MORPHEUS: BENCHMARKING PHYSICAL REASONING OF VIDEO GENERATIVE MODELS WITH REAL PHYSICAL EXPERIMENTS

Chenyu Zhang[1,2,*], **Daniil Cherniavskii**[1,*], **Antonios Tragoudaras**[1,*],
Antonios Vozikis[1,†], **Thijmen Nijdam**[1,†],
Derck W. E. Prinzhorn[1], **Mark Bodracska**[1], **Nicu Sebe**[2],
Andrii Zadaianchuk[1,‡], **Stratis Gavves**[1,3,‡]

University of Amsterdam, the Netherlands[1]
University of Trento, Italy[2]
Archimedes, Athena Research Center, Greece[3]
chenyu.zhang@unitn.it, antonios.tragoudaras@student.uva.nl
{d.cherniavskii, a.zadaianchuk, e.gavves}@uva.nl

ABSTRACT

Recent advances in image and video generation raise hopes that these models possess world modeling capabilities—the ability to generate realistic, physically plausible videos. This could revolutionize applications in robotics, autonomous driving, and scientific simulation. However, before treating these models as world models, we must ask: Do they adhere to physical laws? Current evaluation methods rely on subjective judgments or trajectory matching, limiting their usage for physical reasoning estimation, where many generations could be physically plausible. Thus, we introduce **Morpheus**, a new benchmark for evaluating video generation models on physical reasoning. It features 130 real-world videos capturing physical phenomena, guided by conservation laws. Using those as conditioning for video generation, we assess physical plausibility using physics-informed metrics evaluated with respect to infallible conservation laws known per physical setting, leveraging advances in physics-informed neural networks and vision-language foundation models. Our findings reveal that even with advanced prompting and video conditioning, current models struggle to encode physical principles despite generating aesthetically pleasing videos. All data, leaderboard, and code are open-sourced at: https://physics-from-video.github.io/morpheus-bench/

1 INTRODUCTION

Video generative models (VGMs) such as SORA (Brooks et al., 2024a), COSMOS (Agarwal et al., 2025), and Veo3 (Veo-Team et al., 2024) have taken the world by storm, building upon remarkable advances in image generative models (Ramesh et al., 2021; Saharia et al., 2022; Podell et al., 2023; Yu et al., 2023), and achieving unprecedented levels of visual fidelity and realism. These developments have not only pushed the boundaries of visual aesthetics but have also inspired the community to envision video generative models as potential *world models* (Cho et al., 2024; Agarwal et al., 2025). A world model, in this context, is more than just a system for generating frames, however; it is a model capable of understanding and predicting the dynamics, causal interactions, and underlying mechanisms of the physical world. Accurately benchmarking the physical dynamics of video generation is a critical requirement —and the focus of this work— toward adopting them potentially as world models. Perhaps the most daring challenge from an AI perspective is the need for physical dynamics evaluation that goes beyond visual verification. Current methods use either

Signs ∗, †, ‡ denote equal first, second, and last author contributions correspondingly.

MORPHEUS：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

张晨宇，丹尼尔·切尔尼亚夫斯基，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·博达奇卡，尼库·塞贝，
安德烈·扎达亚恩丘克，斯特拉提斯·加夫斯

荷兰阿姆斯特丹大学，意大利特伦托大学，希
腊阿基米德，雅典研究中心

chenyu.zhang@unitn.it, antonios.tragoudaras@student.uva.nl {d.cherniavskii, a.zadaianchuk, e.gavves}@uva.nl

摘要

近年来图像和视频生成技术的进步让人期待这些模型具备世界建模能力——即生成逼真且符合物理规律的视频。这可能会彻底改变机器人、自动驾驶和科学模拟等领域的应用。然而，在将这些模型视为世界模型之前，我们必须问：它们是否遵循物理定律？当前的评估方法依赖于主观判断或轨迹匹配，这限制了它们在物理推理评估中的应用，因为在物理推理评估中，许多生成结果可能是符合物理规律的。因此，我们引入了 Morpheus，一个用于评估视频生成模型物理推理能力的新基准。它包含 130 个捕捉物理现象的真实世界视频，并遵循守恒定律。利用这些视频作为生成条件，我们通过物理信息指标评估物理合理性，这些指标基于每个物理场景中已知的、不可谬的守恒定律进行评估，并借助物理信息神经网络和视觉语言基础模型的进步。 我们的研究发现，即使使用先进的提示和视频条件，当前模型在生成美观视频的同时，仍然难以编码物理原理。 所有数据、排行榜，以及代码均在以下地址开源：

https://physics-from-video.github.io/morpheus-bench/

1 引言

视频生成模型（VGMs）如 SORA（Brooks 等人，2024a）、COSMOS（Agarwal 等人，2025）和 Veo3（Veo-Team 等人，2024）风靡全球，它们基于图像生成模型的显著进步（Ramesh 等人，2021；Saharia 等人，2022；Podell 等人，2023；Yu 等人，2023）构建，实现了前所未有的视觉保真度和逼真度。这些发展不仅拓展了视觉美学的边界，还激励社区将视频生成模型视为潜在的世界模型（Cho 等人，2024；Agarwal 等人，2025）。在此语境下，世界模型不仅是一个生成帧的系统；它是一个能够理解和预测物理世界的动态、因果关系和潜在机制的系统。准确基准测试视频生成的物理动态是一项关键要求——也是本工作的重点——将其可能作为世界模型采用。从 AI 的角度来看，最大胆的挑战或许是需要超越视觉验证的物理动态评估。当前方法使用要么

符号 ∗、†、‡ 分别表示第一、第二和最后一位作者贡献相同。

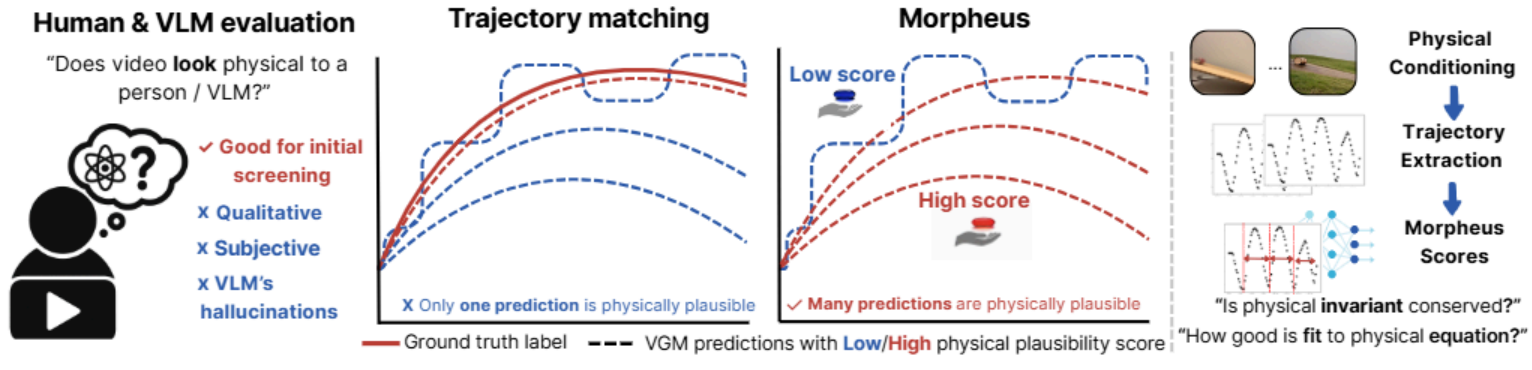


Figure 1: Comparison of evaluation methods for video generative models. a) Human or VLM-based judgments provide only qualitative and subjective assessments of physical plausibility. b) Trajectory matching compares generated and ground-truth paths but may misclassify physically valid trajectories. For example, for projectile motion, many parabolic trajectories are physically plausible when VGMs are conditioned only on an image, as object velocity cannot be estimated from it. c) Our proposed framework, Morpheus, evaluates generated videos via physics-informed scores, testing both conservation of physical invariants and consistency with governing equations of motion.

a) human or VLM judgement (e.g., *"does this video of an object falling look legitimate?"*) (Bansal et al., 2024; Meng et al., 2024), or b) visual or geometric plausibility (e.g., *"is the generated scene visually consistent through time?"*) (Agarwal et al., 2025) or c) comparsion with one possible future using trajectory matching (e.g., comparing the generated locations of a projectile with one possible ground-truth locations) (Kang et al., 2024; Motamed et al., 2025)). While human-based evaluations (potentially distilled in VLMs) are useful for initial screening of generated videos, they do not provide sufficient quantitative and objective evidence of physical plausibility. Trajectory matching evaluation can yield false negatives in cases where the generated video is physically plausible, but the object's trajectory deviates from a particular ground-truth trajectory due to unobserved initial conditions or other visually hidden factors, such as an object's mass or friction, that affect the object's motion.

A deeper physical understanding requires assessing whether a generated video preserves *physical invariants* and *physical dynamics* that govern the underlying system. For instance, in many systems, quantities such as total energy must remain constant as the system evolves, providing opportunities for quantitative benchmarks of physical plausibility. In addition, we can test whether an object's trajectory is consistent with the set of trajectories permitted by the governing physical laws by evaluating the fit of the observed dynamics to the governing *equations of motion*. Using these quantitative evaluations, we can design systematic benchmarks that reveal whether video generative models (VGMs) truly capture the dynamics of the physical world or simply produce visually plausible approximations.

We propose **Morpheus**, a novel physics-informed benchmarking framework designed to evaluate the physical reasoning capabilities of video generative models using real-world physical experiments. The key idea behind **Morpheus** is to map video recordings of physical events—whether generated by models or recorded from real experiments—into a common physical representation that can be analyzed and compared. Leveraging advances in zero-shot object segmentation, object tracking, and physics-informed neural networks (PINNs) (Cuomo et al., 2022; He et al., 2023), our framework a) fits to the video dynamics the ODE that governs the underlying system, and b) extracts standardized physical measurements, such as velocity and acceleration from video data, which should conform to conservation laws. By collecting measurements from both real physical videos and generated ones, and comparing their summary statistics with respect to governing ODEs and physical invariants, **Morpheus** enables fair and systematic benchmarking of physical invariants, such as the conservation of energy or momentum, without requiring or relying on explicit ground truth data.

This work makes three contributions toward rigorous evaluation of physical reasoning in video generative models. First, we introduce **Morpheus**, the first benchmark to systematically evaluate physical reasoning based explicitly on physical invariants (Section 3) using real-world physical experiments to ground VGM's generations in a controllable setting. Second, we propose a novel framework that combines physics-informed deep learning with advanced computer vision techniques to enable coarse- and fine-grained analysis of physical phenomena (Section 4). Third, we evaluate state-of-the-art video generative models on **Morpheus** generating over 9000 videos, including CogVideoX (Yang et al., 2024b), PyramidalFlow (Jin et al., 2024), LTX-Video (HaCohen et al., 2024), Wan2.1 (Wan et al., 2025b), COSMOS (Agarwal et al., 2025) and Veo3 (Veo-Team et al.,

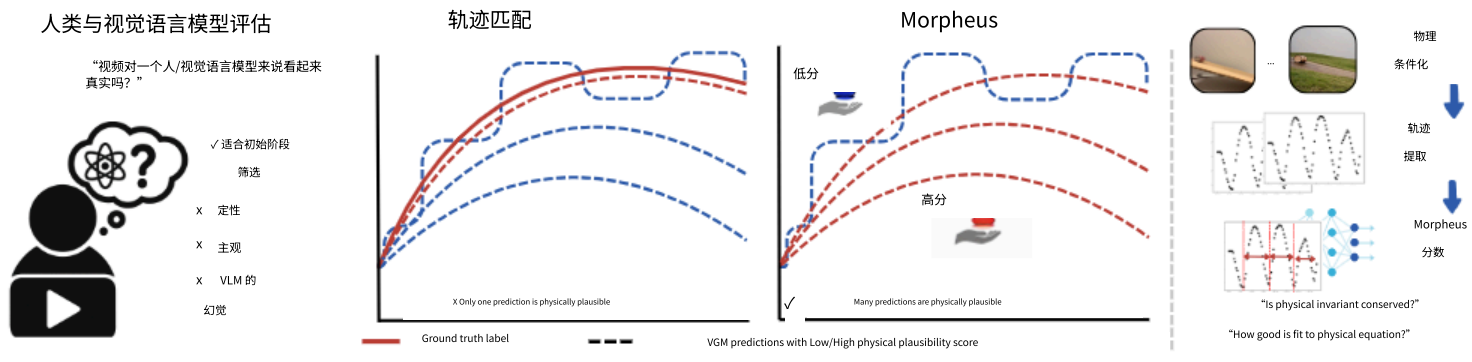


图 1：视频生成模型评估方法的比较。a) 人类或基于视觉语言模型的判断仅提供定性和主观的物理合理性评估。b) 轨迹匹配比较生成和真实路径，但可能错误分类物理上有效的轨迹。例如，对于抛体运动，当视觉生成模型仅基于图像进行条件化时，许多抛物线轨迹在物理上是合理的，因为无法从图像中估计物体速度。c) 我们提出的框架 Morpheus 通过物理信息评分评估生成视频，测试物理不变量的守恒以及与运动方程的一致性。

a) 人类或视觉语言模型（VLM）的判断（例如，“这个物体下落的视频看起来真实吗？”）（Bansal 等人，2024；Meng 等人，2024），或 b) 视觉或几何合理性（例如，“生成的场景在时间上是否视觉一致？”）（Agarwal 等人，2025），或 c) 通过轨迹匹配与一个可能的未来进行比较（例如，将抛射体的生成位置与一个可能的真实位置进行比较）（Kang 等人，2024；Motamed 等人，2025）。虽然基于人类的评估（可能被提炼到 VLM 中）对于生成视频的初步筛选很有用，但它们并不能提供足够的定量和客观的物理合理性证据。轨迹匹配评估在生成视频在物理上是合理的情况下，可能会产生假阴性，因为物体的轨迹由于未观察到的初始条件或其他视觉隐藏因素（如物体的质量或摩擦力）而偏离了特定的真实轨迹，这些因素会影响物体的运动。

更深层次的理解物理现象需要评估生成的视频是否保留了控制底层系统的物理不变量和物理动态。例如，在许多系统中，总能量等量在系统演化过程中必须保持恒定，这为物理合理性的定量基准测试提供了机会。此外，我们可以通过评估观测到的动态与运动控制方程的匹配程度，来测试物体的轨迹是否与由支配物理定律允许的轨迹集一致。利用这些定量评估，我们可以设计系统性的基准测试，以揭示视频生成模型（VGMs）是否真正捕捉了物理世界的动态，还是仅仅生成了视觉上合理的近似。

我们提出了 Morpheus，一个新颖的物理信息基准测试框架，旨在通过真实物理实验评估视频生成模型的物理推理能力。Morpheus 背后的关键思想是将物理事件的视频记录——无论是模型生成的还是真实实验记录的——映射到一个通用的物理表示形式，以便进行分析和比较。利用零样本目标分割、目标跟踪和物理信息神经网络（PINNs）的进展（Cuomo 等人，2022 年；He 等人，2023 年），我们的框架 a) 将控制底层系统的常微分方程拟合到视频动态中，以及 b) 从视频数据中提取标准化的物理测量值，如速度和加速度，这些测量值应符合守恒定律。通过收集真实物理视频和生成视频的测量值，并将它们关于控制常微分方程和物理不变量的汇总统计数据进行比较，Morpheus 能够在不要求或依赖显式真实数据的情况下，对物理不变量（如能量或动量守恒）进行公平和系统的基准测试。

这项工作在严格评估视频生成模型中的物理推理方面做出了三个贡献。首先，我们介绍了 Morpheus，这是首个基于物理不变量（第 3 节）系统地评估物理推理的基准，利用真实世界的物理实验将 VGM 的生成置于可控环境中。其次，我们提出了一种结合物理信息深度学习与先进计算机视觉技术的创新框架，以实现物理现象的粗粒度和细粒度分析（第 4 节）。第三，我们在 Morpheus 上评估了最先进的视频生成模型，生成了超过 9000 个视频，包括 CogVideoX（Yang 等人，2024b）、PyramidalFlow（Jin 等人，2024）、LTX-Video（HaCohen 等人，2024）、Wan2.1（Wan 等人，2025b）、COSMOS（Agarwal 等人，2025）和 Veo3（Veo-Team 等人，

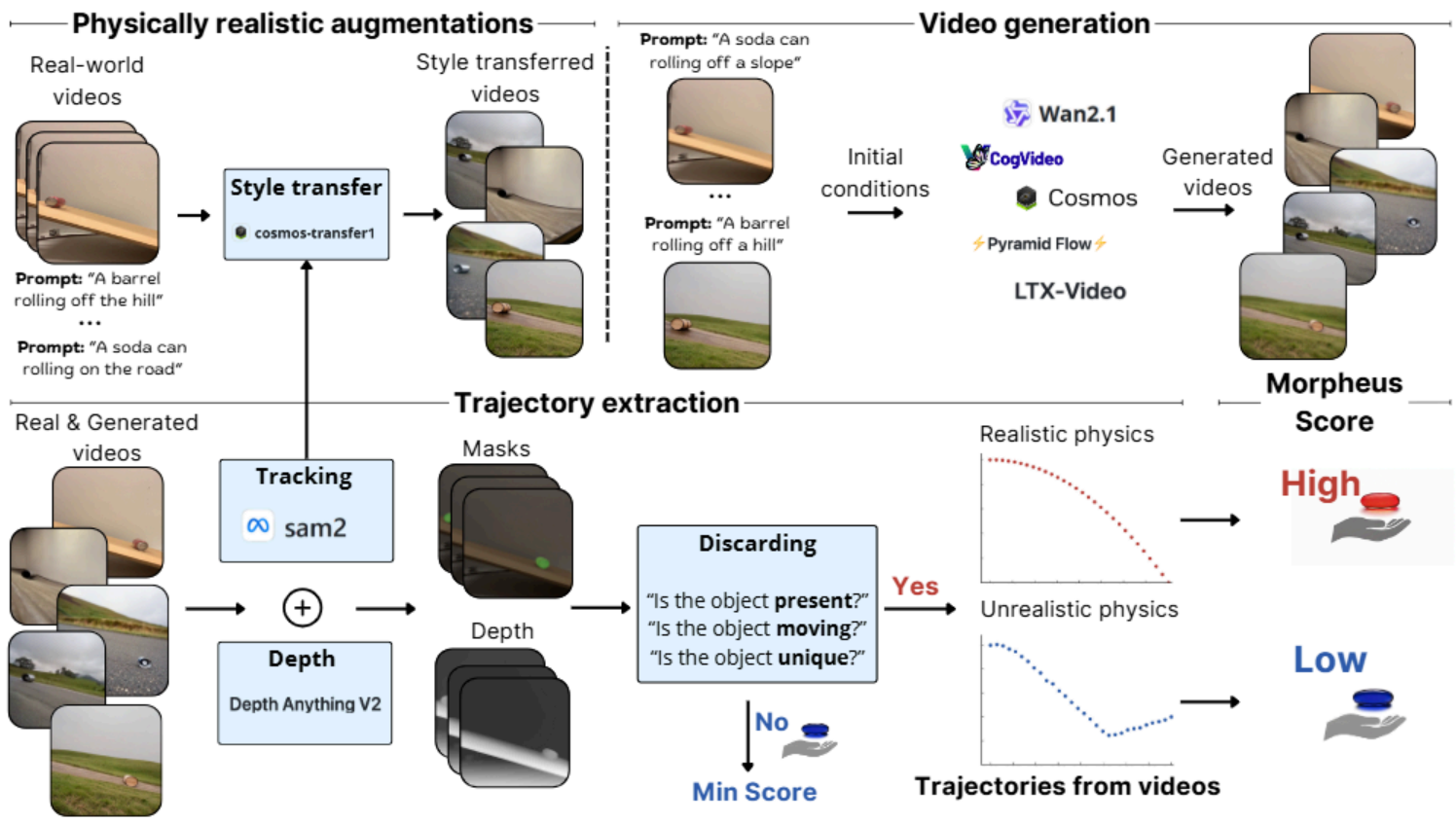


Figure 2: The overview of **Morpheus** benchmark. Video augmentation and generation (upper) and the trajectory extraction pipelines (lower). We start with augmenting recorded videos with realistic style transfer, based on object masks. Next, we use the first frame (or multiple frames in case of video conditioning) of the obtained videos, as well as the textual description, as a prompt for a VGM. After this, we extract object trajectories for both real-world and generated videos using the trajectory extraction pipeline, including trajectory tracking and discarding unreliable trajectories. Finally, we evaluate **Morpheus** scores for all videos with valid trajectories.

2024), and show that while the best of these models excel in visual aesthetics, they fall short in modeling real-world physical dynamics (Section 5).

2 Related work

Evaluation of VGMs. Benchmarking video generation models have evolved to include comprehensive evaluation frameworks that assess multiple dimensions of video quality, temporal coherence, and alignment with prompts. Approaches like EvalCrafter (Liu et al., 2024a), VBench (Huang et al., 2024a), VBench++ (Huang et al., 2024b), AIGCBench (Fan et al., 2024), and TC-Bench (Feng et al., 2024) emphasize diverse metrics to evaluate visual fidelity, motion smoothness, spatial consistency, and temporal dynamics. For example, EvalCrafter (Liu et al., 2024a) uses metrics like Motion-Aware Consistency (MAC) and Scene Change Consistency (SCC) to assess the smoothness and natural progression of motion, while VBench introduces metrics for spatial relationships and subject identity consistency to evaluate logical scene composition. Despite the breadth of these benchmarks, they primarily concentrate on perceptual and semantic aspects of video generation, whereas **Morpheus** focuses on physical plausibility of the generated videos.

Physical reasoning and plausibility in VGMs. Recent advances in evaluating physical plausibility in video generation have employed both human assessments (Bansal et al., 2024) and automated approaches leveraging vision-language models (VLMs) (Bansal et al., 2024; Meng et al., 2024) as well as object tracking metrics (Wang et al., 2024b; Motamed et al., 2025; Agarwal et al., 2025) (see Table 2 for comparison). Notable frameworks include VideoCon-Physics (Bansal et al., 2024), PhyGenBench (Meng et al., 2024) and PhysBench (Chow et al., 2025), which utilize VLMs to assess adherence to physical law prompts; VAMP (Wang et al., 2024b), which quantifies motion characteristics through acceleration and velocity variance; and Physics-IQ (Motamed et al., 2025) and COSMOS (Agarwal et al., 2025), which compare object masks between generated and real-world videos. Kang et al. (Kang et al., 2024) used the PHYRE simulator (Bakhtin et al., 2019) to fine-tune VGM on synthetic 2D data, facilitating out-of-distribution and combinatorial generalization evaluation.

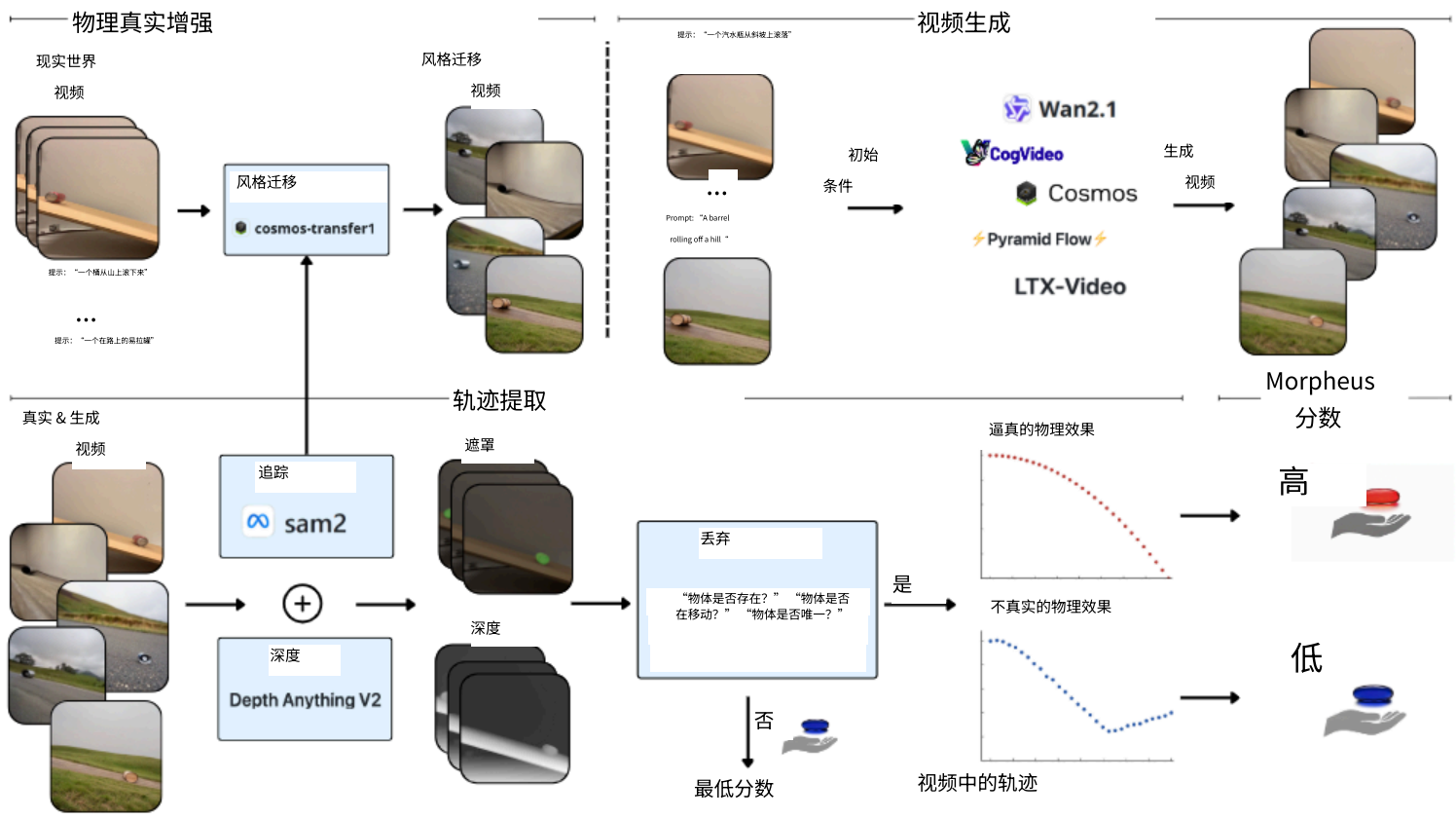


图 2：Morpheus 基准测试的概述。视频增强和生成（上部）以及轨迹提取流程（下部）。我们从基于对象掩码的真实风格迁移开始增强记录的视频。接下来，我们使用获得视频的第一帧（或视频条件化情况下的多帧）以及文本描述，作为视频生成模型（VGM）的提示。之后，我们使用轨迹提取流程提取真实世界和生成视频的对象轨迹，包括轨迹跟踪和丢弃不可靠的轨迹。最后，我们对所有具有有效轨迹的视频评估 Morpheus 分数。

2024 年），并表明这些模型中的最佳模型在视觉美学方面表现出色，但在模拟真实世界的物理动态方面存在不足（第 5 节）。

2 相关工作

VGMs 的评估。视频生成模型的基准测试已发展到包含综合评估框架，这些框架评估视频质量、时间一致性和与提示的匹配度等多个维度。EvalCrafter（Liu 等人，2024a）、VBench（Huang 等人，2024a）、VBench++（Huang 等人，2024b）、AIGCBench（Fan 等人，2024）和 TC-Bench（Feng 等人，2024）等方法强调使用多种指标来评估视觉保真度、运动平滑性、空间一致性和时间动态性。例如，EvalCrafter（Liu 等人，2024a）使用运动感知一致性（MAC）和场景变化一致性（SCC）等指标来评估运动的平滑性和自然进展，而 VBench 引入了用于空间关系和主体身份一致性的指标来评估逻辑场景构图。尽管这些基准测试范围广泛，但它们主要集中于视频生成的感知和语义方面，而 Morpheus 则专注于生成视频的物理合理性。

物理推理和视频生成模型（VGM）中的合理性。近年来，评估视频生成中物理合理性的进展采用了人类评估（Bansal 等人，2024）和利用视觉语言模型（VLM）（Bansal 等人，2024；Meng 等人，2024）以及目标跟踪指标（Wang 等人，2024b；Motamed 等人，2025；Agarwal 等人，2025）（比较见表 2）的自动化方法。值得关注的框架包括 VideoCon-Physics（Bansal 等人，2024）、PhyGenBench（Meng 等人，2024）和 PhysBench（Chow 等人，2025），这些框架利用 VLM 来评估对物理法提示的遵循情况；VAMP（Wang 等人，2024b），通过加速度和速度方差量化运动特征；以及 Physics-IQ（Motamed 等人，2025）和 COSMOS（Agarwal 等人，2025），这些框架比较生成视频和真实世界视频之间的目标掩码。Kang 等人（Kang 等人，2024）使用 PHYRE 模拟器（Bakhtin 等人，2019）在合成 2D 数据上微调 VGM，以促进分布外和组合泛化评估。

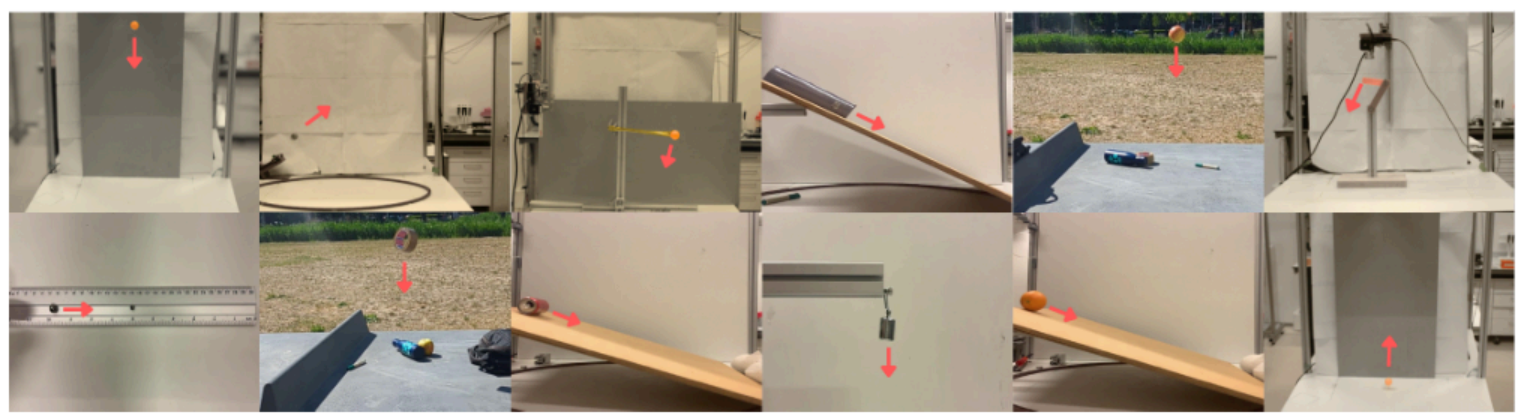


Figure 3: Examples of physical experiments included in the **Morpheus** benchmark, illustrating both different dynamics and variations in object types. Top row (left to right): falling ball, projectile motion, holonomic pendulum, sliding, falling apple, and double pendulum. Bottom row (left to right): collision, falling tape, rolling can, spring, rolling orange, and bouncing.

Despite addressing diverse physical phenomena, these benchmarks have significant limitations. VLM and human evaluations often identify physical deviations categorically, like noting gravity violations without quantifying them. Moreover, VLM can hallucinate (Li et al., 2023) and miss subtle physical inconsistencies (Chow et al., 2025). On the other side, object tracking metrics are often based on simulated data (Agarwal et al., 2025; Bakhtin et al., 2019) and assume that modeled processes should be deterministic and predictable (Kang et al., 2024; Motamed et al., 2025). These limitations highlight the critical need for more robust, interpretable benchmarks that can quantitatively evaluate physical realism by precisely measuring how well-generated videos preserve physical invariants and adhere to specific physical laws.

Learn physical invariants and equations from data. There is progress for learning conservation laws from trajectories (Liu & Tegmark, 2021), and equation discovery in hybrid dynamic systems (Liu et al., 2024b). Mechanistic Neural Networks (MechNN) (Pervez et al., 2024) are able to learn governing ODEs from data, while Mechanistic PDE Networks (Pervez et al., 2025) can learn Partial Differential Equations (PDEs). On the other hand, to compare the theoretical prediction with input data, PINNs (Cuomo et al., 2022; He et al., 2023), which integrate physical equations in the loss function, help identify possible physical factors causing errors (such as unmodeled friction, air drag, etc.) because it is able to learn corrections to make the predictions closer to the actual observed values.

