompt guides an object to travel towards the observer, creating the effect of increasing object dimensions, while also having a 2D motion. We approximate this effect as having a constant vertical velocity $v_y$, and a constant increment rate in the long-axis of the object $\Delta l$.

**Slope Sliding**: An object is prompted to slide down a constant slope. Assuming negligible effects from friction, we can expect accelerations $a_x, a_y$ to be constant.

**Circular Motion**: Objects are guided to orbit in a circular path, we assume the angular velocity about the orbital center $\omega$ is constant.

**Rotation**: When objects are prompted to "spin" or "rotate about their axes", we assume that the brief duration of the video, that they rotate in a constant angular velocity $\omega$.

**Parabolic Motion with Rotation**: Videos under this category should describe a superposition of a projectile motion under gravity, and a rotation about the object's axis. Therefore the metrics used in these two motions $(v_x, a_y, \omega)$ are used for $C$.

**Damped Oscillation** is simulated through various instances of pendulums, hinged at the top. We assume small angles $(\theta)$ for the stride. This leads to the vertical force varying with $cos(\theta)$, and we assume it to be a constant. Thereby we use $C = a_y$.

**Size Changing**: We prompt videos where it's natural to increase an object's overall size, while maintaining it's aspect ratio.(e.g., an inflating balloon). Assuming a constant rate of inflation, we set $C = \Delta r$; the rate of increasing the radius of the object.

**Deformation**: Objects under this category should expand, stretch, or spread-out over time. The aspect ratios may change. (e.g., the spread of a thick viscous liquid). We assume that the object increases its dimensions at a constant rate, and track this rate along it's long axis $\Delta l$.

After selecting $C$ for a motion type, $C_\sigma, C_\mu$ is calculated for every video. This lends to a PIS score per video. The final score reported in table 1 show the median PIS score after generating 12 different videos for each motion. In our case, temporal derivatives are formed from successive frames with $\Delta t = 1/(\text{FPS value})$, then mapped to physical units using a constant of 0.00625 meters per pixel. Each video is preprocessed with a 5-frame moving-average filter to reduce noise in derivative estimates.

Some feature assumptions are idealizations that may not hold in the real world. Accordingly, we report a **Reference** PIS computed directly from the guidance mask used to drive the video generation. For example, in 3D motion, $\Delta l$ need not be constant, so expecting $C_\sigma = 0 \implies \text{PIS} = 1$ is unrealistic. The mask-based PIS instead, serves as a practical upperbound– the score that would be achieved if the generator perfectly followed the guidance mask (which itself has $C_\sigma \neq 0$).

## D.2 Testing Prompts

Samples of text prompts [1] used for evaluation are listed in the table 3.

[1] The reader may refer the supplementary materials for an exhaustive list.

## D 评估细节

### D.1 PIS

方程式 7 中描述的物理不变性分数（PIS）表示某个量 C 是否随时间保持不变。如果物理定律被完美复制，C 将保持恒定，C→0⇔PIS→1。对于每种类型的运动，都应选择一个合适的 C。

匀速运动：物体在每个场景中均被提示以匀速水平运动。
因此，我们选择水平速度 vas 不变特征。

匀加速和匀减速：在这些运动中，我们检查物体是否遵循以恒定速率加速（或减速）的规律。引导参数和提示指定水平运动。因此，我们设置 C = a 为加速，C = −a 为减速。

抛物线运动：在这种运动中，没有水平加速度。因此，预期 v 是恒定的。此外，由于重力引起的垂直加速度应该是恒定的。

三维运动：提示引导物体朝向观察者移动，产生物体尺寸增大的效果，同时也有二维运动。我们将这种效果近似为具有恒定垂直速度 v，以及物体长轴的恒定增量率 Δl。

斜坡滑行：一个物体被提示沿恒定斜坡滑下。假设摩擦力的影响可以忽略不计，我们可以预期加速度 a, at 将保持不变。

圆周运动：物体被引导在圆形路径上运行，我们假设围绕轨道中心的角速度 ω 是恒定的。

旋转：当物体被提示"旋转"或"绕轴旋转"时，我们假设在视频的短暂持续时间中，它们以恒定角速度 ω旋转。

抛物线运动与旋转：属于此类别的视频应描述在重力作用下抛体运动的叠加，以及绕物体轴的旋转。因此，在这两种运动中使用的指标（v, a, ω）被用于 C。

阻尼振荡通过多个顶部铰接的摆锤进行模拟。我们假设步幅的角度（θ）很小。这导致垂直力随 cos(θ)变化，我们假设它是一个常数。因此我们使用 C = a。

尺寸变化：我们提示视频中自然增加物体整体尺寸的场景，同时保持其长宽比（例如，一个正在充气的气球）。假设充气速率恒定，我们设 C = Δr；即物体半径增加的速率。

形变：这一类物体应该随着时间的推移而膨胀、拉伸或展开。长宽比可能会改变。（例如，粘稠液体的扩散）。我们假设物体以恒定的速率增加其尺寸，并沿着其长轴 Δl 跟踪这一速率。

在选择了运动类型 C 后，对每个视频计算 C 和 Cis，从而得到每个视频的 PIS 分数。表 1 中报告的最终分数是每项运动生成 12 个不同视频后的中位数 PIS 分数。在我们的案例中，时间导数由连续帧形成，其中Δt = 1/(FPS 值)，然后使用每像素 0.00625 米的常数映射到物理单位。每个视频都使用 5 帧移动平均滤波器进行预处理，以减少导数估计中的噪声。

一些特征假设是理想化的，在现实世界中可能并不成立。因此，我们直接使用用于驱动视频生成的引导掩码计算参考 PIS。例如，在 3D 运动中，Δl 不必保持恒定，因此期望 C=0⇒PIS=1 是不切实际的。基于掩码的 PIS 则作为一个实用的上限——如果生成器完美地遵循引导掩码（该掩码本身有 $C\!=\!0$）时将获得的分数。

### D.2 TP

用于评估的文本提示样本列在表 3 中。

[1]读者可参考补充材料获取详尽列表。

Table 3: Samples of Testing Prompts.

| **Motion** | **Testing Prompts** |
|---|---|
| Uniform Motion | A small metal cube sliding steadily along a smooth laboratory bench, reflections visible on the surface, scattered tools in the background, captured from a fixed side camera. |
| | A red rubber ball rolling at constant speed on a polished wooden floor, pulled by a thin string, with scattered papers and books in the background, observed from a fixed side camera. |
| Acceleration | A red sedan accelerating in a straight line on a clean highway, the road flat and clear, with only a pale sky and distant horizon in the background, captured from a fixed roadside camera. |
| | A black off-road SUV accelerating in a straight line on sandy terrain, with continuous sand dunes in the background, a few white clouds in the sky, sunlight slanting, kicking up fine sand particles, viewed from a stationary side-angle camera. |
| Deceleration | A yellow bus decelerates in a straight line in front of a traffic light on a city street, with pedestrians crossing nearby, and the wet road reflecting the sky, captured by a fixed side-view camera. |
| | A red coach brakes and decelerates in a straight line on a highway, with road signs and streetlights nearby and the city skyline visible in the distance, captured by a fixed side-view camera. |
| Parabolic Motion | A golf ball is hit at an angle with an initial speed. The camera captures its parabolic trajectory from the side. The scene takes place on a sunny golf course with manicured fairways, sand bunkers, and distant trees, adding depth and realism. |
| | A volleyball is served at an angle, captured from the side by a stationary camera. The scene is set on an outdoor beach volleyball court, with sand texture, net, and distant palm trees in view. |
| 3D Motion | A fighter jet accelerates slowly from the distance along the runway towards the camera, hangars and runway lights visible in the background, captured from a fixed oblique side camera. |
| | A cardboard box slides from the distance along a warehouse floor towards the camera, shelves and crates visible in the background, captured from a fixed oblique side camera. |
| Slope Sliding | A hardcover book accelerating down a carpeted inclined board in a classroom, chalkboard and desks in the background, captured from a fixed side camera parallel to the ramp. |
| | A small metal cube sliding down a laboratory ramp, shiny reflections on its surface, scattered tools and wires in the background, captured from a fixed side camera parallel to the ramp. |
| Circular Motion | A tiny moonlet orbits a gas giant along a smooth, circular path. The top-down view shows the consistent motion without motion trails.. |
| | A comet with a glowing tail orbits a distant star along a stable circular path. A top-down perspective emphasizes the symmetrical orbit and the stationary central star. |
| Rotation | A metal rod spinning on a concrete floor, faint scratches and dust visible, captured from a fixed top-down camera. |
| | A wooden dowel rotating gently on a tiled kitchen floor, soft shadows from ceiling lights, viewed from a stationary overhead camera. |
| Parabola +Rotation | A pen is thrown at an angle, rotating as it falls. Captured from a side camera, the notebook and desk provide background details and depth. |
| | A thin cylindrical rod gently tossed, rotating along its long axis, fixed side camera, realistic reflections, ground shadows visible, subtle motion blur. |
| Damped Oscillation | A small decorative bell hanging from a fine chain. The fixed camera captures realistic material and shadows. |
| | A realistic pendulum with a spherical bob swinging from a fixed pivot. The fixed camera captures the entire motion. |
| Size Changing | A red helium balloon gradually inflating in a sunny park, children playing in the background, trees casting soft shadows, captured from a stationary side camera. |
| | A transparent water balloon expanding in a laboratory, scientific instruments and glassware around, bright fluorescent lights overhead, captured from a fixed top-down camera. |
| Deformation | A long strip of yogurt slowly spreads into a smooth layer, captured by a fixed overhead camera. |
| | A long strip of jelly gradually deforms and flattens on a plate, captured by a fixed overhead camera. |

表 3：测试提示样本。

| 运动测试提示 | |
|---|---|
| 均匀运动 | 一个小的金属立方体在光滑的实验室台面上稳定地滑动，表面可见反射，背景中散落着工具，由一个固定的侧面相机拍摄。一个红色的橡胶球在抛光的木地板上以恒定速度滚动，被一根细绳拉动，背景中散落着纸张和书籍，由一个固定的侧面相机观察。 |
| | 一辆红色轿车在干净的高速公路上直线加速，路面平坦清晰，背景只有淡蓝色的天空和遥远的地平线，由固定路边的摄像机拍摄。 |
| | 一辆黑色越野 SUV 在沙地地形上直线加速，背景是连绵的沙丘，天空中有几朵白云，阳光斜照，扬起细小的沙粒，由固定侧角度摄像机拍摄。 |
| | 减速一辆黄色公交车在城市街道上直线减速，前方是交通信号灯，附近有行人横穿，湿漉漉的路面反射着天空，由固定侧角度摄像机拍摄。 |
| | 一辆红色教练车在高速公路上直线刹车减速，附近有路标和路灯，远处可见城市天际线，由固定侧角度摄像机拍摄。 |
| 抛物线运动 | 一个高尔夫球以一定角度被击出，初始速度为 v。摄像机从侧面捕捉其抛物线轨迹。场景发生在阳光明媚的高尔夫球场，有修剪整齐的球道、沙坑和远处的树木，增添了深度和真实性。一个排球以一定角度被发球，由一个固定摄像机的侧面捕捉。场景设置在户外沙滩排球场上，可以看到沙滩纹理、球网和远处的棕榈树。 |
| 3D 运动 | 一架战斗机从远处沿着跑道缓慢加速朝向摄像机，背景可见机库和跑道灯光，由一个固定的斜向侧面摄像机捕捉。 |
| | 一个纸箱从远处沿着仓库地板滑向摄像机，背景中可见货架和板条箱，由一个固定的斜向侧面摄像机拍摄。 |
| 斜坡滑动 | 一本硬壳书在教室的地毯倾斜板上加速下滑，背景可见黑板和课桌，由一个与斜坡平行的固定侧面摄像机拍摄。一个小金属方块在实验室斜坡上滑下，表面有闪亮反光，背景散落着工具和电线，由一个与斜坡平行的固定侧面摄像机拍摄。 |
| 圆形运动 | 一个微小的卫星沿着平滑的圆形轨道环绕一颗气态巨行星。俯视图显示了持续的运动，没有运动轨迹。一颗带有发光尾巴的彗星沿着稳定的圆形轨道环绕一颗遥远的恒星。俯视视角强调了对称的轨道和静止的中心恒星。 |
| 旋转 | 一根金属杆在混凝土地板上旋转，可以看到微弱的划痕和灰尘，由一个固定的俯视摄像头捕捉。 |
| | 一根木制圆棒在瓷砖厨房地板上轻轻旋转，天花板灯光投下柔和的阴影，由一个固定的俯视摄像头拍摄。 |
| 抛物线+旋转 | 一支笔以角度抛出，在坠落时旋转。由侧面摄像头拍摄，笔记本和桌子提供背景细节和深度。一根细长的圆柱形棒轻轻抛出，沿其长轴旋转，固定侧面摄像头，逼真的反射，地面阴影可见，轻微的运动模糊。 |
| 阻尼振荡 | 一个由细链条悬挂的小装饰铃铛。固定相机捕捉逼真的材质和阴影。一个带有球形摆锤的摆，从固定支点摆动。固定相机捕捉整个运动。 |
| 尺寸变化 | 一个红色的氦气球在阳光明媚的公园里逐渐充气，背景中有孩子们在玩耍，树木投下柔和的阴影，由一个固定的侧面相机拍摄。一个透明的气球在实验室中膨胀，周围是科学仪器和玻璃器皿，头顶是明亮的荧光灯，由一个固定的俯视相机拍摄。 |
| 变形 | 一长条酸奶慢慢散开成一层平滑的薄膜，由一个固定的俯视相机拍摄。 |
| | 一长条果冻逐渐变形并在盘子上变平，由一个固定的俯视相机拍摄。 |

# E More Visual Results

## E.1 More General Comparison Results

From Figure. 18 to Figure. 29, we provide additional visual results and comparisons with other methods.

