PISA Experiments: Exploring Physics Post-Training for Video Diffusion Models by Watching Stuff Drop

PISA 实验：通过观察物体下落探索视频扩散模型的物理后训练

Chenyu Li [*1] Oscar Michel [*1] Xichen Pan [1] Sainan Liu [2] Mike Roberts [2] Saining Xie [1]

Abstract

Large-scale pre-trained video generation models excel in content creation but are not reliable as physically accurate world simulators out of the box. This work studies the process of post-training these models for accurate world modeling through the lens of the simple, yet fundamental, physics task of modeling object freefall. We show state-of-the-art video generation models struggle with this basic task, despite their visually impressive outputs. To remedy this problem, we find that fine-tuning on a relatively small amount of simulated videos is effective in inducing the dropping behavior in the model, and we can further improve results through a novel reward modeling procedure we introduce. Our study also reveals key limitations of post-training in generalization and distribution modeling. Additionally, we release a benchmark for this task that may serve as a useful diagnostic tool for tracking physical accuracy in large-scale video generative model development. Code is available at this repository: https://github.com/vision-x-nyu/pisa-experiments.

大规模预训练视频生成模型在内容创作方面表现出色，但作为物理精确的世界模拟器，它们在开箱即用时并不可靠。本文通过一个简单但基础的物理任务——物体自由落体建模——来研究对这些模型进行后训练以实现精确世界建模的过程。我们展示了最先进的视频生成模型在这一基本任务上表现不佳，尽管它们的输出在视觉上令人印象深刻。为了解决这个问题，我们发现对相对少量的模拟视频进行微调可以有效诱导模型中的下落行为，并且我们通过引入一种新颖的奖励建模程序进一步改善了结果。我们的研究还揭示了后训练在泛化和分布建模方面的关键局限性。此外，我们发布了该任务的基准，该基准可能作为跟踪大规模视频生成模型开发中物理准确性的有用诊断工具。代码可在以下仓库获取：https://github.com/vision-x-nyu/pisa-experiments。

[*]Equal contribution, alphabetical order. [1]New York University [2]Intel Labs.

1. Introduction

在过去的一年中，视频生成模型取得了显著进展，激发了人们对这些模型未来可以作为逼真的世界模型的愿景 (?????)。最先进的视频生成模型在内容创作方面表现出令人印象深刻的结果 (????)，并且已经在广告和电影制作中得到应用 (??)。这些进展引发了一系列研究，试图将这些模型从内容创作者发展为具身代理的世界模拟器 (???)。然而，准确的世界建模比创意内容创作更具挑战性，因为仅仅看起来“足够好”是不够的：生成的像素必须忠实地代表一个按照物理定律和视觉透视演化的世界状态。

我们发现，尽管最先进模型的生成结果在视觉上令人印象深刻，但这些模型在生成物理上准确的结果方面仍然存在困难，即使这些模型是在展示了各种复杂物理交互的互联网规模视频数据上进行预训练的。未能将视觉生成与物理定律对齐表明，预训练是不够的，还需要一个后训练阶段。就像预训练的大型语言模型（LLMs）需要通过后训练才能成为有用的对话助手一样，预训练的视频生成模型也需要通过后训练才能部署为物理上准确的世界模拟器。

在这项工作中，我们通过专注于简单但基础的物理任务——**物体自由落体建模**，严格研究了视频生成模型的后训练过程，我们发现这对最先进的模型来说极具挑战性。具体来说，我们研究了一个图像到视频[1]（I2V）场景，目标是从物体悬停在空中的初始图像开始，生成物体下落并可能与地面其他物体碰撞的视频。我们选择研究这个单一任务，而不是整体的一般物理能力，因为它的简单性使我们能够进行受控实验，从而深入了解后训练过程的优势和局限性，我们相信这将成为生成世界建模研究中越来越重要的组成部分。此外，下

[1]我们讨论了为什么选择在图像到视频设置中而不是视频到视频设置中制定此任务的原因，详见 Appendix A。

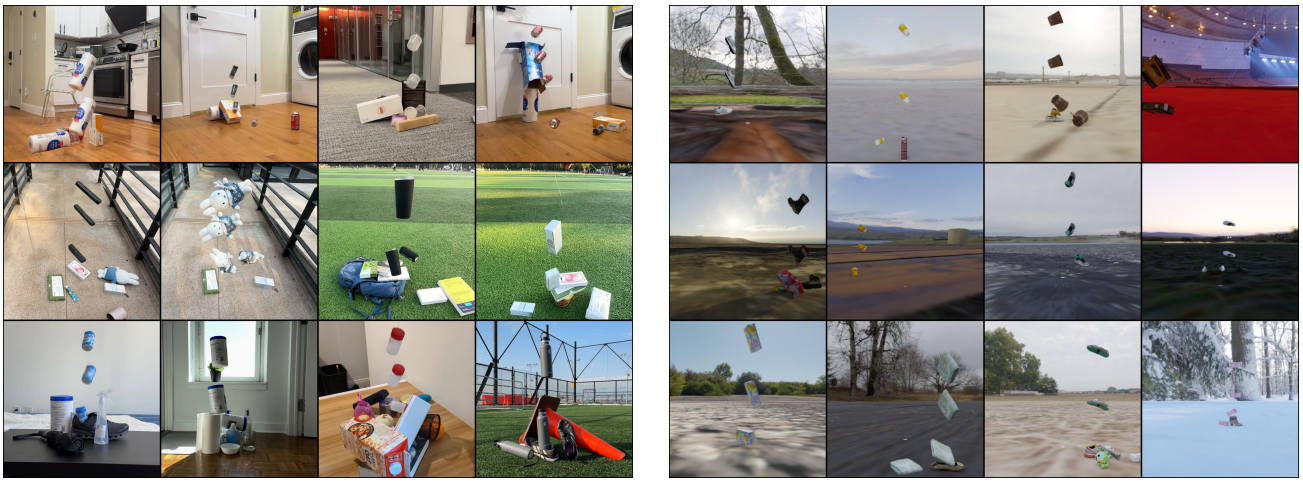


Figure 1. Our PISA (Physics-Informed Simulation and Alignment) evaluation framework includes a new video dataset, where objects are dropped in a variety of real-world (left) and synthetic (right) scenes. For visualization purposes, we depict object motion by overlaying multiple video frames in each image shown above. Our real-world videos enable us to evaluate the physical accuracy of generated video output, and our synthetic videos enable us to improve accuracy through the use of post-training alignment methods.

落任务的简单性使其可以在模拟中实现，这是可取的，因为它使我们能够轻松测试数据集扩展的特性，获得用于评估的地面真实注释，并能够精确操纵模拟环境以进行受控实验。

以伽利略著名的下落实验命名，我们引入了 PISA（Physics-Informed Simulation and Alignment）框架，用于在下落任务的背景下研究物理后训练。PISA 包括新的真实和模拟视频数据集，如 Figure 1所示，包含各种下落场景。PISA 还包括一组特定任务的指标，专注于测量物理准确性。我们的真实世界视频和指标使我们能够评估生成视频输出的物理准确性，而我们的合成视频使我们能够通过引入的后训练过程提高准确性。

我们的研究表明，当前最先进的视频生成模型在物理上准确的物体下落任务中表现不佳。生成的物体经常表现出不可能的行为，例如悬停在半空中、违背重力或在自由落体期间未能保持真实的轨迹。然而，我们发现简单的微调可以非常有效：在一个只有几千个样本的小数据集上微调一个开源模型，使其在物理准确性方面大大优于最先进的模型。我们进一步观察到，预训练模型对成功至关重要；没有利用大规模视频数据集进行预训练的随机初始化模型无法达到可比的结果。我们还引入了一种新的奖励建模框架，进一步提高了性能。我们展示了我们的奖励学习系统具有高度的灵活性，可以选择不同的奖励函数来针对不同的物理改进方向。

我们的分析还揭示了关键的局限性。首先，我们发现当模型面临训练分布之外的场景时，例如从未见过的深度或高度下落的物体，模型性能会下降。此外，尽管我们的后训练模型生成的物体运动在 3D 一致性和物理准确性上表现良好，但我们观察到生成的下落时间分布与地面真实分布之间存在不一致。

这些发现表明，后训练可能成为未来世界建模系统的重要组成部分。我们在这个相对简单的任务中发现的挑战在建模更复杂的物理现象时可能会持续存在。通过引入 PISA 框架和基准，我们为研究人员提供了一个有用的诊断工具，用于测试模型是否正在获得一般物理能力的道路上，以及识别研究人员在通过后训练将新能力集成到模型中时应注意的关键限制。

2. Related Work

Modeling Intuitive Physics. Intuitive physics refers to the innate or learned human capacity to make quick and accurate judgments about the physical properties and behaviors of objects in the world, such as their motion, stability, or interactions. This ability, present even in infancy (???), is crucial for navigating and understanding everyday life. Replicating intuitive physics is a foundational step toward creating systems that can interact effectively and safely in dynamic, real-world environments (?). Gravity, as a core component of intuitive physics, plays a pivotal role in both domains. It is one of the most universal and observable physical forces, shaping our expectations about object motion, stability, and interaction (??). Many studies in cognitive science (?) and AI (??) have relied on physics engines to evaluate and model intuitive physics. Our work uses the Kubric engine (?) to generate training videos.

直觉物理建模。直觉物理指的是人类天生或通过学习

获得的能力，能够快速准确地判断世界中物体的物理属性和行为，例如它们的运动、稳定性或相互作用。这种能力甚至在婴儿期就已经存在 (???)，对于日常生活的导航和理解至关重要。复制直觉物理是创建能够在动态、现实环境中有效且安全交互的系统的基础步骤 (?)。重力作为直觉物理的核心组成部分，在这两个领域中起着关键作用。它是最普遍且可观察的物理力之一，塑造了我们对物体运动、稳定性和相互作用的期望 (??)。认知科学 (?) 和人工智能 (??) 中的许多研究都依赖于物理引擎来评估和建模直觉物理。我们的工作使用 Kubric 引擎 (?) 生成训练视频。

Video Generation Models as World Simulators. Video generation has long been an intriguing topic in computer vision, particularly in the context of predicting future frames (??). More recently, as large-scale generative models have become prominent, Yang et al. explored how a wide range of real-world dynamics and decision-making processes can be expressed in terms of video modeling (??). The introduction of the Sora model (?) marked a leap in the quality of generated videos and ignited interest in leveraging such models as "world simulators." Over the past year, numerous video generation models have emerged, some open-source (????) and others commercially available (????). Related to our work, Kang et al. (?) study the extent to which video generation models learn generalizable laws of physics when trained on 2D data from a synthetic environment.

视频生成模型作为世界模拟器。视频生成长期以来一直是计算机视觉中一个引人入胜的话题，尤其是在预测未来帧的背景下 (??)。最近，随着大规模生成模型的兴起，Yang 等人探索了如何通过视频建模表达广泛的现实世界动态和决策过程 (??)。Sora 模型 (?) 的引入标志着生成视频质量的飞跃，并激发了利用此类模型作为 "世界模拟器" 的兴趣。在过去的一年中，涌现了许多视频生成模型，其中一些是开源的 (????)，另一些则是商业化的 (????)。与我们的工作相关的是，Kang 等人 (?) 研究了视频生成模型在合成环境的 2D 数据上训练时，学习可推广的物理定律的程度。

Evaluating Video Generation Models. Traditional image-based metrics for generative modeling, such Fréchet inception distance (FID) (?) or inception score (IS) (?), can be incorporated into the video domain, either by applying them on a frame-by-frame basis or by developing video-specific versions, such as Fréchet video distance (FVD) (?). Going beyond distribution matching measures, several benchmarks have developed suites of metrics that aim to better evaluate the semantic or visual quality of generated videos. For example, V-Bench (?) offers a more granular evaluation by measuring video quality across multiple dimensions, such as with respect to subject consistency or spatial relationships. In physics, some works, such as VideoPhy (?) and PhyGenBench (?), evaluate in the T2V setting by utilizing multimodal large language models (MLLM) to generate a VQA-based score. More recently, Cosmos (?) and Physics-IQ (?), evaluate physics in the image-to-video and video-to-video settings.