3 MORPHEUS BENCHMARK

To rigorously examine discrepancies in adherence to physical laws within generated videos, we propose the **Morpheus** benchmark, consisting of a dataset for controlled conditioning, evaluation scores, and an analysis of state-of-the-art models' performance.

Dataset methodology We created a dataset of real-world videos of specific physical phenomena, focusing on capturing fundamental aspects of Newtonian mechanics. Videos were recorded under controlled laboratory conditions, allowing us to systematically vary initial parameters and capture repeatable scenarios. By operating in this rigorously controlled setting, we can isolate and test adherence to specific physical laws – such as the periodic dynamics of a harmonic pendulum – rather than merely assessing overall visual plausibility. This sets our dataset apart from previous works that often focus on uncontrolled, general-purpose video data (Motamed et al., 2025; Kang et al., 2024), allowing for a more precise and targeted evaluation of physical consistency.

We recorded a set of nine core physical experiments with 10-20 videos in each experiment, each highlighting different physical principles such as gravity (falling, projectile motion, and bouncing), periodic movements (spring and holonomic pendulum), friction and normal forces (incline sliding and rolling), and more complex dynamical systems such as multi-body collisions and double pendulum (See Fig. 3 and App. A for dataset an illustration and description). For each experiment, we varied the initial conditions (recording 5–7 videos per condition), such as speed for falling, angle and speed for projectile motion, and angle for pendulums. This diverse collection ensures robust coverage of dynamic behaviors, enabling thorough evaluation of generated videos against real-world physical

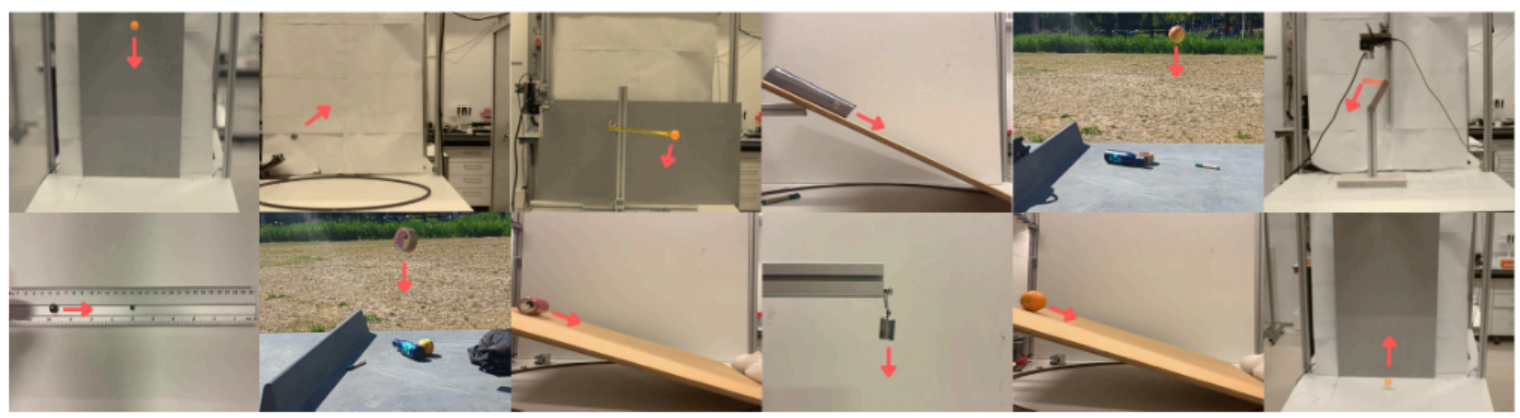


图 3：Morpheus 基准测试中包含的物理实验示例，展示了不同的动力学和物体类型的变化。最上面一行（从左到右）：下落的球、抛体运动、完整摆、滑动、下落的苹果和双摆。最下面一行（从左到右）：碰撞、下落的胶带、滚动的罐子、弹簧、滚动的橙子和弹跳。

尽管这些基准测试涵盖了多种物理现象，但它们存在显著局限性。视觉语言模型（VLM）和人工评估通常以分类方式识别物理偏差，例如指出重力违规现象而不进行量化。此外，VLM 可能会产生幻觉（Li 等人，2023 年）并忽略细微的物理不一致性（Chow 等人，2025 年）。另一方面，目标跟踪指标通常基于模拟数据（Agarwal 等人，2025 年；Bakhtin 等人，2019 年），并假设模型化过程应该是确定性和可预测的（Kang 等人，2024 年；Motamed 等人，2025 年）。这些局限性凸显了迫切需要更稳健、更易于解释的基准测试，这些基准测试能够通过精确测量生成视频如何保持物理不变性并遵循特定物理定律，来量化评估物理真实性。

从数据中学习物理不变量和方程。在从轨迹中学习守恒定律（Liu & Tegmark, 2021）和混合动态系统中的方程发现方面已有进展（Liu et al., 2024b）。机制神经网络（MechNN）（Pervez et al., 2024）能够从数据中学习控制常微分方程（ODEs），而机制偏微分方程网络（MechPDE）（Pervez et al., 2025）可以学习偏微分方程（PDEs）。另一方面，为了将理论预测与输入数据进行比较，将物理方程集成到损失函数中的物理信息神经网络（PINNs）（Cuomo et al., 2022; He et al., 2023）有助于识别可能造成误差的物理因素（如未建模的摩擦、空气阻力等），因为它能够学习修正项，使预测更接近实际观测值。

3 MORPHEUS 基准测试

为了严格检验生成视频中遵循物理定律的差异，我们提出了 Morpheus 基准，包括用于控制条件的数据集、评估分数以及对最先进模型性能的分析。

数据集方法 我们创建了一个包含特定物理现象真实世界视频的数据集，重点关注捕捉牛顿力学的基本方面。视频在受控的实验室条件下录制，使我们能够系统地改变初始参数并捕捉可重复的场景。通过在这个严格受控的环境下操作，我们可以隔离并测试对特定物理定律的遵循——例如谐振摆的周期动力学——而不仅仅是评估整体视觉合理性。这使我们的数据集区别于以往通常关注不受控、通用视频数据的工作（Motamed 等人，2025 年；Kang 等人，2024 年），从而允许对物理一致性进行更精确和有针对性的评估。

我们记录了一系列九个核心物理实验，每个实验包含 10-20 个视频，每个视频突出不同的物理原理，如重力（下落、抛体运动和弹跳）、周期运动（弹簧和完整摆）、摩擦力和法向力（斜坡滑动和滚动），以及更复杂的动力学系统，如多体碰撞和双摆（见图 3 和附录 A，用于数据集的图示和描述）。对于每个实验，我们改变了初始条件（每个条件记录 5-7 个视频），例如下落的速度、抛体运动的角和速度，以及摆的角。这个多样化的集合确保了对动态行为的全面覆盖，从而能够对生成的视频进行真实的物理评估。

phenomena. The initial video frame(s) are used as conditioning to guide VGM's sampling/generation process, ensuring the generated sequences start from the same conditions as the real experiments.

Visually diverse conditioning for robust VGMs evaluation. To obtain more robust evaluations, it is important to study the performance of VGMs under diverse natural initializations of the same physical process. This reduces sensitivity to the specific setup of the originally recorded experiments. Thus, we augment the initial videos with visually diverse yet reliably generated variants. In particular, we use the video-to-video transfer method (Alhaija et al., 2025) with object masks extracted from recorded videos to obtain diverse and visually realistic videos following an additional text prompt for adaptation. While changing the semantics and appearance of the objects, the generated videos strictly follow the provided object masks and are reliable augmentations for conditioning in image-to-video and video-to-video generation. To ensure their quality, we filter augmented initializations through manual inspection for visual and physical plausibility. Thus, the original initializations are expanded 10-fold by generating 3 variations with 3 stylization prompts. For more details, we refer the reader to App. B.

Metrics validation. We use real-world videos as a *"gold standard"* to validate that our evaluation metrics work as intended. By analyzing the metrics on these ground-truth recordings, we demonstrate the reliability of our approach and establish an upper bound on performance, representing the precision with which we measure adherence to physical laws. In essence, the metrics computed on real-world videos provide a baseline for how closely any generative model can align with physical principles.

To structurally analyze the videos, we extract the trajectories of the objects by applying promptable video segmentation. These trajectories comprise 2D coordinates of the recognized objects through time and are used for further analysis with our physical metrics (see Sec. 3.2 for details).

3.1 Prompting methods

The physical dynamics of a scene are determined by its initial conditions: object positions, velocities, and geometric constraints (e.g., shapes, rigid connections). In generative models, these conditions are set through prompting, which can take three forms: *a)* textual prompts, *b)* single-image prompts, and *c)* video (multi-frame) prompts, providing a different level of control over generation. *Textual prompts* offer only broad control, suggesting behaviors (e.g., rolling, falling) without precise states. *Single-image prompts* improve precision by fixing initial locations, but lack motion detail. Only *video prompts* specify both positions and velocities, providing the highest control.

With this gradation in mind, we investigate how different levels of control affect the physical realism. We explore both textual prompt enhancement and various multi-frame prompting for models capable of leveraging these features (e.g. (Agarwal et al., 2025; Yang et al., 2024b)), allowing us to examine the relationship between input precision and output physical fidelity. Following (Yang et al., 2024b), we use a VLM (Zeng et al., 2024) to expand simple scene descriptions into richer prompts via instruction templates in zero- or few-shot settings. As not all VGMs provide their own prompt upsampler, we rely on the ChatGLM family of models (Zeng et al., 2024), while for COSMOS-variants (Agarwal et al., 2025) and WAN-2.1 (Wan et al., 2025a) we use their own devised (NVIDIA, 2024; Agrawal et al., 2024; Wang et al., 2024a) upsamplers respectively. In our evaluation, we create descriptive prompts with an emphasis on physical motion, and the upsampler brings the inference-time prompt distribution closer to that used during training.

3.2 Trajectory extraction

While generated videos could be directly evaluated in terms of 3D consistency (Liu et al., 2024a) or other pixel-level generation properties (Agarwal et al., 2025), such evaluations are limited to visual and geometric realism of the generated videos. Instead, we are interested in how well these videos conform to physical laws. This means that we need to extract the relevant physical state variables, such as positions of objects, velocities, accelerations, masses, and so on. Thus, it is essential to transform the generated videos into perfect state variables of the depicted objects and their trajectories, which can then be further analyzed.

As we need to track objects in both real-world and generated videos, Segment Anything 2 (SAM-2) (Ravi et al., 2024) serve as a reliable way to generate 2D masks for any type of objects. We annotate the first frame of our videos in the dataset with positive and negative labels. Given the

能够对生成的视频进行全面评估，以检验其与真实世界物理现象的符合程度。初始视频帧被用作条件，指导视频生成模型（VGM）的采样/生成过程，确保生成的序列从与真实实验相同的条件开始。

视觉多样性条件化用于鲁棒视频生成模型（VGMs）评估。为了获得更鲁棒的评估，它研究 VGM 在不同自然初始化下的性能非常重要，这可以减少对最初记录实验的具体设置的敏感性。因此，我们通过添加视觉多样但可靠生成的变体来增强初始视频。具体来说，我们使用视频到视频转换方法（Alhaija 等人，2025 年），并结合从记录视频中提取的对象掩码，在提供额外文本提示进行适应后，获得多样且视觉逼真的视频。在改变对象的语义和外观时，生成的视频严格遵循提供对象掩码，并是图像到视频和视频到视频生成中条件化的可靠增强。为确保其质量，我们通过人工检查进行视觉和物理合理性过滤，以增强初始设置。因此，通过使用 3 个风格化提示生成 3 个变体，原始初始化被扩展了 10 倍。更多细节，我们建议读者参考附录 B。

指标验证。我们使用真实世界视频作为“黄金标准”来验证我们的评估指标是否按预期工作。通过分析这些真实记录的指标，我们展示了我们方法的可靠性，并建立了性能的上限，代表了我们对遵循物理定律的测量精度。本质上，在真实世界视频上计算的指标为任何生成模型与物理原理的契合程度提供了一个基准。

为了结构化分析视频，我们通过应用可提示视频分割提取物体的轨迹。这些轨迹包含通过时间的识别物体的二维坐标，并用于进一步使用我们的物理指标进行分析（详细内容见第 3.2 节）。

3.1 提示方法

场景的物理动力学由其初始条件决定：物体位置、速度和几何约束（例如形状、刚性连接）。在生成模型中，这些条件通过提示设置，可以采取三种形式：a) 文本提示，b) 单图像提示，c) 视频（多帧）提示，为生成提供不同级别的控制。文本提示仅提供粗略控制，暗示行为（例如滚动、下落）而不精确状态。单图像提示通过固定初始位置提高精确度，但缺乏运动细节。只有视频提示同时指定位置和速度，提供最高控制。

考虑到这种渐变，我们研究了不同控制水平如何影响物理真实性。我们探索了文本提示增强和多种多帧提示，这些功能可用于利用这些特性的模型（例如 (Agarwal et al., 2025; Yang et al., 2024b)），使我们能够检验输入精度与输出物理保真度之间的关系。遵循 (Yang et al., 2024b)，我们使用一个 VLM（Zeng et al., 2024）通过指令模板在零或少样本设置中将简单场景描述扩展为更丰富的提示。由于并非所有 VGM 都提供自己的提示上采样器，我们依赖于 ChatGLM 系列模型（Zeng et al., 2024），而对于 COSMOS 变体（Agarwal et al., 2025）和 WAN-2.1（Wan et al., 2025a），我们分别使用它们各自设计的（NVIDIA, 2024; Agrawal et al., 2024; Wang et al., 2024a）上采样器。在我们的评估中，我们创建了强调物理运动的描述性提示，而上采样器将推理时的提示分布更接近于训练时使用的分布。

3.2 轨迹提取

虽然生成的视频可以根据三维一致性（Liu 等人，2024a）或其他像素级生成属性（Agarwal 等人，2025）直接进行评估，但此类评估仅限于生成视频的视觉和几何真实性。相反，我们感兴趣的是这些视频如何遵循物理定律。这意味着我们需要提取相关的物理状态变量，例如物体的位置、速度、加速度、质量等。因此，将生成的视频转换为所描绘物体的完美状态变量及其轨迹至关重要，然后可以进一步分析。

由于我们需要在真实世界和生成视频中跟踪物体，Segment Anything 2 (SAM2) (Ravi 等人，2024) 成为生成任何类型物体二维掩码的可靠方法。我们在数据集视频的第一帧上使用正面和负面标签进行标注。给定

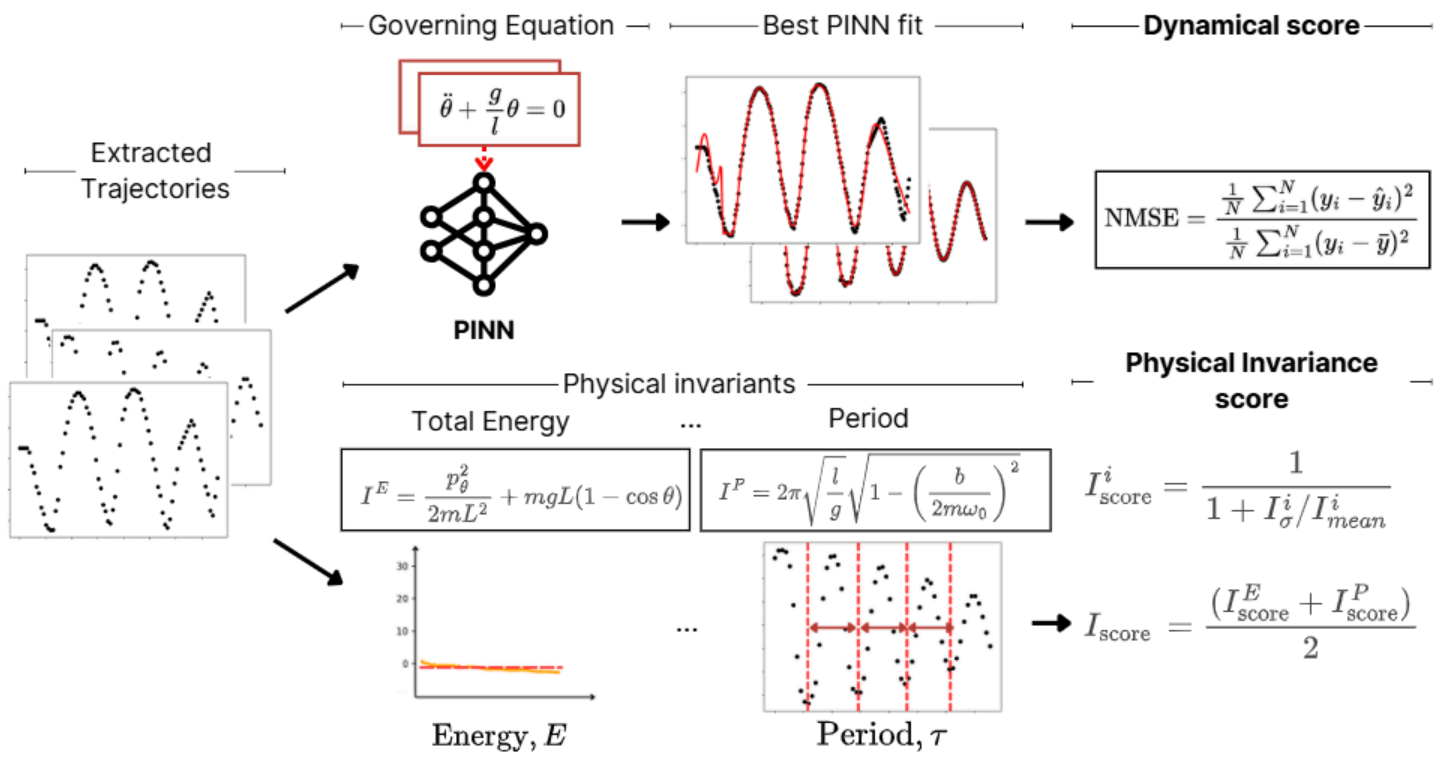


Figure 4: Evaluation of trajectories extracted from real and VGMs videos using our Dynamical (upper) and Physical Invariance (lower) scores. For the Dynamical score, trajectories from real-world or generated videos are fitted to a PINN with the corresponding equation of motion for the particular physical law as an extra loss term. For the Physical Invariance score, using the same trajectories, we estimate physical quantities that should be invariant in the systems, such as total energy and oscillation period, and use their variance as a measure of physical plausibility.

masks generated by SAM-2's we extarct the centroid of the object(s) (center-mass) in the video, at each frame of the video. In addition, we employ Depth Anything V2 (Yang et al., 2024a) to verify that objects have consistent depth through the sequence of movement (See physical score pipeline App. D.2). The depth values are calculated using the correspoding mask generated by SAM-2.

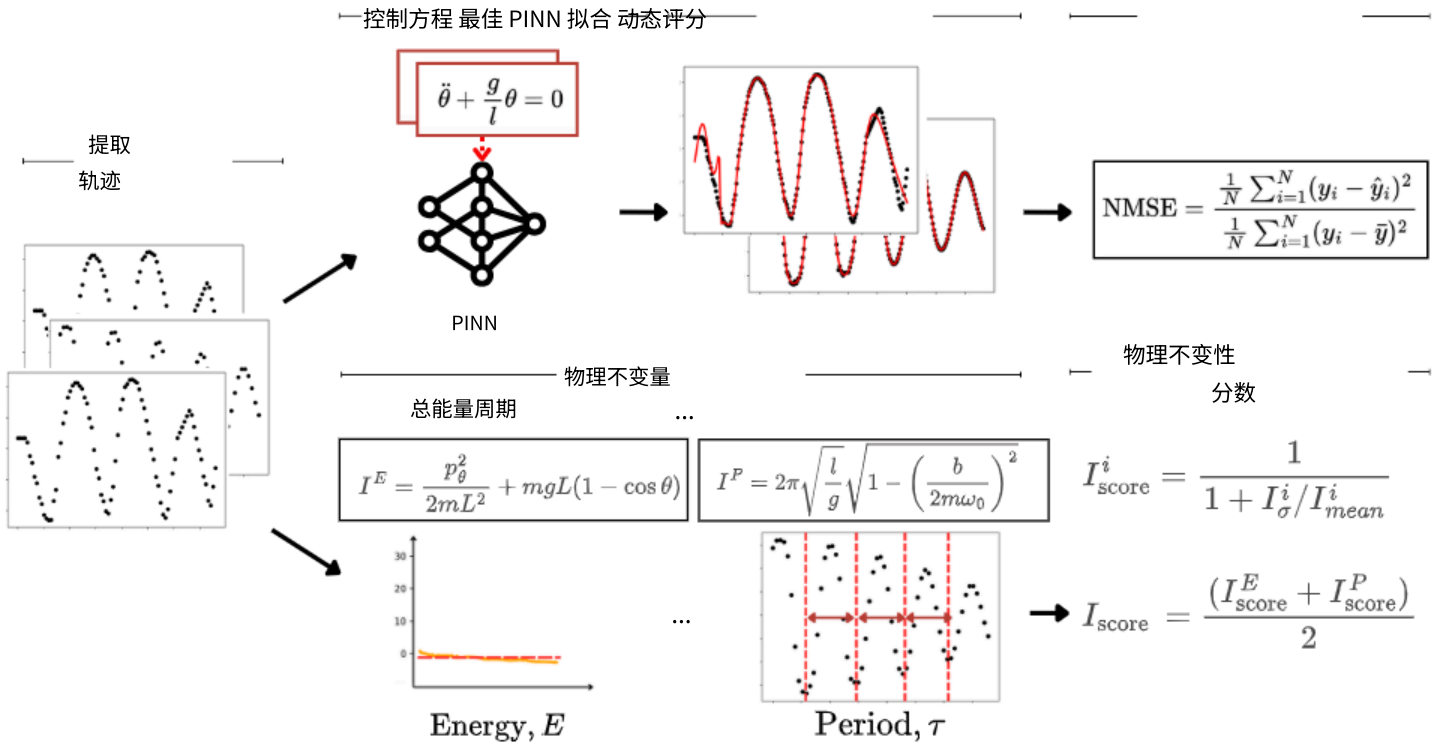
For velocity, acceleration and angular velocity, we employ the *central difference method* (Swanson & Turkel, 1992). To further reduce the noise, generated by the imperfections in the tracking pipeline, we follow with a series of smoothing operations, such as learning a linear regression with a sliding window and applying the Savitzky-Golay smoothing. The details can be found in the App. C.

4 Physics-informed evaluation metrics

To assess the alignment of the generated video trajectories with physical laws, we propose a hierarchical evaluation framework for analyzing physical experiments in both real-world and generated videos.

Discard rate As a first metric, we compute the *discard rate*, which reflects the proportion of model-generated samples that must be discarded to ensure reliable trajectory extraction needed for Physical Invariances and Dynamical scores. The discard filtering is automatic and consists of three criteria: First, we discard generated videos where objects lack sufficient permanence throughout the video. Second, we discard generated videos which do not have a consistent number of objects. Finally, we discard generated videos if there is little motion detected, as such videos are not suitable for physical analysis. The overall discard rate represents the proportion of generated videos that fail at least one of these criteria. In addition, we verify that none of the real-world extracted trajectories are discarded, showing that our discard criteria are effective in distinguishing physical from non-physical videos. We provide further details on our filtering methodology in App. D.1. For the videos that pass filtering, we employ two metrics: *Dynamical score*, which measures adherence to the governing equation of motion, and *Physical Invariance score*, which quantifies invariance of conserved quantities such as energy or angular momentum (see Table 3 for all invariances).

Physics beyond trajectory matching Previous benchmarks such as PhysicsIQ (Motamed et al., 2025) and COSMOS (Alhaija et al., 2025) use trajectory matching by comparing generated trajectories with the ground-truth trajectories. As ground-truth trajectories themselves vary due to noise and hidden physical parameters that are not fully observable from visual image/video conditioning (e.g.,



如图 4 所示：使用我们的动力学（上）和物理不变性（下）分数评估从真实视频和 VGM 视频提取的轨迹。对于动力学分数，真实世界或生成的视频轨迹被拟合到一个 PINN，并将对应于特定物理定律的运动方程作为额外的损失项。对于物理不变性分数，使用相同的轨迹，我们估计系统应该保持不变物理量，例如总能量和振荡周期，并使用它们的方差作为物理合理性的度量。

SAM-2 生成的掩码中，我们提取视频中每个帧内物体（质心）的位置。此外，我们采用 Depth Anything V2（Yang 等人，2024a）来验证物体在运动序列中具有一致的深度（参见物理评分流程附录 D.2）。深度值是使用 SAM-2 生成的相应掩码计算得出的。

对于速度、加速度和角速度，我们采用中心差分法（Swanson & Turkel，1992）。为了进一步减少由跟踪流程中的不完美性产生的噪声，我们进行了一系列平滑操作，例如使用滑动窗口学习线性回归并应用 Savitzky-Golay 平滑。详细信息可在附录 C 中找到。

4 物理信息评估指标

为了评估生成的视频轨迹与物理定律的一致性，我们提出了一种分层评估框架，用于分析真实世界和生成视频中物理实验。

丢弃率 作为第一个指标，我们计算丢弃率，它反映了为确保物理不变性和动态分数所需的可靠轨迹提取，必须丢弃的模型生成样本的比例。丢弃过滤是自动的，包括三个标准：首先，我们丢弃视频中物体缺乏足够持久性的生成视频。其次，我们丢弃物体数量不一致的生成视频。最后，如果检测到的运动很少，我们也丢弃这些生成视频，因为它们不适合物理分析。总体丢弃率表示至少有一个标准失败的生成视频的比例。此外，我们验证了所有从真实世界提取的轨迹都没有被丢弃，这表明我们的丢弃标准在区分物理视频和非物理视频方面是有效的。我们在附录 D.1 中提供了关于我们过滤方法的更多细节。 对于通过筛选的视频，我们采用两个指标：动力学分数，用于衡量对运动控制方程的遵循程度，以及物理不变性分数，用于量化能量或角动量等守恒量的不变性（所有不变性请参见表 3）。

超越轨迹匹配 之前的基准测试如 PhysicsIQ（Motamed 等人，2025 年）和 COSMOS（Alhaija 等人，2025 年）通过将生成的轨迹与真实轨迹进行比较来进行轨迹匹配。由于真实轨迹本身会因噪声和未完全从视觉图像/视频条件中可观察到的隐藏物理参数而变化（例如，

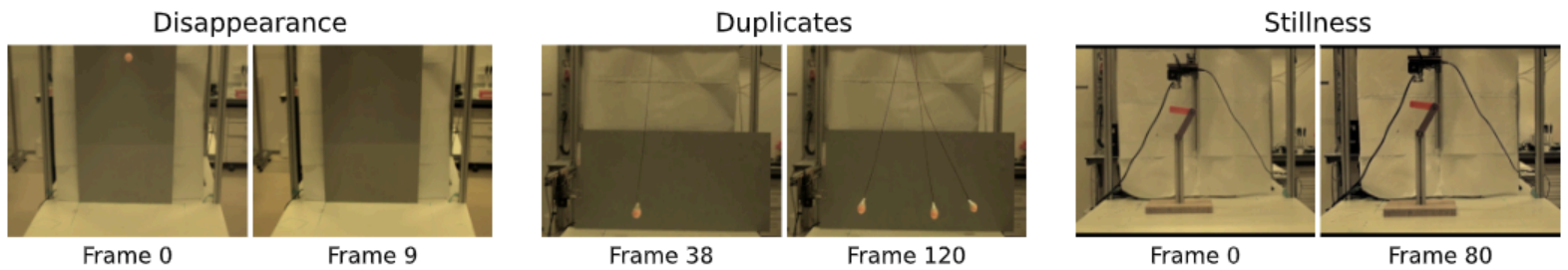


Figure 5: Different types of discarded generated videos: (left) A video showing the disappearance of the orange ball during fall; (middle) A video illustrating generation of multiple objects in pendulum experiment; (right) A video in which the double pendulum does not move.

object mass or friction), PhysicsIQ proposed to compare the obtained trajectory matching score with the variance in the real-world trajectories. However, such variance depends on the studied /recorded variations and can be arbitrarily large in cases where hidden parameters vary significantly.

Given the visual nature of the conditioning in VGMs (i.e., initial image or video), there are often parameters of the physical system that cannot be estimated from the provided *visual* conditioning. In the image conditioning case, such unobserved parameters exist for almost all conditions/experiments, limiting the trajectory matching metric for state-of-the-art image-based VGM, such as Veo3 (Veo-Team et al., 2024) and WAN2.1 (Wan et al., 2025a), which allow only image-based conditioning. Moreover, even in the video conditioning case there are often hidden parameters not observed in the original conditioning, limiting the applicability of trajectory matching only to simple systems. In contrast, **Morpheus** scores overcome those limitations by evaluating results of video generation on the level of physical laws instead of a particular trajectory provided by ground-truth videos.

4.1 Dynamical score

To calculate the Dynamical score, we use physics-informed neural networks (PINNs) (Cuomo et al., 2022), which directly incorporate physical laws as a prior. This setting allows us to learn the physical trajectory that fits the data the most, independent of the initial conditions. Fig. 4 illustrates our approach. A PINN is a neural network that receives a timestep i of the trajectory as input and outputs the trajectory coordinates $\hat{T}_i$, velocity $\hat{\dot{T}}_i$, and acceleration $\hat{\ddot{T}}_i$. The model is typically trained with a loss function, comprising two components L_{data} and L_{physics}: $L_{\text{total}} = L_{\text{data}} + \lambda L_{\text{physics}}$, where the L_{data} is responsible for fitting the model to the datapoints, $L_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{T}_i - T_i\|^2$, and L_{physics} enforces following the physical law. For each experiment, we explicitly implement the equation of motion in the form of an ordinary differential equation as PINN loss functions. E.g., for the falling ball, the equations of motion are: $\dot{x} = 0; \quad \ddot{y} + g = 0$, where y is the vertical position, $\ddot{y}$ is the acceleration and g is the gravitational constant. The L_{physics} is calculated as: $L_{\text{physics}} = \frac{1}{M}\sum_{j=1}^{M} \left\|\hat{\ddot{y}}_j + g\right\|^2 + \left\|\hat{\dot{x}}\right\|^2$, where $\hat{\ddot{y}}_j$ is the predicted acceleration derived from the PINN at the j-th time step.

Computing the Dynamical score. We use PINNs to assess the physical plausibility of trajectories from generated videos by computing the normalized mean square error (NMSE) of the model-learned trajectory derived from videos. We normalize by inverting the error, so 1 marks the best dynamical score and 0 a worse-than-constant PINN fit. A higher Dynamical score implies higher physical plausibility. For more details, please see App. D.3.

4.2 Physical Invariance score

To calculate a more fine-grained Physical Invariance score, we accompany each of our experiments with a list of physical invariances, i.e. values that we can derive from trajectories that stay constant in time. We make a series of reasonable assumptions about the setting and test them on the real-world trajectories. As invariances vary for different experiments, we present here one case study for the falling ball experiments, while describing all the other settings in App. D.4 and Table 3.

Case study: Falling ball. In the falling ball experiments, we have the following physical invariants.