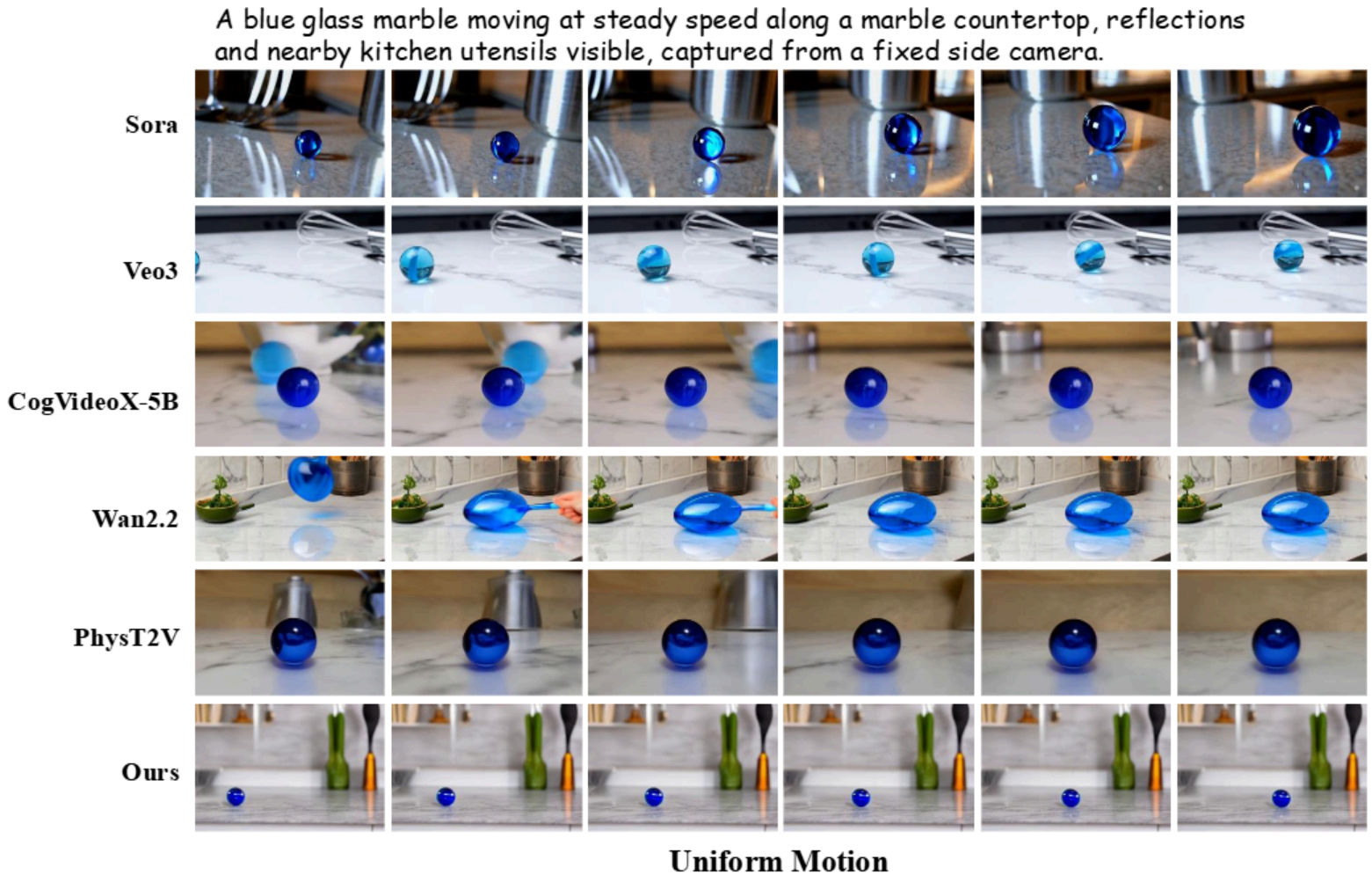



Figure 18: Visual comparisons on uniform motion.


Figure 19: Visual comparisons on acceleration.

## E 更多视觉结果

### E.1 MGCR

从图 18 到图 29，我们提供了其他方法的补充视觉结果和比较。

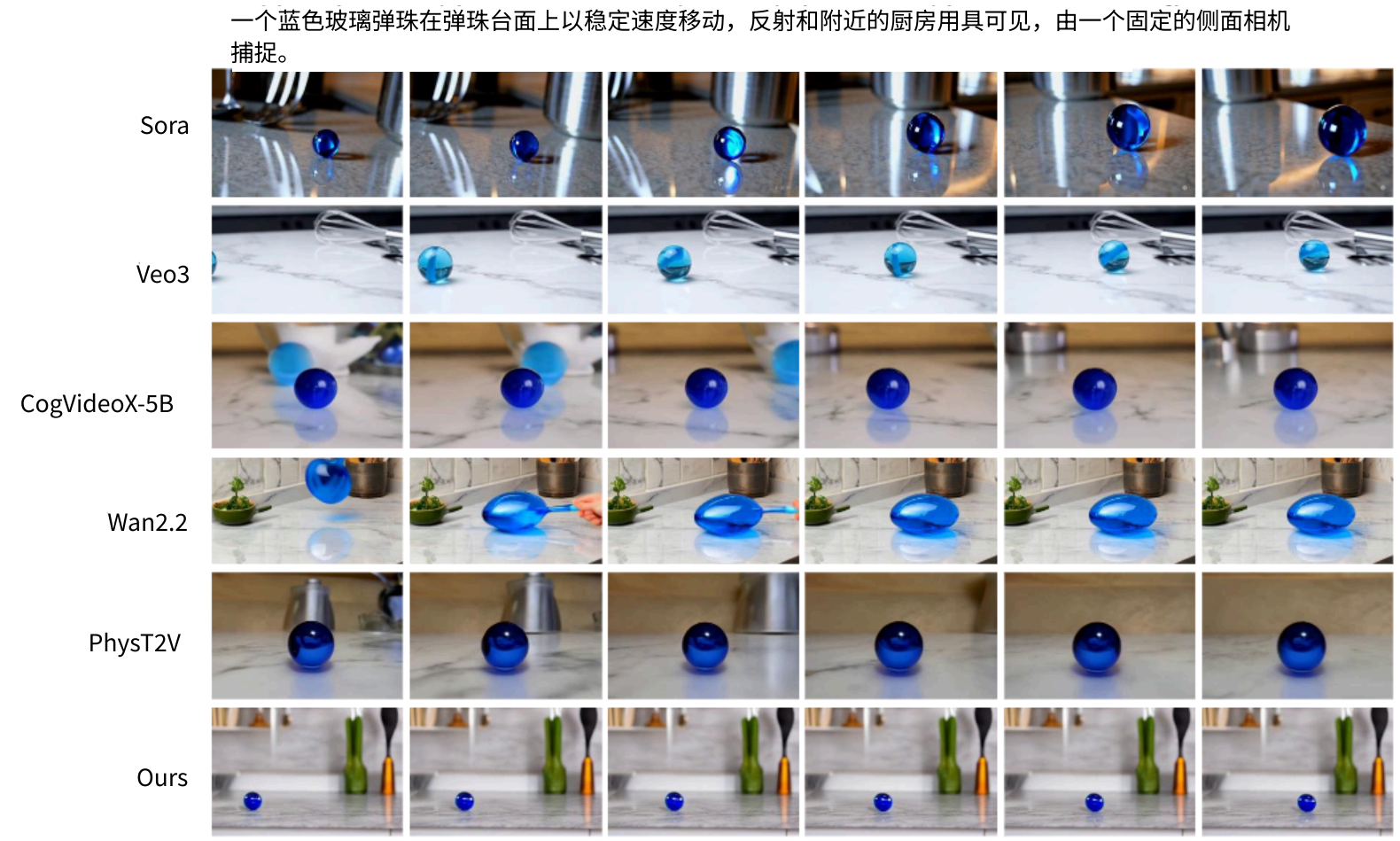



图 18：匀速运动的视觉比较。


图 19：加速时的视觉比较。

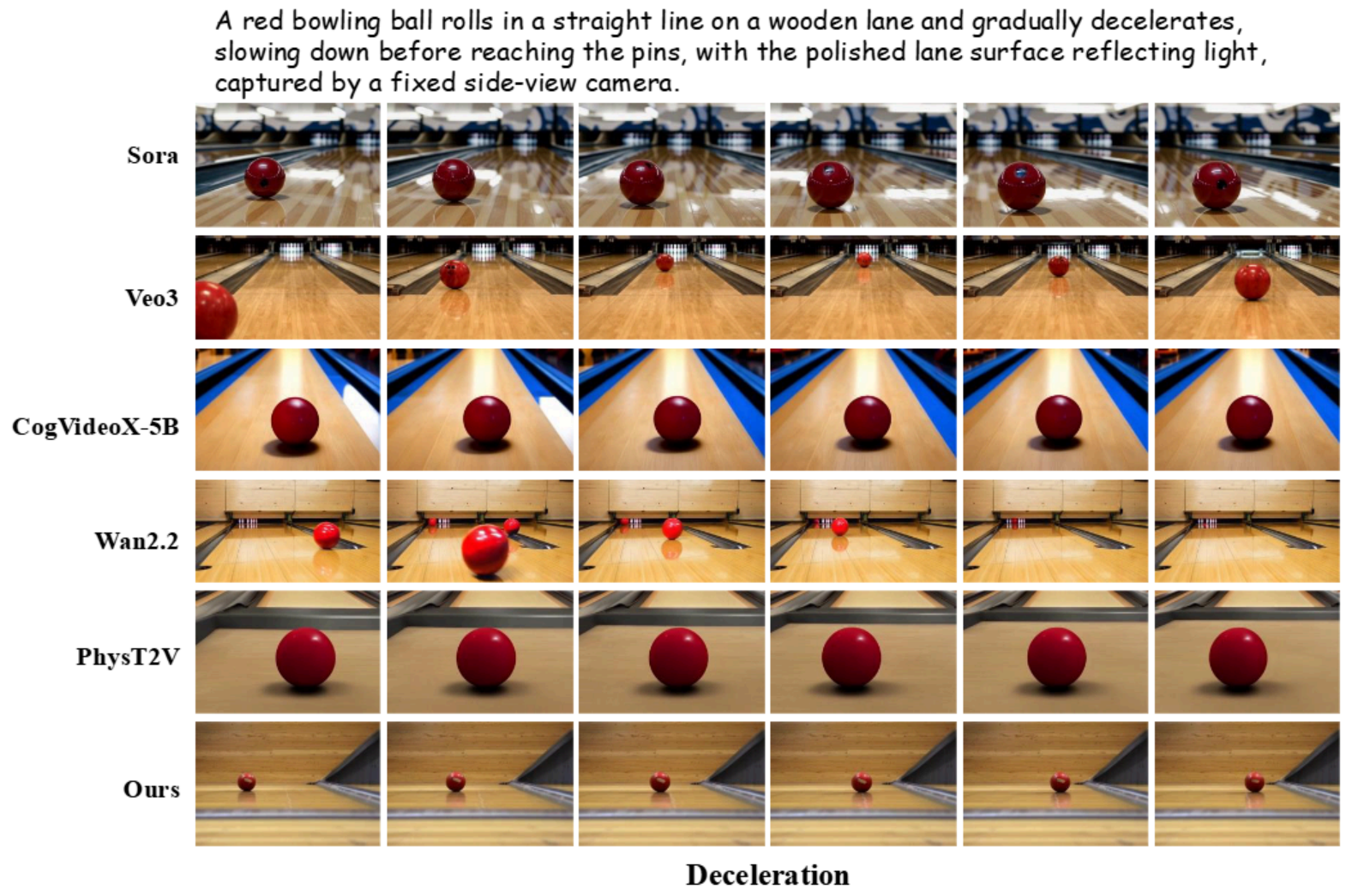



Figure 20: Visual comparisons on deceleration.