评估视频生成模型。传统的基于图像的生成模型评估指标，例如 Fréchet inception distance (FID) (?) 或 inception score (IS) (?)，可以通过逐帧应用或开发视频专用版本（如 Fréchet video distance (FVD) (?)）引入视频领域。超越分布匹配度量，一些基准测试开发了旨在更好地评估生成视频的语义或视觉质量的指标套件。例如，V-Bench (?) 通过测量视频在多个维度上的质量（如主题一致性或空间关系）提供了更细粒度的评估。在物理学方面，一些工作（如 VideoPhy (?) 和 PhyGenBench (?)）在 T2V 设置中利用多模态大语言模型（MLLM）生成基于 VQA 的分数进行评估。最近，Cosmos (?) 和 Physics-IQ (?) 在图像到视频和视频到视频的设置中评估物理特性。

3. PisaBench

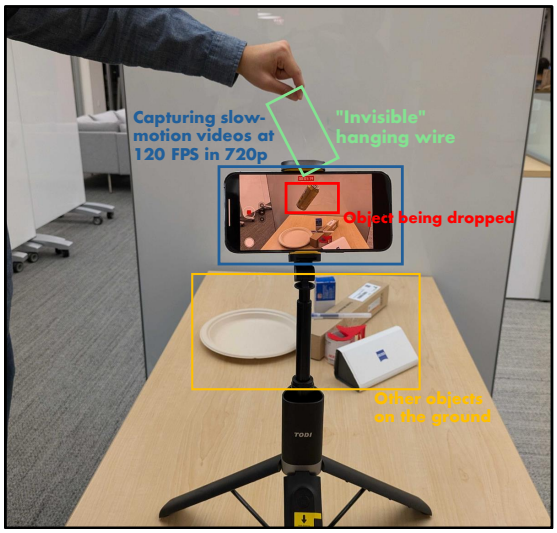


Figure 2. The setup for collecting real-world videos.

我们的基准测试，PisaBench，通过专注于一个简单的掉落任务，检验视频生成模型生成准确物理现象的能力。我们的基准测试，PisaBench，通过专注于一个简单的掉落任务，检验视频生成模型生成准确物理现象的能力。

3.1. 任务定义与假设

我们的任务可以总结如下：给定一个物体悬停在半空中的图像，生成一个物体掉落并与地面以及其他潜在物体碰撞的视频。由于视频是四维世界的不完整部分观察，我们做出了一些假设来约束任务空间。这些假

设对于确保我们的度量标准是物理准确性的可靠信号至关重要，因为它们只是从单一真实视频和生成视频中计算出的任务成功近似值。我们的任务可以总结如下：给定一个物体悬停在半空中的图像，生成一个物体掉落并与地面以及其他潜在物体碰撞的视频。由于视频是四维世界的不完整部分观察，我们做出了一些假设来约束任务空间。这些假设对于确保我们的度量标准是物理准确性的可靠信号至关重要，因为它们只是从单一真实视频和生成视频中计算出的任务成功近似值。

具体来说，我们假设掉落物体在初始帧中完全静止，只有重力作用在物体上，并且相机不移动。前两个假设对于图像到视频的设置是必要的。由于我们不提供多帧作为输入，没有这些假设，就无法确定掉落物体的初始速度或加速度。最后一个假设是必要的，因为我们的度量标准源自分割掩码的运动，如果相机移动，这些掩码将受到影响。具体来说，我们假设掉落物体在初始帧中完全静止，只有重力作用在物体上，并且相机不移动。前两个假设对于图像到视频的设置是必要的。由于我们不提供多帧作为输入，没有这些假设，就无法确定掉落物体的初始速度或加速度。最后一个假设是必要的，因为我们的度量标准源自分割掩码的运动，如果相机移动，这些掩码将受到影响。

3.2. 真实世界数据

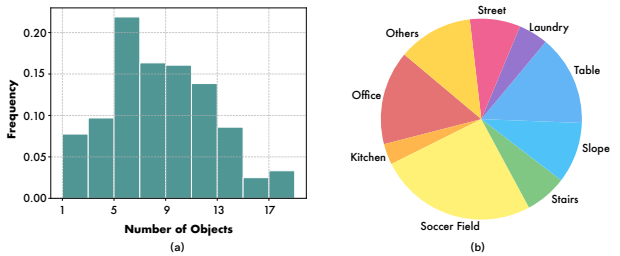


Figure 3. Statistics of the real-world data: (a) number of objects in each video, (b) the proportions of different scenes in the videos.

真实世界视频。我们收集了一组 361 个真实世界视频，展示了掉落任务以供评估。如??所示，数据集包括各种形状和大小的物体，拍摄于各种环境中，如办公室、厨房、公园等（见 Figure 3）。我们收集了一组 361 个真实世界视频，展示了掉落任务以供评估。如??所示，数据集包括各种形状和大小的物体，拍摄于各种环境中，如办公室、厨房、公园等（见 Figure 3）。

每个视频的第一帧中，物体由一根看不见的线悬挂，这是为了强制执行物体在视频开始时静止的假设。这个假设在我们的图像到视频设置中是必要的；否则，物体的初始速度是不明确的。我们在线释放后立即开始剪辑视频。我们以 120 帧每秒（fps）的慢动作录制视频，使用安装在三角架上的手机摄像头以消除相机运动。我们的视频收集设置的一个示例如 Figure 2所示。有关我们收集系统的更多详细信息，请参见 Appendix H。每个视频的第一帧中，物体由一根看不见的线悬挂，这是为了强制执行物体在视频开始时静止的假设。这个假设在我们的图像到视频设置中是必要的；否则，物体的初始速度是不明确的。我们在线释放后立即开始剪辑视频。我们以 120 帧每秒（fps）的慢动作录制视频，使用安装在三角架上的手机摄像头以消除相机运动。我们的视频收集设置的一个示例如 Figure 2所示。有关我们收集系统的更多详细信息，请参见 Appendix H。

模拟测试视频。由于我们的训练后过程使用了模拟视频数据集，我们还创建了一个包含 60 个视频的模拟测试集，以理解模拟到真实的转换。我们创建了两个各 30 个视频的分割：一个包含训练期间见过的物体和背景，另一个包含未见过的物体和背景。有关我们模拟数据创建的详细信息，请参见 Section 4.1。由于我们的训练后过程使用了模拟视频数据集，我们还创建了一个包含 60 个视频的模拟测试集，以理解模拟到真实的转换。我们创建了两个各 30 个视频的分割：一个包含训练期间见过的物体和背景，另一个包含未见过的物体和背景。有关我们模拟数据创建的详细信息，请参见 Section 4.1。

注释。如 ??所示，我们为每个视频添加了注释和从 SAM 2 (?) 视频分割模型估计的分割掩码。我们为每个物体创建了一个描述性标题，格式为“{物体描述}掉落”。当支持文本输入时，此标题用于提供任务上下文。如 ??所示，我们为每个视频添加了注释和从 SAM 2 (?) 视频分割模型估计的分割掩码。我们为每个物体创建了一个描述性标题，格式为“{物体描述} 掉落”。当支持文本输入时，此标题用于提供任务上下文。

3.3. 度量标准

我们提出了三个度量标准来评估轨迹的准确性、形状保真度和物体持久性。我们的每个度量标准都将真实视频的帧与生成视频的帧进行比较。有关度量标准的更多详细信息，包括其公式和我们用于处理 fps 差异的重采样过程，请参见 Appendix B。我们提出了三个度量标准来评估轨迹的准确性、形状保真度和物体持久性。我们的每个度量标准都将真实视频的帧与生成视频的帧进行比较。有关度量标准的更多详细信息，包括其公式和我们用于处理 fps 差异的重采样过程，请参见 Appendix B。

轨迹 L2。对于生成视频和真实视频中的每一帧，我们计算掩码区域的质心。然后，我们计算对应帧质心之间的平均 L_2 距离。对于生成视频和真实视频中的每一帧，我们计算掩码区域的质心。然后，我们计算对

应帧质心之间的平均 L_2 距离。

Chamfer **距离（CD）**。为了评估物体的形状保真度，我们计算生成视频和真实视频掩码区域之间的 Chamfer 距离（CD）。为了评估物体的形状保真度，我们计算生成视频和真实视频掩码区域之间的 Chamfer 距离（CD）。

交并比（IoU）。我们使用交并比（IoU）度量标准来评估物体持久性。IoU 测量生成视频和真实视频之间物体的重叠程度。我们使用交并比（IoU）度量标准来评估物体持久性。IoU 测量生成视频和真实视频之间物体的重叠程度。

3.4. 评估结果

我们评估了 4 个开放模型，包括 CogVideoX-5B-I2V(?)、DynamiCrafter(?)、Pyramid-Flow(?) 和 Open-Sora-V1.2(?)，以及 4 个专有模型，包括 Sora (?)、Kling-V1(?)、Kling-V1.5(?) 和 Runway Gen3 (?)。我们还评估了通过监督微调（PSFT）和物体奖励优化（ORO）过程后训练的 OpenSora；有关详细信息，请参见 Section 4。4 个开放模型，包括 CogVideoX-5B-I2V(?)、DynamiCrafter(?)、Pyramid-Flow(?) 和 Open-Sora-V1.2(?)，以及 4 个专有模型，包括 Sora (?)、Kling-V1(?)、Kling-V1.5(?) 和 Runway Gen3 (?)。我们还评估了通过监督微调（PSFT）和物体奖励优化（ORO）过程后训练的 OpenSora；有关详细信息，请参见 Section 4。

在基准测试上运行基线模型的结果表明，尽管生成的帧具有视觉真实性，但它们一致未能生成物理上准确的掉落行为。定性地，我们在 Figure 4中看到了常见的失败案例，如不合理的物体变形、漂浮、新物体的幻觉和不真实的特效。我们进一步在 ??的左侧可视化了一组随机生成的轨迹。在许多情况下，物体保持完全静止，有时甚至向上移动。当下落运动存在时，它通常很慢或包含不真实的水平运动。在基准测试上运行基线模型的结果表明，尽管生成的帧具有视觉真实性，但它们一致未能生成物理上准确的掉落行为。定性地，我们在 Figure 4中看到了常见的失败案例，如不合理的物体变形、漂浮、新物体的幻觉和不真实的特效。我们进一步在 ??的左侧可视化了一组随机生成的轨迹。在许多情况下，物体保持完全静止，有时甚至向上移动。当下落运动存在时，它通常很慢或包含不真实的水平运动。

4. Physics Post-Training

We present a post-training process to address the limitations of current models described in Section 3.4. We utilize simulated videos that demonstrate realistic dropping behavior. Our approach for post-training is inspired by the two-stage pipeline consisting of supervised fine-tuning followed by reward modeling commonly used in LLMs. We find that our pipeline improves performance on both real and simulated evaluations, with greater gains observed in simulation. This is due to the sim-to-real gap, though our approach still shows substantial gains in transferring to real-world data.

我们提出了一种后训练过程，以解决当前模型在 Section 3.4中描述的局限性。我们利用展示真实掉落行为的模拟视频。我们的后训练方法受到 LLM 中常用的两阶段管道的启发，该管道包括监督微调后接奖励建模。我们发现，我们的管道在真实和模拟评估中都提高了性能，在模拟中观察到了更大的增益。这是由于模拟到现实的差距，尽管我们的方法在转移到现实世界数据时仍然显示出显著的增益。

4.1. Simulated Adaptation Data

The first stage of our approach involves supervised fine-tuning. We use Kubric (?), a simulation and rendering engine designed for scalable video generation, to create simulated videos of objects dropping and colliding with other objects on the ground. Each video consists of 1-6 dropping objects onto a (possibly empty) pile of up to 4 objects underneath them. The videos are 2 seconds long, consisting of 32 frames at 16 fps. The objects are sourced from the Google Scanned Objects (GSO) dataset (?), which provides true-to-scale 3D models created from real-world scans across diverse categories (examples shown in ??). The camera remains stationary in each video and is oriented parallel to the ground plane. To introduce variability, we randomly sample the camera height between 0.4 and 0.6 meters and position objects between 1 and 3 meters away from the camera, which corresponds to the distributions observed in the real-world dataset. More information about the dataset can be found in Appendix K.