$\Rightarrow$ *Total energy.* Assuming negligible air resistance, the total energy —the sum of the kinetic and potential energy— of the ball is conserved. The kinetic energy of the ball is: $T = \frac{1}{2}m(v_x^2 + v_y^2)$,

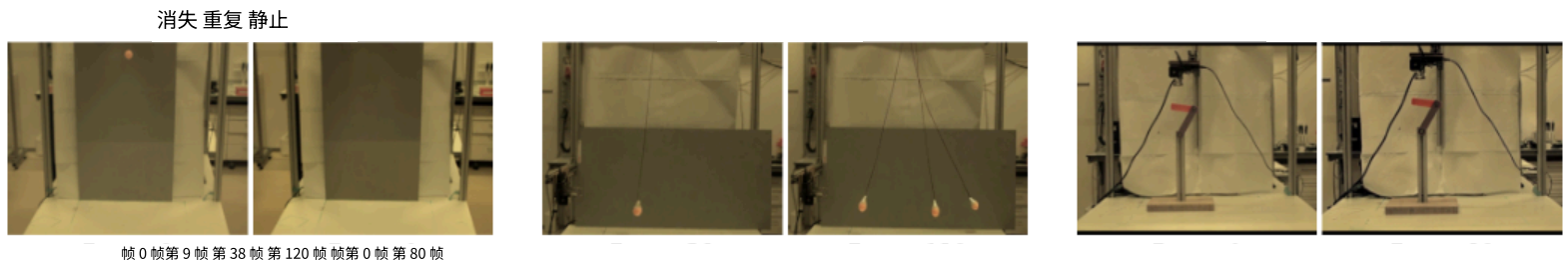


图 5：被丢弃的不同类型的生成视频：（左）一个显示橙色球在坠落过程中消失的视频；（中）一个展示摆动实验中生成多个物体的视频；（右）一个双摆没有移动的视频。

（物体质量或摩擦力等），PhysicsIQ 建议将获得的轨迹匹配分数与真实世界轨迹的方差进行比较。然而，这种方差取决于所研究/记录的变化，并且在隐藏参数变化显著的情况下可能会任意大。

鉴于 VGMs（视频生成模型）的条件化具有视觉特性（即初始图像或视频），通常存在一些物理系统的参数无法从提供的视觉条件中估计出来。在图像条件化情况下，几乎所有条件/实验中都存在此类未观察到的参数，这限制了基于图像的轨迹匹配度量，例如 Veo3（VeoTeam 等，2024）和 WAN2.1（Wan 等，2025a），这些模型仅允许基于图像的条件化。此外，即使在视频条件化情况下，也常常存在原始条件化中未观察到的隐藏参数，这限制了轨迹匹配仅适用于简单系统。相比之下，Morpheus 评分通过在物理定律层面而非由真实视频提供的特定轨迹层面评估视频生成结果，克服了这些限制。

4.1 D

为了计算动态分数，我们使用物理信息神经网络（PINNs）（Cuomo 等人，2022 年），它将物理定律直接作为先验知识整合。这种设置使我们能够学习最符合数据的物理轨迹，而与初始条件无关。图 4 说明了我们的方法。PINN 是一种神经网络，它将轨迹的时间步 i 作为输入，并输出轨迹坐标ˆT 和速度

T. 该模型通常使用包含两个部分的损失函数进行训练，即 和 L：L = L + λL，其中 L 负责将模型拟合到数据点，L =
T−T⁄⁄，并且 Lenforces 遵循物理定律。对于每个实验，我们将运动方程显式地实现为常微分方程形式作为 PINN 损失函数。例如，对于下落的球，运动方程为：•x = 0； ¨y + g = 0，其中 y 是垂直位置，¨y 是加速度和

g 是万有引力常数。L 的计算公式为：L=

其中ˆ¨y 是 PINN 在 j-th 时间步预测的加速度。

计算动态分数。我们使用 PINNs 来评估从生成视频中得到的轨迹的物理合理性，通过计算模型从视频中学习到的轨迹的归一化均方误差（NMSE）。我们通过反转误差进行归一化，因此 1 代表最佳的动态分数，0 代表比常数更差的 PINN 拟合。更高的动态分数意味着更高的物理合理性。更多细节请参见附录 D.3。

4.2 PI

为了计算更细粒度的物理不变性分数，我们对每个实验都附上一份物理不变性清单，即我们可以从随时间保持不变的轨迹中推导出的值。我们对场景做出一系列合理的假设，并在真实世界轨迹上测试这些假设。由于不变性因不同实验而异，我们在附录 D.4 和表 3 中描述了其他所有设置，而在此处展示了一个下落球实验的案例研究。

案例研究：下落球。在下落球实验中，我们有以下物理不变量。
⇒ 总能量。假设空气阻力可忽略不计，球的总能量——动能和势能之和——是守恒的。球的动能是：
T = m(v+ v),

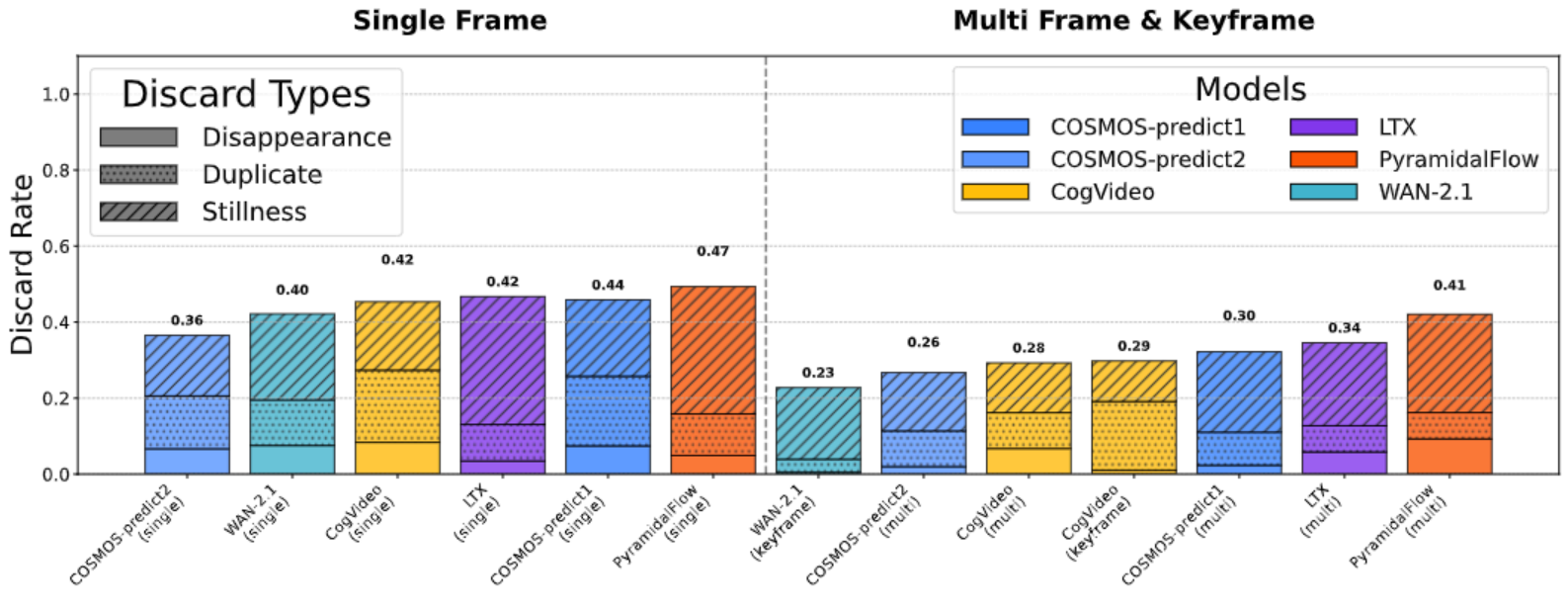


Figure 6: Average discard rates across all physical experiments (lower is better).

where $v = (v_x, v_y)$ is the speed of the ball and m is it's mass. Also, the potential energy is $V = mgy$ where g is the gravitational acceleration constant and y is the vertical coordinate. So, as the total energy is the sum of kinetic and potential, we get: $E = T + V = \frac{1}{2}m(v_x^2 + v_y^2) + mgy$.

$\Rightarrow$ *Energy-to-mass ratio.* Assuming that the mass of the ball is constant, we derive the following invariant: $\frac{E}{m} = \frac{1}{2}(v_x^2 + v_y^2) + gy = \text{const}$, which we can estimate with the data from our trajectory.

$\Rightarrow$ *Acceleration.* As no external forces are acting on the ball except for gravity, which is uniform and is directed downwards, the acceleration of the ball is also constant: $a_y = g = \text{const}$.

$\Rightarrow$ *Horizontal momentum-to-mass ratio.* As with acceleration, the horizontal momentum, $p_x = mV_x$, is also preserved given no external forces. Thus, the horizontal velocity is conserved: $v_x = \text{const}$.

Computing the Physical Invariance score. To convert the invariant into an actual score, like the Energy score, we calculate the standard deviation of the invariant time series and normalize it into the range of $(0, 1)$, with 1 indicating a perfect Physical Invariance score. As invariants must be by nature constant, a high standard deviation of these invariants (and thus a lower physical invariance score), indicates poor modeling of the respective physical invariants. In addition, for discarded trajectories we assign minimal Physical Invariance score equal to 0. A detailed score calculation procedure is described in the App. D.5. For the derivations of each invariant we used, please refer to the App. D.4.

5 Analysis

In this section, we analyze the results of the experiments. We present aggregated results for each model and conditioning type (single frame, multi-frame, or key frame interpolation) in Fig. 6 and Fig. 7. For each model, we average the results across all experiments. We select the best scores between using the *enhanced* (upgraded description, see Sec. 3.1) and *plain* (simple) versions of the textual prompt to represent the best model's ability.

Real-world videos consistently deliver optimal results across all experiments, as demonstrated by their minimal discard rates, high Dynamical scores (0.98-0.99), and consistently high Physical Invariance scores (above 0.93). These metrics confirm the reliability of real-world videos as benchmarks for physically accurate and realistic motion, and validate the correctness of our experimental setups, providing the upper boundary for the performance of the video generation models.

Enhanced prompts typically improve performance metrics compared to plain prompts (see Fig. 23), although this trend varies depending on the specific model. Enhanced prompting leads to higher physical invariance and dynamical scores and lower discard rates in many cases (e.g., COSMOS-predict2 and CogVideo), yet occasionally decreases performance in models such as COSMOS-predict1, indicating that prompt enhancement effectiveness is context-dependent.

Multi-frame prompting and key frame interpolation generally outperform single-frame prompting, achieving lower discard rates and higher dynamical and physical invariance scores, thereby demon-

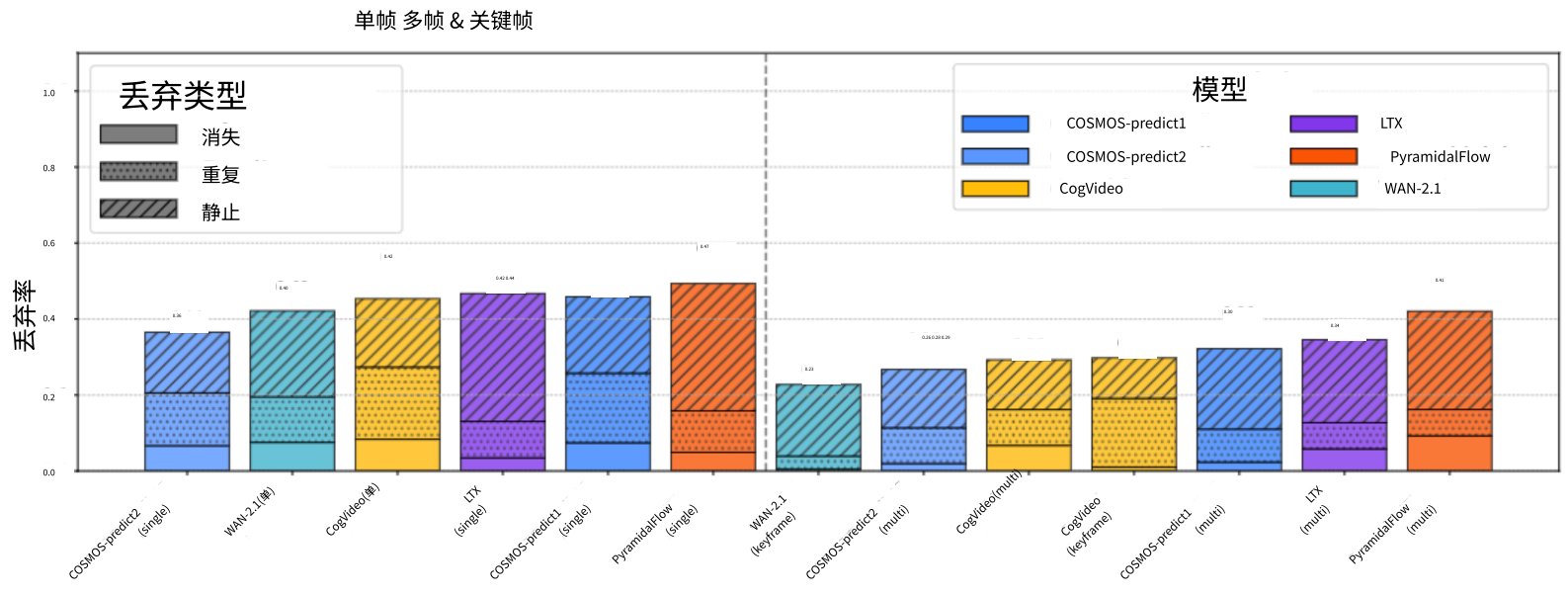


图 6：所有物理实验中的平均丢弃率（越低越好）。

其中 v = (v, v) 是球的速度，m 是它的质量。此外，势能为 V = mgy，其中 g 是重力加速度常数，y 是垂直坐标。因此，由于总能量是动能和势能之和，我们得到：E = T + V = m(v+ v) + mgy。

⇒ 能量-质量比。假设球的质量保持不变，我们推导出以下不变量：= (v+ v) + gy = const，我们可以用轨迹数据来估计它。

⇒ 加速度。由于除了重力之外没有其他外力作用于球，而重力是均匀且指向下方的，因此球的加速度也是恒定的：a= g = const。

⇒ 水平动量与质量比。与加速度类似，在无外力的情况下，水平动量 p= mV 也保持不变。因此，水平速度守恒：v= const。

计算物理不变性分数。为了将不变量转换为实际分数，类似于能量分数，我们计算不变量时间序列的标准差，并将其归一化到 (0, 1) 范围内，其中 1 表示完美的物理不变性分数。由于不变量本质上必须是恒定的，这些不变量（因此物理不变性分数较低）的高标准差，表明对相应物理不变量的建模质量较差。此外，对于被丢弃的轨迹，我们分配最小的物理不变性分数，等于 0。详细的分数计算步骤在附录 D.5 中描述。对于我们使用的每个不变量的推导，请参考附录 D.4。

5 分析

在本节中，我们分析了实验结果。我们分别以图 6 和图 7 展示了每个模型和条件类型（单帧、多帧或关键帧插值）的汇总结果。对于每个模型，我们通过所有实验的平均结果来呈现。我们选取使用增强版（升级描述，见第 3.1 节）和普通版文本提示的最佳得分，以代表最佳模型的能力。

真实世界视频在所有实验中始终提供最佳结果，这体现在其极低的丢弃率、高动态分数（0.98-0.99）以及持续保持的高物理不变性分数（高于 0.93）。这些指标证实了真实世界视频作为物理准确性和真实运动基准的可靠性，验证了我们实验设置的正确性，并为视频生成模型的表现提供了上限。

增强提示通常比普通提示能提升性能指标（见图 23），尽管这一趋势因具体模型而异。增强提示在很多情况下能带来更高的物理不变性和动态分数，以及更低的丢弃率（例如 COSMOSpredict2 和 CogVideo），但偶尔会降低 COSMOSpredict1 等模型的性能，这表明提示增强的有效性取决于具体模型。

多帧提示和关键帧插值通常优于单帧提示，实现更低的丢弃率和更高的动力学与物理不变性得分，从而

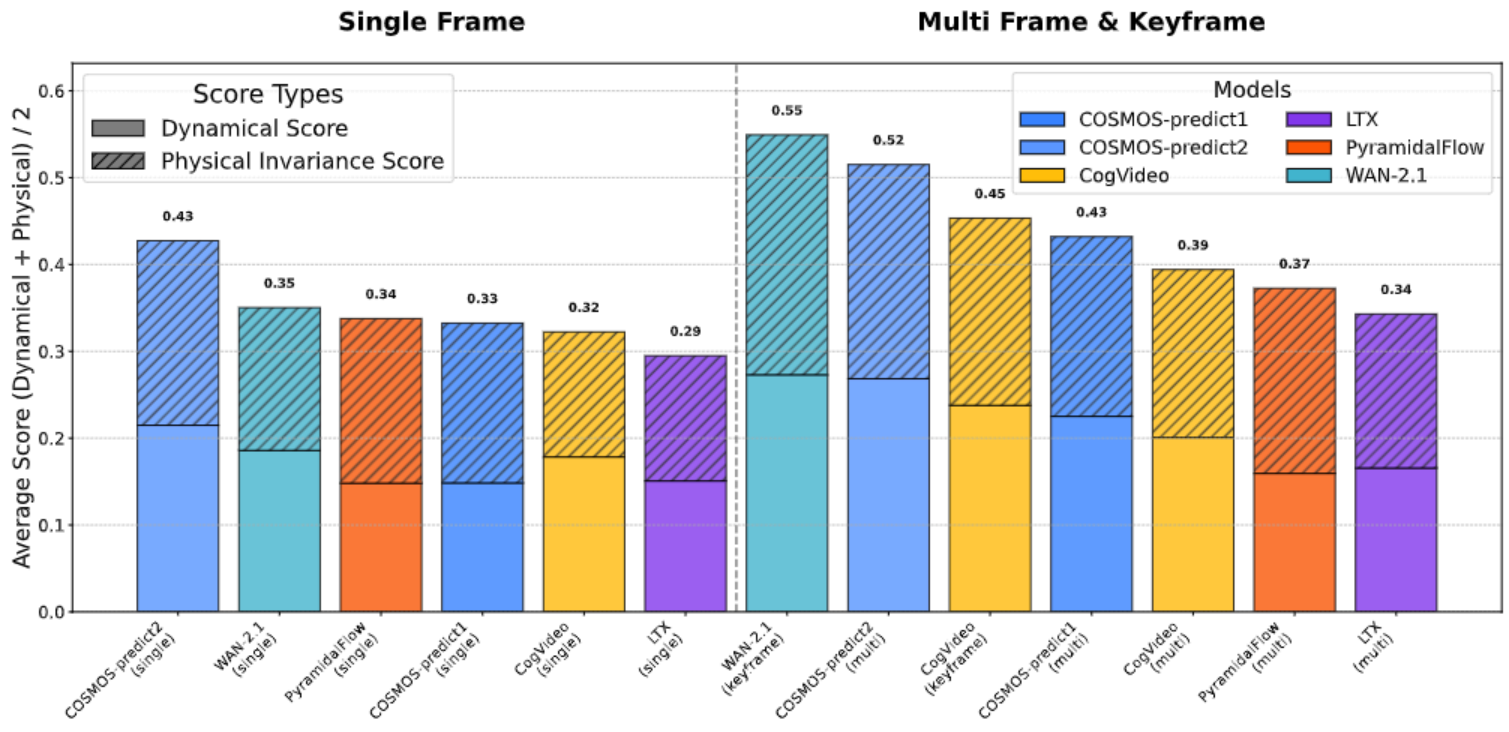


Figure 7: Aggregated scores across all physical experiments (higher is better).

strating the advantage of increased temporal context. The prompting with first and last frames, performs notably well in specific experiments (e.g., holonomic pendulum with WAN-2.1, see Fig. 24), suggesting a promising direction for improving temporal coherence and physical realism, though limiting the ability to generate from scratch.

Among the evaluated models, WAN-2.1 with key frame interpolation demonstrates superior performance, with an average total score of 0.55. However, since the last frame is provided, such prompting should be considered as an easier interpolation case vs. hard extrapolation for single and multi-frame conditioning. COSMOS-predict2 shows great results, leading in single frame conditioning category with the score of 0.43, and benefiting greatly from multi-frame conditioning both in reduced discard rates (-9%) and overall scores ($0.43 \rightarrow 0.52$), and excelling in experiments like bouncing and rolling (see Fig. 18). PyramidalFlow, while occasionally excelling in specific scenarios (e.g., spring, 0.65), exhibits inconsistent performance, with high variability and notably high discard rates. Other models show intermediate results. LTX can be considered the worst-performing model in the list.

The variability of discard rates across setups reflects the reliability of different models in generating physically plausible videos. Between models, average discard rates vary significantly, ranging from as low as 23.0% (WAN-2.1, key frame interpolation) to as high as 47% (PyramidalFlow, single frame). The variance across experiments is also noticeable, as shown in Fig. 19, Fig. 20, and Fig. 21. Some experiments tend to be prone to only certain types of errors, e.g., stillness in sliding experiments. The analysis of the major reasons (see Fig. 6) behind high discard rates reveals the absence of motion (i.e. *stillness*) and the presence of duplicate objects, as well as, to a lesser extent, the disappearance of the object from the video. These persistent shortcomings in the models' abilities to produce consistent and realistic videos are well-known (Huang et al., 2024b).

Overall, all generated models exhibit substantial limitations compared to real-world performance (see Fig. 8), underscoring the significant gaps remaining in simulating physically accurate dynamics.

6 Limitations

Morpheus is restricted to Newtonian physics under controlled settings, which ensures reproducibility but limits coverage of broader physical processes. To simplify evaluation, we assume negligible air resistance and friction; however, this reduces realism and can penalize physically plausible generations when these assumptions are violated. Next, some invariances are inherently easier to satisfy, making purely observational evaluation less comprehensive. In the future, the benchmark could be extended to cover controllable VGMs, where it would be possible to measure how robust the invariance score is under perturbations. Finally, evaluation remains confined to short videos and simple scenes due to the scope of the vision foundational models' applicability and the static camera requirement.

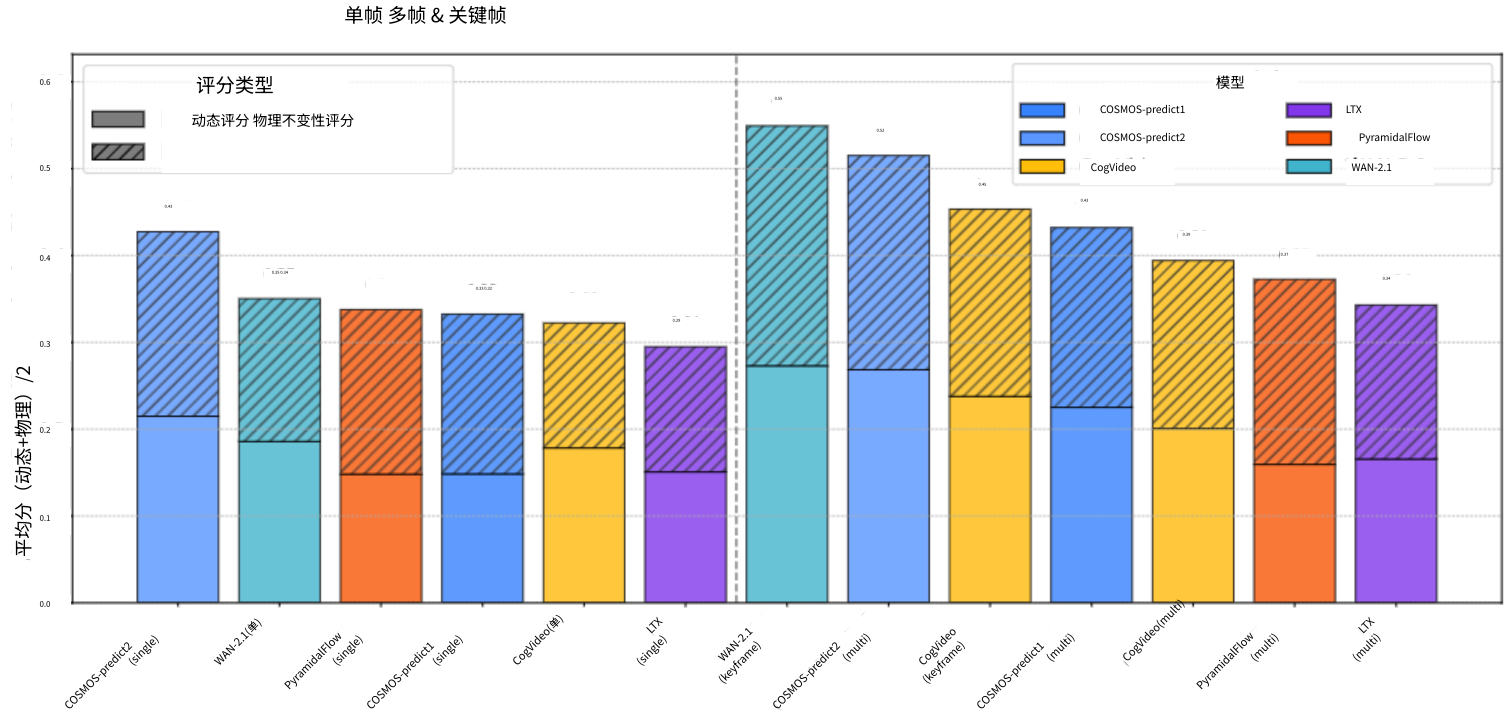


图 7：所有物理实验的汇总分数（越高越好）。

这表明了增加时间上下文的优势。使用第一帧和最后一帧的提示，在特定实验中表现突出（例如，使用 WAN-2.1 的完整摆，见图 24），这为提高时间连贯性和物理真实性指明了有前景的方向，尽管这限制了从零开始生成的能力。

在评估的模型中，使用关键帧插值的 WAN-2.1 表现最佳，平均总得分为 0.55。然而，由于最后帧是已知的，这种提示应被视为比单帧和多帧条件下的困难外推更简单的插值案例。COSMOS-predict2 表现出色，在单帧条件类别中得分最高，为 0.43，并且从多帧条件中受益匪浅，丢弃率降低了 9%，总体得分从 0.43 提升至 0.52，并在弹跳和滚动等实验中表现出色（见图 18）。PyramidalFlow 虽然偶尔在特定场景中表现优异（例如弹簧，得分为 0.65），但表现不稳定，变异性高，丢弃率显著偏高。其他模型表现中等。LTX 可被视为列表中最表现不佳的模型。

不同设置中丢弃率的差异反映了不同模型在生成物理上合理的视频时的可靠性。在模型之间，平均丢弃率差异显著，从低至 23.0%（WAN-2.1，关键帧插值）到高至 47%（PyramidalFlow，单帧）。实验之间的差异也很明显，如图 19、图 20 和图 21 所示。一些实验倾向于只出现特定类型的错误，例如滑动实验中的静止。对高丢弃率主要原因（见图 6）的分析揭示了运动缺失（即静止）和重复对象的存在，以及较小程度上对象从视频中消失。模型在生成连贯和逼真的视频方面的持续缺陷是众所周知的（Huang 等人，2024b）。

总体而言，与实际性能相比，所有生成的模型都表现出重大局限性（见图 8），突显了在模拟物理上准确的动力学方面仍存在巨大差距。

6 局限性

Morpheus 在受控环境下仅限于牛顿物理学，这确保了可重复性但限制了更广泛物理过程的研究范围。为简化评估，我们假设空气阻力和摩擦力可忽略不计；然而，这降低了真实性，当这些假设被违反时，可能会惩罚物理上合理的生成结果。接下来，某些不变性本质上更容易满足，使得纯观察性评估不够全面。未来，基准测试可以扩展到可控的视频生成模型，届时将能够测量不变性得分在扰动下的鲁棒性。最后，由于视觉基础模型的应用范围和静态摄像机要求，评估仍然局限于短视频和简单场景。

7 CONCLUSION

Our study highlights a fundamental limitation in current video generation models: despite their impressive realism, they fail to consistently adhere to physical laws. To address this gap, we introduced Morpheus, a benchmark designed to assess the physical reasoning capabilities of these models. Through a curated dataset of real-world physics experiments and physics-informed evaluation metrics, we demonstrate that even with advanced prompting techniques, existing models struggle to capture fundamental physical principles. In general, all models perform poorly, with significant violations of physical principles, though multi-frame prompting provides some improvement. This underscores the need for future research in integrating physical constraints into generative models.

CONTRIBUTIONS

CZ came up with the idea of measuring the physical reasoning ability of video generation models. EG, DC, AZ and CZ set up the initial meeting to structure the project into three parts: experiments recording, trajectory tracking and extracting, and scores calculation. DC and AT collected all the real-world videos with the help of DP. AT, TN and DP set up video generation code for all models with the help from AZ. AZ set up object segmentation and tracking and integrated all parts of the benchmark into automated pipeline with the help from TN and AT. AV completed the Dynamical scores evaluation, while CZ designed the Physical Invariant Scores. The paper was written by CZ, DC, AZ and EG with DC and DP contributing to Fig. 1, 3, 4, 5, AV created Fig. 9, CZ created Fig. 8 and TN contributed to Fig. 2. AZ created Table 1 and made all the experimental results aggregation and visualization in Fig. 6, 7 with the help from MB. TN, AV and DC worked on the final version of the appendix. The website development was coordinated by AZ with contributions from CZ (leaderboard), DC (videos), MB (overall structure and plots).

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai, 2025.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024a. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024b.

7 结论

我们的研究突显了当前视频生成模型的一个基本局限性：尽管它们表现出令人印象深刻的逼真度，但它们无法始终遵循物理定律。为了弥补这一差距，我们引入了 Morpheus，这是一个旨在评估这些模型物理推理能力的基准。通过一个精心策划的真实世界物理实验数据集和物理信息评估指标，我们证明，即使使用先进的提示技术，现有模型也难以捕捉基本的物理原理。总的来说，所有模型的性能都不佳，存在明显的物理原则违反，尽管多帧提示提供了一些改进。这突显了未来研究将物理约束集成到生成模型中的必要性。

贡献

CZ 提出了测量视频生成模型物理推理能力的主意。EG、DC、AZ 和 CZ 召开初始会议，将项目分为三个部分：实验记录、轨迹跟踪与提取，以及评分计算。DC 和 AT 在 DP 的帮助下收集了所有真实世界视频。AT、TN 和 DP 在 AZ 的帮助下为所有模型搭建了视频生成代码。AZ 搭建了目标分割与跟踪，并在 TN 和 AT 的帮助下将基准测试的所有部分整合到自动化流程中。AV 完成了动态评分评估，而 CZ 设计了物理不变性评分。论文由 CZ、DC、AZ 和 EG 撰写，DC 和 DP 贡献了图 1、3、4、5，AV 创建了图 9，CZ 创建了图 8，TN 贡献了图 2。AZ 创建了表 1，并在 MB 的帮助下完成了图 6、7 中的所有实验结果汇总与可视化。TN、AV 和 DC 负责附录的最终版本。网站开发由 AZ 协调，CZ（排行榜）、DC（视频）、MB（整体结构与图表）均有贡献。

参考文献

尼基特・阿格拉瓦尔，阿斯拉南・阿里，马切伊・巴拉，约格什・巴拉吉，埃里克・巴克，蒂芙尼・蔡，普里特维吉特・查特帕德亚，陈永新，崔寅，丁一帆，等。宇宙世界基础模型平台，用于物理人工智能，2025。

Pravesh Agrawal、Szymon Antoniak、Emma Bou Hanna、Baptiste Bout、Devendra Chaplot、Jessica Chudnovsky、Diogo Costa、Baudouin De Monicault、Saurabh Garg、Theophile Gervet 等。Pixtral 12b。arXiv 预印本 arXiv:2410.07073，2024 年。

Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, 等人. Cosmos-transfer1：条件世界生成与自适应多模态控制. arXiv 预印本 arXiv:2503.14492, 2025.

安东・巴赫京、劳伦斯・范・德・马滕、贾斯汀・约翰逊、劳拉・古斯塔夫森和罗斯・吉尔希克。Phyre：物理推理的新基准。神经信息处理系统进展，32，2019。

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.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, 等. Lumiere: 一种用于视频生成的时空扩散模型. 在 SIGGRAPH Asia 2024 会议论文集中, 第 1-11 页, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, Aditya Ramesh. 视频生成模型作为世界模拟器. 2024a. URL https://openai.com/research/ video-generation-models-as-world-simulators.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, 等. 视频生成模型作为世界模拟器, 2024b.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023.

Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.

Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, pp. 100152, 2024.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

GaoYuan He, YongXiang Zhao, and ChuLiang Yan. Mflp-pinn: A physics-informed neural network for multiaxial fatigue life prediction. *European Journal of Mechanics-A/Solids*, 98:104889, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024b.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

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. *arXiv preprint arXiv:2411.02385*, 2024.

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

Kling AI. Kling ai. https://klingai.com/, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Junkun Yuan, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yanxin Long, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2024.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, 和 Joon-Young Lee. Tracking anything with decoupled video segmentation. In ICCV, 2023.

约瑟夫·乔，法赫丽娜·德维·普斯皮塔萨里，郑胜，郑景瑶，李立杭，金泰浩，洪钟善，张超宁. Sora 作为通用人工智能世界模型？关于文本到视频生成的全面综述. arXiv 预印本 arXiv:2403.05131, 2024.