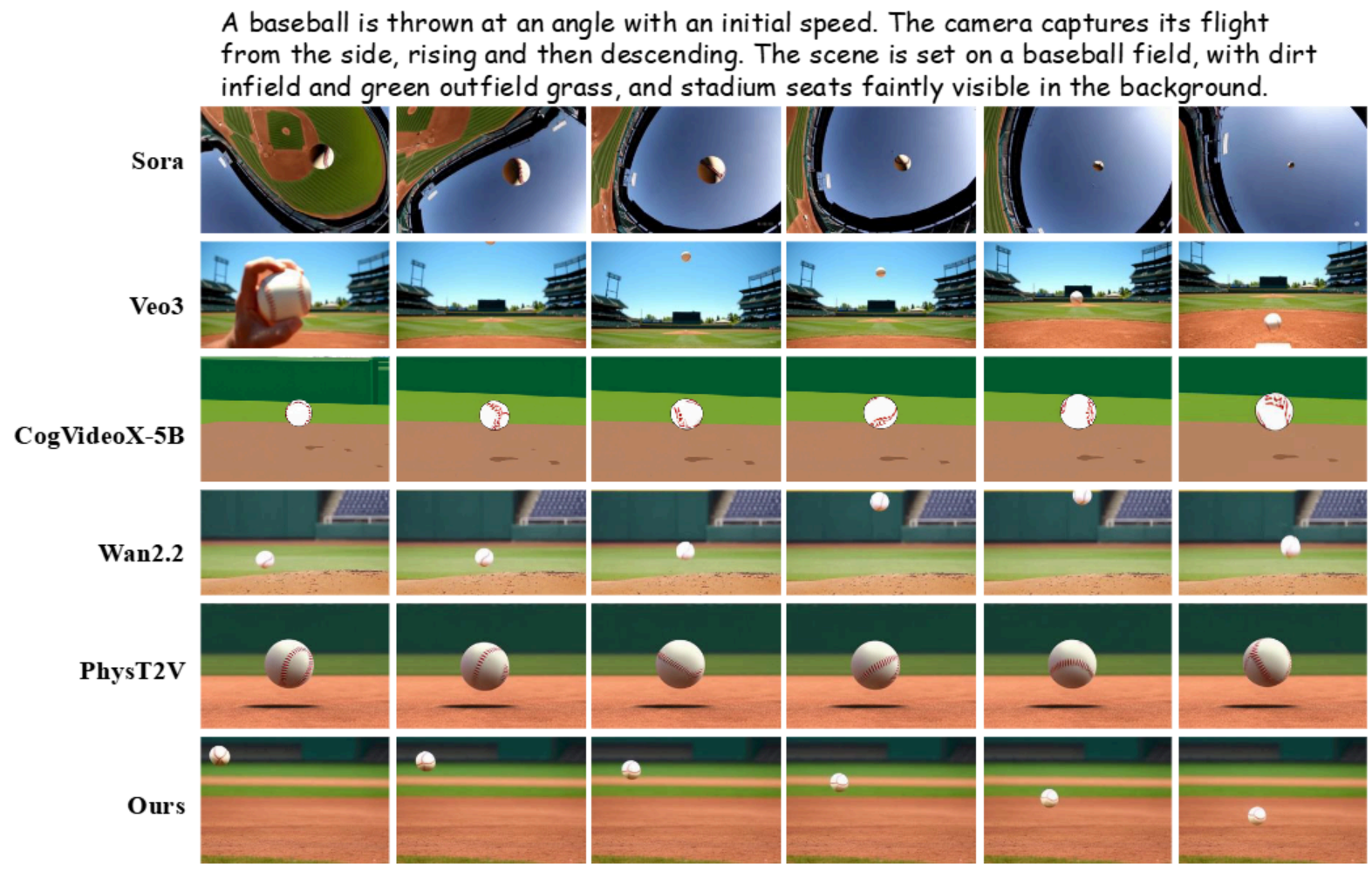



Figure 21: Visual comparisons on parabolic motion.