我们方法的第一阶段涉及监督微调。我们使用 Kubric (?)，一个为可扩展视频生成设计的模拟和渲染引擎，来创建物体掉落并与地面上的其他物体碰撞的模拟视频。每个视频包含 1-6 个掉落的物体，落在下方最多 4 个物体（可能为空）的堆上。视频长度为 2 秒，由 16 帧每秒的 32 帧组成。物体来自 Google Scanned Objects (GSO) 数据集 (?)，该数据集提供了从现实世界扫描中创建的真实比例的 3D 模型，涵盖多种类别（示例如 ??所示）。在每个视频中，摄像机保持静止，并与地平面平行。为了引入变化，我们随机采样摄像机高度在 0.4 到 0.6 米之间，并将物体放置在距离摄像机 1 到 3 米的位置，这对应于在现实世界数据集中观察到的分布。关于数据集的更多信息可以在 Appendix K中找到。

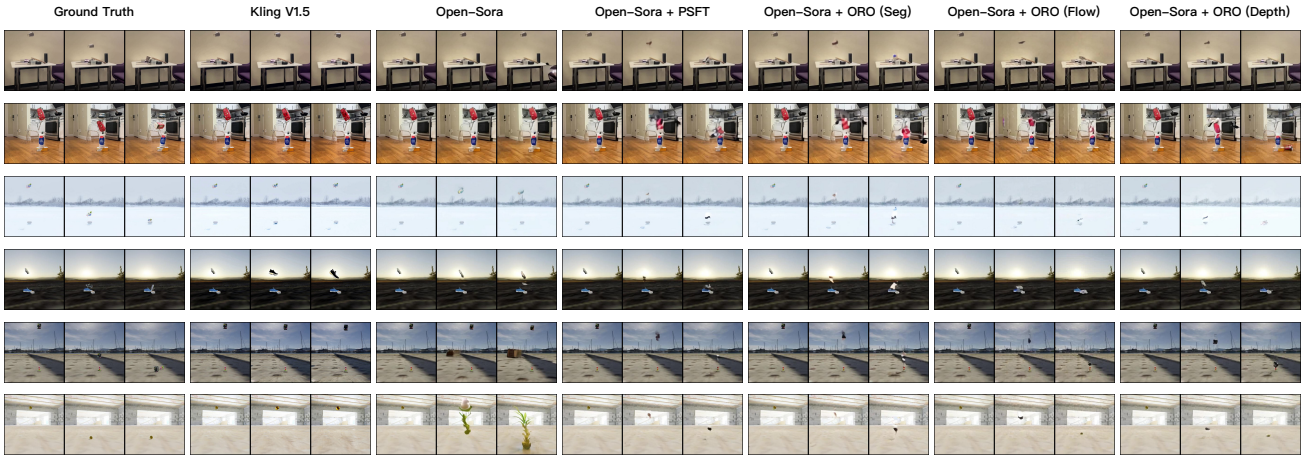


Figure 4. Qualitative comparison of results on real test set (row 1-2), simulated seen test set (row 3-4) and simulated unseen test set (row 5-6). We present the results of popular open-source and commercially available models alongside those of models fine-tuned through our method. Existing models often struggle to generate videos depicting objects falling, whereas our PSFT method effectively introduces knowledge of free-fall into the model. ORO enables the model to more accurately learn object motion and shape.

4.2. Physics Supervised Fine-Tuning (PSFT).

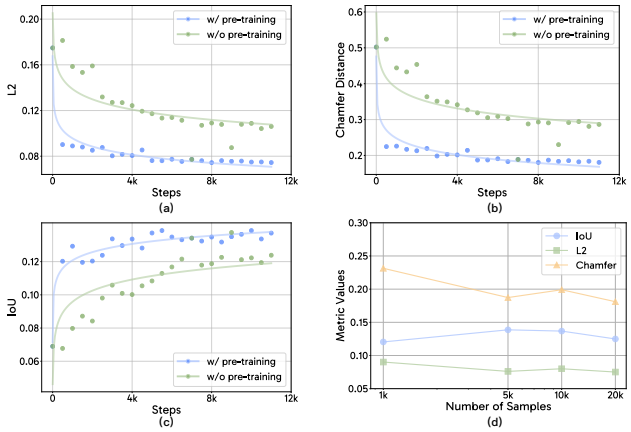


Figure 5. Plots (a), (b), and (c) demonstrate that our metrics tend to improve with further training and that leveraging a pre-trained video diffusion model enhances performance compared to random initialization. In plot (d), the size of the training dataset varies in each training run (each consisting of 5k steps). With only 5k samples, we can achieve optimal results.

We use the pretrained Open-Sora v1.2 (?) model as our base model and fine-tune it on our simulated video dataset. We employ Open-Sora v1.2's rectified flow training objective without modification (?). Each fine-tuning experiment is conducted with a batch size of 128 and a learning rate of 1e−4 on two 80GB NVIDIA A100 GPUs. As shown in Figure 4, fine-tuning with this data alone is sufficient to induce realistic dropping behavior in the model. Quantitatively, our PSFT model substantially improves on both our simulated and real-world benchmark, as shown in Table 1. Dataset size. We conduct an ablation study on the number of training samples to understand the amount of data required for optimal performance on our benchmark. We create random subsets from 500 to 20,000 samples and train our model for 5,000 gradient steps on each subset. Notably, as shown in Figure 5, only 5,000 samples are needed to achieve optimal results. Effect of pre-training. Additionally, we investigate the impact of Open-Sora's pre-training on adaptation. We randomly initialize the Open-Sora's denoising network while keeping the pre-trained initialization of the compressor network and train the model on a dataset of 5k training samples. As shown in ??, the learned knowledge from Open-Sora's pre-training plays a critical role in our task.

我们使用预训练的 Open-Sora v1.2 (?) 模型作为基础模型，并在我们的模拟视频数据集上对其进行微调。我们采用 Open-Sora v1.2 的校正流训练目标，未作修改 (?)。每个微调实验在两个 80GB 的 NVIDIA A100 GPU 上进行，批量大小为 128，学习率为 1e−4。如 Figure 4所示，仅使用这些数据进行微调足以在模型中引发真实的掉落行为。定量上，我们的 PSFT 模型在模拟和现实世界基准测试中都显著改进，如 Table 1所示。**数据集大小**。我们对训练样本数量进行了消融研究，以了解在我们的基准测试中达到最佳性能所需的数据量。我们从 500 到 20,000 个样本中创建随机子集，并在每个子集上训练模型 5,000 个梯度步。值得注意的是，如 Figure 5所示，仅需要 5,000 个样本即可达到最佳结果。**预训练的影响**。此外，我们研究了 Open-Sora 的预训练对适应的影响。我们随机初始化 Open-Sora 的去噪网络，同时保持压缩网络的预训练初始化，并在 5k 训练样本的数据集上训练模型。如 ??所示，从 Open-Sora 的预训练中学到的知识在我们的任务中起

	Method	Real			Sim (Seen)			Sim (Unseen)		
		L2 (↓)	CD (↓)	IoU (↑)	L2 (↓)	CD (↓)	IoU (↑)	L2 (↓)	CD (↓)	IoU (↑)
Proprietary	Sora (?)	0.174	0.488	0.065	0.149	0.446	0.040	0.140	0.419	0.031
	Kling-V1 (?)	0.157	0.425	0.056	0.142	0.415	0.032	0.145	0.437	0.028
	Kling-V1.5 (?)	0.155	0.424	0.058	0.137	0.396	0.033	0.132	0.405	0.029
	Runway Gen3 (?)	0.187	0.526	0.042	0.170	0.509	0.040	0.149	0.460	0.038
Open	CogVideoX-5B-I2V (?)	0.138	0.366	0.080	0.112	0.315	0.020	0.101	0.290	0.020
	DynamiCrafter (?)	0.187	0.504	0.021	0.157	0.485	0.039	0.136	0.430	0.033
	Pyramid-Flow (?)	0.175	0.485	0.062	0.126	0.352	0.059	0.130	0.381	0.048
	Open-Sora (?)	0.175	0.502	0.069	0.139	0.409	0.036	0.130	0.368	0.034
Ours	Open-Sora + PSFT (base)	0.076	0.188	0.139	0.036	0.088	0.165	0.028	0.058	0.129
	base + ORO (Seg)	0.075	0.183	0.142	0.033	0.076	0.170	0.032	0.063	0.145
	base + ORO (Flow)	0.067	0.164	0.136	0.026	0.062	0.122	0.022	0.045	0.071
	base + ORO (Depth)	0.067	0.159	0.129	0.031	0.072	0.124	0.022	0.046	0.096

Table 1. PisaBench Evaluation Results. This table compares the performance of four proprietary models, four open models, and the models fine-tuned with PSFT and PSFT + ORO on our real-world and simulated test set which is decomposed into seen and unseen object splits. Across all metrics, our PSFT models outperform all other baselines, including proprietary models like Sora. Reward modeling further enhances results, with segmentation rewards improving the shape-based IoU metric and optical rewards and depth rewards enhancing the motion-based L2 and CD metrics. This suggests that rewards can be flexibly adjusted to target specific aspects of performance.

着关键作用。

Overall, using PSFT on only 5k samples is sufficient to push Open-Sora's performance past all other evaluated models, including state-of-the-art commercial video generators, by a wide margin. This is made possible by leveraging the knowledge from the sufficiently pre-trained base model.
总体而言，仅使用 5k 样本进行 PSFT 足以使 Open-Sora 的性能远远超过所有其他评估模型，包括最先进的商业视频生成器。这是通过利用充分预训练的基础模型的知识实现的。

4.3. Object Reward Optimization (ORO)

In the second stage, we propose Object Reward Optimization (ORO) to use reward gradients to guide the video generation model toward generating videos where the object's motion and shape more closely align with the ground truth. We follow the VADER framework from (?) and introduce three reward models. The differences between our approach and VADER include: (1) our reward model utilizes both generated videos and ground truth instead of generated videos and conditioning. (2) gradients propagate through all denoising time steps in fine-tuning. Consequently, the VADER objective is modified as follows:

$$J(\theta) = \mathbb{E}_{(x_0,c)\sim\mathcal{D},x_0'\sim p_\theta(x_0'|c)}[R(x_0', x_0)] \qquad (1)$$

where $\mathcal{D}$ is the ground truth dataset, $p_\theta(.)$ is a given video diffusion model, $x_0', x_0 \in \mathbb{R}^{H\times W\times 3}$ are generated video and ground truth, and $c \in \mathbb{R}^{H\times W\times 3}$ is the initial image.
在第二阶段，我们提出了物体奖励优化（ORO），利用奖励梯度来指导视频生成模型生成物体的运动和形状更接近真实情况的视频。我们遵循 VADER 框架 (?)，并引入了三个奖励模型。我们的方法与 VADER 的区别包括：（1）我们的奖励模型利用生成的视频和真实情况，而不是生成的视频和条件。（2）在微调过程中，梯度通过所有去噪时间步传播。因此，VADER 的目标修改如下：

$$J(\theta) = \mathbb{E}_{(x_0,c)\sim\mathcal{D},x_0'\sim p_\theta(x_0'|c)}[R(x_0', x_0)] \qquad (2)$$

其中 $\mathcal{D}$ 是真实情况数据集，$p_\theta(.)$ 是给定的视频扩散模型，$x_0', x_0 \in \mathbb{R}^{H\times W\times 3}$ 是生成的视频和真实情况，$c \in \mathbb{R}^{H\times W\times 3}$ 是初始图像。

Segmentation Reward. We utilize SAM 2 (?) to generate segmentation masks across frames for generated videos. We define segmentation reward as the IoU between the dropping object's mask in generated video and the mask from the groud truth simulated segmentation.
分割奖励。我们使用 SAM 2 (?) 为生成的视频生成跨帧的分割掩码。我们将分割奖励定义为生成视频中掉落物体的掩码与真实模拟分割的掩码之间的 IoU。

Optical Flow Reward. We utilize RAFT (?) to generate generated video's optical flow V^{gen} and ground

truth's optical flow V^{gt}. We define the optical flow reward as $R(x_0', x_0) = -|V^{\text{gen}} - V^{\text{gt}}|$.
光流奖励。我们使用 RAFT (?) 生成生成视频的光流 V^{gen} 和真实情况的光流 V^{gt}。我们将光流奖励定义为 $R(x_0', x_0) = -|V^{\text{gen}} - V^{\text{gt}}|$。

Depth Reward. We utilize Depth-Anything-V2 (?) to generate generated video's depth map D^{gen} and ground truth's depth map D^{gt}. We define the optical flow reward as $R(x_0', x_0) = -|D^{\text{gen}} - D^{\text{gt}}|$.
深度奖励。我们使用 Depth-Anything-V2 (?) 生成生成视频的深度图 D^{gen} 和真实情况的深度图 D^{gt}。我们将光流奖励定义为 $R(x_0', x_0) = -|D^{\text{gen}} - D^{\text{gt}}|$。

Details on implementation can be found in Appendix C.
实现细节可以在 Appendix C中找到。

We begin from the checkpoint of the first stage, which is trained on 5,000 samples trained over 5,000 gradient steps. We then fine-tune the model with ORO on the simulated dataset, using a batch size of 1 and two 80GB NVIDIA A100 GPUs for each fine-tuning experiment. We set a learning rate of 1e−6 for segmentation reward and depth reward and 1e−5 for optical flow.
我们从第一阶段的检查点开始，该检查点在 5,000 个样本上训练了 5,000 个梯度步。然后，我们在模拟数据集上使用 ORO 对模型进行微调，每个微调实验使用批量大小为 1 和两个 80GB 的 NVIDIA A100 GPU。我们为分割奖励和深度奖励设置了 1e−6 的学习率，为光流设置了 1e−5 的学习率。