魏超，毛嘉庚，李伯毅，Daniel Seita，Vitor Guizilini 和王越。Physbench：用于物理世界理解的视觉语言模型的基准测试和增强。arXiv 预印本 arXiv:2501.16411，2025。

萨尔瓦托雷·库莫, 文琴佐·希阿诺·迪科拉, 法比奥·吉安帕奥洛, 吉安卢吉·罗扎, 马兹亚尔·拉伊西,

以及弗朗切斯科·皮奇亚利。基于物理信息神经网络的科学机器学习：我们当前的位置和未来展望。《科学计算杂志》，92(3):88，2022。

邓浩歌，潘婷，貂海文，罗正雄，崔宇峰，陆胡川，山时光，戚永刚，王新龙. 无向量量化的自回归视频生成.
arXiv 预印本 arXiv:2412.14169, 2024.

Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. Aigcbench: Comprehensive evaluation of 由人工智能生成的图像到视频内容。 BenchCouncil 基准、标准和评估交易，第 100152 页，2024 年。

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, 和 William Yang Wang。 Tc-bench：测试文本到视频和图像到视频生成中的时间组合性。 arXiv 预印本 arXiv:2406.08656，2024 年。

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, 等人。 Ltx-video：实时视频潜在扩散。
arXiv 预印本 arXiv:2501.00103, 2024.

高元和, 赵永祥, 和严楚良. Mflp-pinn: 一种物理信息神经网络
用于多轴疲劳寿命预测. 欧洲力学-A/固体, 98:104889, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video 生成模型. 在 IEEE/CVF 计算机视觉与模式识别会议论文集中, pp. 21807–21818, 2024a.

黄子琪, 张帆, 许晓杰, 何一南, 余嘉硕, 董子越, 马前力, Chanpaisit Nattapol, 史晨阳, 姜宇明, 等. Vbench++：视频生成模型的全面且通用的基准测试套件. arXiv 预印本 arXiv:2411.13503, 2024b.

杨金，孙志成，李宁源，徐坤，江浩，庄南，黄曲哲，宋阳，穆亚东，以及林卓辰。金字塔流匹配用于高效视频生成建模。
arXiv preprint arXiv:2410.05954, 2024.

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. arXiv preprint arXiv:2411.02385, 2024.

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

Kling AI. Kling ai. https://klingai.com/, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Junkun Yuan, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yanxin Long, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2025.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024a.

Yongtuo Liu, Sara Magliacane, Miltiadis Kofinas, and Efstratios Gavves. Amortized equation discovery in hybrid dynamical systems. *arXiv preprint arXiv:2406.03818*, 2024b.

Ziming Liu and Max Tegmark. Machine learning conservation laws from trajectories. *Physical Review Letters*, 126(18):180604, 2021.

Jianwen Luo, Kui Ying, and Jing Bai. Savitzky–golay smoothing and differentiation filter for even number data. *Signal Processing*, 85(7):1429–1434, 2005. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2005.02.002. URL https://www.sciencedirect.com/science/article/pii/S0165168405000654.

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

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025.

NVIDIA. Mistral and nvidia. mistral-nemo-12b-instruct: A 12b parameter large language model,. https://huggingface.co/nvidia/Mistral-NeMo-12B-Instruct, 2024.

Adeel Pervez, Francesco Locatello, and Efstratios Gavves. Mechanistic neural networks for scientific machine learning. *arXiv preprint arXiv:2402.13077*, 2024.

Adeel Pervez, Efstratios Gavves, and Francesco Locatello. Mechanistic pde networks for discovery of governing equations. *arXiv preprint arXiv:2502.18377*, 2025.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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*, pp. 8821–8831. Pmlr, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

李一帆，杜一帆，周坤，王金鹏，赵文新，文继荣. 评估大型视觉语言模型中的物体幻觉. arXiv 预印本 arXiv:2305.10355, 2023.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. arXiv 预印本 arXiv:2412.00131, 2024.

石龙 刘, 曾昭阳, 任天鹤, 李峰, 张浩, 杨杰, 姜庆, 李春元, 杨建伟, 苏航, 等. Grounding dino: 将 dino 与有监督预训练结合用于开放集目标检测. 在欧洲计算机视觉会议, 第 38-55 页. 斯普林格出版社, 2025.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large 视频生成模型。在 IEEE/CVF 计算机视觉与模式识别会议论文集中，第 22139–22149 页，2024a。

Yongtuo Liu, Sara Magliacane, Miltiadis Kofinas, 和 Efstratios Gavves。混合动力系统的摊销方程发现。arXiv 预印本 arXiv:2406.03818，2024b。

刘志明和马克斯·泰格马克. 基于轨迹的机器学习守恒定律. 物理评论快报, 126(18):180604, 2021.

罗建文，英奎，白静。Savitzky–Golay 平滑和微分滤波器用于偶数数据。信号处理，85(7):1429–1434，2005。ISSN 0165-1684。doi: https://doi。org/10.1016/j.sigpro.2005.02.002. URL https://www.sciencedirect.com/science/article/pii/S0165168405000654.

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

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, 和 Robert Geirhos. 生成式视频模型是否从观看视频中学习物理原理？arXiv 预印本 arXiv:2501.09038, 2025.

NVIDIA. Mistral 和 nvidia. mistral-nemo-12b-instruct: 一个 12b 参数的大型语言模型，https://huggingface.co/nvidia/Mistral-NeMo-12B-Instruct, 2024.

Adeel Pervez, Francesco Locatello, 和 Efstratios Gavves. 机制神经网络用于科学机器学习. arXiv 预印本 arXiv:2402.13077, 2024.

Adeel Pervez, Efstratios Gavves, 和 Francesco Locatello. 机制偏微分方程网络用于发现控制方程. arXiv 预印本 arXiv:2502.18377, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, 和 Robin Rombach. Sdxl: 改进用于高分辨率图像合成的潜在扩散模型. arXiv 预印本 arXiv:2307.01952, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, 和 Peter J Liu。使用统一的文本到文本 Transformer 探索迁移学习的极限。机器学习研究杂志，21(140):1–67，2020。

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，第 8821-8831 页。Pmlr，2021。

Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman Rädle、Chloe Rolland、Laura Gustafson 等。Sam 2：图像和视频中分割一切。arXiv 预印本 arXiv:2408.00714，2024。

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

R Charles Swanson and Eli Turkel. On central-difference and upwind schemes. *Journal of computational physics*, 101(2):292–306, 1992.

Veo-Team, :, Agrim Gupta, Ali Razavi, Andeep Toor, Ankush Gupta, Dumitru Erhan, Eleni Shaw, Eric Lau, Frank Belletti, Gabe Barth-Maron, Gregory Shaw, Hakan Erdogan, Hakim Sidahmed, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jeff Donahue, José Lezama, Kory Mathewson, Kurtis David, Matthieu Kim Lorrain, Marc van Zee, Medhini Narasimhan, Miaosen Wang, Mohammad Babaeizadeh, Nelly Papalampidi, Nick Pezzotti, Nilpa Jha, Parker Barnes, Pieter-Jan Kindermans, Rachel Hornung, Ruben Villegas, Ryan Poplin, Salah Zaiem, Sander Dieleman, Sayna Ebrahimi, Scott Wisdom, Serena Zhang, Shlomi Fruchter, Signe Nørly, Weizhe Hua, Xinchen Yan, Yuqing Du, and Yutian Chen. Veo 2. 2024. URL https://deepmind.google/technologies/veo/veo-2/.

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*, 2025a.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Zihan Wang, Songlin Li, Lingyan Hao, Bowen Song, and Xinyu Hu. What you see is what matters: A novel visual and physics-based metric for evaluating video generation quality. *arXiv preprint arXiv:2411.13609*, 2024b.

Dirk Weissenborn, Jakob Uszkoreit, and Oscar Täckström. Scaling autoregressive video models. In *ICLR*, 2020.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7395–7405, 2024.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024a.

Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-net：用于生物医学的卷积网络图像分割。在《医学图像计算与计算机辅助干预–MICCAI 2015：第 18 届国际会议，德国慕尼黑，2015 年 10 月 5-9 日，会议录，第 III 部分 18》，第 234-241 页。斯普林格出版社，2015 年。

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, 等. 具有深度语言理解的逼真文本到图像扩散模型. 神经信息处理系统进展, 35:36479–36494, 2022.

R Charles Swanson 和 Eli Turkel. 关于中心差分和迎风格式. 计算物理杂志, 101(2):292–306, 1992.

Veo-Team, :, Agrim Gupta, Ali Razavi, Andeep Toor, Ankush Gupta, Dumitru Erhan, Eleni Shaw, Eric Lau, Frank Belletti, Gabe Barth-Maron, Gregory Shaw, Hakan Erdogan, Hakim Sidahmed, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jeff Donahue, José Lezama, Kory Mathewson, Kurtis David, Matthieu Kim Lorrain, Marc van Zee, Medhini Narasimhan, Miaosen Wang, Mohammad Babaeizadeh, Nelly Papalampidi, Nick Pezzotti, Nilpa Jha, Parker Barnes, Pieter-Jan Kindermans, Rachel Hornung, Ruben Villegas, Ryan Poplin, Salah Zaiem, Sander Dieleman, Sayna Ebrahimi, Scott Wisdom, Serena Zhang, Shlomi Fruchter, Signe Nørly, Weizhe Hua, Xinchen Yan, Yuqing Du, 和 Yutian Chen. Veo 2, 2024. URL https://deepmind.google/technologies/veo/veo-2/.

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, 2025a.

团队 Wan、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang 等。Wan：开放和高级大规模视频生成模型。arXiv 预印本 arXiv:2503.20314, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024a.

Zihan Wang, Songlin Li, Lingyan Hao, Bowen Song, 和 Xinyu Hu. 你看到的就是重要的：一种基于视觉和物理的新颖度量标准，用于评估视频生成质量。arXiv 预印本 arXiv:2411.13609, 2024b.

Dirk Weissenborn, Jakob Uszkoreit, 和 Oscar Täckström. 自回归视频模型的扩展。在 ICLR，2020 年。

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, 等人. Art-v：自回归文本到视频生成扩散模型。在 IEEE/CVF 计算机视觉与模式识别会议论文集中，第 7395–7405 页，2024 年。

金博翔，夏梦涵，张勇，陈浩新，余望博，刘汉源，刘公业，王新涛，山英，黄天赐。Dynamicrafter：用视频扩散先验动态生成开放域图像。在《欧洲计算机视觉会议》，第 399-417 页。斯普林格出版社，2025 年。

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. arXiv:2406.09414, 2024a.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

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. *CoRR*, 2024b.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Zhihan Liu, Zhiyuan Liu, Jialiang Tang, Qiang Liu, and Jie Tang. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, 和 Ming-Hsuan Yang。扩散模型：方法和应用的综合综述。ACM Computing Surveys，56(4):1–39，2023 年。

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, 等。Cogvideox：具有专家变压器的文本到视频扩散模型。CoRR，2024b。

李丽，石 Bowen，帕苏努拉 Ramakanth，穆勒 Benjamin，戈洛夫纳娃 Olga，王天路，阿伦 Babu, Binh Tang, Brian Karrer, Shelly Sheynin, 等. 自回归多模态模型的扩展：预训练和指令微调. arXiv 预印本 arXiv:2309.02591, 2(3), 2023.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Zhihan Liu, Zhiyuan Liu, Jialiang Tang, Qiang Liu, 和 Jie Tang. Chatglm：从 glm-130b 到 glm-4 的全功能大型语言模型家族. arXiv 预印本 arXiv:2406.12793, 2024.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018, 2022.

APPENDIX

A DATASET

Overall, we conducted a set of 9 core physical experiments, highlighting different physical principles, including:

1. Falling: Objects dropped from rest until they make impact with the surface, used to test uniform gravitational acceleration and energy conservation.
2. Projectile motion: A ball launched at various initial velocities and angles, testing the preservation of momentum and energy, as well as the uniformity of gravity.
3. Rolling: A metal can rolling from a slope, with energy conservation.
4. Sliding: A book sliding from a slope, with energy conservation.
5. Holonomic pendulum: A ball affixed to a rigid rod, with periodic motion and energy conservation.
6. Double pendulum: A more complex system with a pendulum attached to another pendulum, illustrating chaotic behavior and conservation laws in nonlinear dynamics.
7. Bouncing: A ball observed from the moment it first impacts the surface until it rebounds and impacts again, testing gravitational acceleration and energy conservation in a more challenging setting.
8. Collision: Two metal balls with the same or different masses collide with each other. One of them is initially stationary and the other one collides with it.
9. Spring: A weight hanging below a vertical spring, perform up and down simple harmonic vibration with energy conservation.

For each system, we recorded multiple times the type of experiment trying to have homogenous videos, while after a few iterations, we varied the initial conditions or configuration parameters. Table 1 below summarizes the number of recordings and configurations for each experiment.

Experiment	**Videos**	**Factors of Variation**	**Configuration Initial Condition Description**
Falling	20	2	Object type and height from which the object was released.
Projectile	15	3	Angle of launch, slingback extension levels, launched ball color.
Bouncing ball	12	1	Heights from which the ball was released before bouncing.
Holonomic Pendulum	22	1	Initial angle from the vertical (zero-degree resting position).
Double Pendulum	10	1	Initial height of the second (top) pendulum bob.
Rolling	15	2	Incline angle of the ramp from which the object was released and object type.
Sliding	10	1	Incline angle of the ramp (slope) from which the object was released.
Collision	12	3	Masses of the colliding objects and their initial velocities before impact.
Spring	7	1	Magnitude of the initial force/impulse used to displace the mass from rest.

Table 1: Summary of the experimental dataset from real-world recorded videos.

For each experiment, we provide representative frames: falling objects in Figure 9; bouncing, projectile motion, holonomic pendulum, and double pendulum in Figure 10; rolling in Figure 11; and sliding, collision, and spring in Figure 12.

附录

一个数据集

总体而言，我们进行了一系列 9 个核心物理实验，涵盖了不同的物理原理，包括：

1. 下落：物体从静止状态下落直到与表面碰撞，用于测试均匀重力加速度和能量守恒。

2. 抛体运动：一个以不同初始速度和角度发射的球，测试动量和能量的保持，以及重力的均匀性。

3. 滚动：一个金属罐从斜坡上滚下，能量守恒。
4. 滑动：一本书从斜坡上滑下，能量守恒。
5. 保守摆：一个固定在刚性杆上的球，具有周期运动和能量守恒。
双摆：一个更复杂的系统，其中一个摆连接到另一个摆上，展示了非线性动力学中的混沌行为和守恒定律。

7. 弹跳：从球第一次撞击表面到它反弹并再次撞击的整个过程，在一个更具挑战性的环境中测试重力加速度和能量守恒。

8. 碰撞：两个质量相同或不同的金属球相互碰撞。其中一个最初静止，另一个与之碰撞。

9. 弹簧：一个重物悬挂在垂直弹簧下方，进行上下简谐振动，并保持能量守恒。

对于每个系统，我们多次记录了实验类型，试图获得均匀的视频，而在几次迭代后，我们改变了初始条件或配置参数。
下表总结了每个实验的记录次数和配置。

实验	视频	因素变化	配置 初始条件 描述
坠落	20	2	物体类型和释放物体的高度。
抛射体	15	3	发射角度、弹弓回弹伸展级别、发射球的颜色。
弹跳的球	12	1	球在弹跳前被释放的高度。
完整摆动器	22	1	初始角度与垂直方向（零度静止位置）。
双摆	10	1	第二个（顶部）摆球的初始高度。
滚动	15	2	物体释放的斜坡倾角和物体类型。
滑动	10	1	物体释放的斜坡（斜率）倾角。
碰撞	12	3	碰撞物体的质量和它们在碰撞前的初始速度。
弹簧	7	1	用于使物体从静止状态移位的初始力/冲量的大小。

表 1：从真实世界记录视频中获取的实验数据集摘要。

对于每个实验，我们提供代表性帧：下落物体在图 9 中；弹跳、抛射运动、完整摆和双摆运动在图 10 中；滚动在图 11 中；以及滑动、碰撞和弹簧在图 12 中。

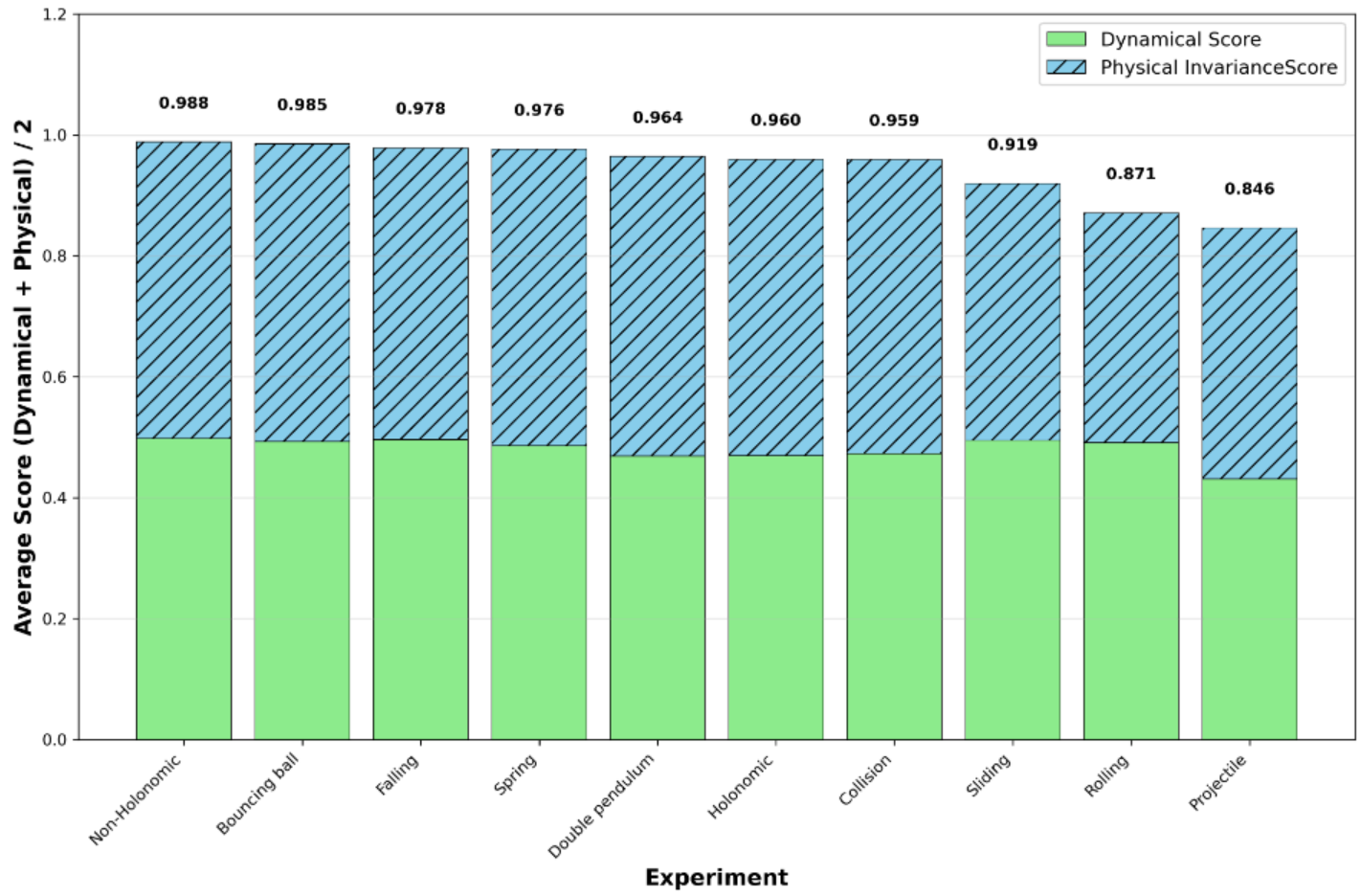


Figure 8: Morpheus scores for each experiment on real-world videos.

Falling For the falling experiment, we started with a standard table tennis orange ball as the simplest object. A mechanical actuator was used to hold the ball at a certain height (initial position) and as a release mechanism to control the moment the ball fell free, before making contact with the surface below. Different height levels from the surface were used as initial positions, resulting in trajectories with different lengths (smaller or larger). In addition to the ball, we conducted extra experiments using different everyday objects, namely a plastic whiteboard marker, an adhesive tape roll, and an apple. Unlike actuator controlled experiments, these objects were released directly from the hand of a person at varying initial heights. This setup introduced additional variability in the conditions and the orientation of the object.

Bouncing ball The bouncing ball experiment begins immediately after the falling ball makes impact with the surface. It focuses on observing the ball during its bounce, capturing its trajectory as it rebounds upwards after contact with the surface.

Projectile For this experiment, a custom 3D printed projectile was built, along with three different balls of the same plastic material but of different colors. The projectile works with string rubber bands following the same principle of a slingback. During our recordings, we varied three different parameters. The angle of the launch for the ball, the force with which the ball was launched into the air, and the color of the ball.

Holonomic pendulum For this setting, a rigid metal structure consisting of a pole, perpendicular to the ground, on which a solid metal stick was mounted. The joint holding the stick was adjusted to allow for a normal friction coefficient, resulting in an intuitive retrogressive back-and-forth movement simulating a typical pendulum oscillatory trajectory. At the end of the metal stick a small table tennis ball was attached, as the SAM2 predictor can confidently track the center of the ball aligning with the central axis at the end of the stick. Using the zero angle as the resting position, we varied the angle at which the pendulum was released resulting in distinct retrogressive trajectories. As in the falling ball experiment the same release mechanism model was employed to manipulate the moment the pendulum was let freely to swing.

Motor Model: T825, Motor serial number: 00362129

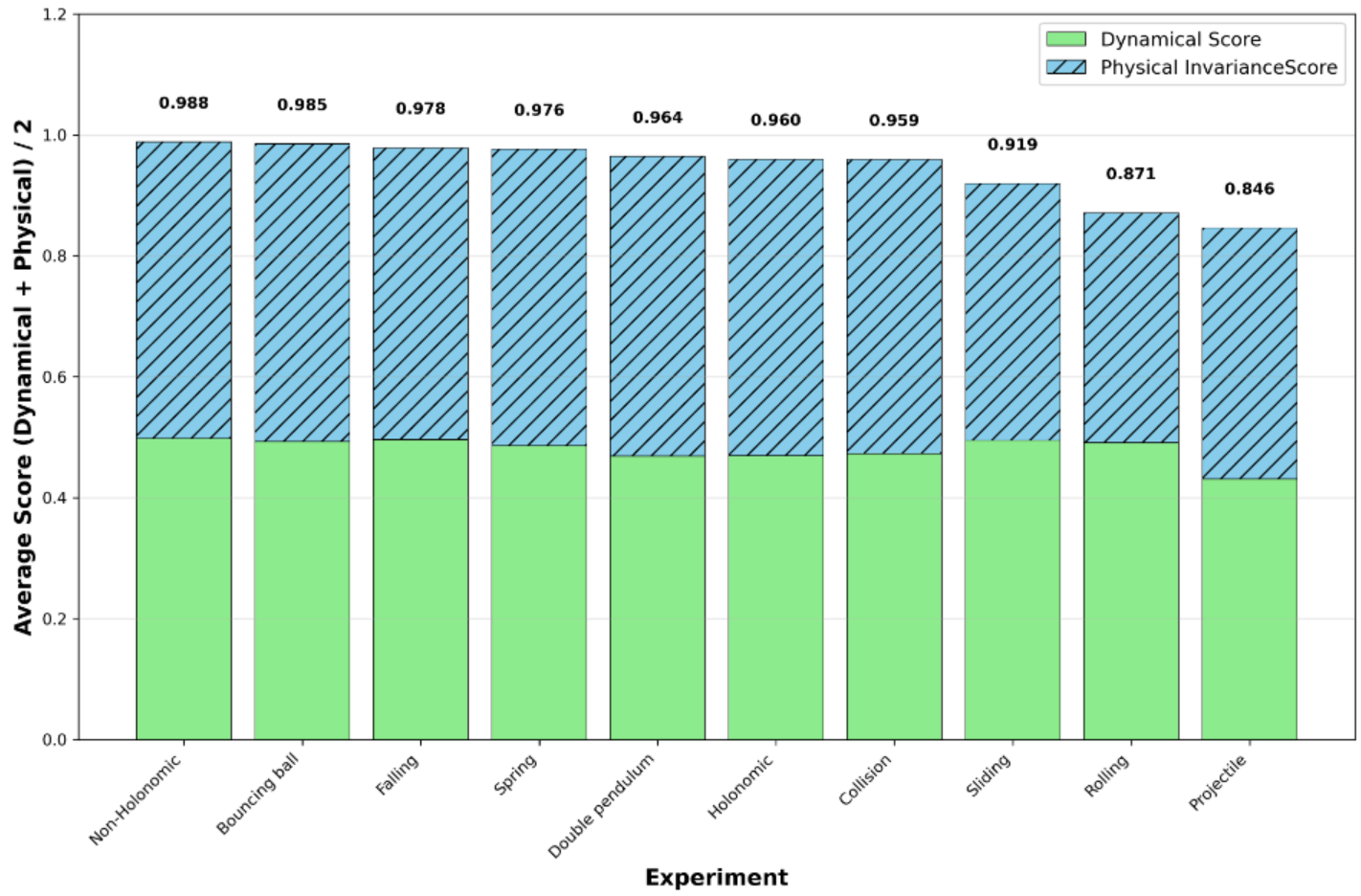


图 8：真实世界视频上每个实验的 Morpheus 分数。

在坠落实验中，我们以一个标准的乒乓球作为最简单的物体。使用机械执行器将球固定在特定高度（初始位置），并作为释放机制来控制球自由落下的时刻，在球接触下方表面之前。使用从表面不同高度作为初始位置，导致轨迹长度不同（较小或较大）。除了球之外，我们还使用不同的日常物品进行了额外实验，即一个塑料白板笔、一个胶带卷和一个苹果。与执行器控制实验不同，这些物体是从不同初始高度的人手中直接释放的。这种设置增加了条件和物体方向上的额外可变性。

弹跳球 弹跳球实验在落球撞击表面后立即开始。它专注于观察球在弹跳过程中的运动，捕捉球在接触表面后向上反弹的轨迹。

投射物 在这个实验中，我们制作了一个定制的 3D 打印投射物，以及三个由相同塑料材料制成但颜色不同的球。投射物使用橡皮筋，遵循与回旋镖相同的原理。在我们的录制过程中，我们改变了三个不同的参数：球的发射角度、球被发射到空中的力量，以及球的颜色。

全向摆 对于这个设置，是一个由一根垂直于地面的杆和一根固定在杆上的实心金属棒组成的刚性金属结构。连接棒关节的摩擦系数调整为正常值，从而产生直观的来回摆动，模拟典型的摆动轨迹。金属棒的末端固定了一个小乒乓球，因为 SAM2 预测器可以自信地跟踪球的中心与棒末端的中心轴对齐。以零角度作为静止位置，我们改变了摆释放的角度，从而产生了不同的来回轨迹。与下落球实验一样，采用了相同的释放机制模型来控制摆被自由摆动的时间。

电机模型：T825，电机序列号：00362129

Double pendulum A custom structure consisting of a wooden base, a metal pole mounted on the top of the base, and a joint mounted at a degrees angle to the center axis of the pole, to keep the longer bob of the pendulum in place. These structures ensure that each 3D printed plastic bobs of the pendulum can rotate freely with normal friction resulting in the typical chaotic motion double pendulum are known for. A double pendulum consists of two bobs attached end-to-end. Each pendulum has its angle relative to the vertical. The same release mechanism as in previous experiments is utilized to define the starting position of each pendulum link. This starting position can be described as the angle each bob makes with the vertical when it is still stationary.

Rolling For this setting, we examined objects that roll down an inclined surface. We used three different objects: a full can, which was sealed, an empty can, and an orange. The slope of the surface was adjustable, allowing us to vary the steepness of the slope. The objects were placed by hand at the top of the slope before being released. Due to different mass distributions and shapes, we had different rolling behaviors across these three objects.

Sliding For this experiment, we investigated the sliding motion on an inclined surface using a flat book. We varied the slope of the surface between trials. The book was placed by hand at the top of the slope before being released. Depending on the angle of the slope, the trajectories exhibited smooth sliding motion.

Collision In the collision experiment, we studied collisions between objects of different sizes. Three setups were recoded: A large object collides with a smaller one, two objects of equal size collide with each other, and a small object collides with a larger one. For all cases, the initial velocity of the moving object (leftmost object with the right one at rest) was introduced with a gentle push at random speeds. These settings produced a diverse outcome for the aforementioned experiment, depending on the relative mass of the objects and the initial velocity.

Spring In this spring experiment, we analyzed the oscillatory motion of a cylindrical metal weight suspended from a vertical spring. The object was initially displaced from its rest position by hand with a small (but random) force and then released. The amplitude of oscillation was defined by this displacement and the restoring force led the object to move in a vertical periodic movement until equilibrium. Initial force was the main reason for the motion, with no actuator being employed, and gradually over time decayed due to damping effects.

B AUGMENTATION WITH COSMOS-TRANSFER1

We augment a subset of experiments with style transfer to diversify object appearances while preserving underlying physical dynamics, creating initializations more representative of typical VGM training data. Specifically, we apply COSMOS-Transfer to five experiment types—Falling ball, Projectile, Holonomic pendulum, Rolling, and Sliding. For each experiment type, we select representative videos and generate style-transferred variants that alter object semantics and appearance while retaining motion consistency. This yields three transferred variants per original video (two for Holonomic pendulum), providing diverse yet physically plausible conditioning scenarios. Per experiment type, this produces $3 \times 3 = 9$ augmented conditioning scenarios for most experiments, and $3 \times 2 = 6$ for Holonomic pendulum.

The transfer process uses per-frame object masks from SAM2 Ravi et al. (2024), reusing the same masks from our trajectory-extraction pipeline to isolate only the object under study. We provide concise, object-focused prompts describing the desired replacement. All transfers undergo manual screening for visual and physical plausibility. When artifacts such as camera motion, temporal drift, hallucinations, shape misalignment, or mask leakage are observed, we enable Canny-edge guidance in COSMOS-Transfer with an edge-conditioning weight of 0.5, with the object mask now accounting for the other 0.5 conditioning weight. This additional constraint preserves scene structure while still allowing complete style transformation.

Figure 13 shows representative examples comparing the first frames of original videos with their style-transferred variants. These augmentations effectively multiply our conditioning scenarios: each original video contributes multiple physically consistent variants, expanding the diversity of initial conditions for VGM evaluation.