一个红色的保龄球沿着木制球道直线滚动，逐渐减速，在到达瓶子的前方时速度变慢，抛光的球道表面反射着光线，由一个固定的侧视摄像机捕捉。

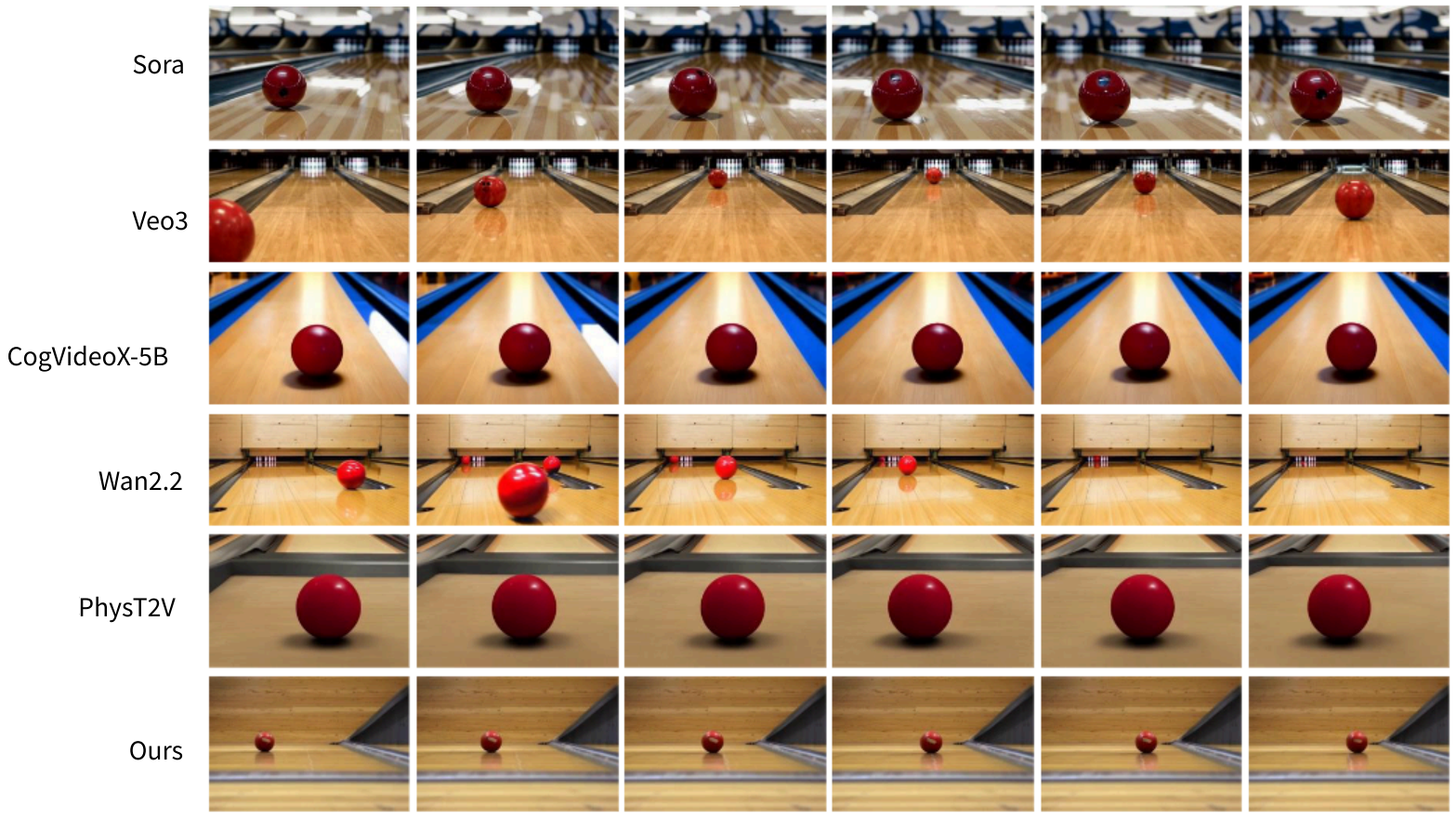



图 20：减速度的视觉比较。

一个棒球以一定角度和初速度被抛出。摄像机从侧面捕捉它的飞行轨迹，先上升后下降。场景设置在棒球场上，内场是泥土，外场是绿色草地，背景中隐约可见体育场座位。

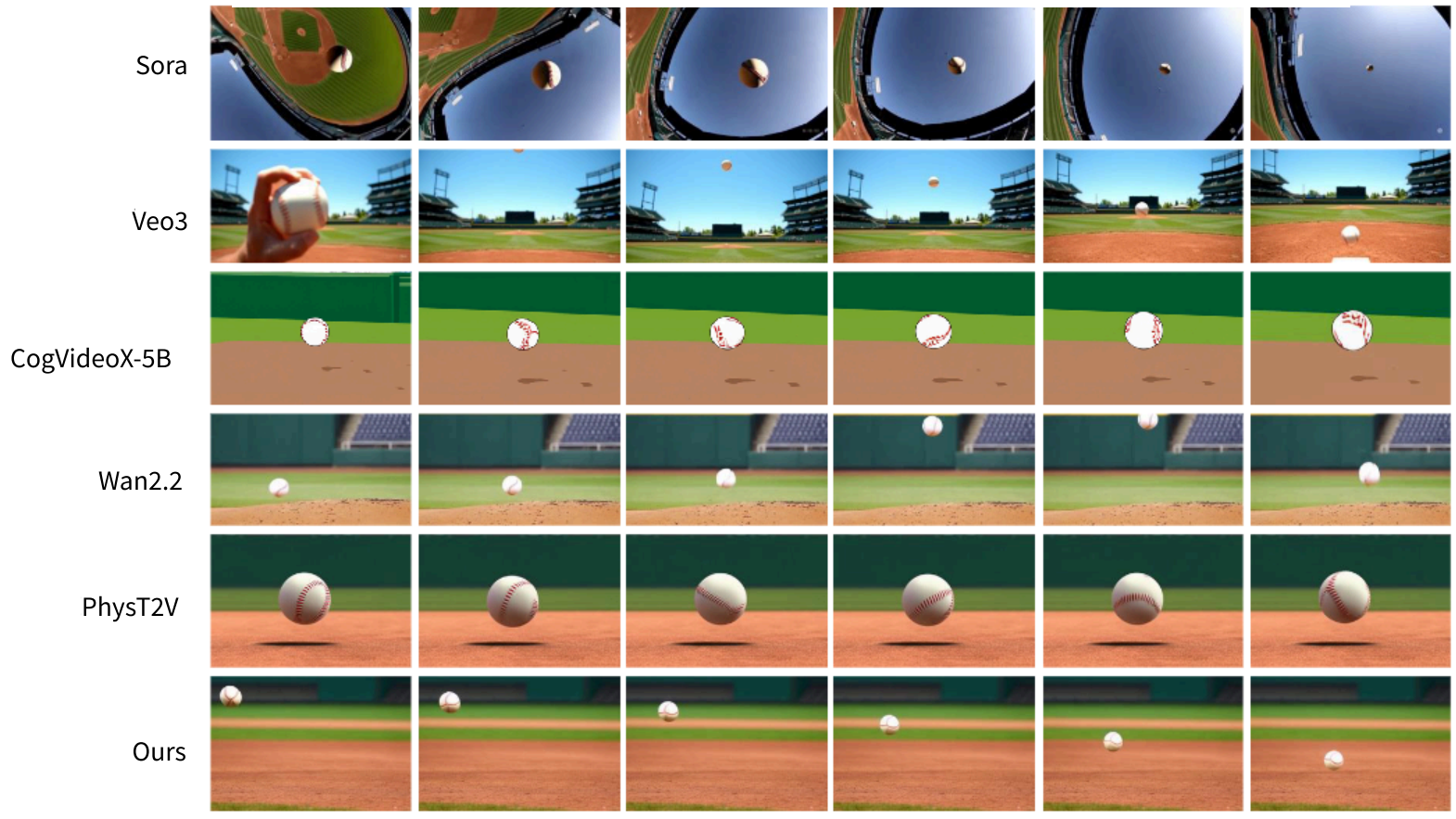



图 21：抛物线运动的视觉比较。

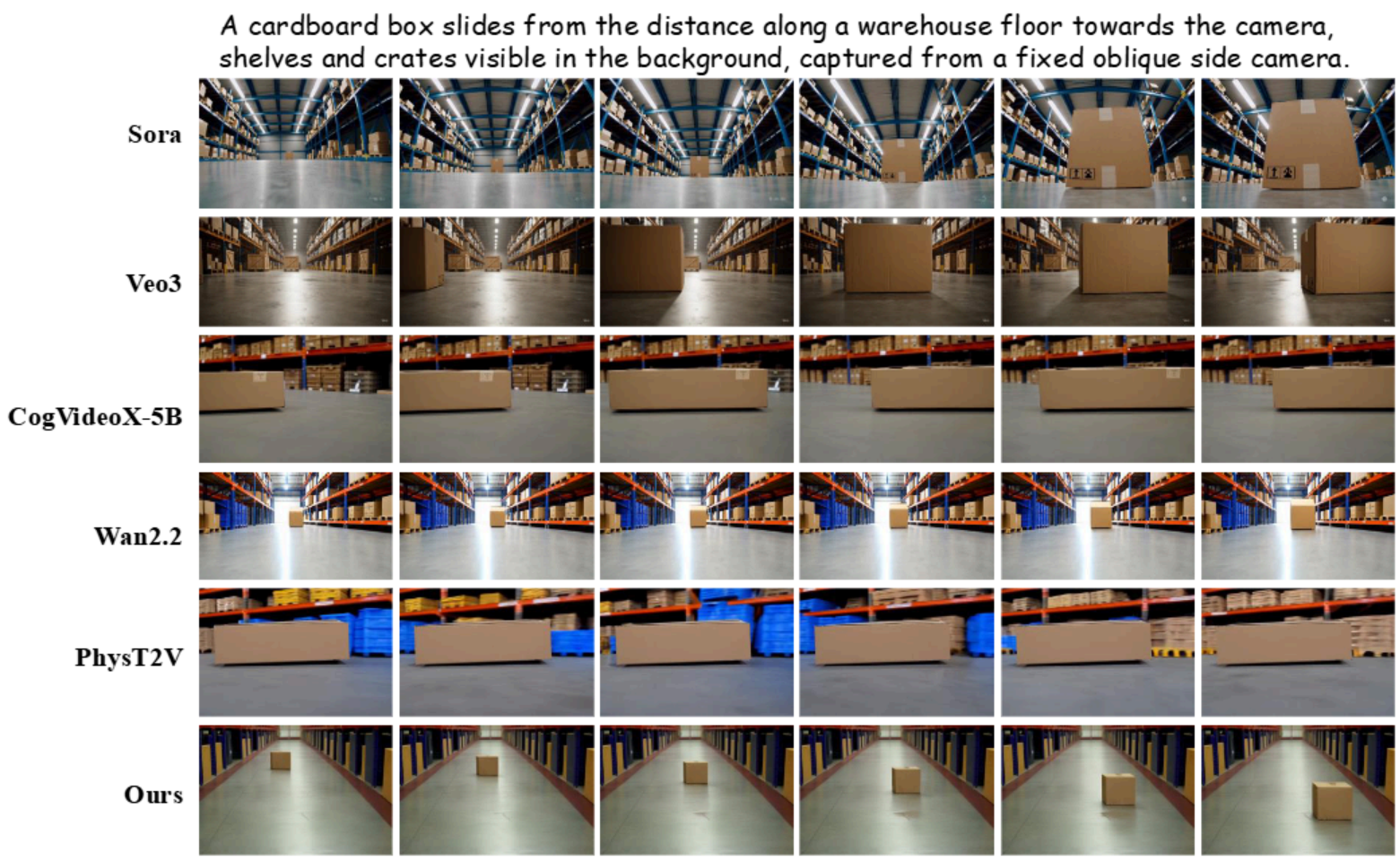



Figure 22: Visual comparisons on 3D motion.

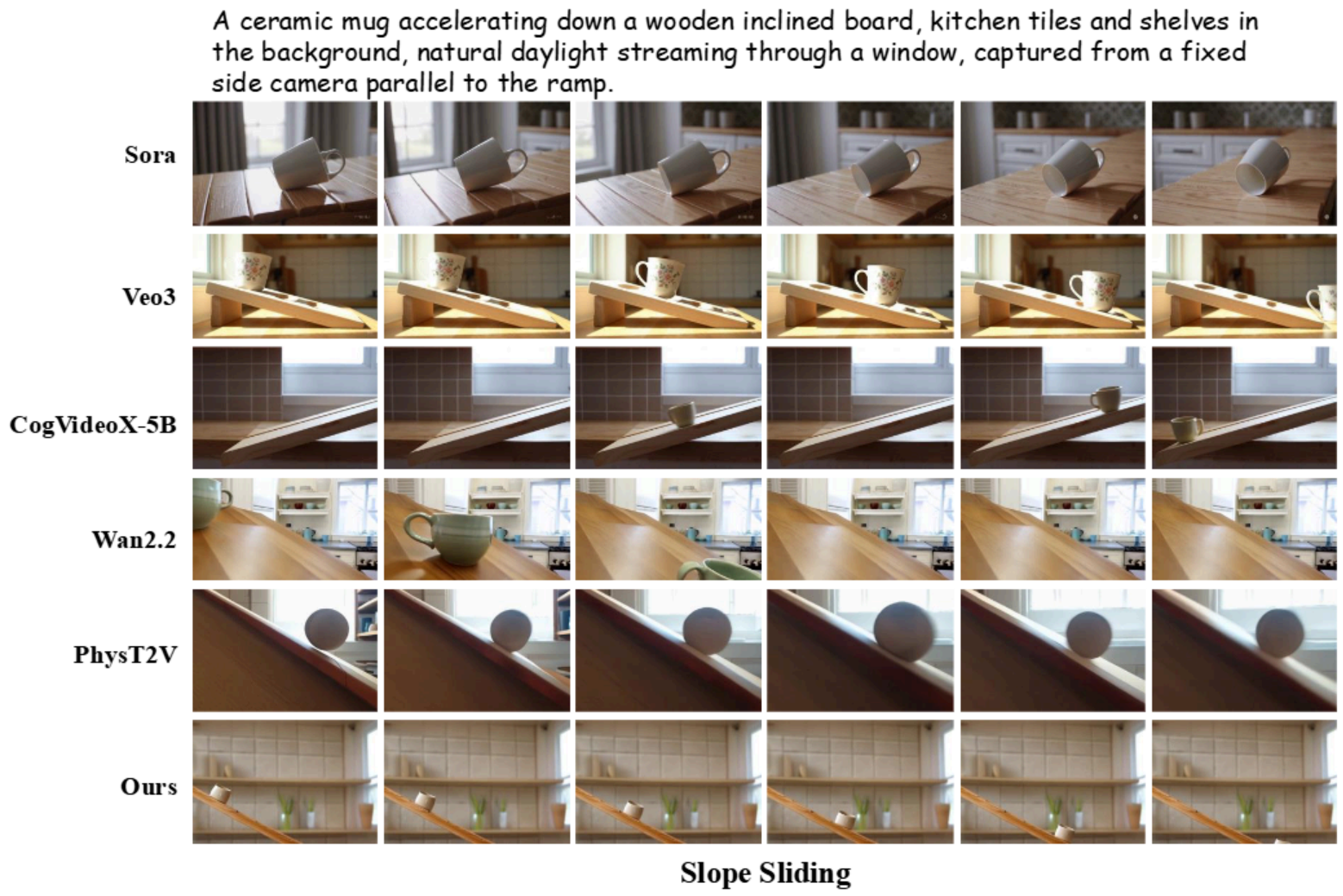
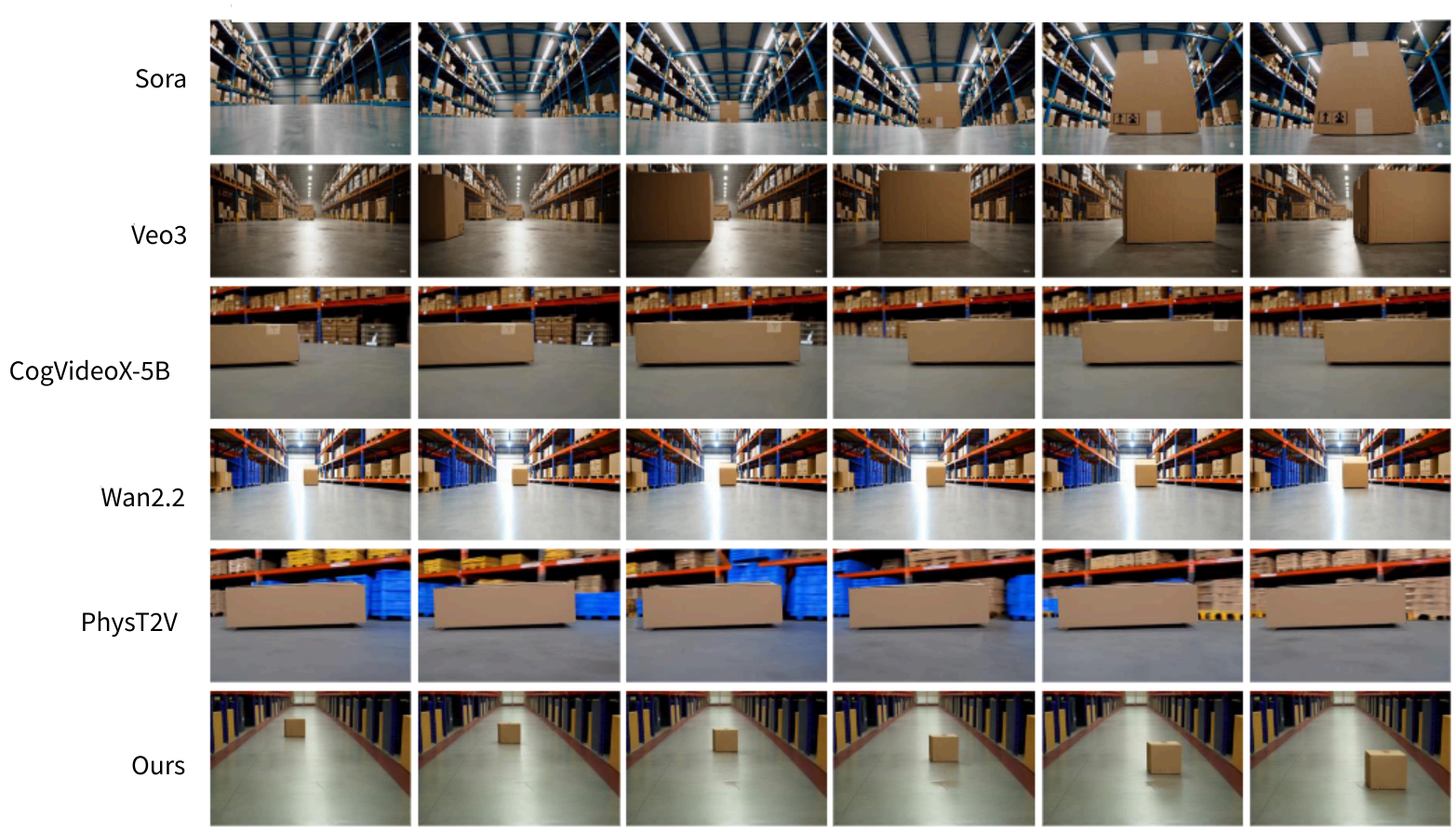



Figure 23: Visual comparisons on slope sliding.

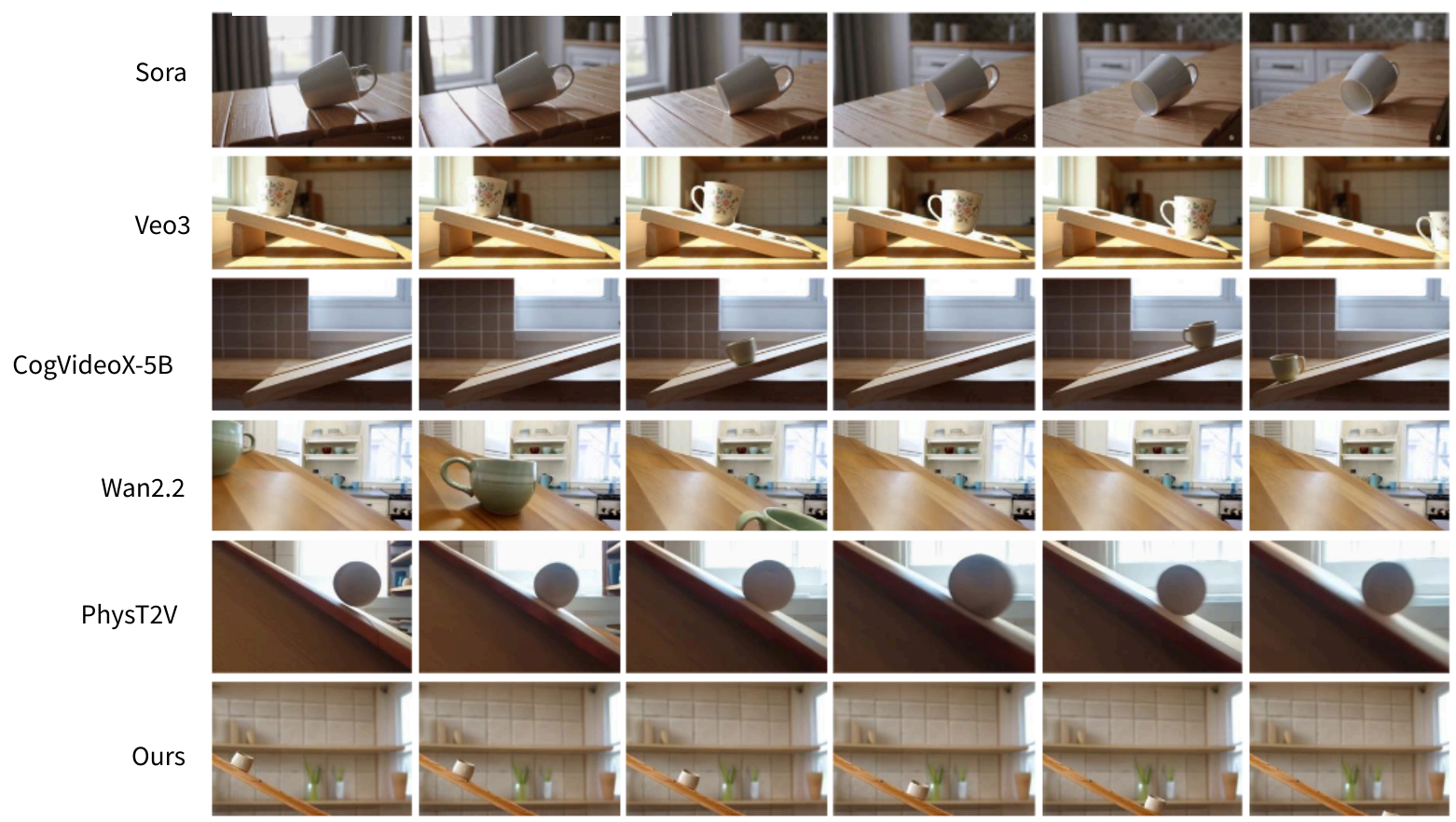
一个纸箱从远处沿着仓库地板滑向摄像机，背景中可见货架和板条箱，由一个固定的斜向侧面摄像机拍摄。




3D 运动

图 22：3D 运动的视觉比较。

一个陶瓷杯在木质斜板上加速下滑，背景是厨房瓷砖和架子，自然日光透过窗户照射进来，由一个与斜坡平行的固定侧相机拍摄。




斜坡滑动

图 23：斜坡滑动的视觉比较。


Figure 24: Visual comparisons on circular motion.