As shown in Table 1, incorporating ORO in reward modeling further improves performance. Additionally, each reward function enhances the aspect of physicality that aligns with its intended purpose—segmentation rewards improve shape accuracy, while flow rewards and depth rewards improve motion accuracy. This demonstrates the process is both modular and interpretable.
如 Table 1所示，在奖励建模中引入 ORO 进一步提高了性能。此外，每个奖励函数都增强了与其预期目的相符的物理性——分割奖励提高了形状准确性，而光流奖励和深度奖励提高了运动准确性。这表明该过程既模块化又可解释。

5. 评估学习到的物理行为

在 Section 4中介绍了我们的训练后方法后，我们深入探讨了模型对重力和透视之间相互作用的理解——这两个法则决定了我们视频的动力学。我们首先测试模型学习到的物理行为是否能够推广到超出其训练分布的掉落高度和深度。然后，我们研究模型学习由透视不确定性引起的概率分布的能力。

5.1. 对未见过的深度和高度的泛化

深度和高度是影响我们视频中物体下落动力学的主要因素。通过结合重力法则和透视法则，在我们的相机假设下，我们可以将物体的图像 y 坐标建模为时间的函数（关于我们坐标系的更多细节在 Appendix G中描述）：

$$y(t) = \frac{f}{Z}\left(Y_0 - \frac{1}{2}gt^2\right). \tag{3}$$

从 Equation (3) 中，我们可以看到影响物体运动的随机变量是 Z（深度）和 Y（高度）（相机焦距 f 是固定的）。因此，我们有兴趣测试在未见过的 Y 和 Z 值上的泛化能力。

我们创建了一个模拟测试集，其中单个物体从不同的深度和高度掉落，使用训练期间未见过的物体和背景。我们分别从 $[1, 5]$ 和 $[0.5, 2.5]$ 的笛卡尔积中均匀采样深度和高度值（以米为单位）。相机高度固定在 $0.5m$，并且丢弃相机视锥之外的深度-高度对。如果样本的掉落深度和高度都落在 $[1, 3]$ 和 $[0.5, 1.5]$ 范围内，则该样本属于分布内（ID）。

由于我们在模拟中可以访问真实的下落时间，我们还使用了下落时间误差，我们在 Appendix B中描述了这个指标。我们在 Table 2中的分析表明，在分布外场景中性能下降。

由于深度和高度是影响下落动力学的主要物理量，这一发现表明我们的模型可能难以学习一个完全可推广的法则来解释透视和重力的相互作用。

Setting	L2 (↓)	Chamfer (↓)	IOU (↑)	Time Error (↓)
ID	0.036	0.088	0.155	0.091
OOD	0.044	0.143	0.049	0.187

Table 2. Results of our metrics on in-distribution (ID) and out-of-distribution (OOD) depth-height combinations. The values used for depth range from 1-5m (ID range $[1, 3]$) and height values range from 0.5-2.5 (ID range $[0.5, 1.5]$).

5.2. 分布分析

物理系统的演化并不是由单个初始图像唯一确定的，因为透视的有损不确定性会导致一系列可能的结果，如 Figure 6所示。一个理想的视频世界模型应该（1）输出与某些合理世界状态演化一致的视频，并且（2）在其条件信号可能的整个世界分布上提供准确的覆盖。在本节中，我们通过研究 $p(t|y)$ 来检验这两个方面：从图像平面中坐标 y 处的物体可能的下落时间分布。为此，我们创建了一个模拟数据集，其 $p(t|y)$ 分布比我们的 PSFT 数据集更广泛。关于其构建的更多细节，请参见 Appendix F。

测试（1）：轨迹的三维一致性。

在这个新数据集上训练我们的模型后，我们测试其

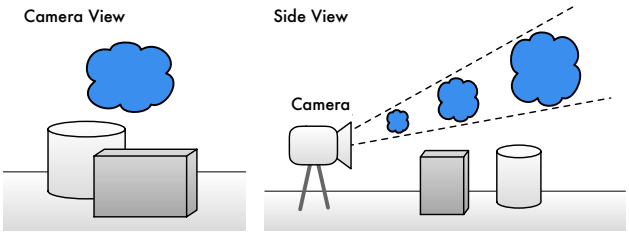


Figure 6. Demonstration of ambiguity in 2D perspective projections. Each of the three clouds appears the exact same in the camera's image. The right side shows how we perform a scale and translation augmentation to generate deliberately ambiguous data.

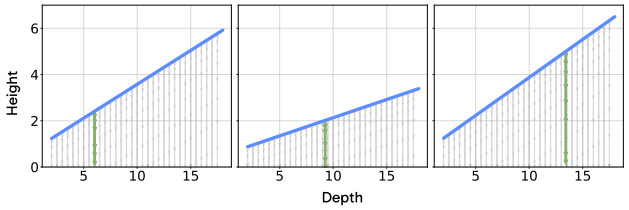


Figure 7. Examples of model trajectories lifted to 3D. The blue line represents the height of the camera ray passing through the bottom of the dropping object as a function of depth. The set of possible dropping trajectories at a given depth are depicted in gray. The lifted trajectory of the model is depicted in green.

轨迹是否与有效的三维世界状态一致。我们首先使用 Section 5.1中描述的过程从生成的视频中获取估计的下落时间。利用相机位置、焦距、传感器宽度和 y 的知识，我们可以获得轨迹的隐含深度和高度。然后，我们可以将视频轨迹反投影到三维空间，并分析它们是否构成物理上准确的轨迹。我们在 Appendix G中提供了关于这个过程的更多细节。如 Figure 7所示，我们发现我们模型的提升轨迹始终与由其下落时间隐含的高度和深度的三维轨迹一致，这证明了模型的视觉输出与某些合理的现实世界状态一致。

测试（2）：分布一致性。

超越单个轨迹的层面，我们研究了模型学习到的条件分布 $p(t|y)$。我们创建了 50 个不同的初始图像，每个图像的 y 值不同，从每个图像生成 128 个不同的视频，并估计每个视频中的下落时间。利用重力法则、透视法则以及我们数据集中均匀深度采样的假设，我们可以解析地推导出概率 $p(t|y)$ 为

$$p(t|y) = \begin{cases} \frac{gt}{(Z_{\max} - Z_{\min})\beta}, & t_{\min} \le t \le t_{\max} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

其中 β 是一个依赖于 f、y 和相机高度的常数。推导过程在 Appendix E中给出。

然后，我们使用 Kolmogorov–Smirnov（KS）检验 (?)

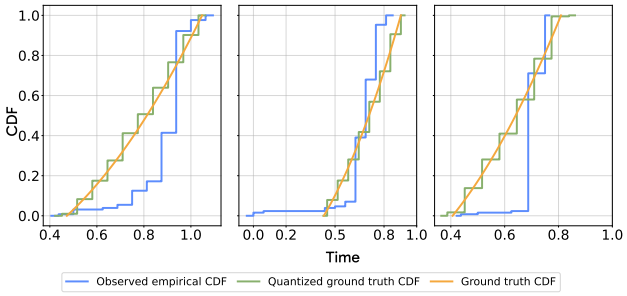


Figure 8. Visualizing $p(t|y)$ misalignment for different images. Green shows the ground-truth CDF, orange is the 32-frame quantized version, and blue is the empirical CDF of 128 different samples of dropping times from the model.

对每个实验的 50 个样本进行拟合优度测量。KS 检验的零假设是两个被比较的分布是相等的，我们认为 p 值小于 0.05 是分布不一致的证据。由于我们测量的时间精度有限，并且只能取 32 个不同的值——由于估计接触帧的原因——我们使用蒙特卡罗方法近似真实分布 $p(t|y)$。我们从真实分布中采样 1000 个值，然后将它们量化为 32 个对应于帧的区间，我们在 KS 检验中使用这些作为真实观测值。我们发现，在所有 50/50 的情况下，检验的 p 值小于 0.05，这提供了模型没有学习到正确下落时间分布的证据。我们在 Figure 8中可视化了模型的经验 CDF 与真实分布之间的不一致。

总之，虽然我们模型的轨迹显示出将自己锚定到合理三维世界状态的有希望的趋势，但模型可能输出的范围与真实分布并不一致。

6. Conclusion

This work studies post-training as an avenue for adapting adapting pre-trained video generator into world models. We introduce a post-training strategy that is highly effective in aligning our model. Our work raises interesting insights into the learned distributions of generative models. Qualitatively, large scale image or video generative models appear to excel at generating likely samples from the data distribution, but this alone does not imply that they match the data distribution well in its entirety. As long as a model is able to generate likely samples, global distributional misalignment is not necessarily a problem for content creation. However, this problem becomes critical for world models, where alignment across the entire distribution is necessary for faithful world simulation. The insights revealed by our study, made possible by our constrained and tractable setting, indicate that although post-training improves per-sample accuracy, general distributional alignment remains unsolved.
这项工作研究了将预训练的视频生成器适应为世界模型的后训练方法。我们提出了一种非常有效的后训练

策略，用于对齐我们的模型。我们的工作对生成模型的学习分布提出了有趣的见解。从定性上看，大规模图像或视频生成模型似乎擅长从数据分布中生成可能的样本，但这本身并不意味着它们在整个数据分布上匹配得很好。只要一个模型能够生成可能的样本，全局分布不对齐并不一定是内容创建的问题。然而，对于世界模型来说，这个问题变得至关重要，因为在整个分布上的对齐对于忠实的世界模拟是必要的。我们的研究揭示的见解表明，尽管后训练提高了每个样本的准确性，但一般的分布对齐问题仍未解决。

致谢

We thank Boyang Zheng, Srivats Poddar, Ellis Brown, Shengbang Tong, Shusheng Yang, Jihan Yang, Daohan Lu, Anjali Gupta and Ziteng Wang for their help with data collection.
我们感谢郑伯阳、Srivats Poddar、Ellis Brown、童圣邦、杨树生、杨继汉、卢道涵、Anjali Gupta 和王子腾在数据收集方面的帮助。

We thank Jiraphon Yenphraphai for valuable assistance in setting up our simulation code.
我们感谢 Jiraphon Yenphraphai 在设置模拟代码方面提供的宝贵帮助。

We thank Runway and Kling AI for providing API credit.
我们感谢 Runway 和 Kling AI 提供的 API 积分。

SX also acknowledges support from Intel AI SRS, Korean AI Research Hub, Open Path AI Foundation, Amazon Research Award, Google TRC program, and NSF Award IIS-2443404.
SX 还感谢 Intel AI SRS、韩国人工智能研究中心、Open Path AI 基金会、亚马逊研究奖、Google TRC 计划以及 NSF 奖项 IIS-2443404 的支持。

References

A. Discussion of Image-to-Video setting.

We note that our choice of single-image input, as opposed to multi-frame input, comes with some trade-offs. We choose the image-to-video setting because it is widely supported among many different models, allowing us to make effective comparisons across the current state-of-the-art. However, only conditioning on a single frame introduces significant ambiguity. Due to the loss of information caused by projecting the 3D world through perspective, it may not be possible to directly infer the size of the object or its height. In practice, we find our metrics are still reliable signals of task success, but we still study the problem of ambiguity more extensively in Section 5.2.
我们注意到，与多帧输入相比，我们选择单图像输入会带来一些权衡。我们选择图像到视频设置，因为它在许多不同的模型中得到广泛支持，使我们能够在当前最先进的技术中进行有效比较。然而，只对单帧进行条件化会带来很大的歧义。由于通过透视投影 3D 世界会导致信息丢失，因此可能无法直接推断物体的大小或高度。在实践中，我们发现我们的指标仍然是任务成功的可靠信号，但我们仍在 Section 5.2 中更广泛地研究歧义问题。