双摆一个定制结构，由木质底座、安装在底座顶部的金属杆以及安装在杆中心轴成一定角度的关节组成，以固定摆锤较长的一端。这些结构确保每个 3D 打印的塑料摆锤可以自由旋转，正常摩擦力导致双摆产生典型的混沌运动。双摆由两个末端相连的摆锤组成。每个摆锤相对于垂直方向的角度。利用与先前实验相同的释放机制来定义每个摆锤链接的起始位置。这个起始位置可以描述为摆锤静止时与垂直方向形成的角度。

滚动 在这个设置中，我们考察了物体沿斜面滚下的情况。我们使用了三种不同的物体：一个密封的完整罐头、一个空罐头和一个橙子。斜面的坡度是可调节的，使我们能够改变斜面的陡峭程度。物体由手放置在斜面的顶部，然后释放。由于质量分布和形状不同，这三种物体表现出不同的滚动行为。

滑动 在这个实验中，我们使用一本平本书研究了在斜面上的滑动运动。我们通过试验改变了表面的坡度。书被手放在斜面的顶部，然后释放。根据斜面的角度，轨迹表现出平滑的滑动运动。

碰撞 在碰撞实验中，我们研究了不同尺寸物体之间的碰撞。重新录制了三种设置：一个大型物体与一个较小物体碰撞，两个相同尺寸的物体相互碰撞，以及一个较小物体与一个较大物体碰撞。在所有情况下，运动物体的初始速度（左侧物体与右侧静止物体）通过随机速度的轻推引入。这些设置产生了多样化的结果，具体取决于物体的相对质量和初始速度。

在这个春季实验中，我们分析了悬挂在垂直弹簧上的圆柱形金属重物的振荡运动。该物体最初被手施加一个小的（但随机的）力从其平衡位置移开，然后释放。振荡的振幅由这个位移定义，恢复力使物体在垂直周期性运动中运动，直到达到平衡。初始力是运动的主要原因，没有使用执行器，并且随着时间的推移由于阻尼效应逐渐衰减。

B 使用 COSMOS-TRANSFER1 进行增强

我们通过在部分实验中添加风格迁移来多样化物体外观，同时保留底层物理动态，从而创建更符合典型 VGM 训练数据的初始化。具体来说，我们将 COSMOS-Transfer 应用于五种实验类型——下落球、抛射体、完整摆、滚动和滑动。对于每种实验类型，我们选择具有代表性的视频，并生成风格迁移的变体，这些变体改变物体语义和外观，同时保持运动一致性。这为每个原始视频生成了三个迁移变体（完整摆为两个），提供了多样化且物理上合理的条件场景。对于每种实验类型，大多数实验产生了 3 × 3 = 9 个增强条件场景，而完整摆产生了 3 × 2 = 6 个。

迁移过程使用来自 SAM2 Ravi 等人（2024）的逐帧对象掩码，重用我们轨迹提取流程中的相同掩码，以仅隔离研究对象。我们提供简洁、以对象为中心的提示来描述所需的替换。所有迁移都经过人工筛选，以检查视觉和物理上的合理性。当观察到相机运动、时间漂移、幻觉、形状错位或掩码泄漏等伪影时，我们在 COSMOS-Transfer 中启用 Canny 边缘引导，边缘调节权重为 0.5，此时对象掩码承担剩余的 0.5 调节权重。这一附加约束在保持场景结构的同时，仍允许完整的风格转换。

图 13 展示了原始视频第一帧与其风格迁移变体的代表性对比示例。这些增强有效地增加了我们的调节场景：每个原始视频都贡献多个物理一致的变体，扩展了 VGM 评估的初始条件多样性。

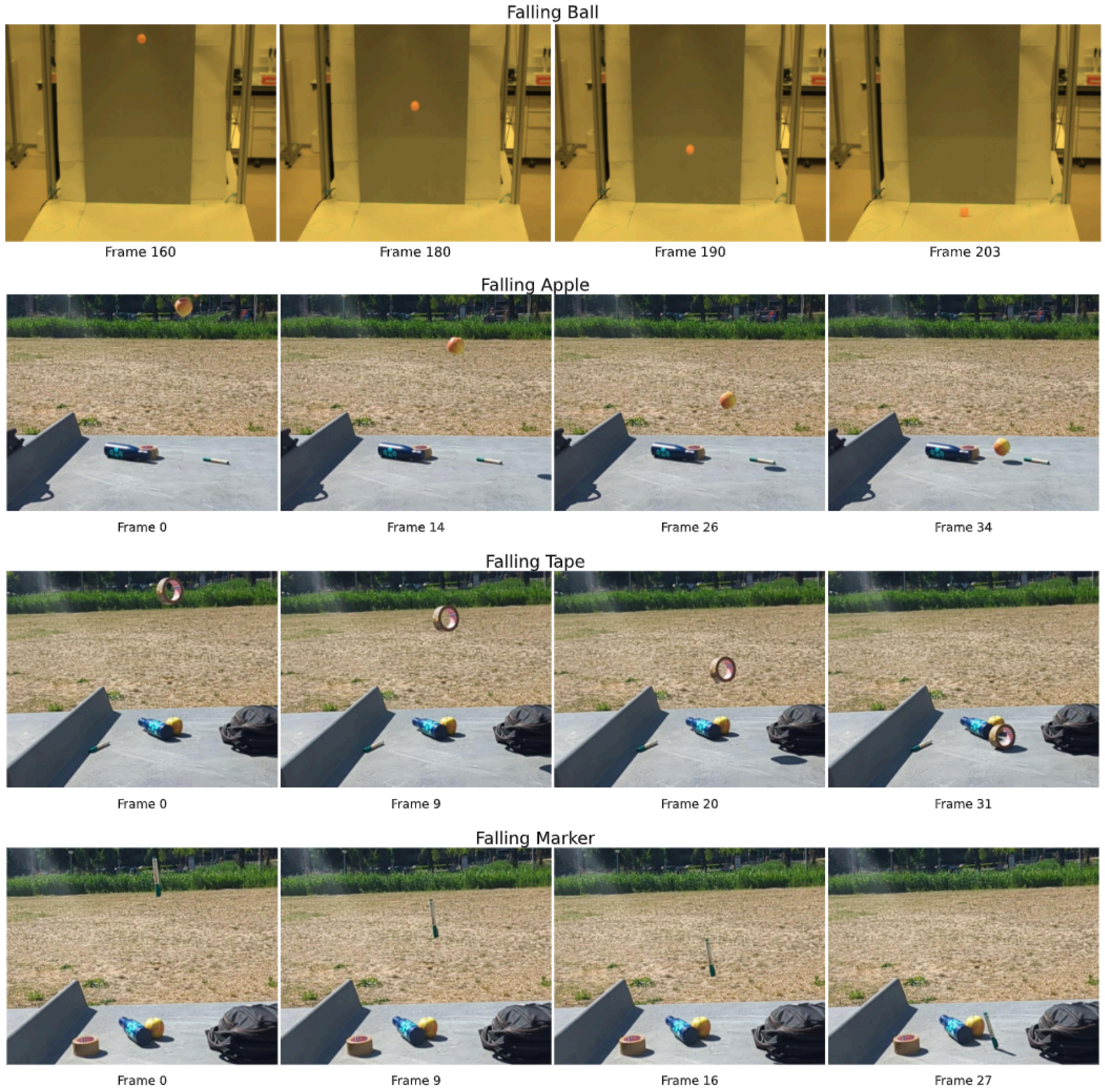


Figure 9: Representative frames from the falling experiments in the **Morpheus** benchmark: ball, apple, tape, and marker.

C VELOCITY AND ACCELERATION ESTIMATION

We estimate objects' velocity and acceleration from the extracted trajectory using multiple stages.

We use the central difference method for most points in the time series. This method computes velocity by considering both forward and backward positions, reducing single-sided differentiation errors.

$$v_i = \frac{x_{i+1} - x_{i-1}}{t_{i+1} - t_{i-1}}, \quad 1 \leq i \leq N-2 \tag{1}$$

Since the central difference is not applicable at endpoints, we use one-sided differences. Forward difference (starting point):

$$v_0 = \frac{x_1 - x_0}{t_1 - t_0}$$

Backward difference (ending point):

$$v_N = \frac{x_N - x_{N-1}}{t_N - t_{N-1}}$$

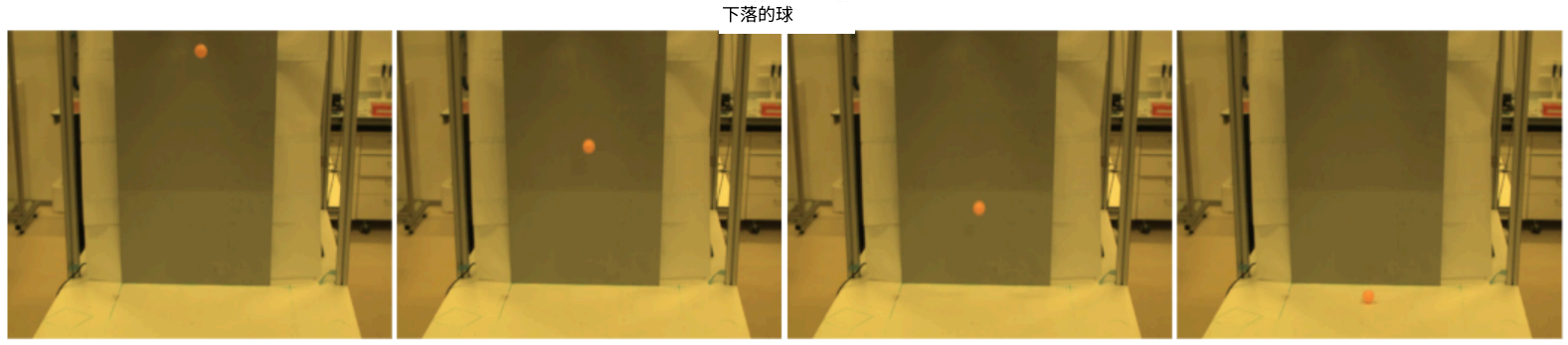


图 9：Morpheus 基准中下落实验的代表性帧：球、苹果、胶带和标记。

C 速度和加速度估计

我们通过多个阶段从提取的轨迹中估计物体的速度和加速度。

我们对时间序列中的大多数点使用中心差分法。该方法通过考虑正向和反向位置来计算速度，减少了单边微分误差。

$$v= \quad \frac{x- x}{t- t} \quad \cdot \quad , 1 \leqslant i \leqslant N - 2 \tag{1}$$

由于中心差分在端点处不适用，我们使用单边差分。前向差分（起点）：

$$v= \quad \frac{x- x}{t- t}$$

向后差分（终点）：

$$v= \quad \frac{x- x}{t- t}$$

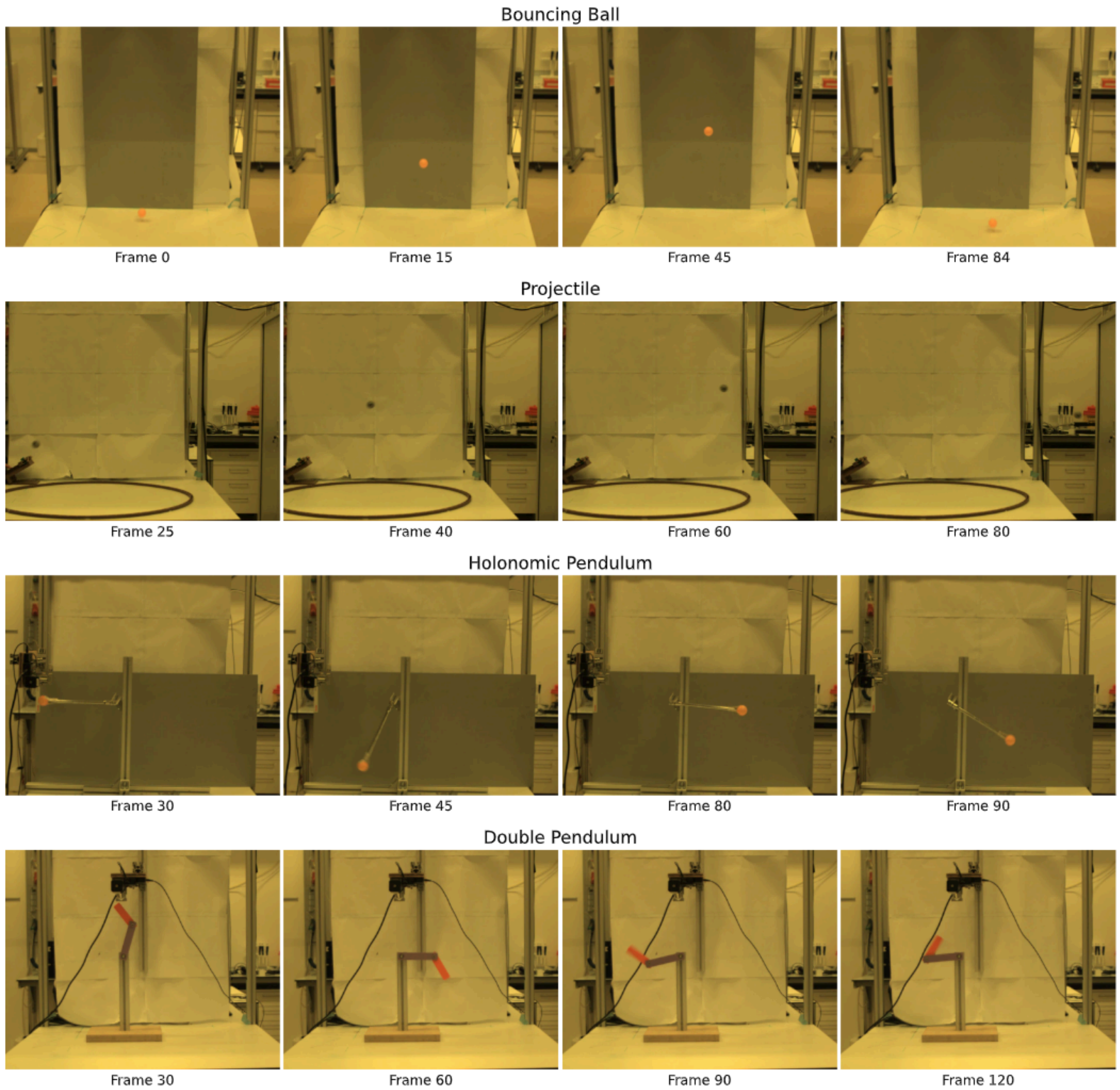


Figure 10: Representative frames from four experiments in the **Morpheus** benchmark: bouncing ball, projectile motion, holonomic pendulum, and double pendulum.

To enhance precision, we perform linear regression within a sliding window.

$$x(t) = vt + b \tag{2}$$

The velocity (slope) is solved using the least squares method with window size w:

$$\begin{bmatrix} v \\ b \end{bmatrix} = \left(A^T A\right)^{-1} A^T x \tag{3}$$

where matrix A contains time information.

$$A = \begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_w & 1 \end{bmatrix} \tag{4}$$

We combine linear regression and central difference results using weighted averages.

$$v_{\text{final}} = \alpha v_{\text{regression}} + (1 - \alpha) v_{\text{central}} \tag{5}$$

Here $\alpha = 0.7$, indicating greater confidence in the regression method. Finally, we apply Savitzky-Golay filtering for smoothing (Luo et al., 2005). This step effectively removes high-frequency noise from velocity calculations.

$$v_{\text{smoothed}} = \text{SG}(v_{\text{final}}, \text{window}, 3) \tag{6}$$

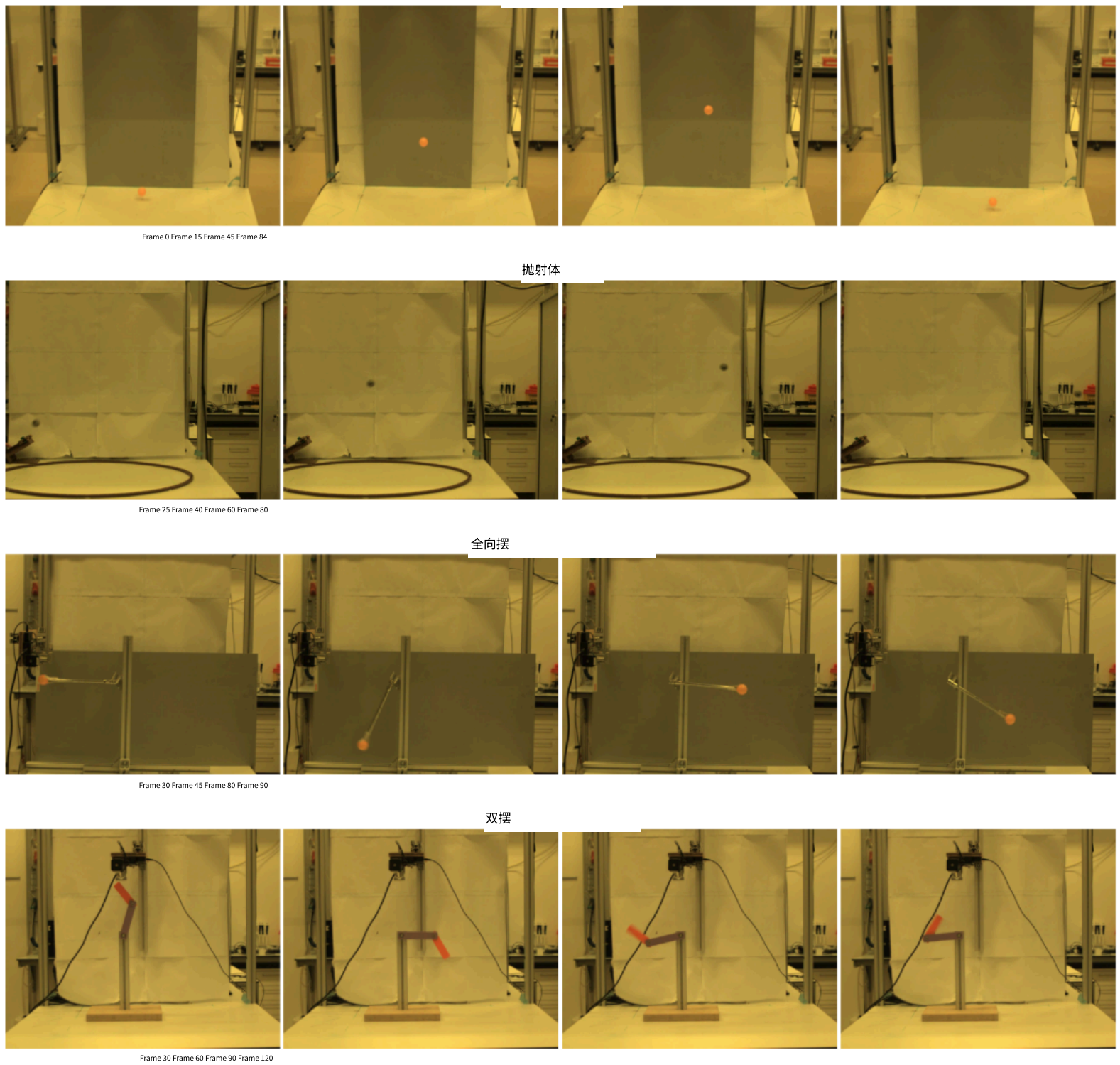


图 10：Morpheus 基准中四个实验的代表性帧：弹跳球、抛体运动、完整约束摆和双摆。

为提高精度，我们在滑动窗口内进行线性回归。

$$x(t) = vt + b \tag{2}$$

速度（斜率）使用窗口大小 w 的最小二乘法求解：

$$\begin{bmatrix} v \\ b \end{bmatrix} = AA \quad Ax \tag{3}$$

其中矩阵 A 包含时间信息。

$$A = \begin{bmatrix} t1 & \\ t1 & \\ \vdots & \vdots \\ t1 & \end{bmatrix} \tag{4}$$

我们使用加权平均法结合线性回归和中心差分结果。

$$v = \alpha v_{reg} + (1-\alpha) v_{cd} \tag{5}$$

这里 α = 0.7，表示对回归方法更有信心。最后，我们应用 Savitzky-Golay 滤波进行平滑（Luo 等人，2005 年）。这一步有效地从速度计算中消除了高频噪声。

$$v = SG(v, window, 3) \tag{6}$$

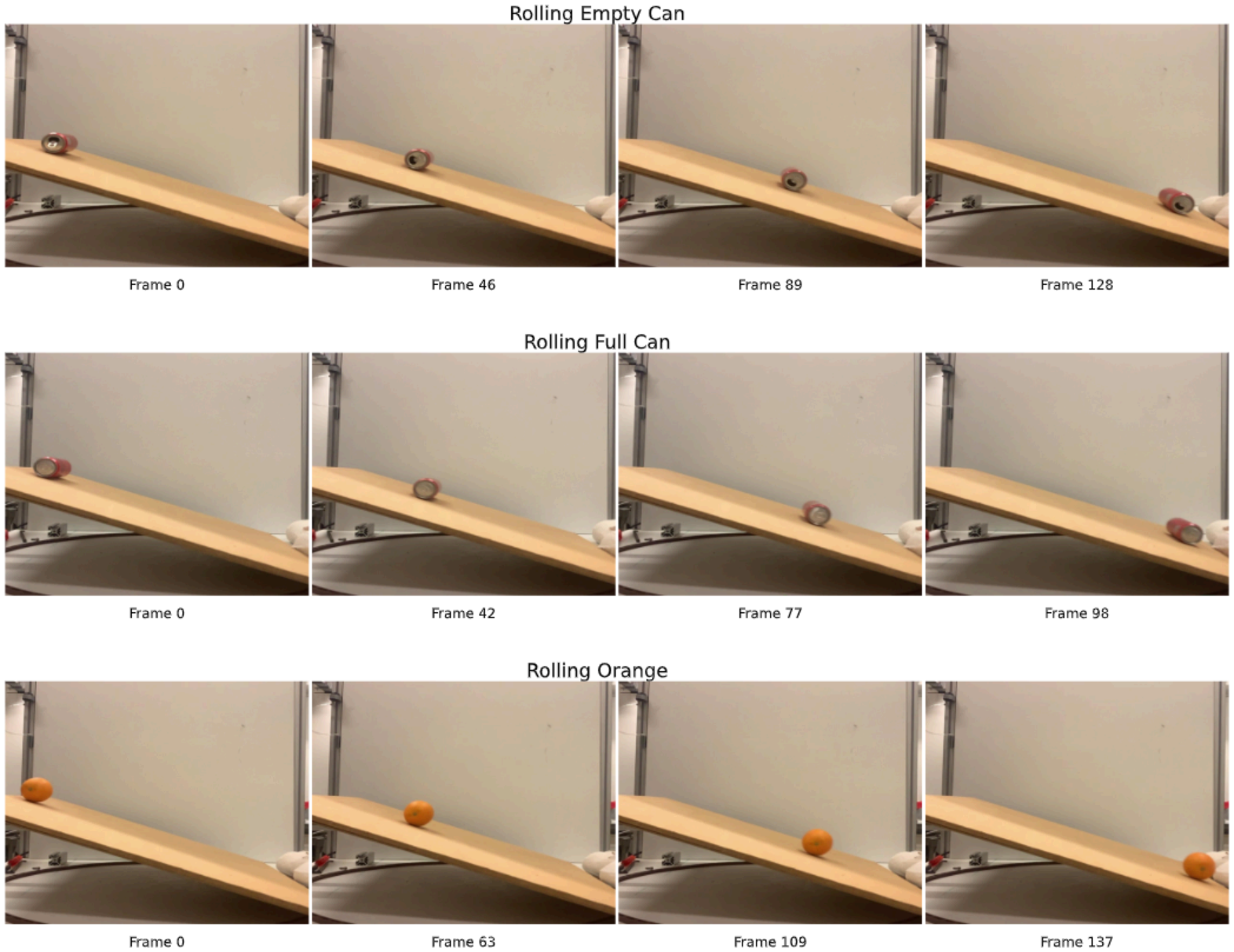


Figure 11: Representative frames of the rolling experiments in the **Morpheus** benchmark: an empty can, a full can, and an orange, each released by hand on an inclined surface of varying slope.

The entire calculation process can be summarized as:

$$v(t) = \mathrm{SG}\left(\alpha v_{\text{regression}}(t) + (1-\alpha)v_{\text{central}}(t), w, 3\right) \tag{7}$$

where w is the window size (odd number for symmetry); $\alpha = 0.7$ is the weighting coefficient; SG represents Savitzky-Golay filter of order 3; Regression window range: $[t - w/2, t + w/2]$.

For the acceleration, we first calculate the acceleration using the central difference. For $1 \leq i \leq N-2$:

$$a_i = \frac{v_{i+1} - v_{i-1}}{t_{i+1} - t_{i-1}} \tag{8}$$

Dealing with the endpoints using the same metric as velocities, we get the final acceleration for the entire trajectory.

$$a_0 = \frac{a_1 - a_0}{t_1 - t_0}$$
$$a_N = \frac{v_N - v_{N-1}}{t_N - t_{N-1}}$$

D Evaluation metrics

D.1 Discard rate

We generate N_{total} videos for each type of experiment. Among these videos, we discard those that do not meet our quality standards, following a three-stage filtering out. First, we discard videos where object are disappearing from the videos the number of such videos is $N_{disappear}$. Second, we analyze the number of objects in each video and discard videos that do not maintain a consistent

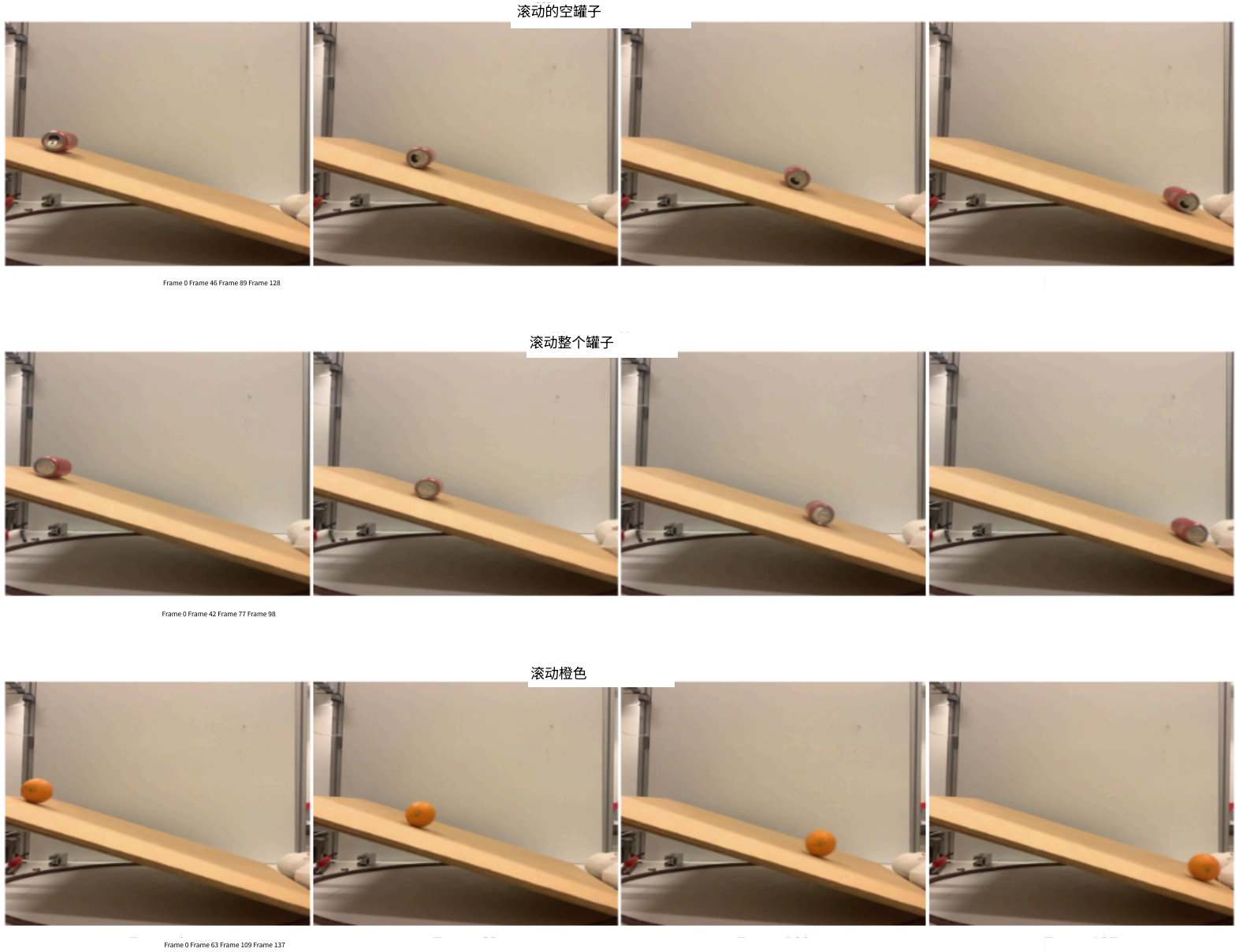


图 11：Morpheus 基准中滚动实验的代表性帧：一个空罐子、一个满罐子和一个橙子，它们被手从一个倾斜度不同的表面上释放。

整个计算过程可以总结为：

$$v(t) = SG\,(\alpha v(t) + (1-\alpha)v(t), w, 3) \tag{7}$$

其中 w 是窗口大小（奇数以保持对称性）；α = 0.7 是权重系数；SG 表示三阶 Savitzky-Golay 滤波器；回归窗口范围：[t − w/2, t + w/2]。

对于加速度，我们首先使用中心差分法计算加速度。对于 $1 \leqslant i \leqslant N-2$：

$$a = \frac{\bar{v} - v}{\bar{t} - t} \tag{8}$$

使用与速度相同的度量方法处理端点，我们得到整个轨迹的最终加速度。

$$a = \frac{\bar{a} - a}{\bar{t} - t}$$

$$a = \frac{\bar{v} - v}{\bar{t} - t}$$

D 评估指标

D.1 D

我们为每种实验类型生成 N 个视频。在这些视频当中，我们通过三阶段过滤去除不符合我们质量标准的视频。首先，我们去除视频中对象消失的视频，这类视频数量为 N。其次，我们分析每个视频中的对象数量，去除那些不保持一致的视频。

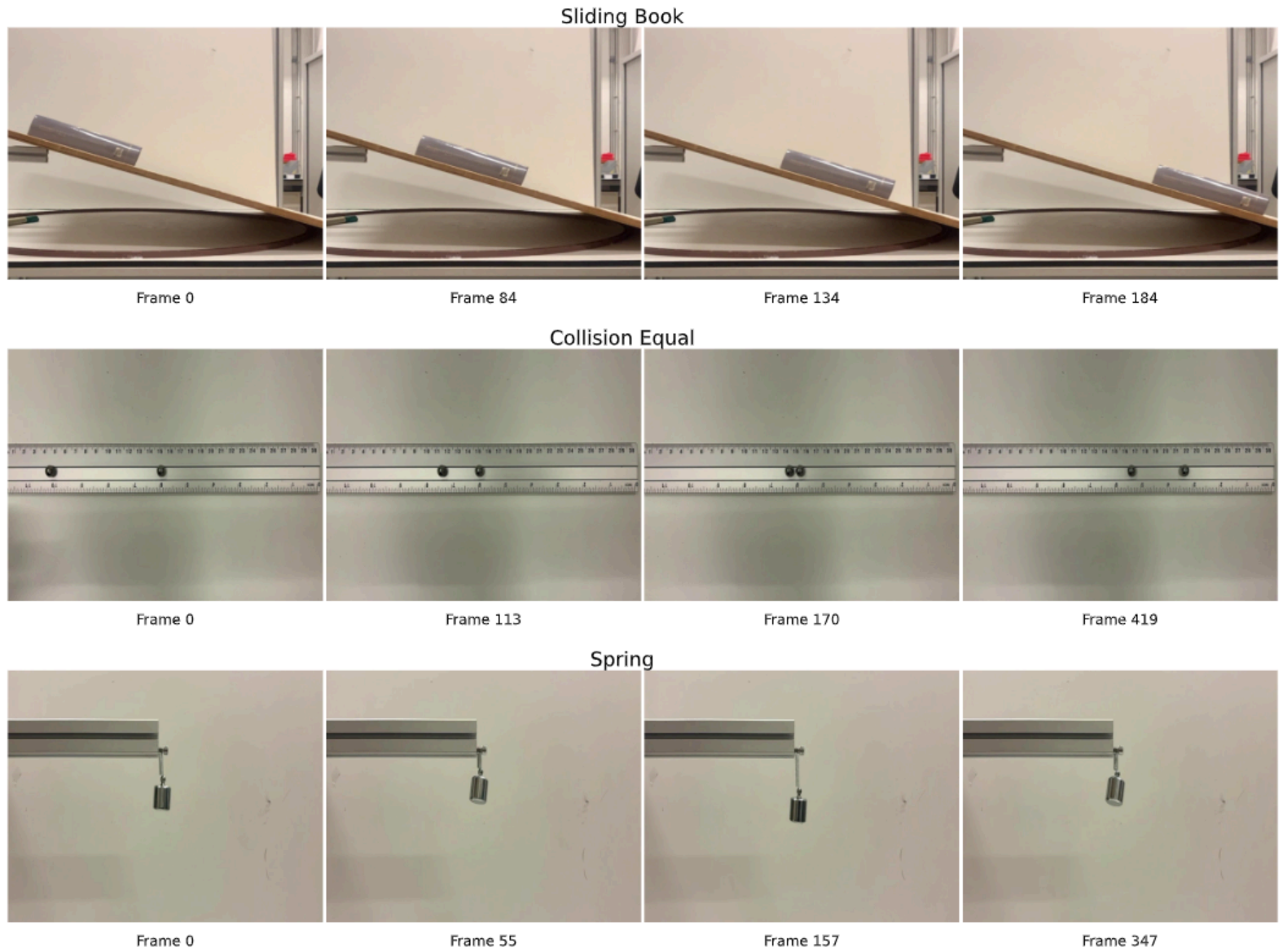


Figure 12: Representative frames from three experiments in the **Morpheus** benchmark: sliding, collision (equal-sized objects), and spring.