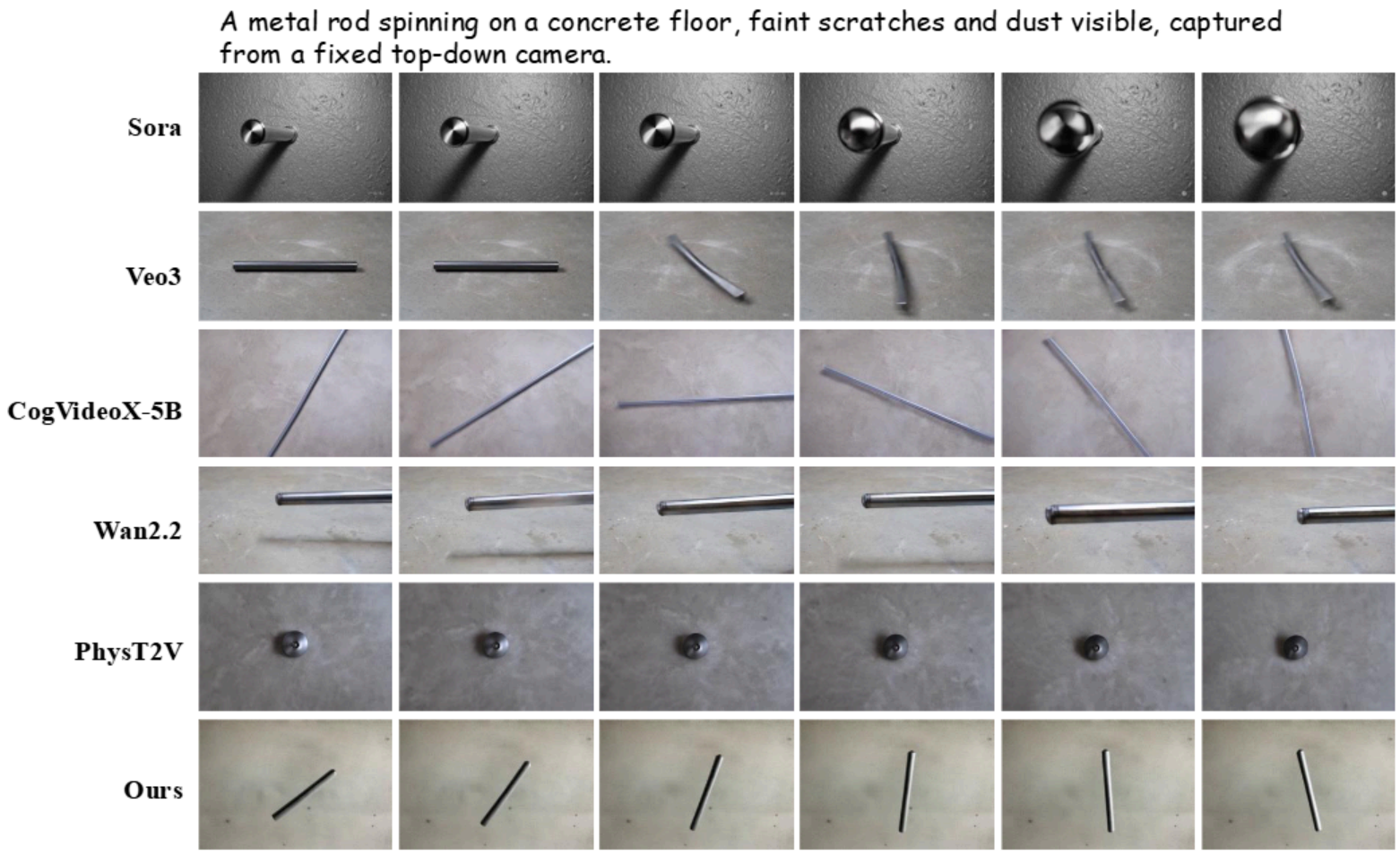


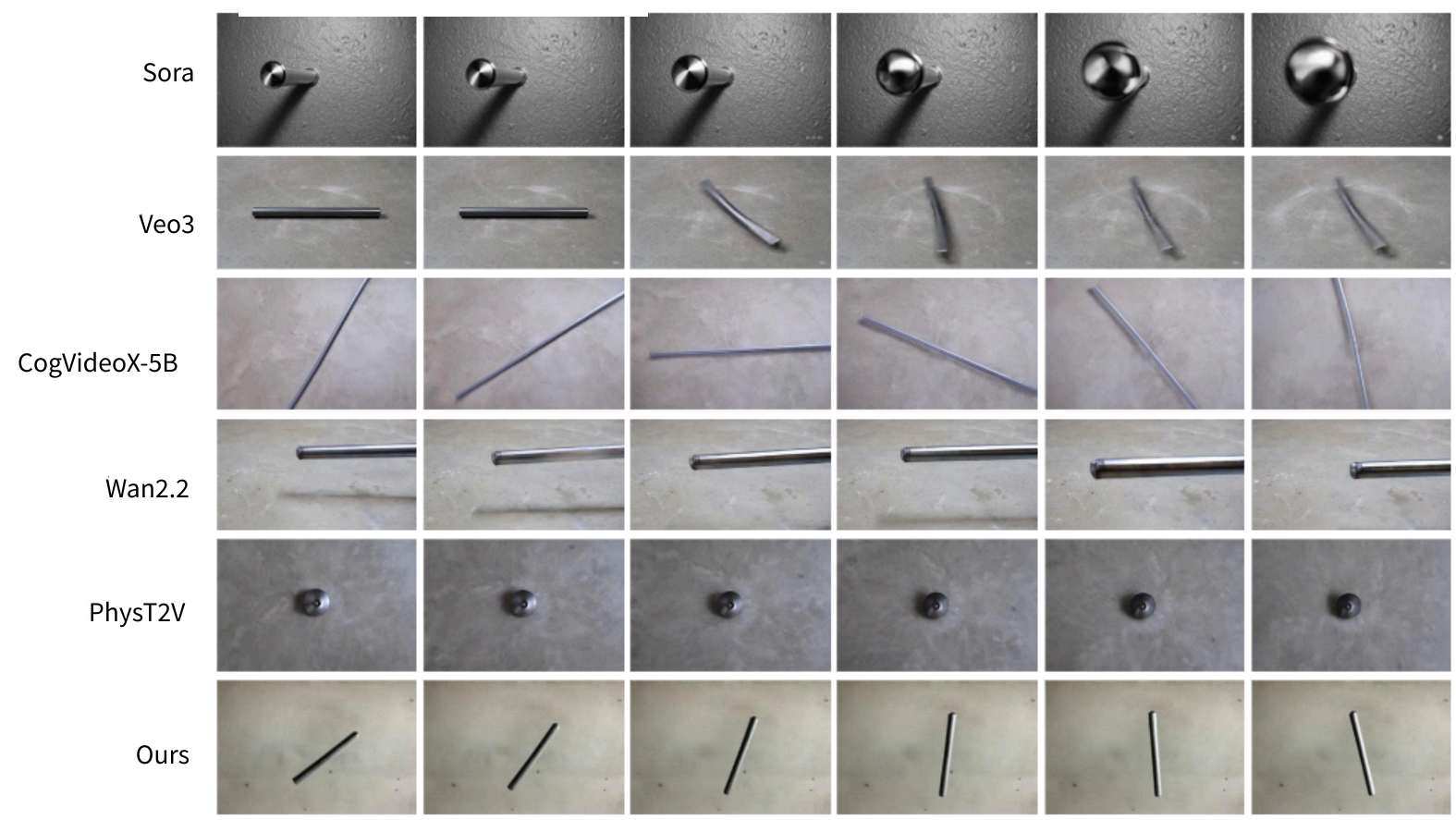

Figure 25: Visual comparisons on rotation.

一个带有发光尾巴的彗星沿着稳定的圆形轨道环绕着一颗遥远的恒星。俯视视角突出了对称的轨道和静止的中央恒星。


圆周运动

图 24：圆周运动的视觉比较

一根金属棒在混凝土地板上旋转，可见微弱的划痕和灰尘，由固定式俯视摄像头拍摄。




旋转

图 25：旋转的视觉比较。

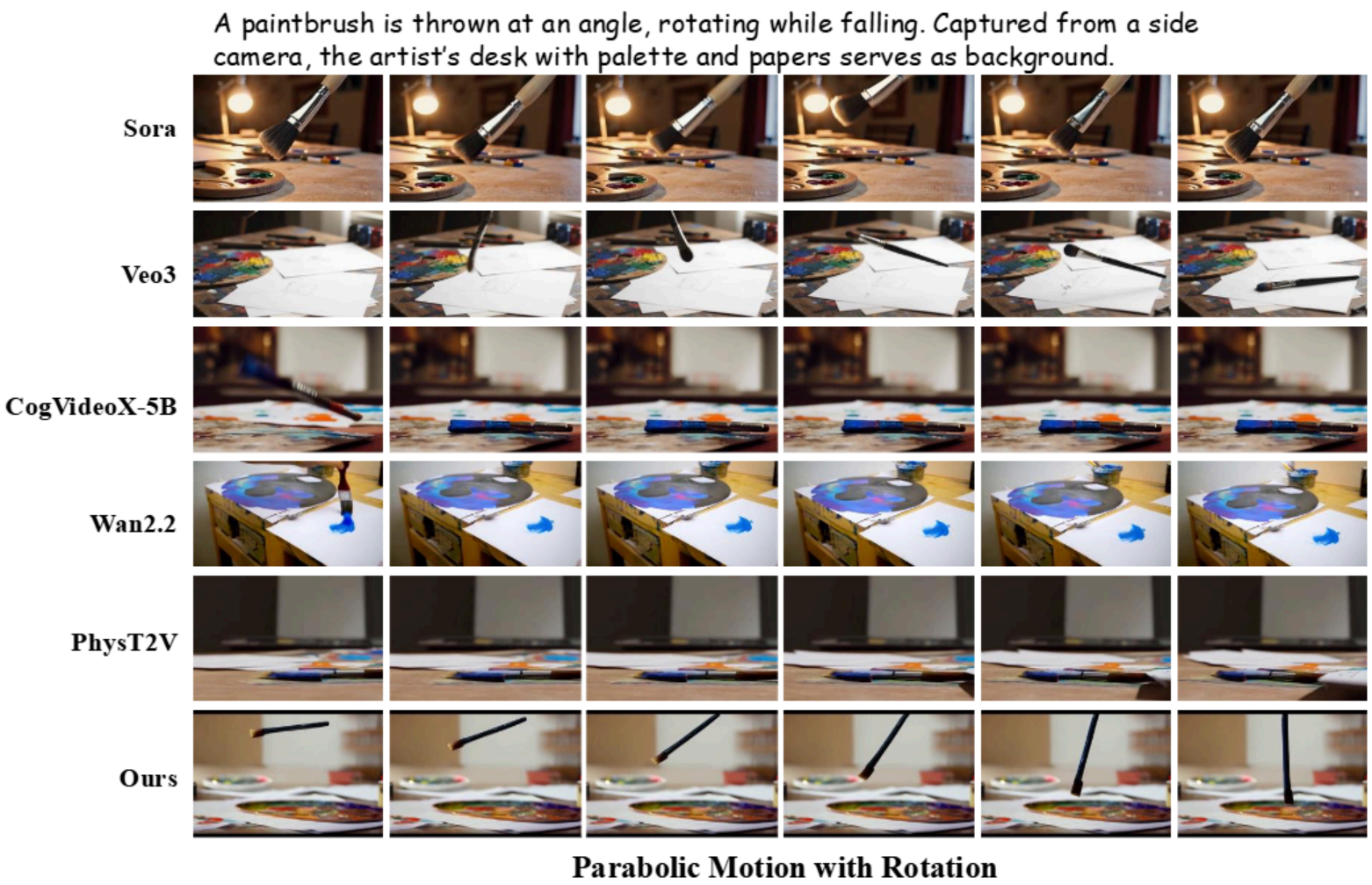



Figure 26: Visual comparisons on parabolic motion with rotation.