B. Metric details.

We propose three metrics to assess the accuracy of trajectories, shape fidelity, and object permanence. Each of our metrics compare frames from the ground-truth video with the generated video. Because different models can operate at different fps, we perform fps alignment as part of our evaluation process. To perform fps alignment, we map each frame index of the generated videos to the ground truth using f_{gen} and f_{gt}, where f_{gen} and f_{gt} are the fps of generated video and ground truth respectively. For i-th frame in the generated video, we find the corresponding aligned frame index j in the ground truth video
我们提出了三个指标来评估轨迹的准确性、形状保真度和物体持久性。我们的每个指标都将来自真实视频的帧与生成的视频进行比较。由于不同的模型可以以不同的 fps 运行，因此我们在评估过程中执行 fps 对齐。为了执行 fps 对齐，我们使用 f_{gen} 和 f_{gt} 将生成的视频的每个帧索引映射到真实视频，其中 f_{gen} 和 f_{gt} 分别是生成的视频和真实视频的 fps。对于生成的视频中的第 i 帧，我们在真实视频中找到相应的对齐帧索引 j：

$$j = \text{round}(i \cdot \frac{f_{\text{gen}}}{f_{\text{gt}}}) \tag{5}$$

Through fps alignment, we downsample the ground truth video to match the frame number of the generated video. We denote the downsampled ground truth as $\{I_i^{\text{gt}}\}_{i=1}^N$ and the generated video as $\{I_i^{\text{gen}}\}_{i=1}^N$, where N is the number of frames in the generated video.
Trajectory L2. For each frame in both the generated video and ground truth, we calculate the centroid of the masked region. We then compute L_2 distance between the centroids of corresponding frames
通过 fps 对齐，我们对真实视频进行下采样，以匹配生成视频的帧数。我们将下采样的真实视频表示为 $\{I_i^{\text{gt}}\}_{i=1}^N$，将生成的视频表示为 $\{I_i^{\text{gen}}\}_{i=1}^N$，其中 N 是生成视频中的帧数。
轨迹 L2。对于生成的视频和真实视频中的每一帧，我们计算掩蔽区域的质心。然后，我们计算相应帧质心之间的 L_2 距离：

$$L_2 = \frac{1}{N}\sum_{i=1}^{N} \|C_i^{\text{gen}} - C_i^{\text{gt}}\|_2 \tag{6}$$

where $C_i^{\text{gen}}, C_i^{\text{gt}} \in \mathbb{R}^2$ are the centroids of the dropping object in the i-th frame of generated video and the ground truth respectively.
Chamfer Distance (CD). To assess the shape fidelity of objects, we calculate the Chamfer Distance (CD) between the mask regions of the generated video and ground truth:

$$\text{CD} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{|P_i|}\sum_{p\in P_i}\min_{q\in Q_i}\|p-q\|_2 + \frac{1}{|Q_i|}\sum_{q\in Q_i}\min_{p\in P_i}\|q-p\|_2\right)$$

where $P_i = \{p_j\}_{j=1}^{|P_i|}$ and $Q_i = \{q_j\}_{j=1}^{|Q_i|}$ are the sets of mask points in the i-th frame of the generated video and ground truth respectively.
其中 $P_i = \{p_j\}_{j=1}^{|P_i|}$ 和 $Q_i = \{q_j\}_{j=1}^{|Q_i|}$ 分别是生成视频和地面实况的第 i 帧中的掩码点集。

Intersection over Union (IoU). We use the Intersection over Union (IoU) metric to evaluate object permanence. IoU measures objects' degree of overlap between the generated video and ground truth. This is formulated as follows
交并比 (IoU)。我们使用交并比 (IoU) 指标来评估物体持久性。IoU 测量生成的视频和地面实况之间物体的重叠程度。其公式如下：

$$\text{IoU} = \frac{1}{|N|}\sum_{i=1}^{N}\frac{|M_i^{\text{gen}} \cap M_i^{\text{gt}}|}{|M_i^{\text{gen}} \cup M_i^{\text{gt}}|} \tag{7}$$

where $M_i^{\text{gen}}, M_i^{\text{gt}} \in \{0,1\}^{H\times W}$ are binary segmentation masks of the falling object in the i-th frame of the generated and ground truth videos respectively.
其中 $M_i^{\text{gen}}, M_i^{\text{gt}} \in \{0,1\}^{H\times W}$ 分别是生成视频和地面真实视频的第 i 帧中下落物体的二元分割蒙版。

Time error. When testing on videos generated in simulation, we can provide a timing error. From the dropping height Y_0 of the ground truth video, which we have access to from the simulator, we can derive $t_{\text{drop}} = \sqrt{Y_0 \frac{2}{g}}$. We then obtain a dropping time from the model's output by estimating the frame of impact as the first frame F whose centroid velocity in the y direction is negative. If t_{drop} occurs in between F and $F-1$, then we define the time error E_{time} as zero. Otherwise, we define the time error as
时间误差。在模拟生成的视频上进行测试时，我们可以提供时间误差。从我们可以从模拟器访问的地面真实视频的下落高度 Y_0，我们可以得出 $t_{\text{drop}} = \sqrt{Y_0 \frac{2}{g}}$。然后，我们通过将撞击帧估计为第一个帧 F（其质心速度在 y 方向上为负）来从模型的输出中获得下落时间。如果 t_{drop} 发生在 F 和 $F-1$ 之间，则我们将时间误差 E_{time} 定义为零。否则，我们将时间误差定义为

$$E_{\text{time}} = \min\left(\left|\frac{F-1}{\text{fps}} - t_{\text{drop}}\right|, \left|\frac{F}{\text{fps}} - t_{\text{drop}}\right|\right). \tag{8}$$

C. ORO implementation details.

In our setting, we do not cut the gradient after step k like VADER. The gradient $\nabla_\theta R(x_0', x_0)$ backpropagates through all diffusion timesteps and update the model weights θ:
在我们的设置中，我们不会像 VADER 那样在步骤 k 之后切断梯度。梯度 $\nabla_\theta R(x_0', x_0)$ 通过所有扩散时间步骤反向传播并更新模型权重 θ：

$$\nabla_\theta(R(x_0', x_0)) = \sum_{t=0}^{T}\frac{\partial R(x_0', x_0)}{\partial x_t}\cdot\frac{\partial x_t}{\partial \theta} \tag{9}$$

where T is the total diffusion timesteps.
其中 T 是总扩散时间步长。

Segmentation Reward. We utilize SAM 2 (?) to generate segmentation masks across frames for generated video:

$$M^{\text{gen}} = \text{SAM-2}(x_0) \tag{10}$$

where M^{gen} denotes the masks of the falling object in the generated video. We obtain ground truth masks M^{gt} using Kubric (?). To avoid non-differentiable reward, we use Sigmoid to normalize mask logits of generated video instead of converting them to binary masks. We use IoU between M^{gen} and M^{gt} as reward function
其中 M^{gen} 表示生成视频中下落物体的掩码。我们使用 Kubric (?) 获得真实掩码 M^{gt}。为了避免不可微分的奖励，我们使用 Sigmoid 来规范化生成视频的掩码对数，而不是将它们转换为二进制掩码。我们使用 M^{gen} 和 M^{gt} 之间的 IoU 作为奖励函数：

$$R(x_0', x_0) = \text{IoU}(M^{\text{gen}}, M^{\text{gt}}) \tag{11}$$

Maximizing objective 2 is equivalent to minimizing the following objective:

$$J(\theta) = \mathbb{E}_{(x_0,c)\sim\mathcal{D}, x_0'\sim p_\theta(x_0'|c)}[1 - \text{IoU}(M^{\text{gen}}, M^{\text{gt}})] \tag{12}$$

This objective constrains the position and shape of the generated object in the video, encouraging a greater intersection with the object region in the ground truth video. The model learns to generate more accurate object

positions and shapes through training with this objective.
此目标限制了视频中生成物体的位置和形状，从而鼓励与地面真实视频中的物体区域有更大的交集。通过使用此目标进行训练，模型可以学习生成更准确的物体位置和形状。

Optical Flow Reward. We utilize RAFT (?) to generate optical flow for both generated videos and ground truth:

$$\begin{aligned} V^{\text{gen}} &= \text{RAFT}(x_0') \\ V^{\text{gt}} &= \text{RAFT}(x_0) \end{aligned} \tag{13}$$

where $V^{\text{gen}}, V^{\text{gt}}$ denote the optical flows of generated videos and ground truth. We define the reward as follows:

$$R(x_0', x_0) = -|V^{\text{gen}} - V^{\text{gt}}| \tag{14}$$

Maximizing objective 2 is equivalent to minimizing the following objective:

$$J(\theta) = \mathbb{E}_{(x_0,c)\sim\mathcal{D}, x_0'\sim p_\theta(x_0'|c)}[|V^{\text{gen}} - V^{\text{gt}}|] \tag{15}$$

This objective constrains the motion of the generated object in the video. The model learns to generate more accurate physical motion through training with this objective.

Depth Reward. We utilize Depth-Anything-V2 (?) to generate optical depth maps for both generated videos and ground truth:

$$\begin{aligned} D^{\text{gen}} &= \text{Depth-Anything-V2}(x_0') \\ D^{\text{gt}} &= \text{Depth-Anything-V2}(x_0) \end{aligned} \tag{16}$$

where $D^{\text{gen}}, D^{\text{gt}}$ denote the depth maps of generated videos and ground truth. We define the reward as follows:

$$R(x_0', x_0) = -|D^{\text{gen}} - D^{\text{gt}}| \tag{17}$$

Maximizing objective 2 is equivalent to minimizing the following objective:

$$J(\theta) = \mathbb{E}_{(x_0,c)\sim\mathcal{D}, x_0'\sim p_\theta(x_0'|c)}[|D^{\text{gen}} - D^{\text{gt}}|] \tag{18}$$

This objective constrains the 3d motion of the generated object in the video. The model learns to generate more accurate 3d physical motion through training with this objective.

D. Coordinate system

We give a visualization of the coordinate system used in this paper in Figure 9. To compute y, we first leverage a segmentation map and find pixel row index that is just below the object. Once this row index is found, y can easily be computed from the camera position, camera sensor size, and image resolution. We note that because our camera is assumed to be in perspective with the XY plane, we can ignore X and x (not shown in figure) in our analyses in Section 5.1 and Section 5.2.
我们在 Figure 9 中给出了本文中使用的坐标系的可视化。为了计算 y，我们首先利用分割图并找到位于物体正下方的像素行索引。一旦找到此行索引，就可以轻松地根据相机位置、相机传感器尺寸和图像分辨率计算出 y。我们注意到，由于我们的相机被假定与 XY 平面成透视关系，因此我们可以在 Section 5.1 和 Section 5.2 中的分析中忽略 X 和 x（图中未显示）。

E. Derivation of $p(t|y)$

In our dataset construction, we assume a uniform distribution for Z, where $Z \sim \mathcal{U}(Z_{\min}, Z_{\max})$, where $Z_{\min} = 2$ and $Z_{\max} = 18$. As shown in Figure 9, the dropping height Y is a linear function of Z, i.e. $Y = y + \beta Z$ for the slope β that can be computed from y, f, the sensor size, and the camera height. This means we can solve for dropping time as $t = \sqrt{\frac{2}{g}Y} = \sqrt{\frac{2}{g}(y + \beta Z)}$. Applying the transformation rule for probability density yields
在我们的数据集构建中，我们假设 Z 服从均匀分布，其中 $Z \sim \mathcal{U}(Z_{\min}, Z_{\max})$，其中 $Z_{\min} = 2$ 且 $Z_{\max} = 18$。

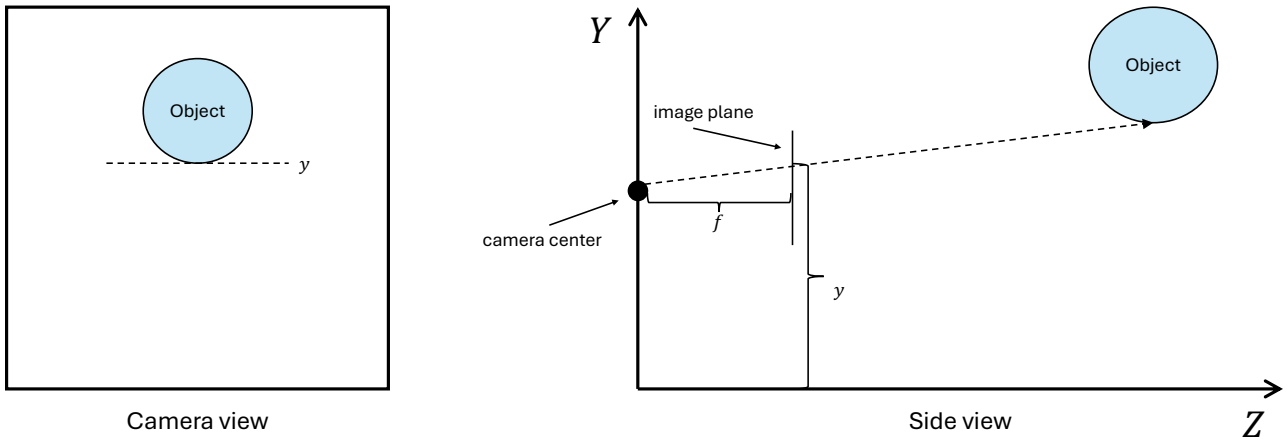


Figure 9. A visualization of the coordinate system used in this paper (not to scale). The image plane height of the object is denoted as y, its actual height in 3D as Y, and its depth as Z. The camera focal length is denoted as f.