Benchmark	Real-world ground-truth	Quantitative evaluation	Initial condition grounding	Physical laws evaluation
VideoCon-Physics	✓	✗	✗ (only text)	✗
VAMP	✓	✓	✗ (only text)	✗
PhyGenBench	✓	✗	✗ (only text)	✗
Kang et al. (2024)	✗	✓	✓ (1 or 3 frames)	✗
COSMOS	✗	✓	✓ (image + video)	✗
Physics-IQ	✓	✓	✓ (image + video)	✗
Morpheus (ours)	✓	✓	✓ (image + video)	✓

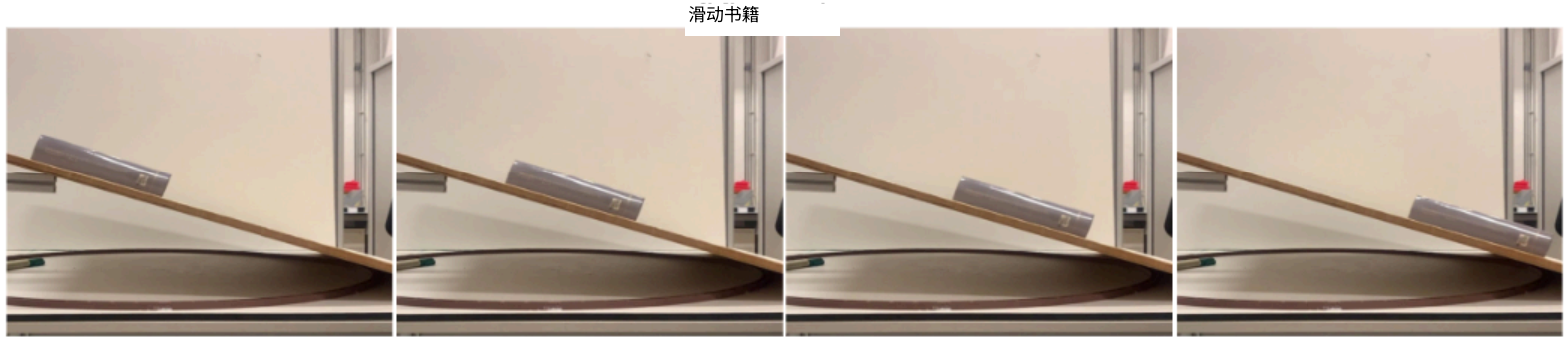
Table 2: Comparison of physics-based video understanding benchmarks. Our benchmark is the first to use real physical laws for evaluation. Symbols: ✓= supported, ✗= not supported

object count in the not discarded yet videos. For this purpose we employ DEVA tracking (Cheng et al., 2023) built on top of Grounded SAM (Ren et al., 2024) (with object names from the prompt as Grounding DINO (Liu et al., 2025) query) for consistent open-vocabulary prediction of 2D object masks. We denote the number of discarded videos in this step as $N_{duplicate}$. Specifically, we evaluate the proportion of frames containing multiple objects. Videos are filtered out if this proportion exceeds a predetermined threshold. Finally, we discard videos where the motion is too small to be meaningful in the not discarded yet videos, the number of such videos is N_{stil}. The overall *discard rate* DR is defined as

$$DR = \frac{N_{disappear} + N_{duplicate} + N_{still}}{N_{total}}.$$

D.2 Depth Consistency Evaluation

In all the studied experiments, the video camera is orthogonal to the object's motion and is fixed. This allows us to compute the Physical Invariance and Dynamical scores using only information extracted



我们分析了每个视频中的物体数量，并丢弃了那些不保持一致性的视频。图 12：Morpheus 基准中三个实验的代表性帧：滑动、碰撞（等大小物体）和弹簧。

基准	现实世界真实值	定量评估	初始条件基础	物理定律评估
VideoCon-Physics	✓	✗	✗ (仅文本)	✗
VAMP	✓	✓	✗ (仅文本)	✗
PhyGenBench	✓	✗	✗ (仅文本)	✗
Kang 等人(2024)	✗	✓	✓ (1 或 3 帧)	✗
COSMOS	✗	✓	✓ (图像+视频)	✗
Physics-IQ	✓	✓	✓ (图像+视频)	✗
Morpheus(我们)	✓	✓	✓ (图像+视频)	✓

表 2：基于物理的视频理解基准测试比较。我们的基准测试是首个使用真实物理定律进行评估的测试。符号：✓=支持，✗=不支持

尚未被丢弃的视频中的物体数量。为此，我们采用基于 Grounded SAM（Ren et al., 2024）构建的 DEVA 跟踪（Cheng et al., 2023）（使用提示中的物体名称作为 Grounding DINO（Liu et al., 2025）查询），以实现一致的开放词汇预测 2D 物体掩码。我们将此步骤中丢弃的视频数量表示为 N。具体来说，我们评估包含多个物体的帧的比例。如果该比例超过预定阈值，则过滤掉视频。最后，在尚未被丢弃的视频中，丢弃运动太小而无法产生意义的视频，此类视频数量为 N。整体丢弃率 DR 定义为

$$\mathrm{DR} = \frac{N + N + N}{N} \quad .$$

D.2 DCE

在所有研究的实验中，视频相机与物体的运动方向垂直且固定。这使我们能够仅使用提取的信息来计算物理不变性和动态分数。

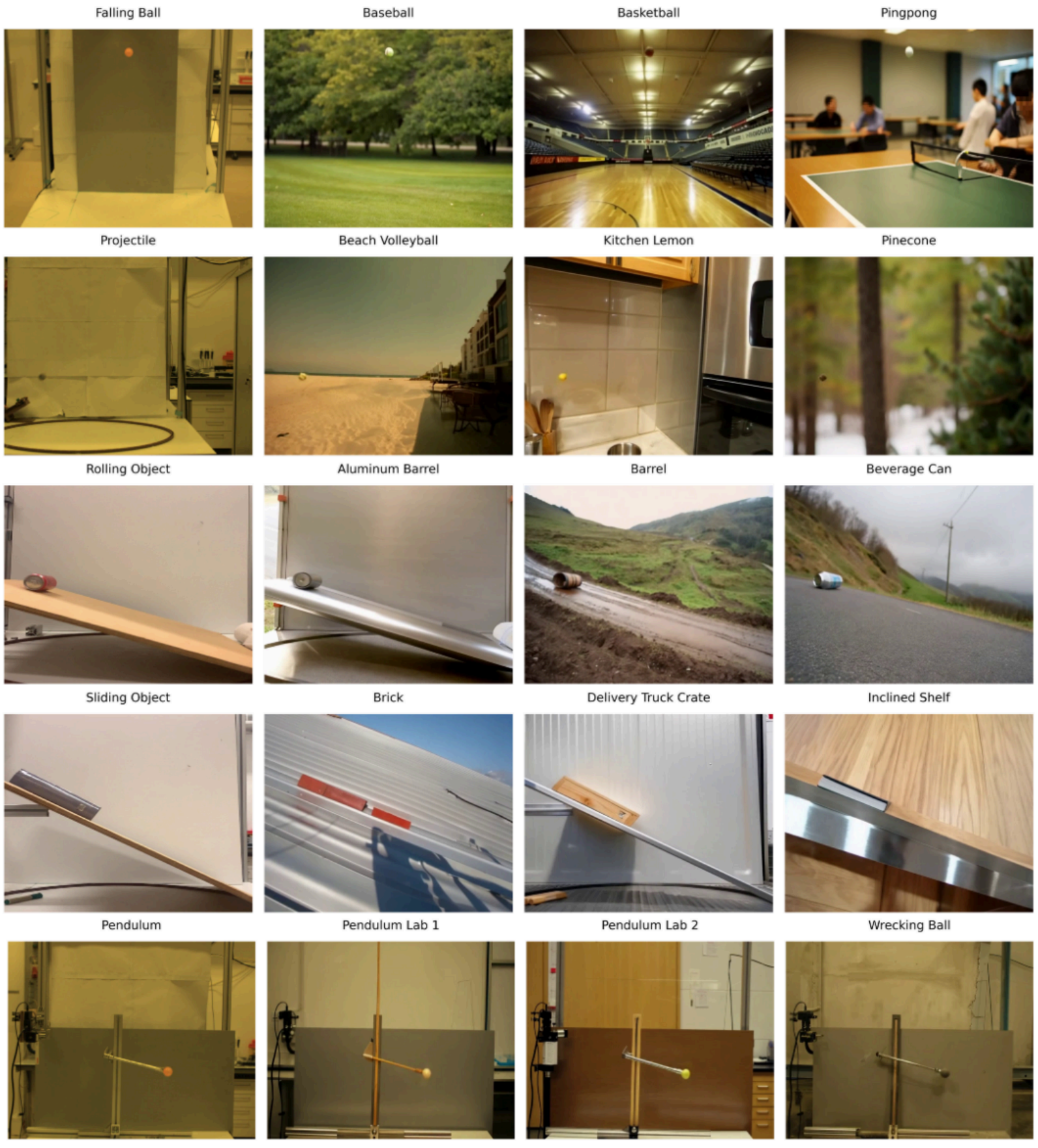
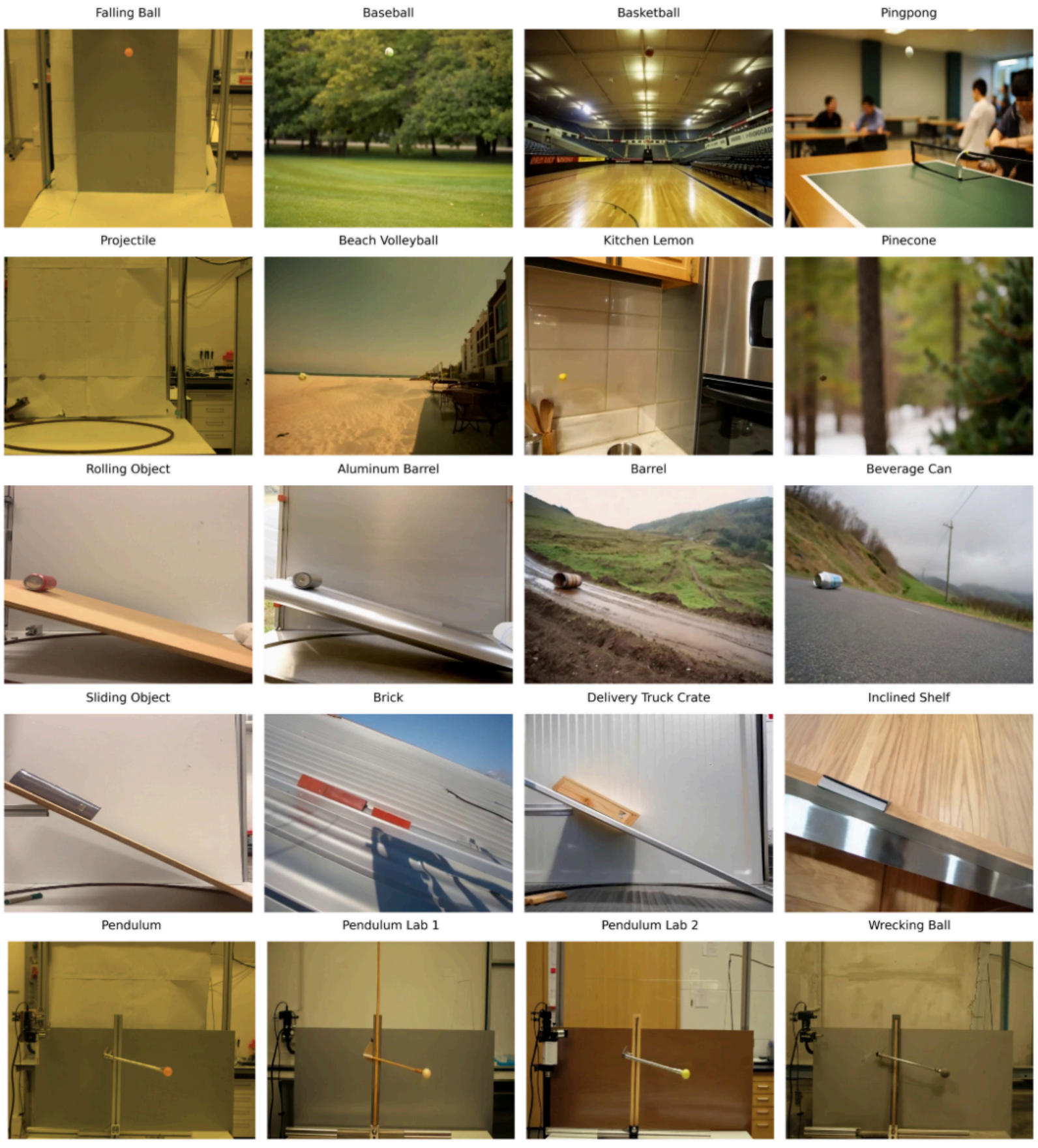


Figure 13: COSMOS-Transfer style augmentations for physics experiments. Each row displays the original experiment (left) with its style-transferred variants (right), showing how object appearance can be altered while preserving physical dynamics. The transfers provide diverse yet physically consistent initializations for VGM evaluation across five fundamental physics scenarios.

from 2D pixel space, available for generated videos. The results are presented in Fig. 14, showing that most of the models are reasonably consistent and thus object properties like energy conservation could be also studied using only 2D coordinates. Proprietary Veo3 and Veo3-fast (Veo-Team et al., 2024), which were also initially investigated, were excluded from the final analysis due to budget limits to compute a similar batch as other models, however, we plan to regularly update our leaderboard with state-of-the-art models's evaluation.

D.3 Physically-informed Neural Networks

Unlike typical neural networks, which are normally trained only on data, prior knowledge about the physical system is integrated into PINNs. This prior knowledge of the physical system, often in governing physical laws such as Newtonian mechanics or energy conservation, is imposed during



这使得我们能够仅使用提取的图 13：COSMOS-Transfer 风格的物理实验增强。每一行显示原始实验（左）及其风格转换变体（右），展示了如何改变物体外观同时保持物理动态。这些转换为 VGM 评估在五个基本物理场景中提供了多样但物理一致的初始化。

从二维像素空间，可用于生成的视频。结果展示在图 14 中，显示大多数模型具有较好的一致性，因此物体属性如能量守恒也可以仅使用二维坐标进行研究。专有的 Veo3 和 Veo3-fast（Veo-Team 等，2024 年）最初也进行了研究，但由于预算限制无法与其他模型计算类似的批次，因此被排除在最终分析之外。然而，我们计划定期更新我们的排行榜，加入最先进的模型评估。

D.3 P-NN

与通常仅在数据上训练的典型神经网络不同，PINNs 将物理系统的先验知识整合了进去。这种物理系统的先验知识，通常体现在支配物理系统的物理定律中，如牛顿力学或能量守恒定律，这些知识在训练过程中被施加。

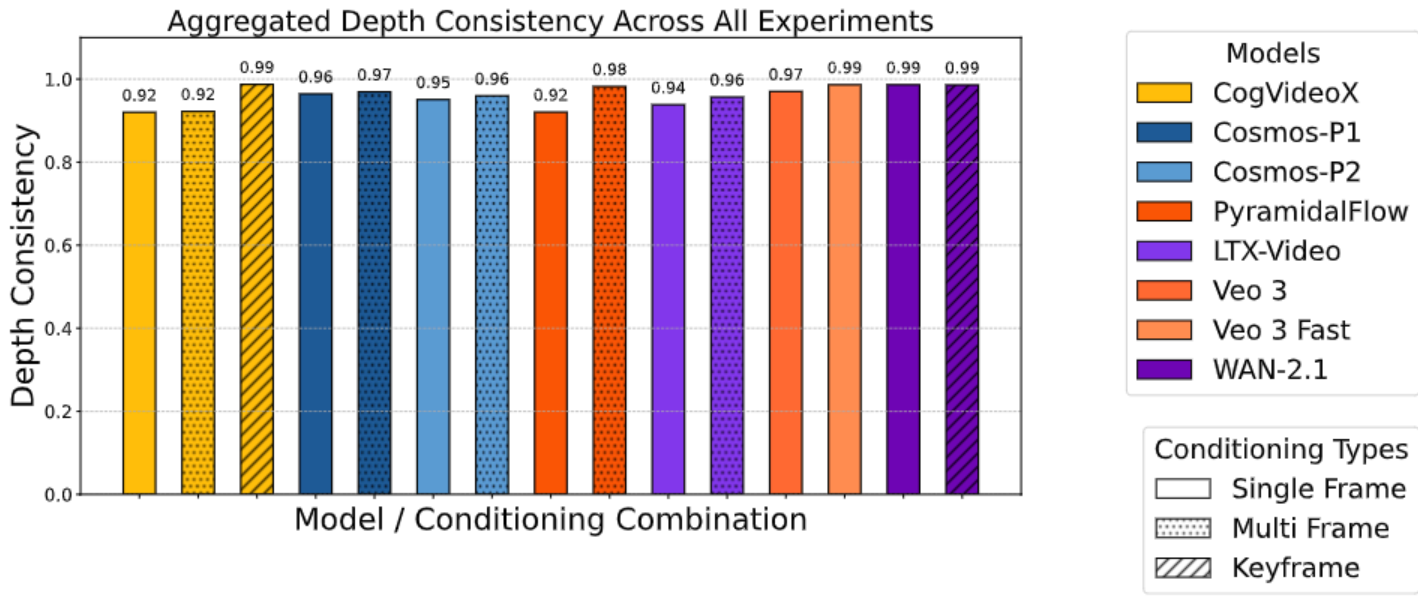


Figure 14: Average depth consistency for different video generation models across all studied experiments.

training. Given that the system modeled from the generated videos is known from the provided prompt, the training process incorporates these laws into the loss function. The total loss for a PINN is defined as:

$$L_{\text{total}} = L_{\text{data}} + \lambda L_{\text{physics}}, \tag{9}$$

where L_{data} ensures that the network's output can match the observed data. At the same time, L_{physics} is penalizing deviations from the governing physical equation and λ is a hyperparameter balancing the contribution of the each loss component. For our own experimentation λ has a value of 1. In this way, PINNs can bring both data and physical laws together during training while being consistent with the underline physical system. For a trajectory T, the data loss is defined as

$$L_{\text{data}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{T}_i - T_i\|^2, \tag{10}$$

, where $\hat{T}_i$ is the trajectory predicted by the network at the i-th timestep and T_i is the corresponding ground truth trajectory at the same timestep. On the other hand, the physics loss is derived separately for each experiment, given the nature of the system's dynamics. The motion of a free-falling object follows:

$$\ddot{y} + g = 0, \tag{11}$$

where y is the vertical position, $\ddot{y}$ is the acceleration and g is the gravitational constant. This means that for this phenomenon, the loss is defined as: The physics loss for free fall is defined as:

$$L_{\text{physics}} = \frac{1}{M} \sum_{j=1}^{M} \left\| \hat{\ddot{y}}_j + g \right\|^2, \tag{12}$$

where $\hat{\ddot{y}}_j$ is the predicted acceleration derived from the PINN at the j-th time step. The motion of a holonomic pendulum is governed by:

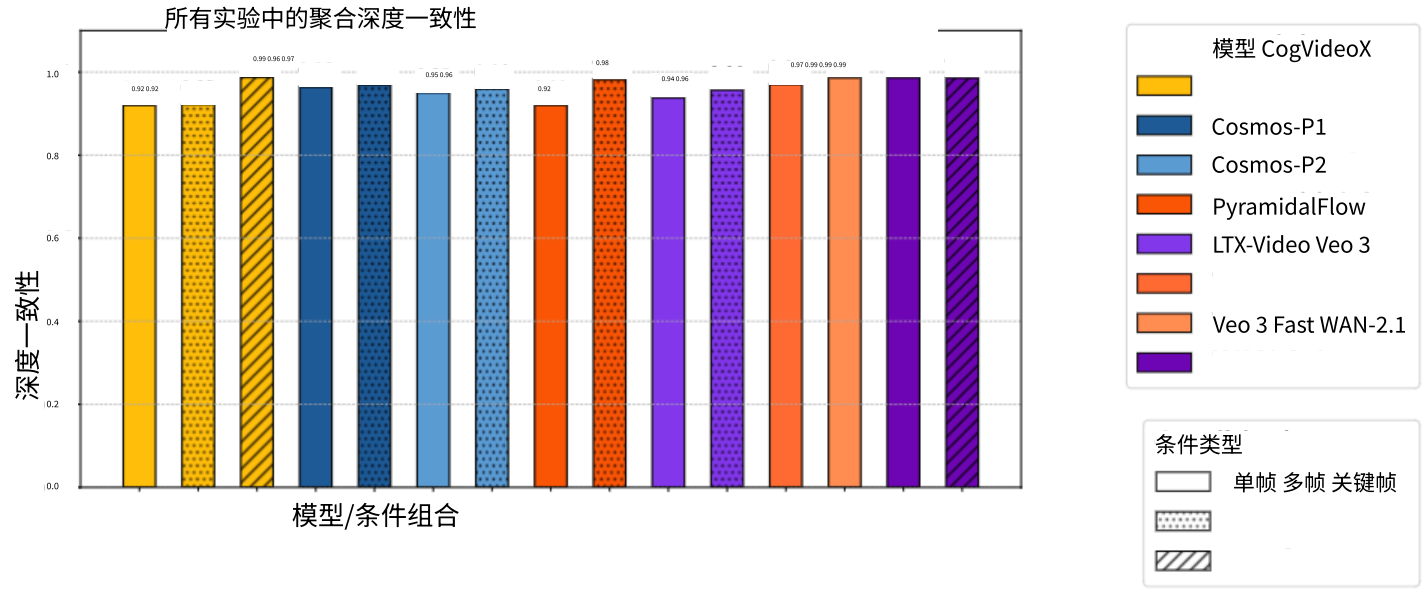
$$\ddot{\theta} + \frac{g}{l} \sin\theta = 0, \tag{13}$$

where θ is the angular displacement, l is the pendulum length and g is the gravitational constant.

The corresponding physics loss is:

$$L_{\text{physics}} = \frac{1}{M} \sum_{j=1}^{M} \left\| \hat{\ddot{\theta}}_j + \frac{g}{l} \sin(\hat{\theta}_j) \right\|^2, \tag{14}$$

where $\hat{\ddot{\theta}}_j$ and $\hat{\theta}_j$ are the network-predicted angular acceleration and displacement, respectively, at the j-th timestep. In the present work, we use the Dynamical score to evaluate how well the does the predicted trajectories align with the ground truth. The Dynamical score is derived from the



如图 14 所示：在所有研究的实验中，不同视频生成模型的平均深度一致性。

在训练过程中施加。鉴于从生成的视频中建模的系统已知来自提供的提示，训练过程将这些定律纳入损失函数。PINN 的总损失定义为：

(9) 其中 Lensures 网络输出能够匹配观测数据，同时，L 惩罚偏离控制物理方程的偏差，λ是平衡每个损失组件贡献的超参数。在我们的实验中，λ的值为 1。通过这种方式，PINNs 在训练过程中能够将数据和物理定律结合起来，同时与底层物理系统保持一致。对于轨迹 T，数据损失定义为

$$L = \frac{1}{N}\sum_{i=1}^{N} \left\| \hat{T} - T \right\|, \tag{10}$$

，其中网络在第 i 个时间步预测的轨迹，T 是同一时间步对应的真实轨迹。另一方面，由于系统动力学的性质，物理损失是针对每个实验单独推导的。自由落体运动遵循：

(11) 其中 y 是垂直位置，¨y 是加速度，g 是重力常数，这意味着对于这种现象，损失定义为：自由落体的物理损失定义为：

$$L = \frac{1}{M}\sum_{j=1}^{M} \left(\hat{\ddot{y}} + g \right)^2, \tag{12}$$

ˆ¨y 是 PINN 在 j-th 时间步预测出的加速度。一个完整约束摆的运动由以下公式控制：

$$\ddot{\theta} + \frac{g}{l}\sin\theta = 0, \tag{13}$$

其中θ是角位移，l 是摆长，g 是重力常数。

相应的物理损失为：

$$L = \frac{1}{M}\sum_{j=1}^{M} \left(\hat{\ddot{\theta}} + \frac{g}{l}\sin(\hat{\theta}) \right)^2, \tag{14}$$

其中ˆ¨θ分别是网络在第 j 个时间步预测的角加速度和位移。在本工作中，我们使用动态分数来评估预测轨迹与真实轨迹的吻合程度。动态分数是从...推导而来的

Table 3: Conserved quantities for each physical experiment in an ideal case.

Experiment Name	Assumption	Conserved Quantities
falling ball	no air resistance	energy, acceleration (gravity), horiz. momentum
projectile	no air resistance	energy, acceleration (gravity), horiz. momentum
bouncing ball	no air resistance	energy, acceleration (gravity), horiz. momentum
holonomic pendulum	low resistance	energy, period, pendulum length
sliding	uniform surface forces	acceleration
rolling	uniform surface forces	acceleration
elastic collision	perfect elasticity	total energy, linear momentum
spring-mass system	ideal Hookean spring	period
double pendulum	low resistance	total energy, two pendulums length

Normalized Mean Squared Error (NMSE), which provides a relative measure of error by normalizing the Mean Squared Error (MSE) with the variance of the ground truth trajectory. Main motivation behind this choice, is to make the evaluation independent of scale. The NMSE is calculated as:

$$\text{NMSE} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{15}$$

where:

- y_i is the true value at timestep i,
- $\hat{y}_i$ is the predicted value at timestep i,
- $\bar{y}$ is the mean of the ground truth values, defined as:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{16}$$

- $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ represents the MSE between the predicted and ground truth trajectories,
- $\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sigma^2$ represents the variance of the ground truth trajectory.

To ensure robustness, the predicted trajectory is compared against the interpolated ground truth values. Depending on the experiment, we address physical consistency by quantifying how well the learned solution adheres to the underlying physical equation. This is quantified using the physics loss, which penalizes deviations from the expected dynamics. For training, each PINN is optimized using the Adam optimizer with a learning rate of 10^{-3} for 200,000 iterations. The network used for all experiments, consists of two hidden layers of 20 neurons, with $\tanh$ as activation functions. The final score is defined as $S_{dyn} = \min(1 - \text{NMSE}, 0)$.. Similarly to the Physical Invariance score, in cases when the original trajectory is discarded, the score is assigned to a minimal value equal to zero.

D.4 Physical Invariances

Falling Ball For falling balls, energy must be conserved between consecutive bouncing points. Additionally, according to Newton's second law:

$$F = ma$$

In free fall, gravity is the sole force acting on the object, resulting in constant acceleration. Assuming that the gravitational field is uniform in space and time, we have $F = mg$, which means that $a = g$, so the acceleration should stay constant. Therefore, in this part, we introduce three quantitative metrics to assess trajectory physics: the Energy Conservation score (ES), which measures energy conservation within a specified time window, and the Acceleration Conservation score (AS), which evaluates the consistency of acceleration during this interval, and the Horizontal Momentum Conservation score (MS), which measures the conservation of momentum.

The Energy Conservation score is calculated as follows. Given the mass of the ball to be m, the g a freefall acceleration constant, kinetic energy:

$$T = \frac{1}{2}m|\vec{v}|^2 = \frac{1}{2}m(v_x^2 + v_y^2)$$

表 3：理想情况下每个物理实验的守恒量。

实验名称	假设守恒量
下落球	无空气阻力能量、加速度（重力）、水平动量
抛射体	无空气阻力能量、加速度（重力）、水平动量
弹跳球	无空气阻力能量、加速度（重力）、水平动量
完整摆动杆	低阻力能量、周期、摆长
滑动均匀表面力	加速度
滚动均匀表面力	加速度
弹性碰撞	完全弹性总能量、线性动量
弹簧质量系统	理想胡克弹簧周期
双摆	低阻力总能量，两个摆的长度

动态分数是从归一化均方误差（NMSE）推导而来的，它通过用真实轨迹的方差对均方误差（MSE）进行归一化来提供误差的相对度量。这一选择背后的主要动机是使评估与尺度无关。NMSE 的计算公式为：

$$\text{NMSE} = \frac{\frac{1}{N}\sum_{i=1}^{P}(y - \hat{y})}{\frac{1}{N}\sum_{i=1}^{P}(y - \bar{y})}, \tag{15}$$

其中：

- y 是第 i 个时间步的真值，
- $\hat{y}$ 是第 i 个时间步的预测值，
- $\bar{y}$ 是真实值的平均值，定义为：

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y, \tag{16}$$

- $\frac{1}{N}\sum_{i=1}^{P}(y - \hat{y})$表示预测轨迹与真实轨迹之间的均方误差，
- $\frac{1}{N}\sum_{i=1}^{P}(y - \bar{y}) = \sigma$表示真实轨迹的方差。

为确保鲁棒性，预测轨迹与插值真实值进行比较。根据实验情况，我们通过量化学习到的解决方案如何遵循底层物理方程来处理物理一致性。这通过物理损失来量化，该损失惩罚与预期动力学的偏差。在训练过程中，每个 PINN 使用 Adam 优化器以 10 的速率进行 200,000 次迭代优化。所有实验使用的网络由两个包含 20 个神经元的隐藏层组成，激活函数为 tanh。最终得分定义为 $S = \min(1 - \text{NMSE}, 0)$。类似于物理不变性得分，在原始轨迹被舍弃的情况下，得分被赋予最小值零。

D.4 PI

对于下落的球，在连续的弹跳点之间必须保持能量守恒。此外，根据牛顿第二定律：

$$F = ma$$

在自由落体中，重力是作用在物体上的唯一力，导致加速度恒定。假设重力场在空间和时间上均匀，我们有 $F = mg$，这意味着 $a = g$，因此加速度应该保持恒定。因此，在这部分，我们引入三个定量指标来评估轨迹物理：能量守恒分数（ES），用于测量指定时间窗口内的能量守恒，加速度守恒分数（AS），用于评估该时间段内加速度的一致性，以及水平动量守恒分数（MS），用于测量动量的守恒。

能量守恒分数的计算方法如下。给定球的质最为 m，自由落体加速度常数为 g，动能为：

$$T = \frac{1}{2}m|\boxtimes v| = \frac{1}{2}m(v + v)$$

and potential energy:

$$V = mgh$$

where $h = y$. Total energy is the sum of two:

$$E = T + V = \frac{1}{2}m(v_x^2 + v_y^2) + mgy$$

From this formula, assuming the mass of the ball is constant in time, we get:

$$\frac{E}{m} = \frac{1}{2}(v_x^2 + v_y^2) + gy = \text{const} \tag{17}$$

The calculation of the Acceleration Conservation score is self-evident:

$$a = \text{const} \tag{18}$$

The conservation of horizontal momentum arises from the fact that the only force acting on the ball is gravity, which is pointed downwards:

$$p_x = mV_x = \text{const}$$

and analogous to the energy, we deduce:

$$\frac{p_x}{m} = V_x = \text{const} \tag{19}$$

We provide some examples of estimated invariants in Fig. 25.

Projectile For projectile motion, we analyze the same physical invariants as in the falling ball experiment. Throughout the projectile's trajectory, neglecting the air resistance, energy, acceleration, and horizontal momentum should be conserved. The calculations for energy and acceleration follow the same methodology used in the falling ball analysis.

Holonomic Pendulum For the holonomic pendulum, let's first examine energy conservation. Energy in the ideal (frictionless) case:

$$H = T + V = \frac{p_\theta^2}{2mL^2} + mgL(1 - \cos\theta)$$

where θ is the angular displacement, l is the pendulum length, g is the gravitational acceleration, and $p_\theta = mL^2\dot{\theta}$ is the momentum.

In this case, the equation that we obtain is:

$$\ddot{\theta} + \frac{g}{l}\sin\theta = 0$$

Since our real-world pendulum experiments were conducted in a laboratory environment, friction causes energy attenuation over time. We quantify this energy loss by measuring both its range and rate of decline, establishing these as upper bounds for evaluating generated videos. To be specific, the holonomic pendulum with friction can be expressed as

$$\ddot{\theta} + \frac{b}{m}\dot{\theta} + \frac{g}{l}\sin\theta = 0 \tag{20}$$

where b is the damping coefficient, m is the bob mass, and $\frac{b}{m}\dot{\theta}$ represents the damping force term. The energy decay over time:

$$\frac{dE}{dt} = -b(\dot{\theta})^2 \tag{21}$$

In our experiments, we assume that the energy loss can be ignored for a short time period, meaning we can apply the Energy Conservation score.