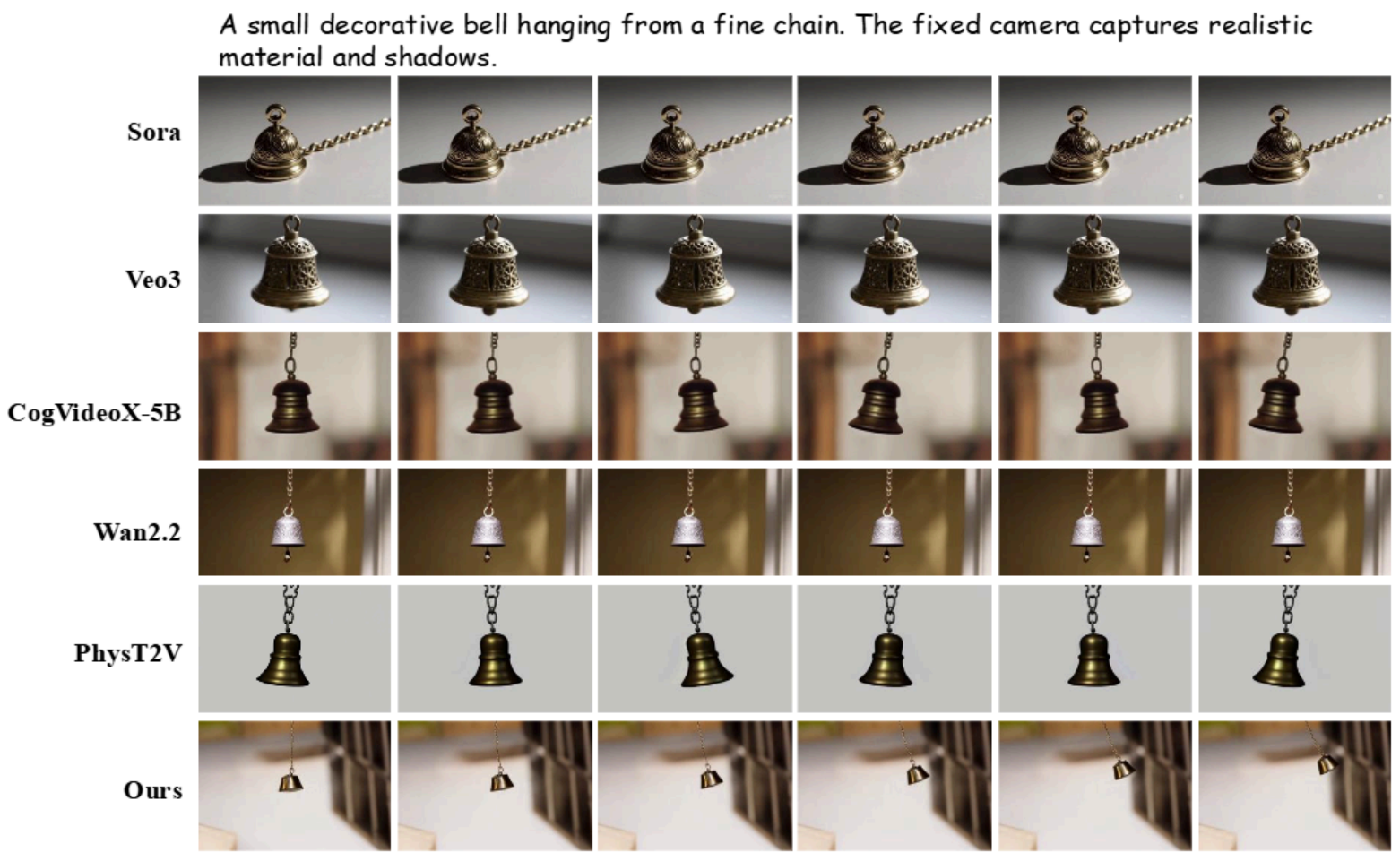



Figure 27: Visual comparisons on damped oscillation.