如 Figure 9 所示，下落高度 Y 是 Z 的线性函数，即 $Y = y + \beta Z$，斜率 β 可根据 y、f、传感器尺寸和相机高度计算得出。这意味着我们可以求解下落时间，即 $t = \sqrt{\frac{2}{g}Y} = \sqrt{\frac{2}{g}(y + \beta Z)}$。应用概率密度变换规则可得出

$$p(t|y) = \begin{cases} \frac{gt}{(Z_{\max} - Z_{\min})\beta}, & t_{\min} \le t \le t_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

where $t_{\min} = \sqrt{\frac{2}{g}(y + \beta Z_{\min})}$ and $t_{\max} = \sqrt{\frac{2}{g}(y + \beta Z_{\max})}$. Plugging in $Z_{\min} = 2$ and $Z_{\max} = 18$ yields Equation (4).

F. Ambiguous dataset

We introduce a new dataset for distributional analysis that broadens $p(t|y)$, in contrast to the PSFT dataset, which prioritizes realism and has a narrower distribution due to limited object depth variability. To create a dataset with $p(t|y)$ that is sufficiently diverse for meaningful analysis, we first set up the initial scenes as before, but then apply an augmentation where a new depth values is sampled uniformly from $[2, 18]$ and the object is scaled and translated such that it appears the same in the original image, as shown in Figure 6. For simplicity, we limit our scenes to a single dropping object with no other objects on the ground. We also disable shadows, preventing the model from using them as cues to infer depth and height. Our dataset contains 5k samples consisting of 1k unique initial scenes each containing 5 different trajectories produced by the augmentation.
我们引入了一个新的数据集用于分布分析，它拓宽了 $p(t|y)$，与 PSFT 数据集不同，PSFT 数据集优先考虑真实性，由于对象深度变化有限，分布较窄。为了创建一个具有足够多样性的 $p(t|y)$ 数据集以进行有意义的分析，我们首先像以前一样设置初始场景，然后应用增强，从 $[2, 18]$ 均匀采样新的深度值，并缩放和平移对象，使其在原始图像中看起来相同，如 Figure 6 所示。为简单起见，我们将场景限制为单个掉落物体，地面上没有其他物体。我们还禁用阴影，以防止模型使用它们作为推断深度和高度的线索。我们的数据集包含 5k 个样本，由 1k 个独特的初始场景组成，每个场景包含由增强生成的 5 条不同轨迹。

G. Lifting trajectories to 3D

To lift trajectories to 3D, we first estimate t_{drop} as described in Section 5.1. Using SAM2 to estimate object masks in the generated video, we can obtain a trajectory of the bottom of the object which we denote as $y_0, y_1, \ldots, y_N$ where $N = t_{\text{drop}} \times \text{fps}$. From t_{drop}, we can solve for an implied depth $Z = \frac{\frac{1}{2}gt^2 - y}{\beta}$. We then compute the lifted 3D trajectory as $y_i \mapsto y_i + \beta Z$
将轨迹提升到 3D。为了将轨迹提升到 3D，我们首先如 Section 5.1 中所述估计 t_{drop}。使用 SAM2 估计生成视频中的物体掩码，我们可以获得物体底部的轨迹，记为 $y_0, y_1, \ldots, y_N$，其中 $N = t_{\text{drop}} \times \text{fps}$。从 t_{drop}，我们可

以求解隐含深度 $Z = \frac{\frac{1}{2}gt^2 - y}{\beta}$。然后我们计算提升后的 3D 轨迹为 $y_i \mapsto y_i + \beta Z$。

H. PisaBench Details

In this section, we discuss the details of our data collection pipeline and annotations. We present more examples of real-world videos and corresponding annotations in Figure 10.
PisaBench 细节。在本节中，我们讨论了数据收集流程和注释的细节。我们在 Figure 10 中展示了更多真实世界视频及其相应注释的示例。

H.1. Data Collection Pipeline

Collecting Real World Videos. We enlist approximately 15 volunteers to participate in the data collection process. We hand out a tripod, tape, and invisible wire for each volunteer. To ensure the quality, diversity, and minimize the ambiguity introduced by the environments, volunteers are provided with detailed guidelines. The key points of the data collection guidelines are shown in Table 3.
数据收集流程。我们招募了大约 15 名志愿者参与数据收集过程。我们为每位志愿者分发了一个三脚架、胶带和隐形线。为了确保质量、多样性并减少环境引入的歧义，志愿者被提供了详细的指南。数据收集指南的关键点如 Table 3 所示。

Raw videos processing. For the collected raw videos, we cut each video into multiple clips and crop their sizes. For each video clip, we annotate its starting position in the original long video and ensure that the duration of each segment does not exceed 12 seconds. Regarding the sizes of the videos, we manually crop each video to an aspect ratio of 1 : 1, ensuring that the falling objects remain fully visible within the frame during the cropping process. The processing interface is shown in Figure 11.
原始视频处理。对于收集到的原始视频，我们将每个视频切割成多个片段并裁剪其大小。对于每个视频片段，我们注释其在原始长视频中的起始位置，并确保每个片段的持续时间不超过 12 秒。关于视频的大小，我们手动将每个视频裁剪为 1:1 的宽高比，确保在裁剪过程中下落物体在帧内完全可见。处理界面如 Figure 11 所示。

H.2. Annotation Details

We present our annotation details Figure 12. For video captions, we present the word cloud figure in (a). For segmentation masks, we annotate all objects in the first frame using positive and negative points, which are then propagated across frames using the SAM 2 (?) model to produce segmentation masks for all objects throughout the video. The annotation interface is shown in (b).
注释细节。我们展示了注释细节 Figure 12。对于视频字幕，我们在 (a) 中展示了词云图。对于分割掩码，我们使用正负点在第一帧中注释所有物体，然后使用 SAM 2 (?) 模型在帧之间传播，以生成整个视频中所有物体的分割掩码。注释界面如 (b) 所示。

In addition to providing the annotated caption "{object description} falls." , we also add information to inform off-the-shelf models of the task's context as much as possible. To further enhance task comprehension, we append an additional description "A video that conforms to the laws of physics." We also employ negative prompts "no camera motion" and "no slow-motion" to ensure environmental stability and impose constraints on the generated videos. These prompts explicitly instruct the models to avoid including camera motion or any non-real-time object motion, thereby maintaining consistency with real-world physics.
除了提供注释的字幕 "{物体描述} 下落。"，我们还添加了尽可能多的信息，以便现成模型了解任务的上下文。为了进一步增强任务理解，我们附加了一个额外的描述 "符合物理定律的视频。" 我们还使用负向提示 "无相机运动" 和 "无慢动作" 来确保环境稳定性并对生成的视频施加约束。这些提示明确指示模型避免包含相机运动或任何非实时的物体运动，从而保持与现实世界物理的一致性。

I. Inference Details

We present the inference configurations of each closed or open model we evaluate in Table 4. For models that do not support generating videos with 1:1 aspect ratio, we pad initial frames with black borders to the resolution supported by these models, and finally remove the black borders from the generated videos.
推理细节。我们在 Table 4 中展示了我们评估的每个封闭或开放模型的推理配置。对于不支持生成 1:1 宽高比视

频的模型，我们用黑色边框填充初始帧到这些模型支持的分辨率，最后从生成的视频中移除黑色边框。

J. More Qualitative Examples

We present more qualitative examples in Figure 13 - Figure 19. Although in some showcases, models can roughly predict the downward trend, models still struggle to predict plausible shape and motion. The defects in the models can be mainly attributed to the following aspects:

- Trajectory correctness: in most videos, models fail to predict even the basic falling trajectory of objects, as shown in Figure 16 (a), despite this being highly intuitive for humans. Even in cases where the falling trajectory is roughly correctly predicted, the models still struggle to accurately predict subsequent events, such as collisions, as illustrated in Figure 13 (f).
- Object consistency: in many generated videos, object consistency is poor. Models struggle to infer the appearance of objects from multiple viewpoints in a physically plausible manner, resulting in unnatural appearances, as shown in Figure 13 (a). Additionally, models perform poorly in maintaining object permanence, causing objects to appear blurry, as illustrated in Figure 17 (f). Furthermore, models sometimes introduce new objects into the video, as depicted in Figure 17 (e).
- Scene consistency: models struggle to maintain scene consistency, leading to abrupt transitions in many videos. These sudden changes make videos appear unnatural, as shown in Figure 15 (f).

更多定性示例。我们在 Figure 13 - Figure 19 中展示了更多定性示例。尽管在某些展示中，模型可以大致预测下降趋势，但模型仍然难以预测合理的形状和运动。模型的缺陷主要归因于以下几个方面：

- 轨迹正确性：在大多数视频中，模型甚至无法预测物体的基本下落轨迹，如 Figure 16 (a) 所示，尽管这对人类来说非常直观。即使在下落轨迹大致正确预测的情况下，模型仍然难以准确预测后续事件，如碰撞，如 Figure 13 (f) 所示。
- 物体一致性：在许多生成的视频中，物体一致性较差。模型难以从多个视角以物理上合理的方式推断物体的外观，导致不自然的外观，如 Figure 13 (a) 所示。此外，模型在保持物体持久性方面表现不佳，导致物体看起来模糊，如 Figure 17 (f) 所示。此外，模型有时会在视频中引入新物体，如 Figure 17 (e) 所示。
- 场景一致性：模型难以保持场景一致性，导致许多视频中出现突然的过渡。这些突然的变化使视频看起来不自然，如 Figure 15 (f) 所示。

K. Simulated Adaption Details

We use the Kubric (?) simulation and rendering engine for creating our simulated videos. Kubric uses PyBullet (?) for running physics simulations and Blender (?) for rendering. We set the simulation rate to 240 steps per second and render 2-second videos at 16 fps, resulting in 32 frames per video. Each scene consists of objects from the Google Scanned Objects (GSO) dataset (?) and uses environmental lighting from HDRI maps provided by Kubric. We use 930 objects and 458 HDRI maps for training and 103 objects and 51 HDRI maps for testing.
模拟适应细节。我们使用 Kubric (?) 模拟和渲染引擎来创建我们的模拟视频。Kubric 使用 PyBullet (?) 进行物理模拟，并使用 Blender (?) 进行渲染。我们将模拟速率设置为每秒 240 步，并以 16 fps 渲染 2 秒的视频，每个视频包含 32 帧。每个场景由来自 Google Scanned Objects (GSO) 数据集 (?) 的物体组成，并使用 Kubric 提供的 HDRI 地图进行环境光照。我们使用 930 个物体和 458 张 HDRI 地图进行训练，使用 103 个物体和 51 张 HDRI 地图进行测试。

For each video, we randomly choose 1-6 objects to drop. These objects are placed at a height uniformly sampled from 0.5m to 1.5m. Below each of these objects, a possibly empty pile of up to 4 objects spawns beneath to create collisions. The objects are placed in a spawn region of size 2m × 2m.
对于每个视频，我们随机选择 1-6 个物体进行下落。这些物体被放置在从 0.5m 到 1.5m 均匀采样的高度。在每个物体的下方，可能会生成最多 4 个物体的空堆以创建碰撞。物体被放置在大小为 2m × 2m 的生成区域中。

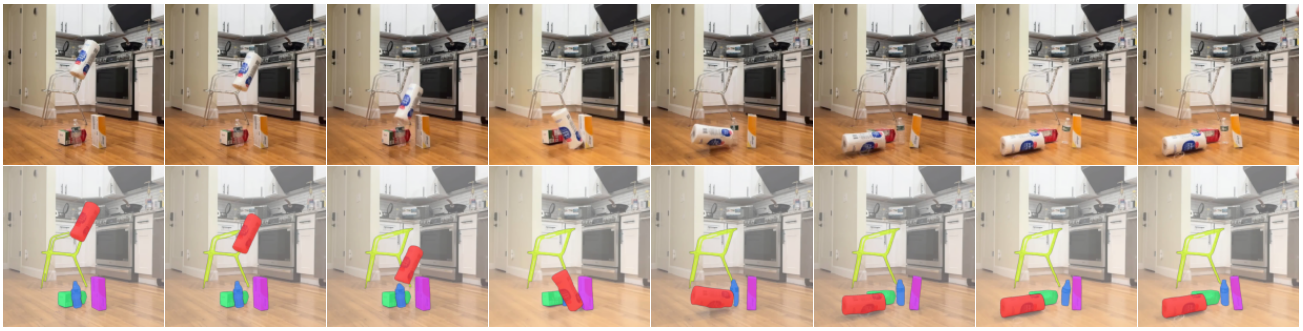
The camera is initially positioned 1m behind this region, with its height varying uniformly between 0.4m and 0.6m. Once all objects are placed, the camera moves back in random increments until all objects are visible

within the camera frame. The camera uses a focal length of 35 mm, a sensor width of 32mm, and an aspect ratio of 1×1.
相机最初位于该区域后方 1m 处，其高度在 0.4m 到 0.6m 之间均匀变化。一旦所有物体放置完毕，相机会以随机增量向后移动，直到所有物体在相机帧内可见。相机使用 35 mm 的焦距，32mm 的传感器宽度，以及 1×1 的宽高比。