和势能：

$$V = mgh$$

其中 h = y。总能量是两个之和：

$$E = T + V = \frac{1}{2}m(v + v) + mgy$$

从这个公式中，假设球的质量随时间保持不变，我们得到：

$$\frac{E}{m} = \frac{1}{2}(v + v) + gy = \text{const} \tag{17}$$

加速度守恒得分的计算显而易见：

$$a = \text{const} \tag{18}$$

水平动量守恒源于这样一个事实：作用在球上的唯一力是重力，它指向下方：

$$p = mV = \text{const}$$

类似于能量，我们推导出：

$$\frac{p}{m} = V = \text{const} \tag{19}$$

我们在图 25 中提供了一些估计不变量的示例。

抛体 对于抛体运动，我们分析了与下落球实验相同的物理不变量。在整个抛体轨迹中，忽略空气阻力，能量、加速度和水平动量应该守恒。能量和加速度的计算遵循下落球分析中使用的相同方法。
完整摆 对于完整摆，让我们首先检查能量守恒。理想（无摩擦）情况下的能量：

$$H = T + V = \frac{p}{2mL} + mgL(1 - \cos\theta)$$

其中 θ 是角位移，l 是摆长，g 是重力加速度，和
$p = mL\dot{\theta}$ 是动量。
在这种情况下，我们得到的方程是：

$$\ddot{\theta} + \frac{g}{l}\sin\theta = 0$$

由于我们的真实世界摆锤实验是在实验室环境中进行的，摩擦会导致能量随时间衰减。我们通过测量其范围和衰减速率来量化这种能量损失，并将其作为评估生成视频的上限。具体来说，带摩擦的完整摆锤可以表示为

$$\ddot{\theta} + \frac{b}{m}\dot{\theta} + \frac{g}{l}\sin\theta = 0 \tag{20}$$

其中 b 是阻尼系数，m 是摆的质量，$-\dot{\theta}$表示阻尼力项。
能量随时间衰减：

$$\frac{dE}{dt} = -b(\dot{\theta}) \tag{21}$$

在我们的实验中，我们假设在短时间内能量损失可以忽略不计，这意味着我们可以应用能量守恒分数。

The period of holonomic pendulum with friction with a small amplitude can be expressed as

$$T = 2\pi\sqrt{\frac{l}{g}}\sqrt{1 - \left(\frac{b}{2m\omega_0}\right)^2} \tag{22}$$

where $\omega_0 = \sqrt{\frac{g}{l}}$ is the natural angular frequency without damping. When the damping is small ($b \ll m\omega_0$), the period approaches that of an undamped pendulum $T_0 = 2\pi\sqrt{\frac{l}{g}}$. We observe this regime in our experiments and propose to use the Period Conservation score (PC).

For the holonomic pendulum, it is obvious that the pendulum length l remains constant throughout the experiment, as the holonomic constraint of the system. Therefore, we also consider the pendulum length as a physical invariant.

D.5 Physical Score Scaling

When we obtain the physical invariant value C, we calculate the relative standard deviation over time:

$$C_{\bar{\sigma}} = C_\sigma / C_{mean} \tag{23}$$

To ensure that the score is within the $[0, 1]$ range, we design the Physical score, derived from the invariant, as follows

$$S = \frac{1}{1 + \alpha * C_{\bar{\sigma}}} \tag{24}$$

Where α is a normalization factor. In the experiment, we set it to 1.0.

Two critical considerations emerge during the score calculation process. First, The time window must be carefully selected. For each trajectory, we partition it using a sliding time window and select the highest score among all segments as the trajectory's overall score. This approach addresses a key challenge in real-world experiments like bouncing balls, where fluctuations near bouncing points create large standard deviations and low scores. By using the highest score across all segments, we effectively capture the most stable portion of the trajectory.

In our experiments, we set the time window length between 10% and 25% of the total trajectory duration. Specifically, for real videos, we use $t_{window} = L_{trajectory}/10$, while for generated videos, we use $t_{window} = L_{trajectory}/4$. This difference in window size is necessary because generated videos have much shorter total durations - using $t_{window} = L_{trajectory}/10$ would result in trajectory segments that are too short for meaningful analysis.

Second, proper scaling is essential: since the trajectory coordinates are recorded in pixel space rather than real-world 3D coordinates, a precise coordinate transformation to physical units is required. Notably, improper scaling can significantly impact the total energy calculations. Third, we need to be careful not to choose the range when the mean of the selected physical invariant is near zero. The absolute value of mean of the selected physical invariant should be equal to or greater than a threshold of 10 times of standard deviation:

$$C_{threshold} = 10 * C_\sigma \tag{25}$$

so it can be neglected that the influence of mean energy/acceleration is near zero. In the experiment, if $|C_{mean}| \geq C_{threshold}$, we calculate the Physical score as defined in Eq. 24. Otherwise, we use the following Eq. 26 that takes the absolute standard deviation rather than the relative standard deviation.

$$S = \frac{1}{1 + \alpha * C_\sigma} \tag{26}$$

This method has two key limitations. First, the scores are highly sensitive to the choice of time window size. Larger time windows tend to yield lower scores as they encompass more fluctuations in the trajectory. Thus, we kept time window constant between evaluating different model generations. Second, the method may fail to detect unphysical behavior in generated videos where objects remain stationary for long periods; however, this is addressed by discarding videos due to stillness.

有摩擦的小振幅完整摆动的周期可以表示为

$$T = 2\pi \sqrt{\frac{l}{g}} \sqrt{1 - \frac{b}{2m\omega}} \tag{22}$$

$\omega = \sqrt{\frac{g}{l}}$ 是无阻尼的自然角频率。当阻尼很小时

(我们在实验期间接近这种状态)，我们使用周期守恒分数（PC）。$q \frac{l}{g}$

对于完整约束摆，很明显摆长 l 在整个实验过程中保持不变，这是系统完整约束的结果。因此，我们也考虑摆长作为物理不变量。

D.5 PSS

当我们获得物理不变量值 C 时，我们计算随时间变化的相对标准偏差：

$$C = C/C \tag{23}$$

为确保分数在[0, 1]范围内，我们基于不变量设计了物理分数，具体如下：

$$S = \frac{1}{1 + \alpha * C} \tag{24}$$

其中 α 是一个归一化因子。在实验中，我们将其设置为 1.0。

在评分计算过程中，有两个关键问题需要考虑。首先，时间窗口必须仔细选择。对于每条轨迹，我们使用滑动时间窗口将其划分，并从所有片段中选择最高分作为轨迹的总体评分。这种方法解决了现实世界实验（如弹跳球实验）中的一个关键挑战，因为在弹跳点附近出现的波动会导致标准偏差增大和评分降低。通过选取所有片段中的最高分，我们有效地捕捉了轨迹中最稳定的部分。

在我们的实验中，时间窗口长度设置为总轨迹持续时间的 10%到 25%。具体来说，对于真实视频，我们使用 t= L/10，而对于生成视频，我们使用 t= L/4。这种窗口大小的差异是必要的，因为生成视频的总持续时间要短得多——如果使用 t= L/10，轨迹片段会过短，无法进行有意义的分析。

其次，适当的缩放至关重要：由于轨迹坐标记录在像素空间而不是真实世界的 3D 坐标中，需要将精确的坐标转换到物理单位。值得注意的是，不适当的缩放会显著影响总能量计算。第三，我们需要小心不要在所选物理不变量的均值接近零时选择范围。所选物理不变量的均值的绝对值应等于或大于标准偏差的 10 倍：

$$C = 10 * C \tag{25}$$

这样就可以忽略均值能量/加速度的影响接近零。在实验中，如果 $|C| \geqslant C$，我们根据公式 24 计算物理分数。否则，我们使用以下公式 26，该公式采用绝对标准偏差而不是相对标准偏差。

$$S = \frac{1}{1 + \alpha * C} \tag{26}$$

这种方法有两个主要局限性。首先，分数对时间窗口大小的选择非常敏感。较大的时间窗口往往会得到较低的分数，因为它们包含了轨迹中更多的波动。因此，我们在评估不同的模型生成时保持时间窗口不变。其次，该方法可能无法检测到生成视频中物体长时间保持静止的非物理行为；然而，这可以通过丢弃因静止而导致的视频来解决这个问题。

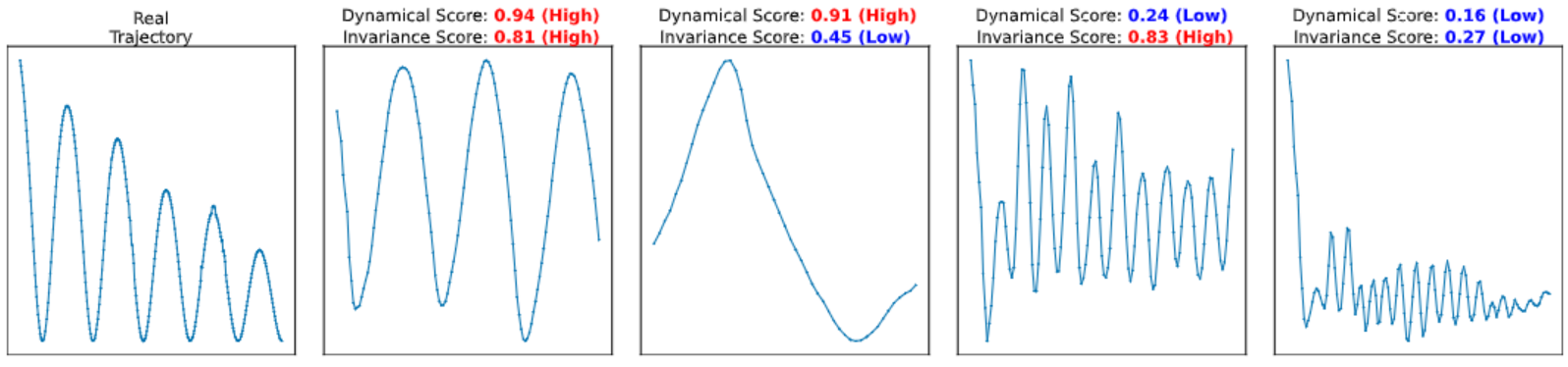


Figure 15: Real pendulum trajectory alongside four cases of generated videos, demonstrating how different combinations of dynamical and physical invariance scores appear in practice.

E Extended Limitations

The main focus of our benchmark is towards a set of Newtonian scenarios, which take place under a controlled environment and static camera conditions. While this enables clean invariants and can be reproducible, it limits a full scale physical coverage. Additionally, ambiguity is introduced (e.g. distortion of lens, unknown scale, etc.) because all of our estimates come from 2D pixel trajectories without any camera calibration. At the same time, we rely on assumptions such as negligible air resistance and friction. Unmodeled forces, such as air drag, friction and rotational kinetic energy could cause violations, which do not reflect generations that are wrong but rather evaluator mismatch. We perform trajectory extraction from generated videos using fixed first/last frame or short multi-frame conditioning that cannot specify initial conditions fully, such as the mass of an object or the coefficient of the spring. Moreover, a physically plausible generation can be penalized heavily when some of the assumptions are violated, leading to misleading results. Similarly, one single invariant can look really "good", but the generated video could violate multiple laws, making it physically implausible.

Errors in segmentation and tracking introduce noise. Additionally, we are only considering short clips, and the model might drift away from physical laws over a longer temporal horizon, which our current metrics would not be capable of capturing. Finally, it should be noted that while our scores remain proxies rather than direct measurements of underlying physical plausibility, that would require falsifiable and controllable VGM generation. Prior benchmarks that rely on judgments from humans or VLMs are prone to hallucinations and often miss slight inconsistencies. In contrast, our scores avoid these issues but still need to be interpreted jointly to give a much more reliable picture. With that in mind, Morpheus should be treated as a focused and reproducible stress test for core physical laws and not a benchmark that can test and evaluate all physical dynamics that might take place in a single video.

F Video Generative Models Details

At their core, latent video generative models often utilize a combination of a 3D Variational Autoencoder (VAE) (Kingma, 2013; Lin et al., 2024) to tokenize individual frames, a text encoder, like T5 (Raffel et al., 2020) to encode frames into latent. During training a noisy latent is produced by the forward diffusion process. This latent is then processed by a parametrized model, either a transformer model (Yang et al., 2023) or a U-Net (Ronneberger et al., 2015; Zhou et al., 2022; Bar-Tal et al., 2024) resulting in a patchified long sequence of visual tokens, in case of the former type of model.

Depending on the model architecture, different input modalities can be handled like text-to-video, image-to-video, text+image-to-video, and sometimes video continuation regimes facilitating both open-domain and controlled generation scenarios (Yang et al., 2024b; Kong et al., 2024; Agarwal et al., 2025; Brooks et al., 2024b). Although some state-of-the-art video generation models adopt an autoregressive framework, predicting frames sequentially based on prior outputs (Weng et al., 2024; Deng et al., 2024; Weissenborn et al., 2020), many others utilize non-autoregressive approaches to generate frames simultaneously (Yang et al., 2024b; Kong et al., 2024; Xing et al., 2025). In Table 4, we specify the parameters of particular models used in our benchmark, along with it's architectural design choices. As to faithfully get the best generation outcome we use the best hyperparameters reported for each model. Due to the high API costs-per five seconds video, along with the impressive

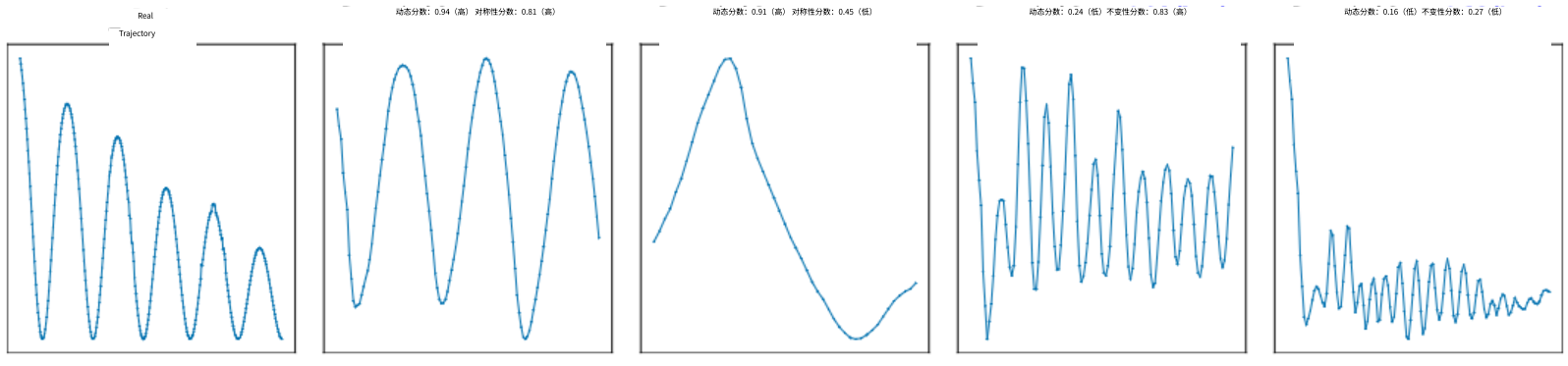


图 15：真实摆动轨迹与四个生成的视频案例，展示了动态和物理不变性分数的不同组合在实际中的表现。

E 扩展局限性

我们的基准测试主要关注一组牛顿场景，这些场景在受控环境和静态摄像机条件下进行。虽然这能够实现干净的等价关系并且可以重复，但它限制了全面的物理覆盖范围。此外，由于所有估计都来自未经相机校准的 2D 像素轨迹，因此引入了模糊性（例如镜头畸变、未知比例等）。同时，我们依赖于诸如空气阻力可忽略和摩擦力可忽略等假设。未建模的力，如空气阻力、摩擦力和旋转动能，可能会导致违规，这并不反映错误的生成，而是评估器不匹配。我们从生成的视频中提取轨迹，使用固定的首/尾帧或短多帧条件，这些方法无法完全指定初始条件，例如物体的质量或弹簧系数。此外，当某些假设被违反时，一个物理上合理的生成可能会受到严重惩罚，导致误导性结果。 同样地，单个不变量可能看起来非常"好"，但生成的视频可能违反多条定律，使其在物理上不可行。

分割和跟踪中的错误会引入噪声。此外，我们只考虑了短片段，模型在更长的时间尺度上可能会偏离物理定律，而我们的当前指标无法捕捉这一点。最后，应该指出的是，虽然我们的分数仍然是潜在物理合理性的代理指标而非直接测量，但这需要可证伪且可控的视频生成模型。之前的基准测试依赖于人类或视觉语言模型（VLM）的判断，容易产生幻觉，并且常常忽略细微的不一致。相比之下，我们的分数避免了这些问题，但仍需要联合解读以提供更可靠的评估。考虑到这一点，Morpheus 应该被视为核心物理定律的聚焦且可复现的压力测试，而不是一个可以测试和评估单个视频中可能发生的所有物理动态的基准测试。

F 视频生成模型细节

在其核心，潜在视频生成模型通常结合使用 3D 变分自动编码器（VAE）（Kingma，2013；Lin 等人，2024）来对单个帧进行分词，以及像 T5（Raffel 等人，2020）这样的文本编码器来将帧编码为潜在表示。在训练过程中，正向扩散过程会产生一个噪声潜在表示。然后，这个潜在表示由一个参数化模型进行处理，该模型要么是一个转换器模型（Yang 等人，2023），要么是一个 U-Net（Ronneberger 等人，2015；Zhou 等人，2022；Bar-Tal 等人，2024），对于前一种类型的模型，结果会生成一个分块的长序列视觉标记。

根据模型架构不同，不同的输入模态可以被处理，如文本到视频、图像到视频、文本+图像到视频，有时还包括视频延续模式，以促进开放域和受控生成场景（Yang 等人，2024b；Kong 等人，2024；Agarwal 等人，2025；Brooks 等人，2024b）。尽管一些最先进的视频生成模型采用自回归框架，根据先前的输出生成帧序列（Weng 等人，2024；Deng 等人，2024；Weissenborn 等人，2020），但许多其他模型利用非自回归方法同时生成帧（Yang 等人，2024b；Kong 等人，2024；Xing 等人，2025）。在表 4 中，我们指定了我们基准测试中使用的特定模型的参数，以及它的架构设计选择。为了忠实地获得最佳生成结果，我们使用每个模型报告的最佳超参数。由于每五秒钟视频的高 API 成本，以及令人印象深刻的

performance of COSMOS (Agarwal et al., 2025) among established benchmarks, we opted to not use any closed-source VGMs like (Brooks et al., 2024a; Kling AI, 2024). Still all the open-source alternatives used for our analysis match the performance of most closed-source one across established benchmarks for VGMs.

Model [Params.]	Resolution	Number of Video Frames	Guidance Scale	Sampling Steps
CogVideoX [5B]	960 x 768	84	6.0	50
Cosmos-Predict 1 [14B]	1280 × 704	121	7.0	35
Cosmos-Predict 2 [14B]	1280 × 704	93	7.0	35
WAN2.1 [14B]	1280 × 720	121	5.0	40
LTX-Video [13B]	960x736	81	3.0	50
PyramidalFlow[12B]	1280 x 768	121	4.0	10

Table 4: Details of video generation models adopted in our benchmark study, including their resolution, number of video frames, guidance scale, and sampling steps.

G Prompts for Video Generation Models

For each experiment, we carefully designed a prompt that describes the physical setup and motion of the experiment being conducted. For example, in the falling ball experiment, the prompt specifies that the ball falls and makes contact with the table below. Similarly, in the projectile experiment, we describe how the ball is launched at a slight upward angle and follows a natural parabolic trajectory. We enhance these prompts using an internally provided upsampler or ChatGLM if no upsampler is provided to incorporate more detailed scene descriptions and contextual elements derived from the reference images. All prompts are shown in Table 5 and Table 6.

随着 COSMOS（Agarwal 等人，2025）在现有基准测试中表现出色，我们选择不使用任何闭源视频生成模型（VGM），例如（Brooks 等人，2024a；Kling AI，2024）。然而，我们用于分析的开放源替代方案在 VGM 的现有基准测试中与大多数闭源模型性能相当。

模型 [参数]	分辨率	数量 视频帧	指导 规模	采样 步骤
CogVideoX [5B]	960 x 768	84	6.0	50
Cosmos-Predict 1 [14B] 1280 × 704 121 7.0 35 Cosmos-Predict 2 [14B] 1280 × 704 93 7.0 35 WAN2.1 [14B]				
	1280 × 720 121		5.0	40
LTX-Video [13B]	960x736 81		3.0	50
PyramidalFlow[12B] 1280 x 768 121			4.0	10

表 4：我们基准研究中所采用的视频生成模型的详细信息，包括它们的分辨率、视频帧数、指导比例和采样步数。

视频生成模型的 G 提示

对于每个实验，我们精心设计了一个描述实验物理设置和运动的提示。例如，在落球实验中，提示指定球体落下并接触下方桌面。同样，在抛射实验中，我们描述球体以微小的向上角度发射并遵循自然的抛物线轨迹。如果没有提供内部提供的上采样器，我们使用 ChatGLM 增强这些提示，以整合更详细的场景描述和从参考图像中得出的上下文元素。所有提示显示在表 5 和表 6 中。

Experiment Name	Base Prompt	Enhanced Prompt
Falling Ball	Orange ping-pong ball falling down and making impact with the table surface below. Fixed camera view, no camera movement.	A ping-pong ball is captured in mid-air, suspended above a laboratory table, poised to make contact with the surface below. The ball's descent is governed by the force of gravity, creating an arc that suggests a controlled experiment in progress. The backdrop is a stark, clinical room with a neutral palette, punctuated by the sterile lines of a metal frame and the functional design of a nearby cabinet. The lighting is subdued, casting a soft glow that highlights the ball's trajectory and the anticipation of impact. The table beneath the ball is marked with faint lines, perhaps indicating measurements or guidelines for the experiment. As the ball continues its downward journey, it will likely bounce off the table, adding a dynamic element to the scene and marking the conclusion of this controlled descent. Fixed camera view, no camera movement.
Projectile	A single, small 3D-printed ball, dark gray in color, is launched from a plastic, small-scale ramp with a slight upward angle. The ball follows a natural, smooth, arcing trajectory upward and then downward, continuing along that arc until it exits the right side of the video frame. The video should accurately simulate the ball's motion under standard earth gravity, showing a clear parabolic arc. The video should emphasize a smooth and realistic physics-based movement of the ball without any sudden changes in speed. The ball should not bounce or collide with any objects in the scene. Fixed camera view, no camera movement.	In a meticulously crafted scene, a solitary, dark gray 3D-printed ball, with its sleek, spherical form, is propelled from a plastic ramp that slopes gently upward. The ball, weighing a mere fraction of a kilogram, is captured in high-definition, showcasing every nuance of its motion. As it leaves the ramp's edge, the ball arcs gracefully into the air, its trajectory a perfect parabola that mirrors the laws of physics under standard earth gravity. The video's frame follows the ball's smooth ascent and descent, highlighting the ball's consistent speed and the absence of any sudden accelerations or decelerations. The scene remains unobstructed, ensuring that the ball's journey is uninterrupted by any external forces, save for the pull of gravity, resulting in a visually stunning and scientifically accurate demonstration of a parabolic motion. Fixed camera view, no camera movement.
Rolling Can	An empty soda can rolls down a wooden surface. Experiment carried out in a laboratory controlled environment. Fixed camera view, no camera movement.	An empty aluminum red soda can rolls steadily down an inclined wooden board in a controlled laboratory environment. This motion should be depicted with physical realism, accurately simulating the key dynamics: the can accelerates under gravity, with its combined rotational and translational movement appearing authentic and governed by the frictional interaction with the wooden surface, ensuring a physically plausible rolling movement. Fixed camera view, no camera movement.
Sliding Book	A book slides down a wooden surface. Experiment carried out in a laboratory controlled environment. Fixed camera view, no camera movement.	A book slides steadily down an inclined wooden board in a controlled laboratory environment. This motion should be depicted with physical realism, accurately simulating the key dynamics: the book accelerates due to the component of gravity acting along the incline, opposed by kinetic friction from the wooden surface. Its movement should be purely translational, maintaining consistent contact with the board and without any significant rotation or tumbling, ensuring a physically plausible descent. Fixed camera view, no camera movement.
Holonomic Pendulum	A single pendulum moving retrogressively back and forth. At the bottom of a pendulum, there is a ball attached to it. The pendulum is holonomic. Fixed camera view, no camera movement.	A pendulum with a spherical ball attached swings back and forth in a controlled manner, its motion captured in a moment of retrograde swing. The pendulum's arm, likely made of metal, extends horizontally from a stand, connected to a pivot point that allows for rotational movement. The ball, positioned at the lower end of the pendulum, appears to be in motion, indicating the pendulum's swing. The environment suggests a laboratory or testing setting, with a backdrop of technical apparatus and equipment, and the lighting is artificial, casting a uniform glow over the scene. The pendulum's movement, while currently in a retrogressive swing, could potentially change direction, continuing its oscillatory motion. Fixed camera view, no camera movement.

Table 5: Base and enhanced textual prompts used for video generation experiments: Falling Ball, Projectile, Rolling Can, Sliding Book, Holonomic Pendulum. Enhanced prompts are generated using corresponding upsampler VLMs such as ChatGLM (Zeng et al., 2024) and incorporate more detailed scene descriptions and contextual elements derived from the reference images. Slight modifications to these prompts were made in case of different variants such as object type.

实验名称	基础提示词	增强提示词
下落的球	橙色乒乓球下落并与下方桌面表面发生碰撞。固定相机视角，无相机移动。	一个乒乓球在空中被捕捉，悬挂在实验室桌子上方，准备与下方表面接触。球的下落受重力影响，形成一条弧线，暗示着正在进行的一项受控实验。背景是一个 stark、临床的房间，色调中性，点缀着金属框架的洁净线条和附近柜子的功能性设计。光线柔和，照亮了球的轨迹和撞击的期待感。球下方的桌子上有淡淡的线条，可能表示测量或实验的指导线。随着球继续其下落过程，它可能会弹起，为场景增添动态元素，并标志着这一受控下落的结束。固定相机视角，无相机移动。
抛射物	一个单一的、小型的 3D 打印球，深灰色颜色，从一个小型塑料斜坡上以微小的上倾角发射。小球沿着自然、平滑的抛物线轨迹向上然后向下运动，继续沿着该弧线直到离开视频帧的右侧。视频应准确模拟小球在标准地球重力下的运动，显示清晰的抛物线轨迹。视频应强调小球平滑且真实的基于物理的运动，速度上不应有任何突然变化。小球不应与场景中的任何物体发生弹跳或碰撞。固定摄像机视角，无摄像机移动。	在一个精心设计的场景中，一个孤独的、深灰色的 3D 打印球体，以其光滑的球形形态，从一条倾斜向上的塑料斜坡上被推动出去。这个仅重几公斤的球体被高清拍摄，展现其运动的每一个细节。当它离开斜坡边缘时，球体优雅地弧线飞向空中，其轨迹是一个完美的抛物线，反映了在标准地球重力下的物理定律。视频的帧跟随球体的平稳上升和下降，突出了球体保持匀速以及没有任何突然的加速或减速。整个场景没有阻碍，确保球体的旅程不会受到任何外部力的干扰，除了重力的作用，从而呈现出视觉上令人惊叹且科学上精确的画面。抛物线运动的演示。固定相机视角，无相机移动。
滚动的罐子	一个空的汽水罐沿着木头滚动表面。实验在受控的实验室环境中进行。固定摄像机视角，无摄像机移动。	一个空的铝制红色汽水罐在受控的实验室环境中稳稳地沿着倾斜的木板滚下。这种运动应该以物理真实性来描绘，准确模拟关键动力学：罐子在重力作用下加速，其旋转和平动结合的运动看起来真实，并由与木板表面的摩擦相互作用所控制，确保一个物理上合理的滚动运动。固定摄像机视角，无摄像机移动。
滑动的书	一本书滑下木质表面。在受控的实验室环境中进行的实验。固定摄像机视角，无摄像机移动。	一本书在受控的实验室环境中稳稳地沿着倾斜的木板滑下。这种运动应该以物理真实性来描绘，准确模拟关键动力学：书由于沿倾斜面的重力分量而加速，受到来自木质表面的动摩擦力阻碍。其运动应为纯平动，保持与木板的持续接触，且无显著旋转或翻滚，确保一个物理上合理的下落。固定摄像机视角，无摄像机移动。
全息的钟摆	一个单摆来回做逆行运动。在钟摆的底部，有一个球附着在它上面。钟摆是全息的。固定相机视角，没有相机运动。	一个带有球形球的钟摆在受控的方式下来回摆动，其运动在逆行摆动的一个瞬间被捕捉。钟摆的臂，可能是由金属制成的，延伸从支架水平悬挂，连接到一个枢轴点，允许旋转运动。摆球位于摆锤的下方，看起来在运动，表明摆锤在摆动。环境暗示是实验室或测试场所，背景是技术设备和装置，照明是人工的，在场景上投下均匀的光芒。摆锤的运动，虽然目前是逆行摆动，但可能会改变方向，继续其振荡运动。固定相机视角，无相机运动。

表 5：用于视频生成实验的基础和增强文本提示：下落的球、抛射物、滚动的罐子、滑动的书、完整约束摆。增强提示使用相应的上采样视觉语言模型（如 ChatGLM（Zeng 等人，2024 年））生成，并包含从参考图像中得出的更详细的场景描述和上下文元素。对于不同变体（如物体类型），对这些提示进行了轻微修改。

Experiment Name	Base Prompt	Enhanced Prompt
Double Pendulum	Double pendulum, consisting of a purple and an orange segment. Each segment moves independently. Fixed camera view, no camera movement.	In a meticulously arranged laboratory setting, a double pendulum setup swings gracefully, each pendulum segment adhering to the immutable laws of physics. The upper pendulum, a sleek purple rod, contrasts strikingly with the lower orange rod, both suspended from a sturdy, metallic frame. The room is bathed in soft, ambient light, casting subtle shadows that accentuate the pendulums' arcs. The scene captures the intricate dance of the pendulums, their movements a mesmerizing testament to the natural order, with each swing a silent symphony of motion and balance. Fixed camera view, no camera movement.
Bouncing Ball	A single orange ping pong ball bounces vertically as a result of making impact with the table after being in free fall. The ball starts in the center of the frame, and moves upwards. Fixed camera view, no camera movement.	A solitary orange ping pong ball, with its vibrant hue standing out against the stark white of the table, plummets from the center of the frame. As the ball bounces upwards, it arcs gracefully, the trajectory a perfect parabola. The frame remains centered, emphasizing the ball's solitary dance of motion and the physics of its rebound. Fixed camera view, no camera movement.
Collision	Generate a realistic video of two metallic spheres colliding. In the first frames the leftmost sphere in the video is moving towards the static one. At the moment of contact the ball in motion transfer its kinetic energy to the ball at rest. The two spheres have identical physical properties, namely material, shapes, masses. The momentum should be conserved. Fixed camera view, no camera movement.	In a high-definition, slow-motion sequence, two identical metallic spheres, each polished to a mirror-like finish, are meticulously positioned on a frictionless surface. The left sphere, propelled by an unseen force, hurtles towards the stationary sphere, its trajectory a perfect parabola. As the oncoming sphere approaches, the static sphere begins to subtly vibrate, a prelude to the impending collision. The moment of contact is captured in stunning detail, revealing the transfer of kinetic energy as the moving sphere's momentum is transferred to the stationary one. The spheres, crafted from the same dense material, maintain their spherical shapes and masses, ensuring the conservation of momentum throughout the impact. The scene is illuminated by a single, soft light source, casting long shadows and emphasizing the physics of the collision. Fixed camera view, no camera movement.
Spring	A single metallic slotted mass attached on a spring moving periodically up and down. Fixed camera view, no camera movement.	In a meticulously crafted scene, a solitary metallic slotted mass, polished to a gleaming silver, is ingeniously attached to a tensioned spring. The mass, weighing several pounds, oscillates with a fluid grace, its movement initiated by the subtle release of the spring. The camera captures the mass as it arcs upwards, the tension in the spring visible, before it gently descends with a rhythmic sway, the sound of its metallic slats clinking softly in the background. The scene is set against a stark white backdrop, emphasizing the stark contrast between the mass and the spring, and the smooth, periodic motion of the mechanical dance. Fixed camera view, no camera movement.

Table 6: Base and enhanced textual prompts used for video generation experiments: Double Pendulum, Bouncing Ball, Collision, Spring. Enhanced prompts are generated using corresponding upsampler VLMs such as ChatGLM (Zeng et al., 2024) and incorporate more detailed scene descriptions and contextual elements derived from the reference images. Slight modifications to these prompts were made in case of different variants such as object type.