一把画笔以角度抛出，在坠落时旋转。从侧面摄像机拍摄，艺术家的工作台与调色板和纸张作为背景。

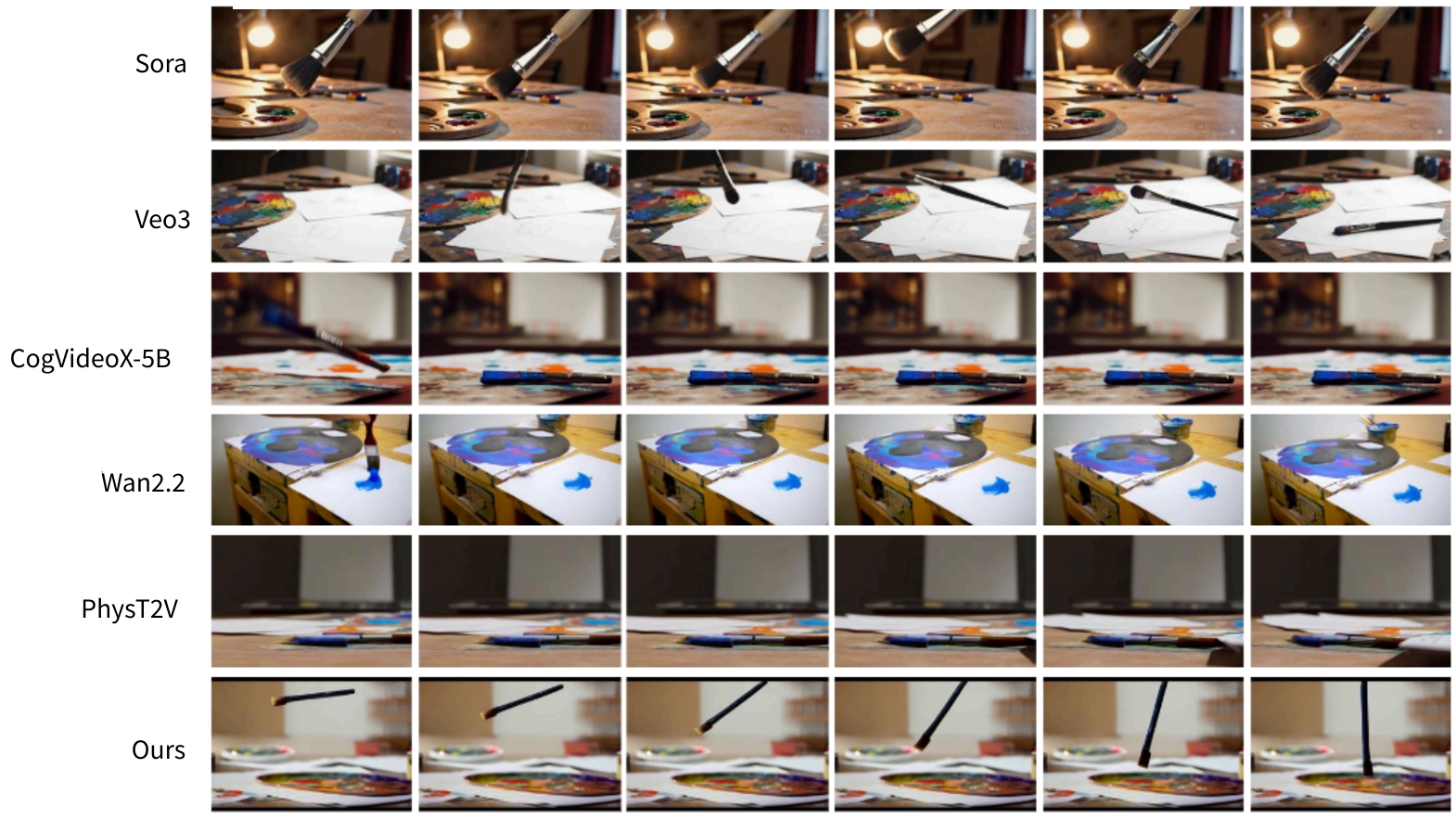



图 26：抛物线运动与旋转的视觉比较。

一个由细链条悬挂的小装饰铃。固定摄像头捕捉到逼真的材质和阴影。

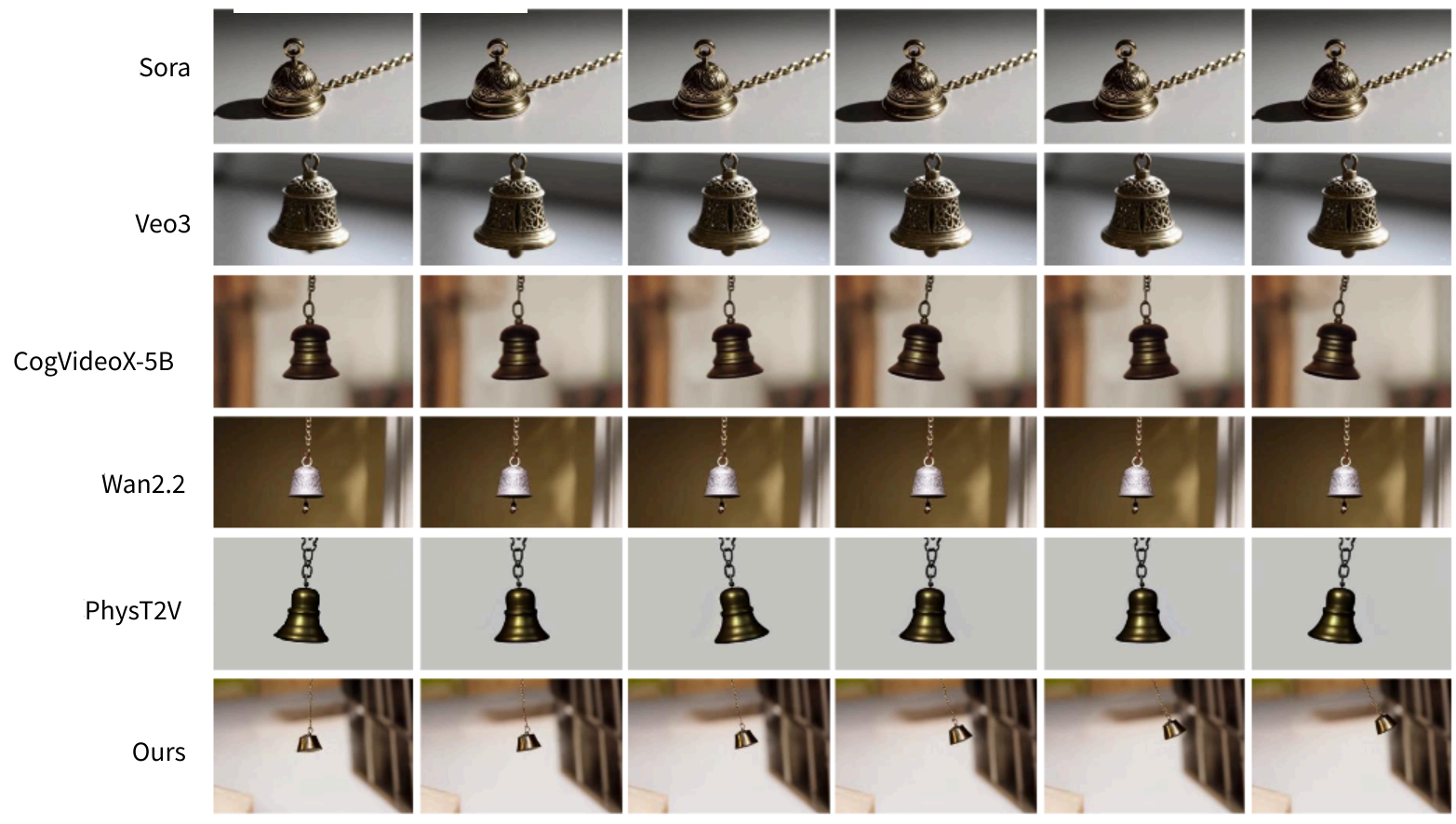



图 27：阻尼振动的视觉比较

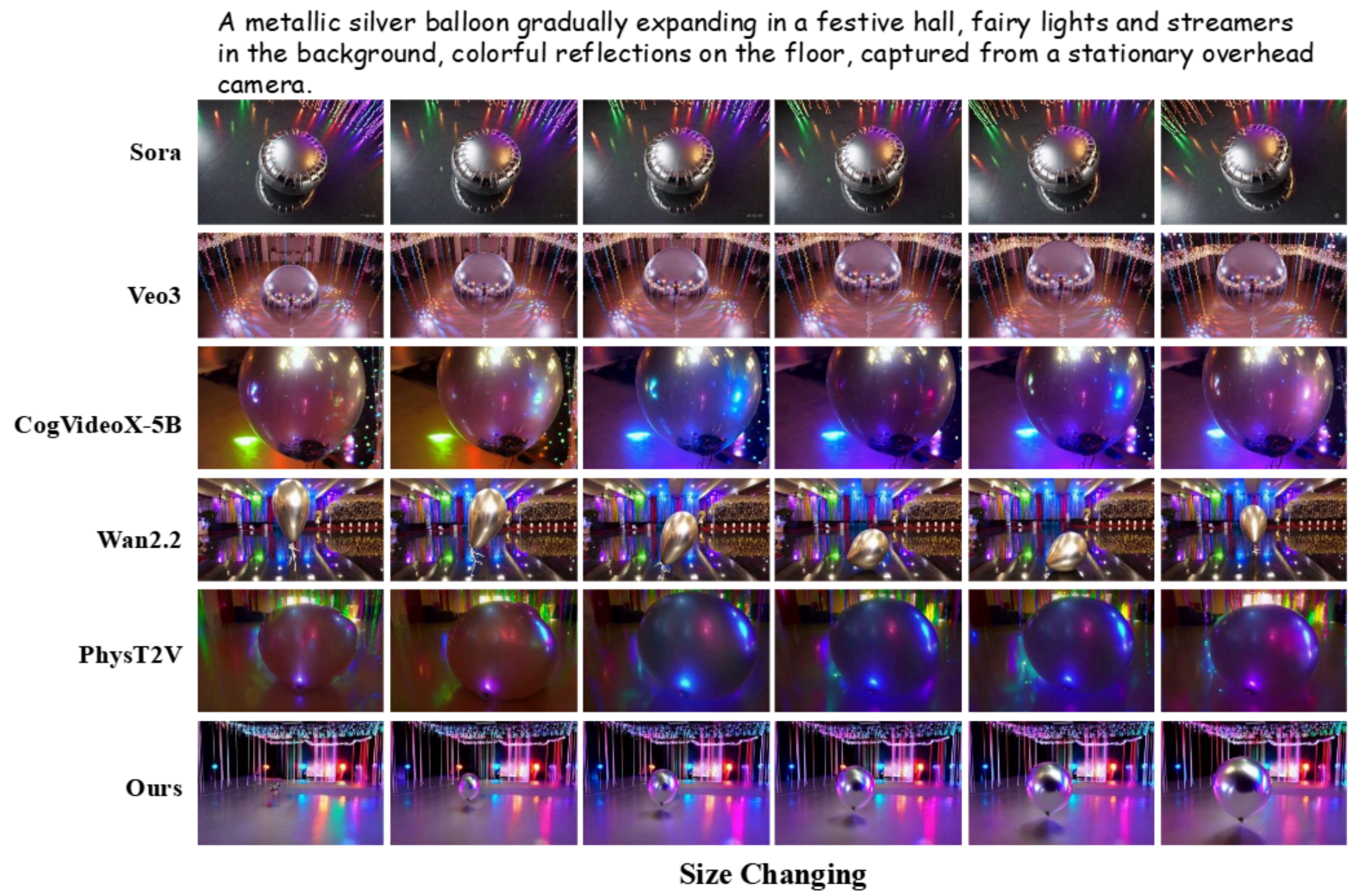



Figure 28: Visual comparisons on size changing.


Figure 29: Visual comparisons on deformation.

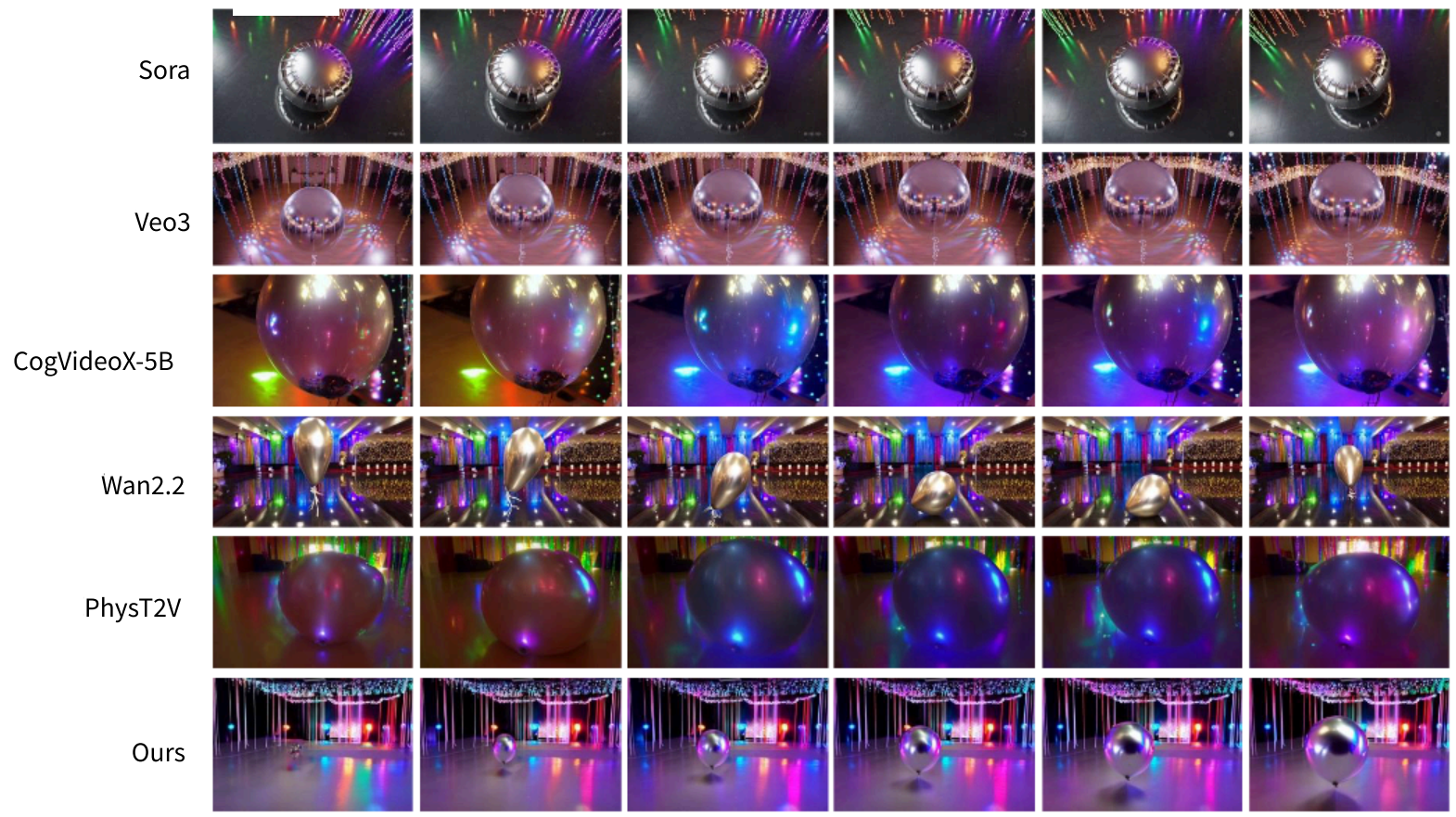



图 28：大小变化的视觉比较。


图 29：变形的视觉比较。

## E.2 More Parameter Controllability Comparison Results

Figure. 30 and Figure. 31 illustrate the physical parameter control capability of NewtonGen.

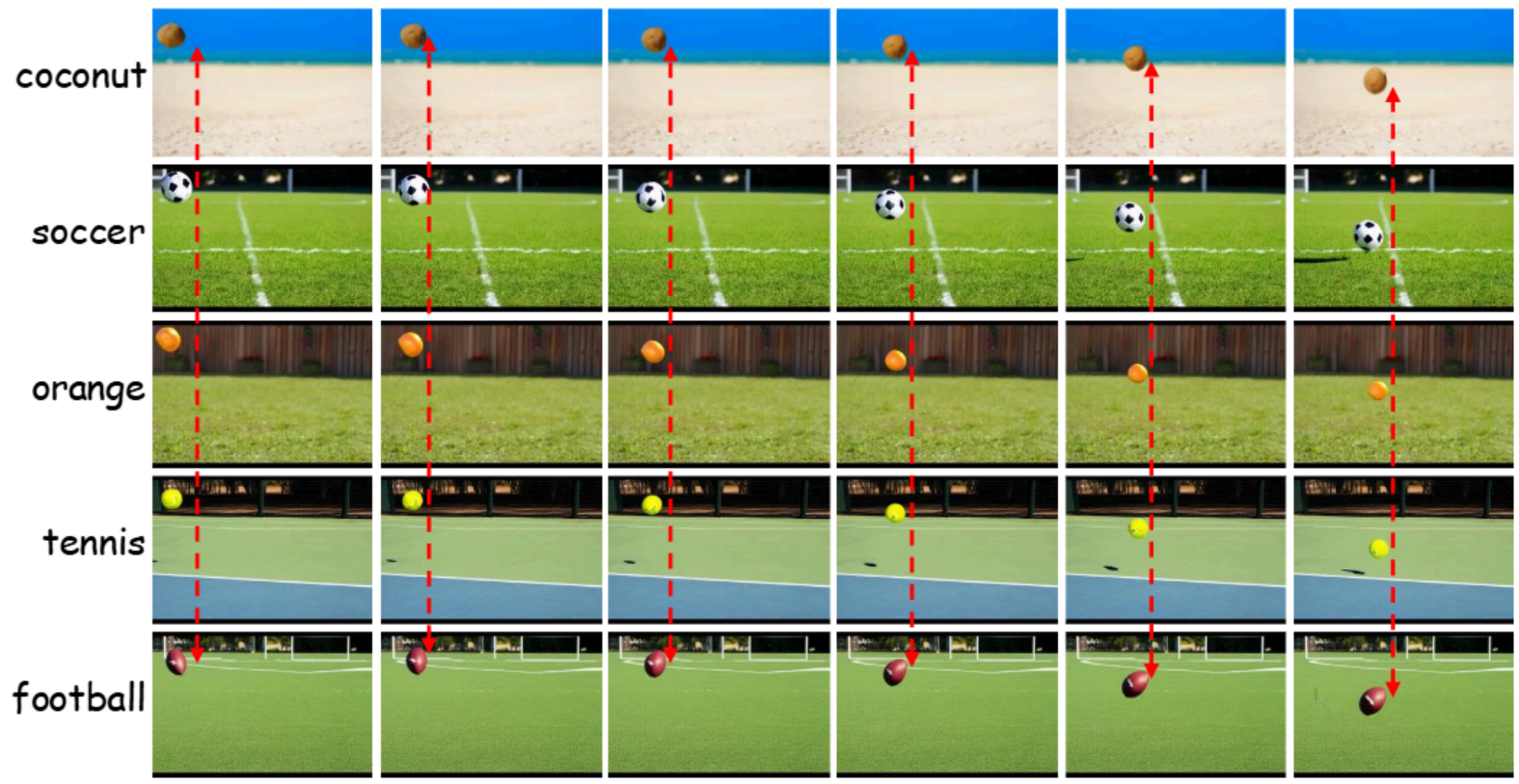



Figure 30: Given the same initial physical states but different scene descriptions, NewtonGen can generate diverse scenes with consistent motion.

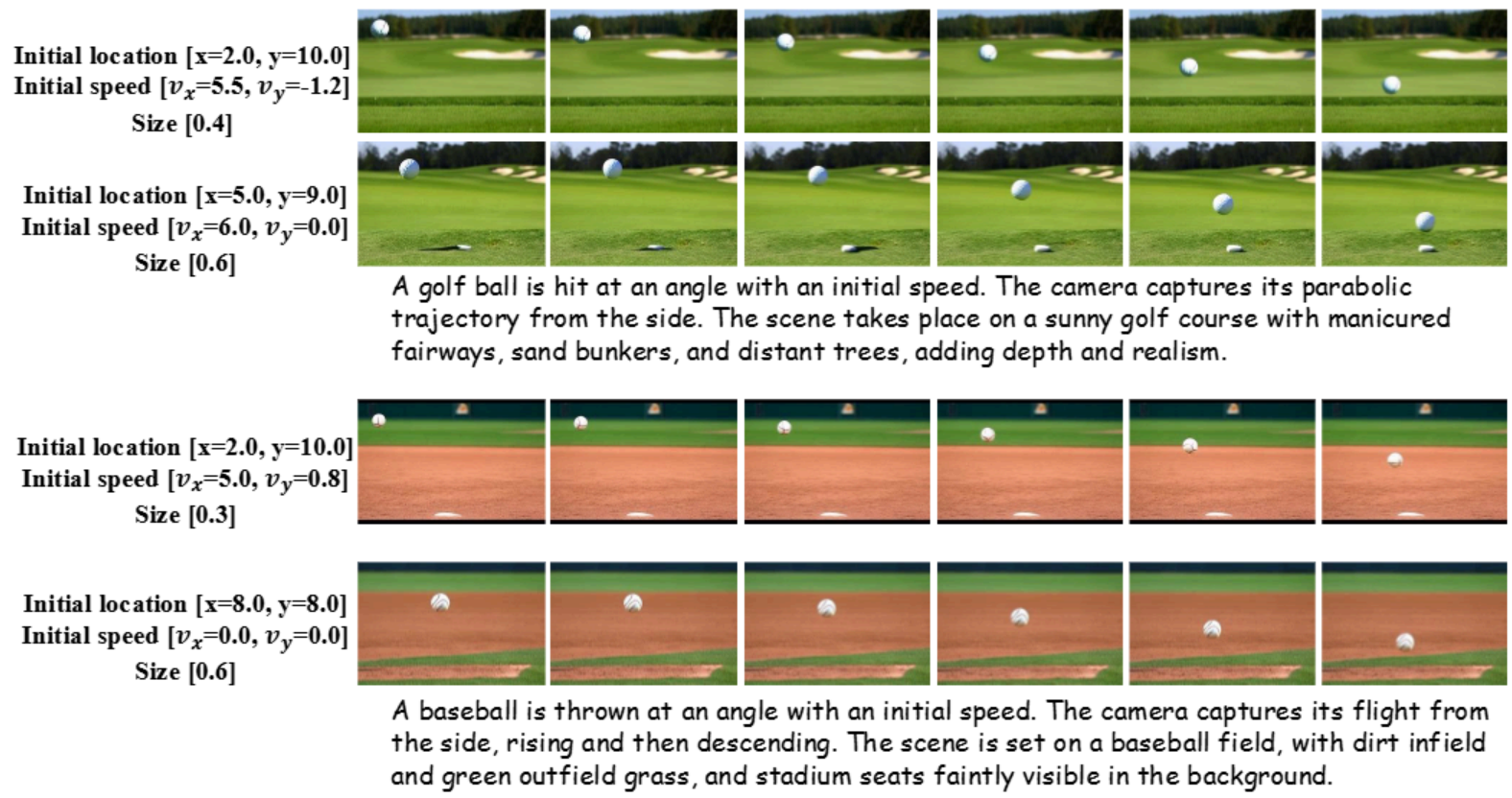



Figure 31: Given different initial physical states but the same scene description, NewtonGen can generate the corresponding motions.