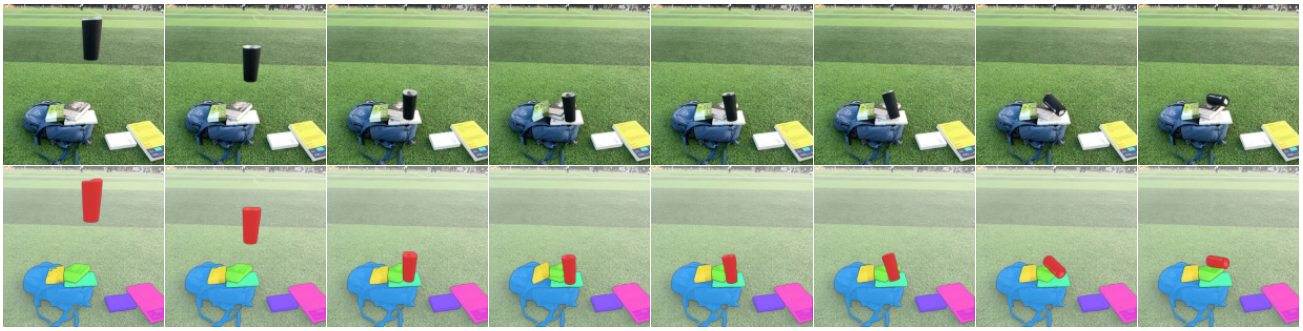
L. Limitations

In this work, we collect and manually annotate a dataset of 361 real-world videos and design three spatial metrics to evaluate the performance of state-of-the-art image-to-video (I2V) models in a fundamental physical scenario: free fall. Our metrics focus solely on spatial positional relationships, excluding object appearance attributes such as color. To enable more fine-grained evaluations of appearance characteristics, we aim to develop metrics based on Multimodal Large Language Models (MLLMs) or pixel-level analysis in future work.
局限性。在这项工作中，我们收集并手动注释了 361 个真实世界视频的数据集，并设计了三个空间度量来评估最先进的图像到视频（I2V）模型在基本物理场景（自由落体）中的表现。我们的度量仅关注空间位置关系，排除了物体外观属性（如颜色）。为了实现对外观特征的更细粒度评估，我们计划在未来工作中开发基于多模态大语言模型（MLLMs）或像素级分析的度量。

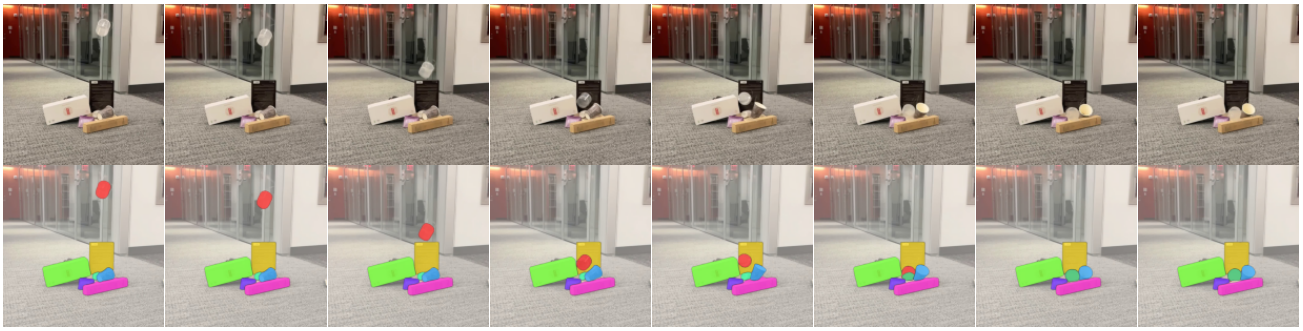
Furthermore, we propose the PSFT and ORO methods to fine-tune the Open-Sora model (?), improving its ability to generate physically plausible videos. Despite these improvements, certain limitations remain, specifically, the generation of blurry objects in some videos. We hope to address these challenges in future research by refining both the dataset and the fine-tuning strategies, aiming to produce videos that better maintain object visuals.
此外，我们提出了 PSFT 和 ORO 方法来微调 Open-Sora 模型 (?)，提高其生成物理上合理视频的能力。尽管有这些改进，仍然存在一些局限性，特别是在某些视频中生成了模糊的物体。我们希望通过在未来研究中改进数据集和微调策略来解决这些挑战，旨在生成更好地保持物体视觉效果的视频。



(a) A white paper roll falls.



(c) A black bottle falls.



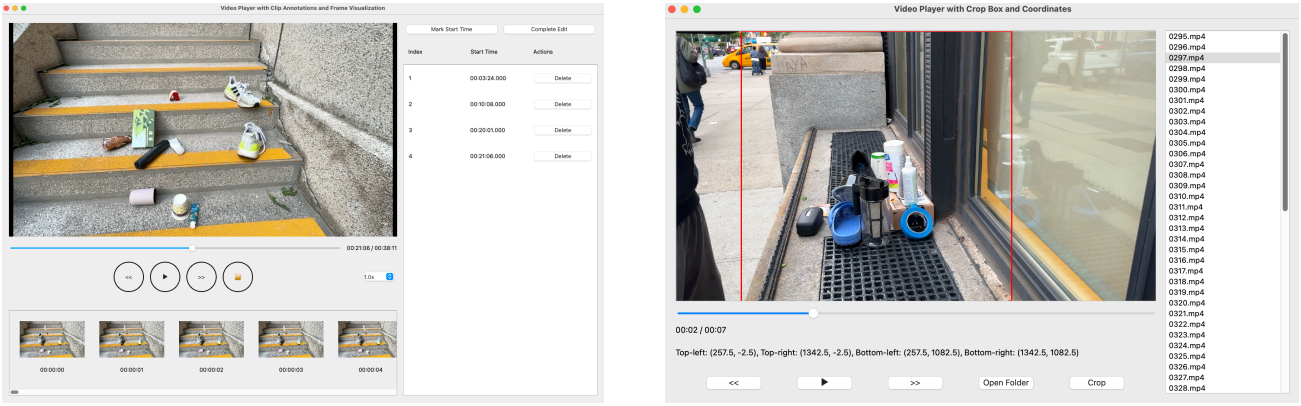
(b) A transparent bottle falls.

(d) A white bottle falls.

Figure 10. Examples of real world videos and annotations. We present video frames in the first row and mask annotations in the second row.

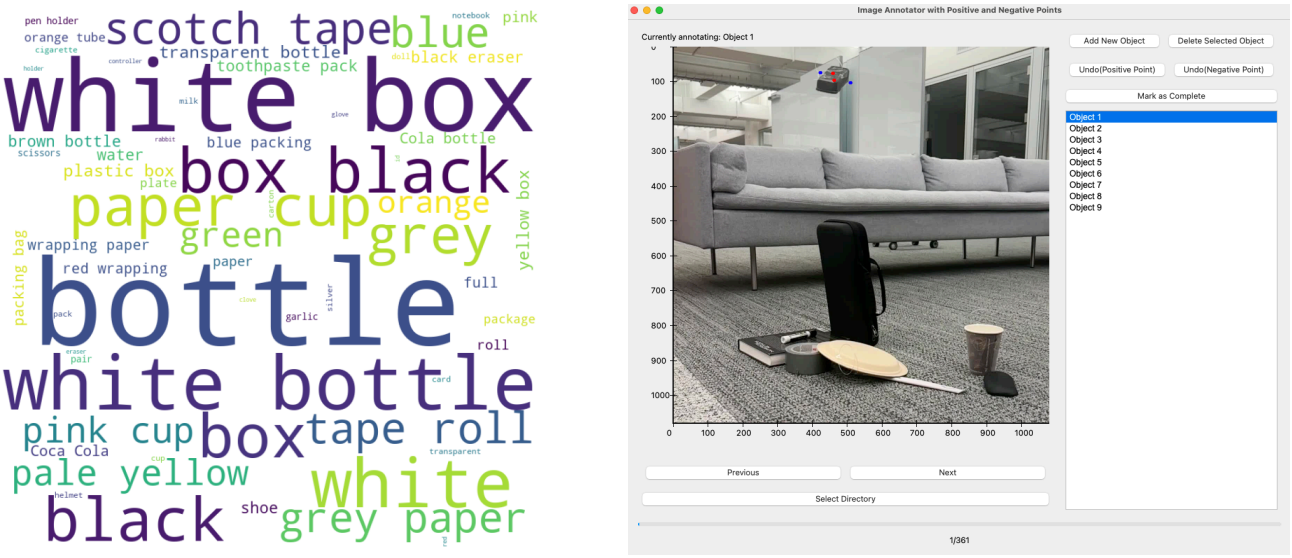
Aspect	Requirements
Camera	• The camera must be stabilized using a tripod. • The dropping object should remain visible throughout the entire fall. • The trajectory of the object should be sufficiently centered in the frame. • Ensure the slow-motion setting is configured to 120 fps. • Avoid a completely top-down perspective; the frame should include both the floor and the wall for spatial context. • It is acceptable to record one long video containing multiple drops at the same location.
Objects	• Most objects should be rigid and non-deformable. • A limited number of flexible or deformable objects may be included, as such data is also valuable.
Dropping Procedure	• Secure the object with a wire using tape, ensuring stability. Multiple tapings may be necessary for proper stabilization. • Visibility of the wire in the video is acceptable. • No body parts should appear in the frame. If this is challenging, consider having a partner monitor the camera or use screen-sharing software to view the camera feed on a laptop for uninterrupted framing. • Record videos in a horizontal orientation to simplify cropping and to help keep the frame free of unnecessary elements. • Use a short wire to enhance object stability. • The object should remain stationary before being dropped.
Scene Composition	• Make the scenes dynamic and engaging. Include interactions with other objects, such as collisions or objects tipping over. Static objects should serve as active elements rather than mere background props. • Avoid filming in classroom or laboratory environments. • Include a variety of dropping heights. • Film in different environments, ensuring at least one setting is outside your apartment. • Minimize human shadows in the frame whenever possible. • Ensure good lighting and maintain strong contrast between the objects and the background.

Table 3. Key points of real world videos collection guideline. We have detailed requirements for camera, objects, dropping procedure and scene composition to ensure the quality, diversity and minimize ambiguity introduced by environments.



(a) (b)

Figure 11. Video processing interface. (a) we annotate starting positions in the original long videos and clip them into multiple clips less than 12 seconds. (b) We drag the cropping box to crop the video size to an aspect ratio of 1:1.

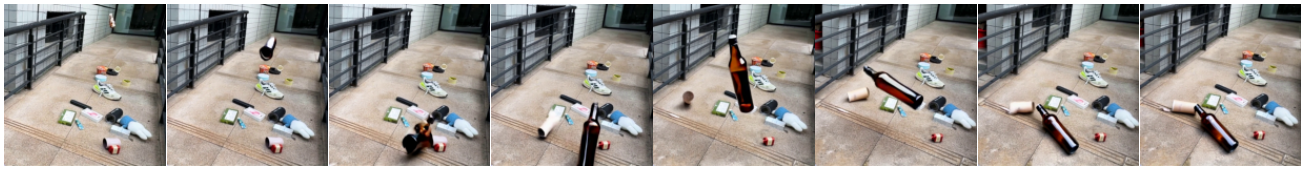


(a) (b)

Figure 12. Annotation details of real world videos. (a) Word cloud of objects in video captions. Our videos contain a variety of daily life objects. (b) Interface for annotating positive and negative points in the first frame. Red and blue dots indicate positive and negative points respectively. We annotate all objects in the midair and ground.

	Model	Resolution	Number of Frames	FPS	Guidance Scale	Sampling Steps	Noise Scheduler
Closed	Sora	720 × 720	150	30	-	-	-
	Kling-V1.5	960 × 960	150	30	1.0	-	-
	Kling-V1	960 × 960	150	30	1.0	-	-
	Runway Gen3	1280 × 768	156	30	-	-	-
Open	CogVideoX-5B-I2V	720 × 480	48	8	6.0	50	DDIM
	DynamiCrafter	512 × 320	90	30	0.7	50	DDIM
	Pyramid-Flow	1280 × 768	120	24	4.0	10	EulerDiscrete
	Open-Sora	512 × 512	90	30	7.0	30	RFLOW

Table 4. Inference details for models we evaluate, where "-" indicates the information is not available.