E.2 MPCCR

图 30 和图 31 展示了 NewtonGen 的物理参数控制能力。

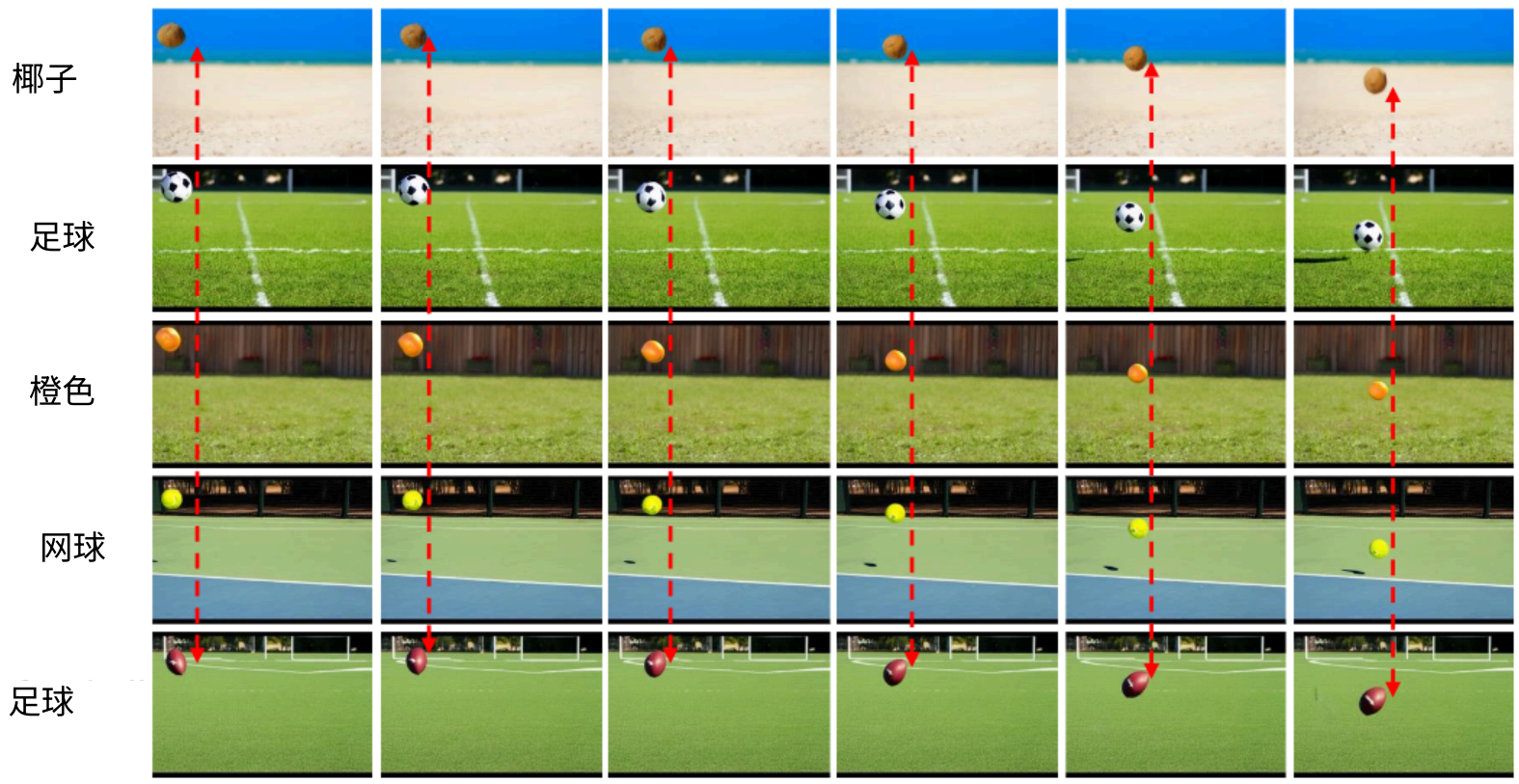



图 30：在相同的初始物理状态下，但具有不同的场景描述时，NewtonGen 能够生成具有一致运动的多样场景。

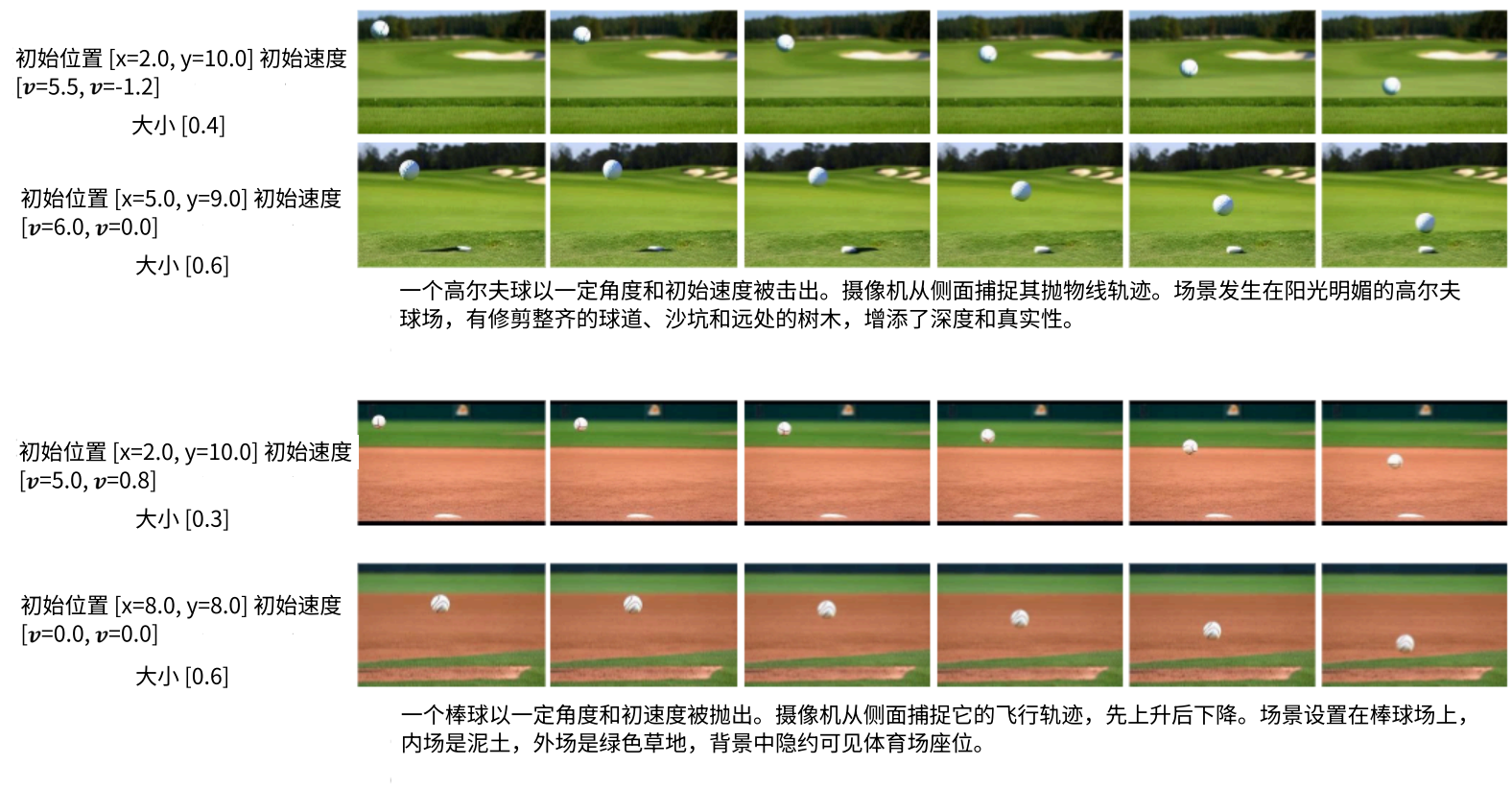



图 31：给定不同的初始物理状态但相同的场景描述，NewtonGen 可以生成相应的运动。

## F Questions and Answers

**Question 1:** Why is a NND (neural ODE) necessary to model/forecast Newtonian motion, and why not train a simple neural network to predict the coefficients of a parabola (for the parabolic trajectory)?
**Answer 1:** Our NND learns the underlying dynamics behind different systems, rather than merely fitting simple kinematics (trajectories) from data. They also provide a unified framework capable of representing diverse types of dynamics.

**Question 2:** For some motions, the underlying physical dynamics equations are already known, so why do we still need neural networks to learn dynamics?
**Answer 2:** Many complex or real-world motions are difficult to capture with simple physical formulas. For example, when rotation, parabolic motion, and even deformation occur simultaneously, it is challenging for humans to explicitly formulate the underlying physical laws. In contrast, our ODE model directly learns the dynamics from video data.

**Question 3:** Does your physical control model compromise the generative model's original physical effects or performance (e.g., shadows)?
**Answer 3:** Empirically, we have not observed any degradation in physical plausibility, such as shadow dynamics, after applying control. Our framework is not train-free in the second stage; instead, it injects physically consistent optical flow as a control condition only during inference, which preserves the model's original capabilities.

**Question 4:** Can NewtonGen (NND) handle video generation tasks involving collisions, rebounds, or explosions?
**Answer 4:** Currently, NewtonGen (NND) does not support such cases, as it is designed for continuous dynamics. These tasks would require additional event-based ODEs or hard-coded implementations.

**Question 5:** Can NewtonGen generate the motions of multiple objects' motion in a video?
**Answer 5:** Yes. NND can independently predict the physical states of multiple objects and then feed them into the motion-controlling video generator. The main bottleneck for video quality lies in the latter.

**Question 6:** Why choose "Go-with-the-Flow" instead of other motion control models as the base model for the second-stage video generation?
**Answer 6:** Other models often control motion through trajectories or bounding boxes, which makes it difficult for them to handle tasks involving deformation or rotation. In contrast, Go-with-the-Flow is based on optical flow control and thus has the potential to address such challenges.

**Question 7:** Is NND fast during training and inference?
**Answer 7:** Yes. NND is trained in the latent space rather than directly on videos, and its learnable parameters are concentrated in a lightweight three-layer MLP. As a result, inference can achieve real-time or faster speeds.

F 问题和答案

问题 1：为什么需要使用 NND（神经常微分方程）来模拟/预测牛顿运动，而不是训练一个简单的神经网络来预测抛物线（抛物线轨迹）的系数？

答案 1：我们的 NND 学习不同系统背后的基本动力学，而不仅仅是从数据中拟合简单的运动学（轨迹）。它们还提供了一个统一的框架，能够表示各种类型的动力学。

问题 2：对于某些运动，其基础物理动力学方程已经已知，那么为什么我们仍然需要神经网络来学习动力学？

回答 2：许多复杂或现实世界的运动难以用简单的物理公式捕捉。例如，当旋转、抛物线运动甚至变形同时发生时，人类很难明确地表述其背后的物理定律。相比之下，我们的 ODE 模型直接从视频数据中学习动力学。

问题 3：您的物理控制模型是否损害了生成模型的原始物理效果或性能（例如，阴影）？

回答 3：经验上，我们在应用控制后没有观察到物理合理性方面的任何退化，例如阴影动态。我们的框架在第二阶段不是免训练的；相反，它仅在推理过程中将物理一致的视差流作为控制条件注入，从而保留了模型的原始能力。

问题 4：NewtonGen（NND）能否处理涉及碰撞、反弹或爆炸的视频生成任务？

回答 4：目前，NewtonGen（NND）不支持这类情况，因为它是为连续动力学设计的。这些任务需要额外的基于事件的 ODE 或硬编码实现。

问题 5：NewtonGen 能否在视频中生成多个物体的运动？

回答 5：是的。NND 可以独立预测多个物体的物理状态，然后将它们输入到运动控制视频生成器中。视频质量的主要瓶颈在于后者。

问题 6：为什么选择“Go-with-the-Flow”而不是其他运动控制模型作为第二阶段视频生成的基准模型？

回答 6：其他模型通常通过轨迹或边界框来控制运动，这使得它们难以处理涉及变形或旋转的任务。相比之下，Go-with-the-Flow 基于光流控制，因此有可能应对这些挑战。

问题 7：NND 在训练和推理时是否快速？

回答 7：是的。NND 是在潜在空间中训练的，而不是直接在视频上训练，其可学习参数集中在一个轻量级的三层 MLP 中。因此，推理可以实现实时或更快的速度。