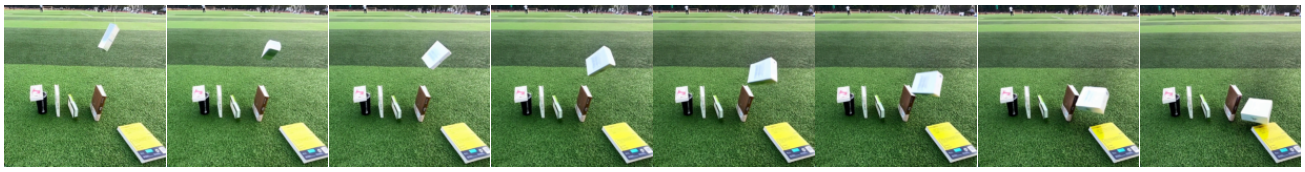
(a) A brown bottle falls.

(b) A grey bottle falls.

(c) A grey paper cup falls.

(d) A paper cup falls.

(e) A white bottle falls.



(f) A white box falls.

Figure 13. Qualitative examples of Kling-V1 (?). In (a) (b) (c) (f), objects have a tendency to fall. (b) (c) are roughly consistent with the laws of physics. In (a) (f), the shape of the object does not match the first frame. In (d), the paper cup is suspended in midair. In (e), new object is introduced. In (e), the model fails to correctly predict the collision that occurs after the white box falls and the chain of events that follows.

(a) A black and grey glove falls.

(b) A black bottle falls.

(c) A blue and white box falls.

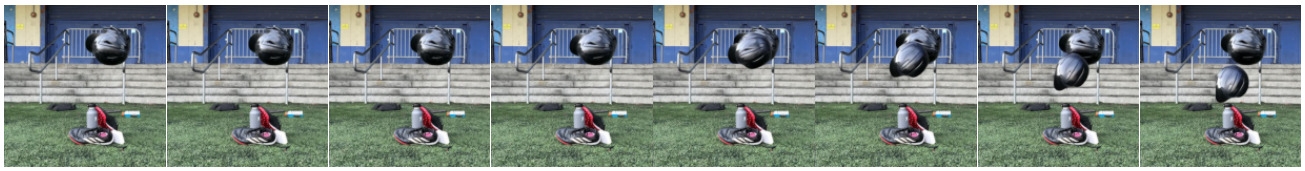
(d) A brown bottle falls.

(e) A Coca–Cola can falls.

(f) A pink box falls.

Figure 14. Qualitative examples of Runway Gen3 (?). In (b) (e), objects have a tendency to fall. In (a) (e) (f), new objects are introduced. In (b) (d), the shape of the object does not match the first frame. In (c), the box is suspended in midair.

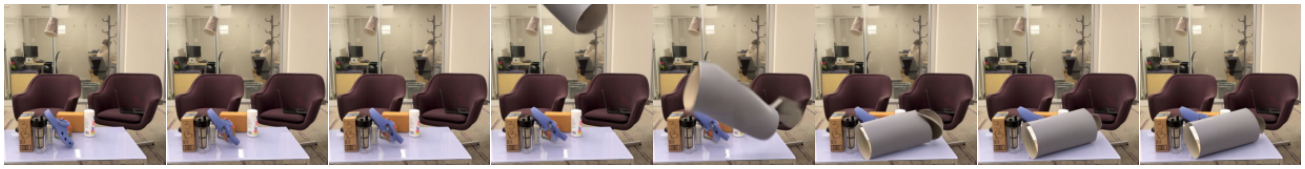
(a) A black bottle falls.



(b) A black helmet falls.

(c) A paper box falls.

(d) A white bottle falls.



(e) A grey paper cup falls.

(f) A white box falls.

Figure 15. Qualitative examples of CogVideoX-5B-I2V (?). In (a) - (f), objects have a tendency to fall. However, in all the videos, there are violations of physics. In (a) (b), the objects are divided into two parts. In (c) (d) (e), the shape of the object does not match the first frame. In (c), the trajectory is not a vertical fall. In (f), scene changes suddenly, which does not match the first frame.

(a) A black box falls.

(b) A card holder falls.

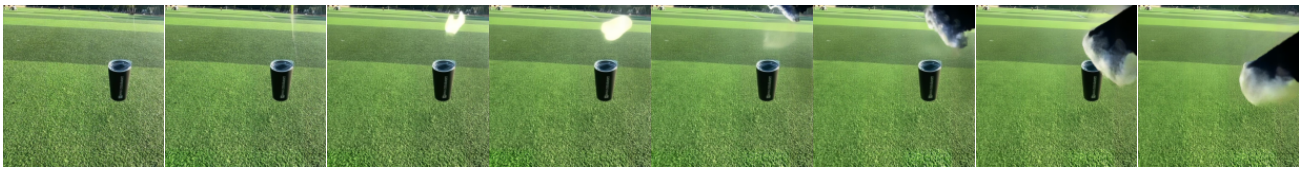
(c) A white bottle falls.

(d) A white box falls.

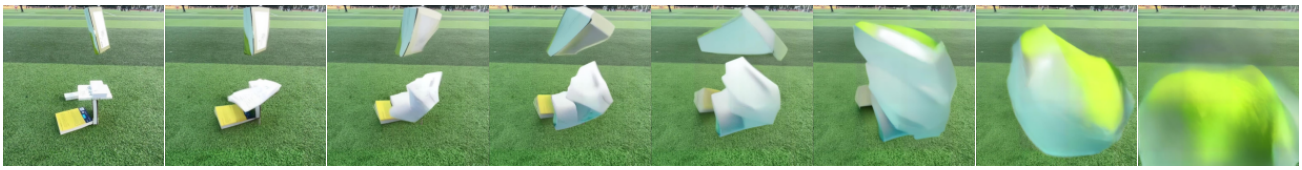
(e) An orange and white box falls.

(f) A shoe falls.

Figure 16. Qualitative examples of DynamiCrafter (?). In all the videos, objects do not have a tendency to fall, suspended in the midair.

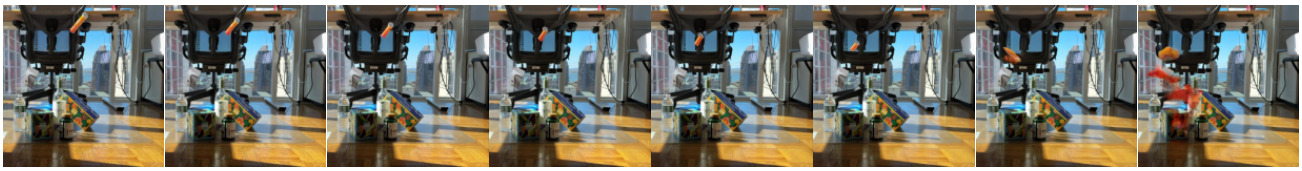


(a) A black bottle falls.



(b) A green and white box falls.

(c) A grey bottle falls.

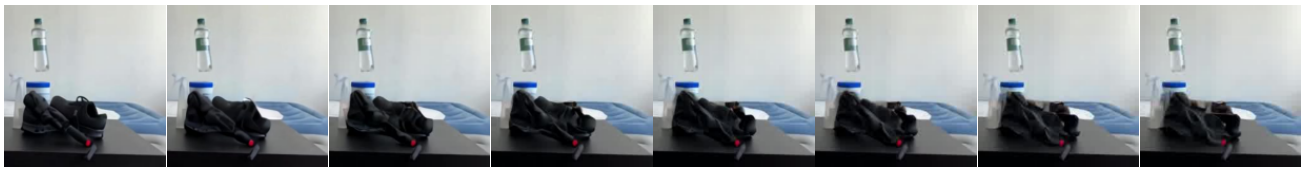


(d) An orange tube falls.

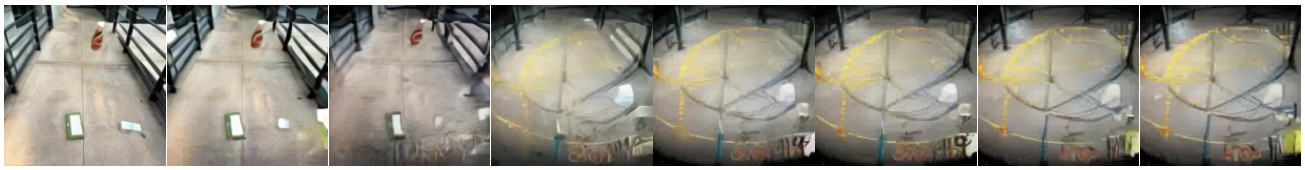
(e) A white bottle falls.

(f) A plastic box falls.

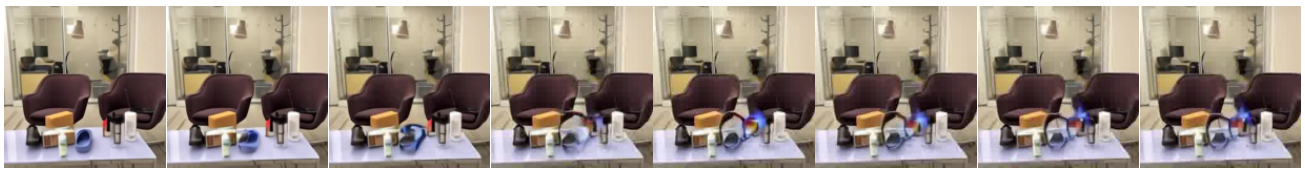
Figure 17. Qualitative examples of Pyramid-Flow (?). In (b) (d) (e), objects have a tendency to fall. In (a) (b) (e) (f), new objects are introduced. In (c), scene changes, which does not match the first frame.. In (d), the tube becomes blurry.



(a) A bottle full of water falls.



(b) A brown bottle falls.



(c) A grey paper cup falls.

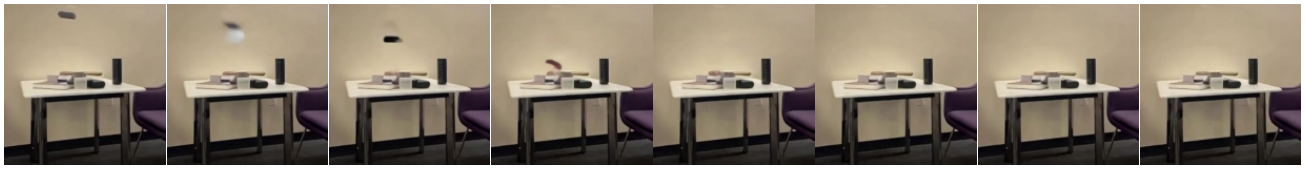
(d) A paper box falls.

(e) A white bottle falls.

(f) A white box falls.

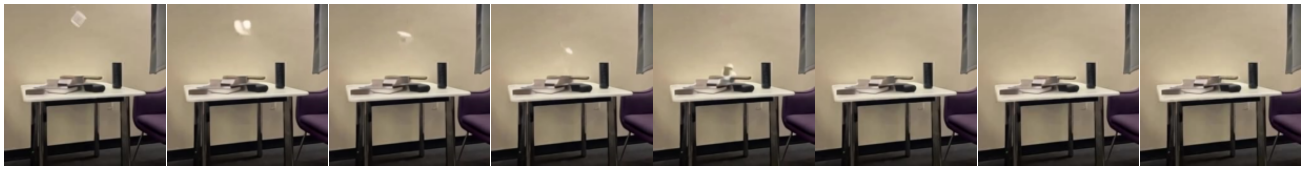
Figure 18. Qualitative examples of Open-Sora (?). In all the videos, objects do not have a tendency to fall, suspended in the midair. In (b) (d), scene changes suddenly, which does not match the first frame. In (e), new object is introduced.

(a) A brown bottle falls.



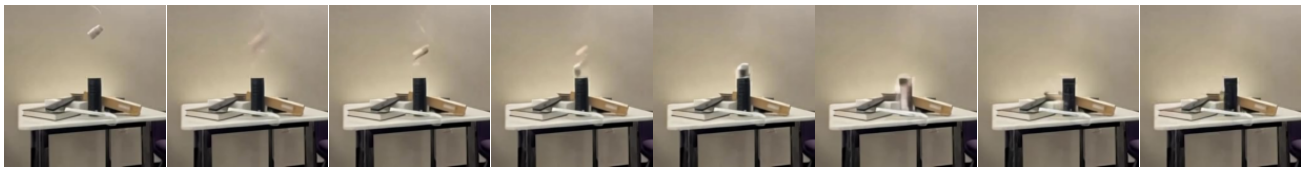
(b) A grey eraser falls.

(c) A grey paper cup falls.



(d) A transparent bottle falls.

(e) A red wrapping paper falls.



(f) A white bottle falls.

Figure 19. Qualitative examples of our method (Open-Sora + PSFT + ORO). In all the videos, objects have a tendency to fall. However, the consistency of objects is still insufficient. In some frames, objects become blurry. Objects sometimes disappear after collision.