实验名称	基础提示词	增强提示词
双笔 dulum	双摆，由一个紫色和一个橙色部分组成。每个部分独立运动。固定相机视角，无相机移动。	在一个精心布置的实验室环境中，一个双摆装置优雅地摆动，每个摆杆段都遵循着不可改变的物理定律。上摆是一个光滑的紫色杆，与下摆的橙色杆形成鲜明对比，两者都悬挂在一个坚固的金属框架上。房间被柔和的环境光照亮，投下微妙的阴影，突显出摆杆的轨迹。这一场景捕捉了摆杆的复杂舞动，它们的运动是对自然秩序的迷人证明，每一次摆动都是运动与平衡的无声交响曲。固定相机视角，无相机移动。
弹跳球	一个橙色的乒乓球在自由落体后撞击桌面而垂直弹跳。球从画面的中心开始，向上移动。固定摄像头视角，无相机移动。	一个孤独的橙色乒乓球，其鲜艳的色彩与桌面的纯白色形成鲜明对比，从画面的中心坠落。当球向上弹跳时，它优雅地划出弧线，轨迹是一个完美的抛物线。画面保持居中， 强调球的独自运动舞蹈及其弹跳的物理原理。固定摄像机视角，无摄像机移动。
碰撞生成一个逼真的金属物体视频	球体碰撞。在最初的几帧中，视频中最左侧的球体正朝着静止的球体移动。在接触的瞬间，运动的球体将其动能传递给静止的球体。这两个球体具有相同的物理属性，即材料、形状和质量。动量应该守恒。固定相机视角，无相机移动。	在一个高清、慢动作的序列中，两个完全相同的金属球，每个都抛光到镜面般的光泽，被精心放置在无摩擦的表面上。左侧的球被一股看不见的力量推动，朝着静止的球飞驰而去，其轨迹是一个完美的抛物线。随着接近的球越来越近，静止的球开始轻微振动，这是即将发生的碰撞的前奏。接触的瞬间被惊人地捕捉到，揭示了动能的传递，即运动球的动量传递给了静止的球。这两个球由相同的致密材料制成，保持了它们的球形和质量，确保在整个碰撞过程中动量守恒。场景由一个单一的柔和光源照亮，投下长长的阴影，突出了碰撞的物理学。固定摄像机视角，无摄像机移动。
弹簧	一个金属槽形质量附着在弹簧上，使其周期性地上下运动。 固定相机视角，无相机运动。	在一个精心设计的场景中，一个单独的金属槽形重物，被抛光至闪亮的银色，巧妙地连接在一张拉紧的弹簧上。这个重物重达数磅，以流畅优雅的姿态摆动，其运动由弹簧的微妙释放所引发。镜头捕捉到重物向上摆动时形成的弧线，弹簧的张力清晰可见，随后它以有节奏的摇摆轻轻下降，背景中传来金属槽轻微的碰撞声。整个场景设置在 stark 白色的背景下，突出了重物与弹簧之间的 stark 对比，以及机械舞蹈的平滑、周期性运动。固定摄像机视角，无摄像机移动。

表 6：用于视频生成实验的基础和增强文本提示：双摆、弹跳球、碰撞、弹簧。增强提示使用相应的上采样 VLM（如 ChatGLM（Zeng 等人，2024））生成，并包含从参考图像中提取的更详细的场景描述和上下文元素。对于不同变体（如物体类型），对这些提示进行了轻微的修改。

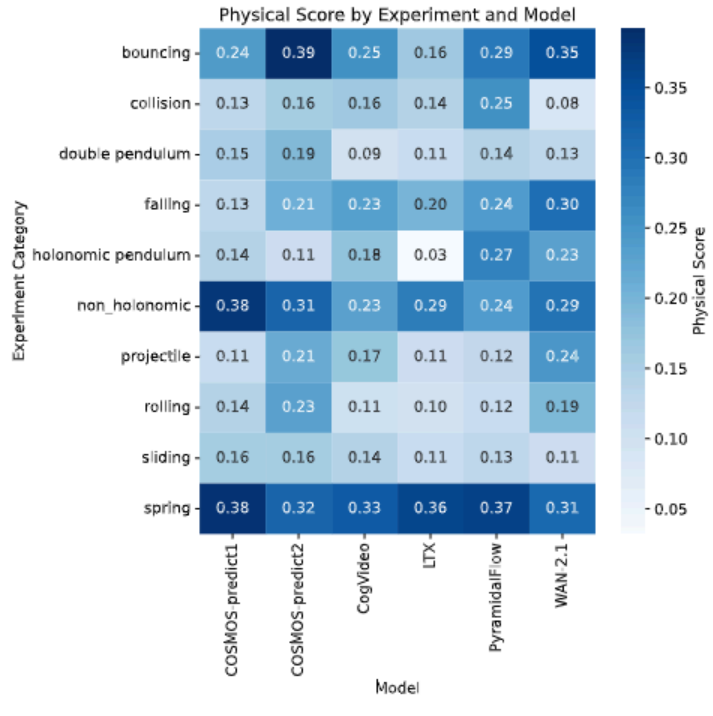


Figure 16: The distribution of the average physical invariance score across models and experiments.

Figure 17: The distribution of the average dynamical score across models and experiments.

H Additional Analysis

In Figure 25, we present an additional visualization of the energy and acceleration conservation. In Figure 15, we visualize the difference between the object trajectories of real-world and generated videos. In Figure 25(a), we present the total, kinetic, and potential energy over time. As expected, the total energy dissipates with every new bounce while remaining nearly constant between bounces.

I Usage of Large Language Models

Large language models were used solely for grammar correction and language polishing of the manuscript. They had no involvement in the research design, experimentation, analysis, or generation of results; all technical contributions are the work of the authors.

Ethics statement. This work involves the collection of benchmark datasets and the evaluation of publicly available models, all used in compliance with their original licenses. No new human data or personal annotations were gathered. Our study focuses exclusively on benchmarking and reproducibility, without any collection or processing of user data.

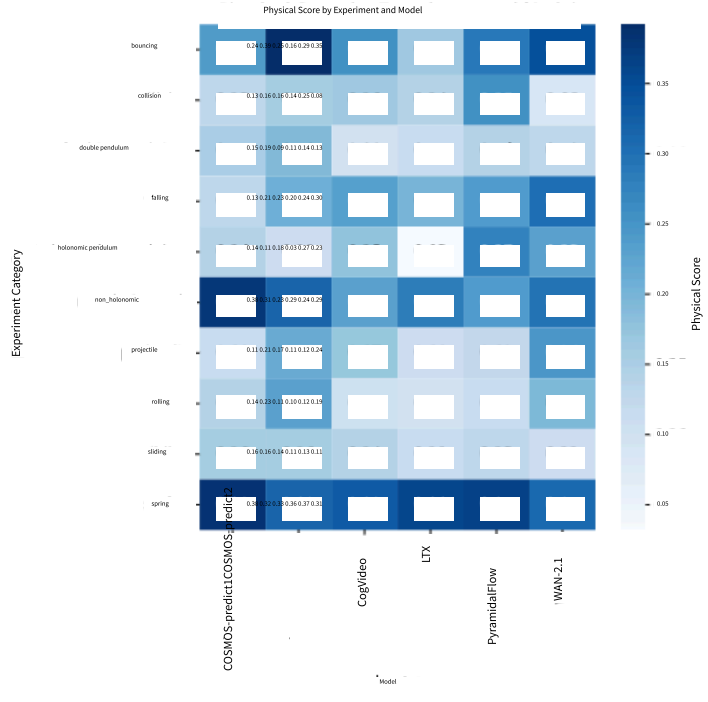


图 16：跨模型和实验的平均物理不变性分数分布

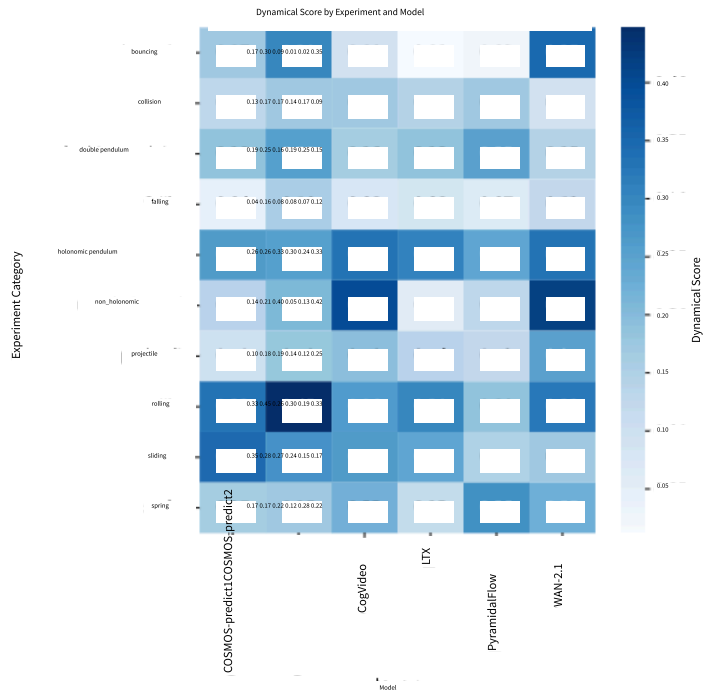


图 17：模型和实验中平均动力学分数的分布。

H 额外分析

在图 25 中，我们展示了能量和加速度守恒的额外可视化。在图 15 中，我们可视化了真实世界视频和生成视频的对象轨迹差异。在图 25(a)中，我们展示了随时间变化的总能量、动能和势能。正如预期，总能量在每个新的弹跳中耗散，而在弹跳之间保持几乎恒定。

I 大型语言模型的应用

大型语言模型仅用于修正手稿的语法和润色语言。它们未参与研究设计、实验、分析或结果生成；所有技术贡献均为作者的工作。

伦理声明。本研究涉及基准数据集的收集和公开可用模型的评估，均符合其原始许可协议。未收集新的个人数据或个人标注。我们的研究专注于基准测试和可复现性，未进行任何用户数据的收集或处理。

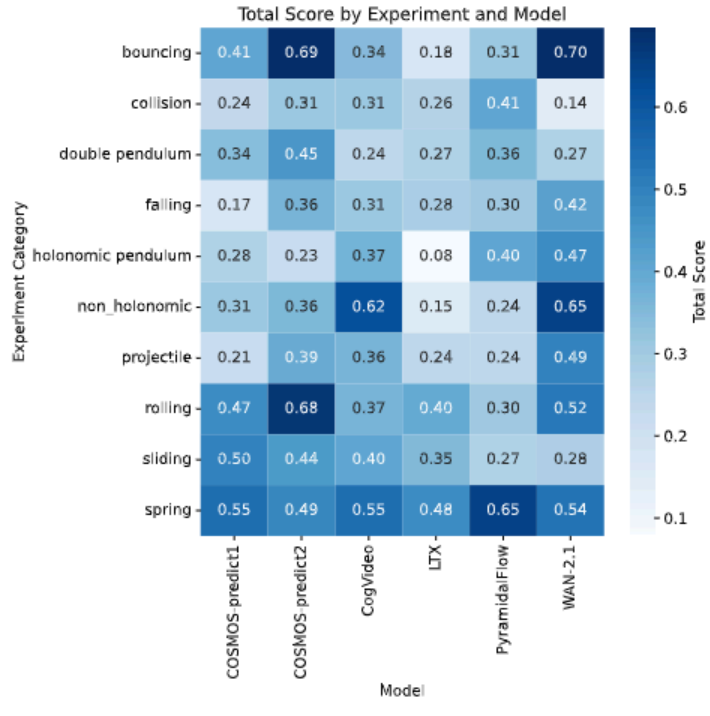


Figure 18: The distribution of the average total score across models and experiments.

Figure 19: The distribution of the average discard rate due to the **objects' disappearance** across models and experiments.

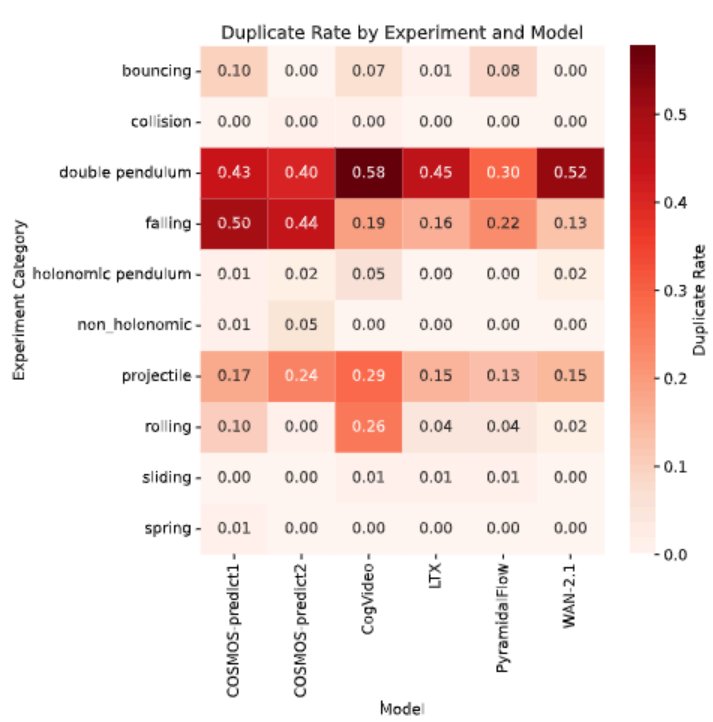


Figure 20: The distribution of the average discard rate due to the **objects' duplicates** across models and experiments.

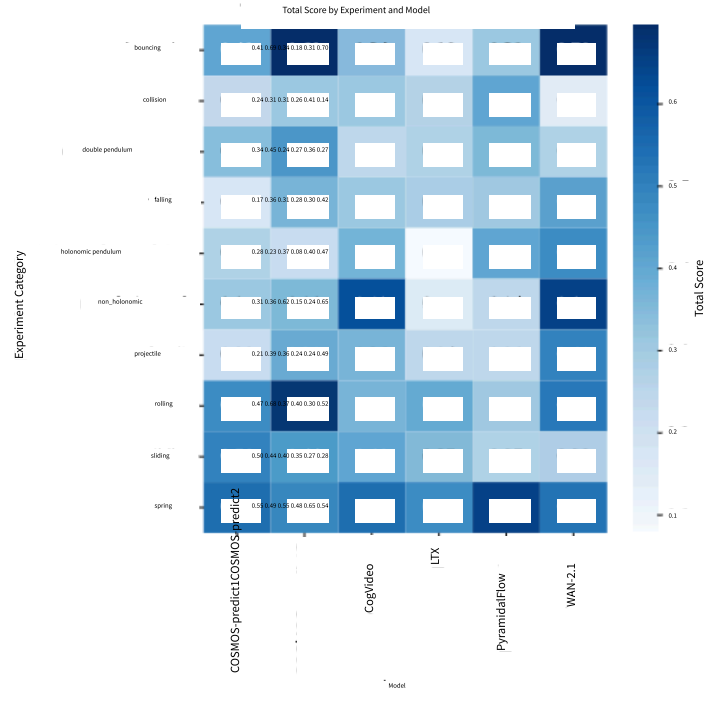


图 18：模型和实验的平均总分分布

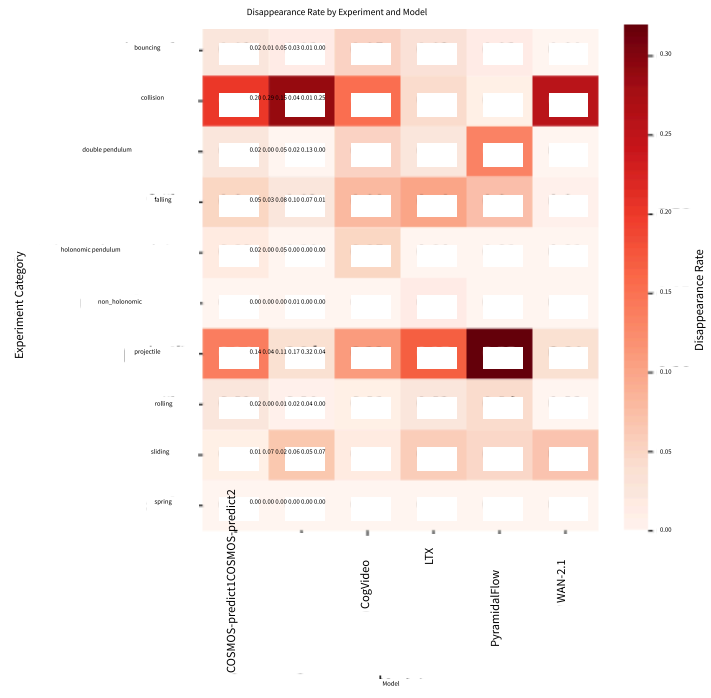


图 19：不同模型和实验中，因物体消失导致的平均丢弃率分布。

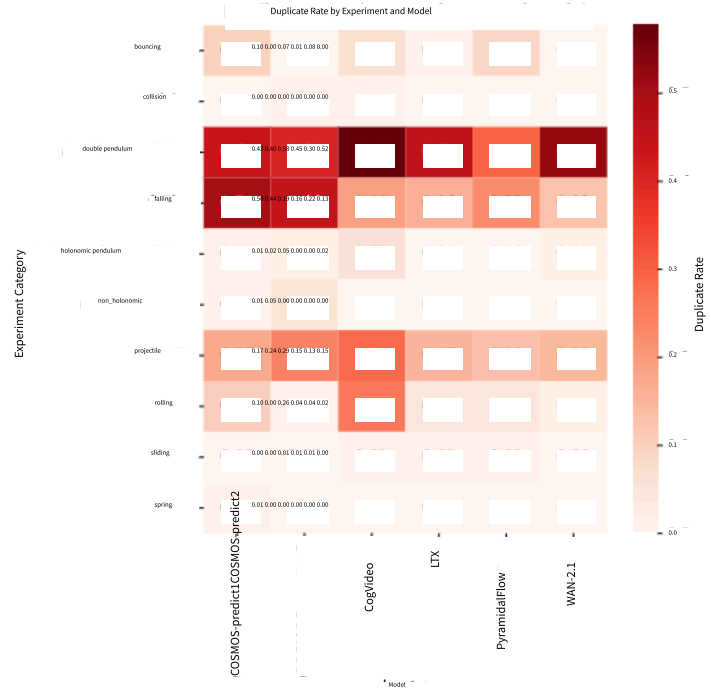


图 20：不同模型和实验中，因物体重复导致的平均丢弃率分布。

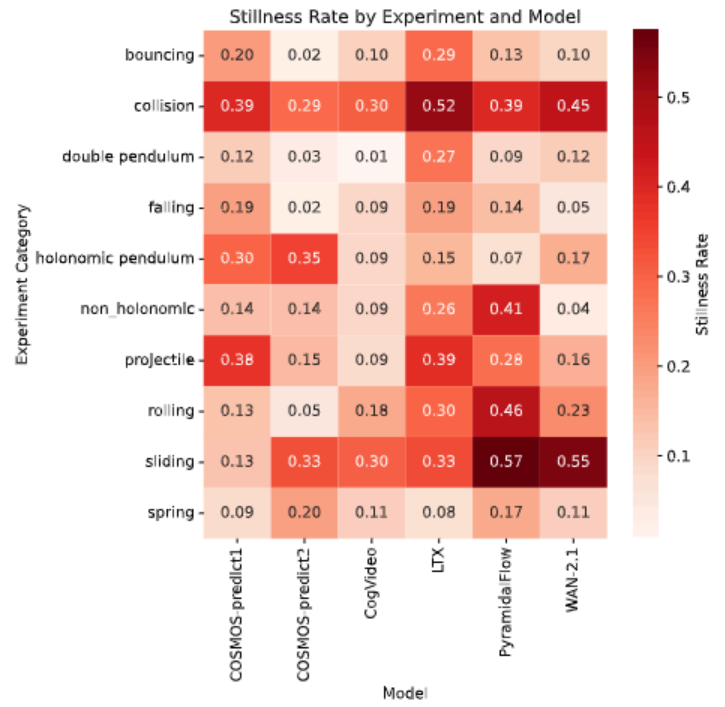


Figure 21: The distribution of the average discard rate due to **objects' stillness** across models and experiments.

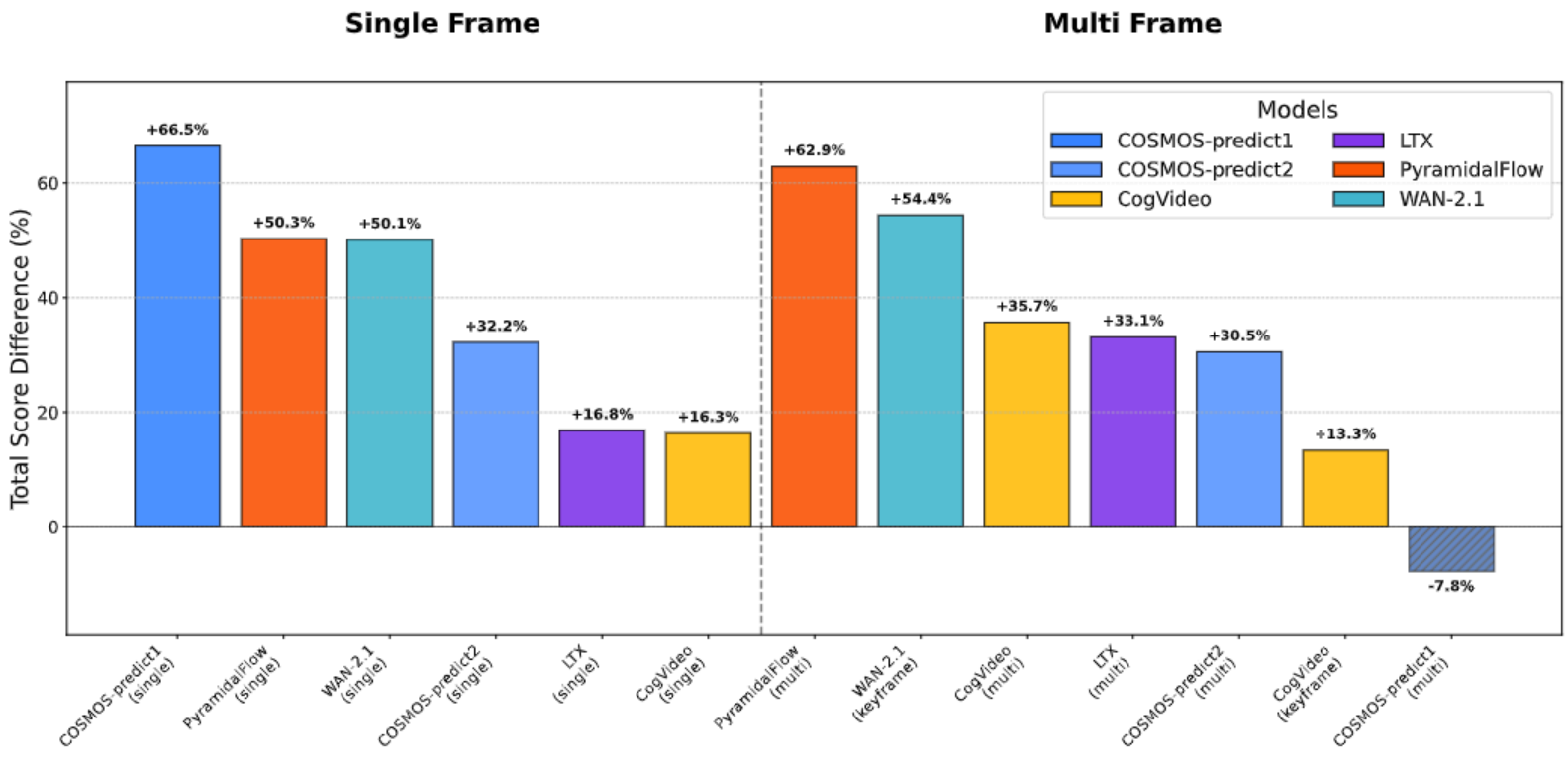


Figure 22: The relative change in scores evaluating on original data vs. on the augmented data. The augmented examples seem to be much harder.

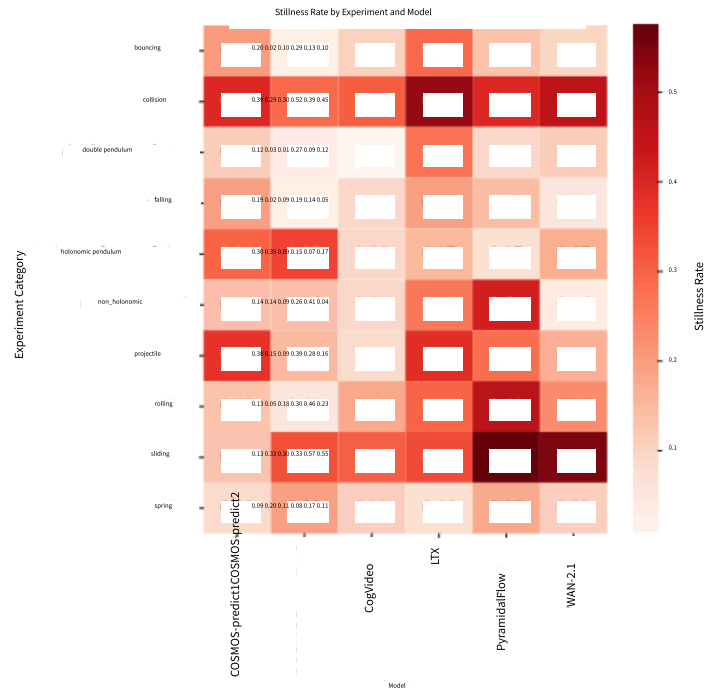


图 21：跨模型和实验的物体静止导致的平均丢弃率分布

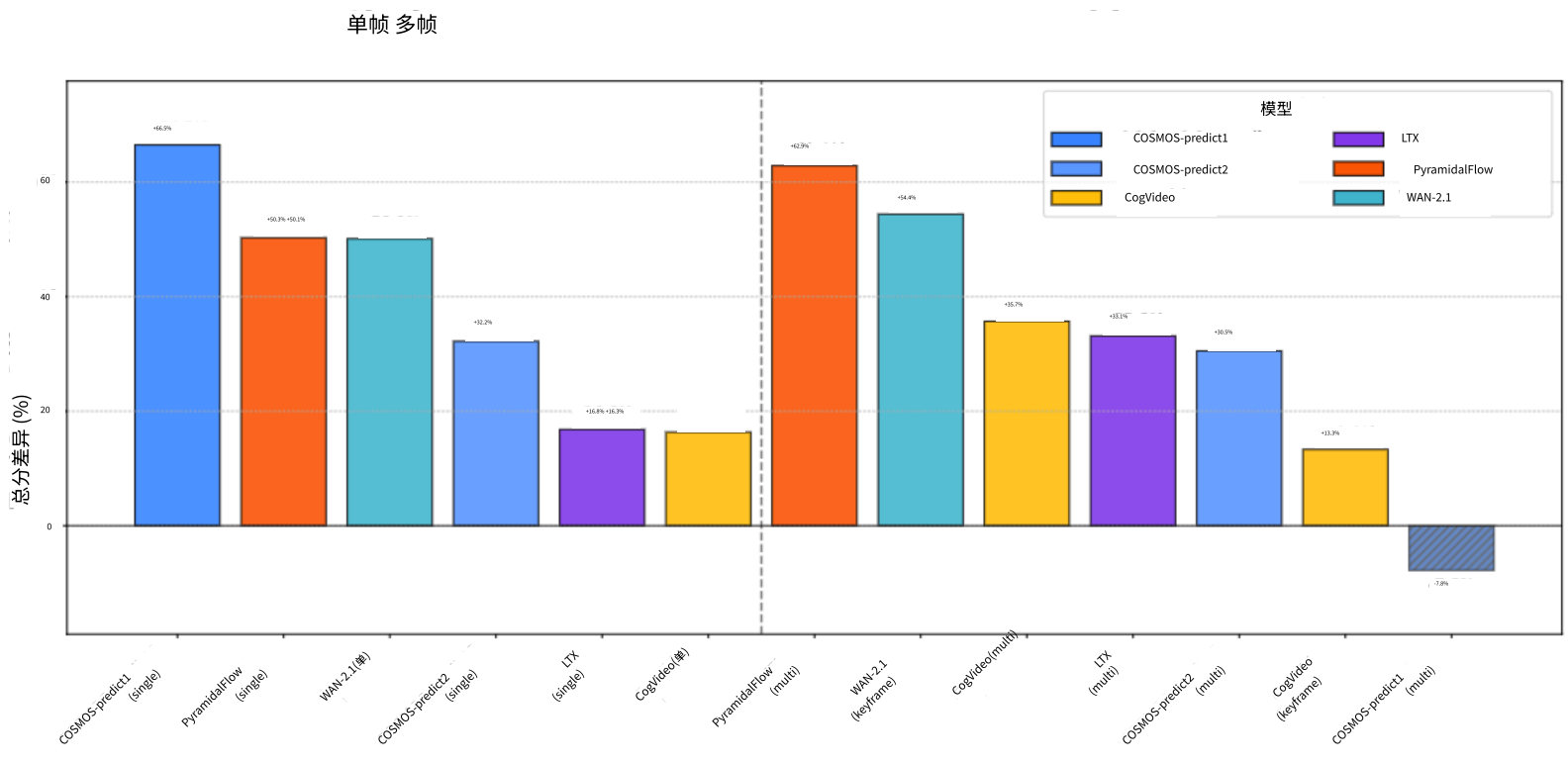


图 22：在原始数据上评估与在增强数据上评估的分数相对变化。增强示例似乎要难得多。

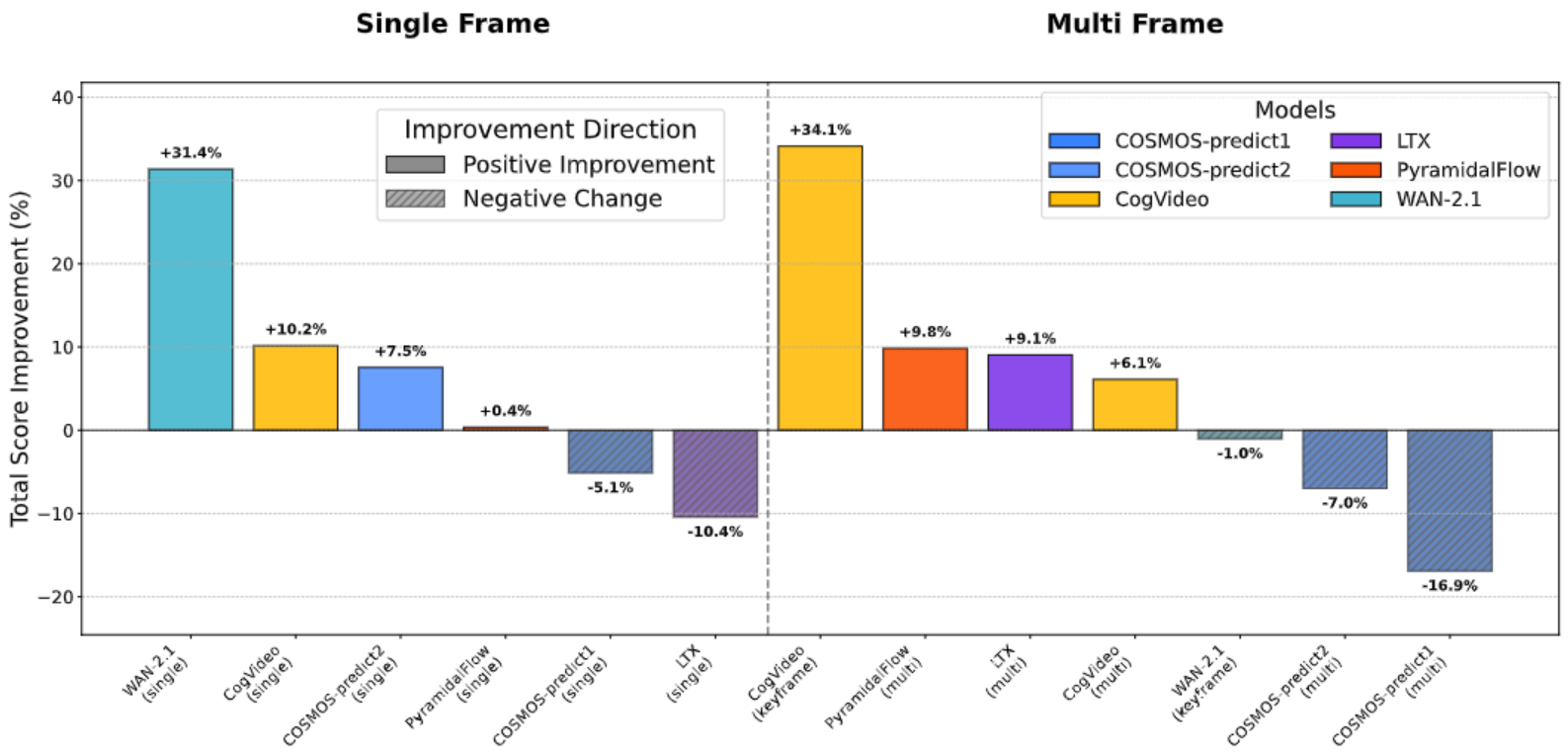


Figure 23: The relative change in scores for using an enhanced prompt vs the plain one.

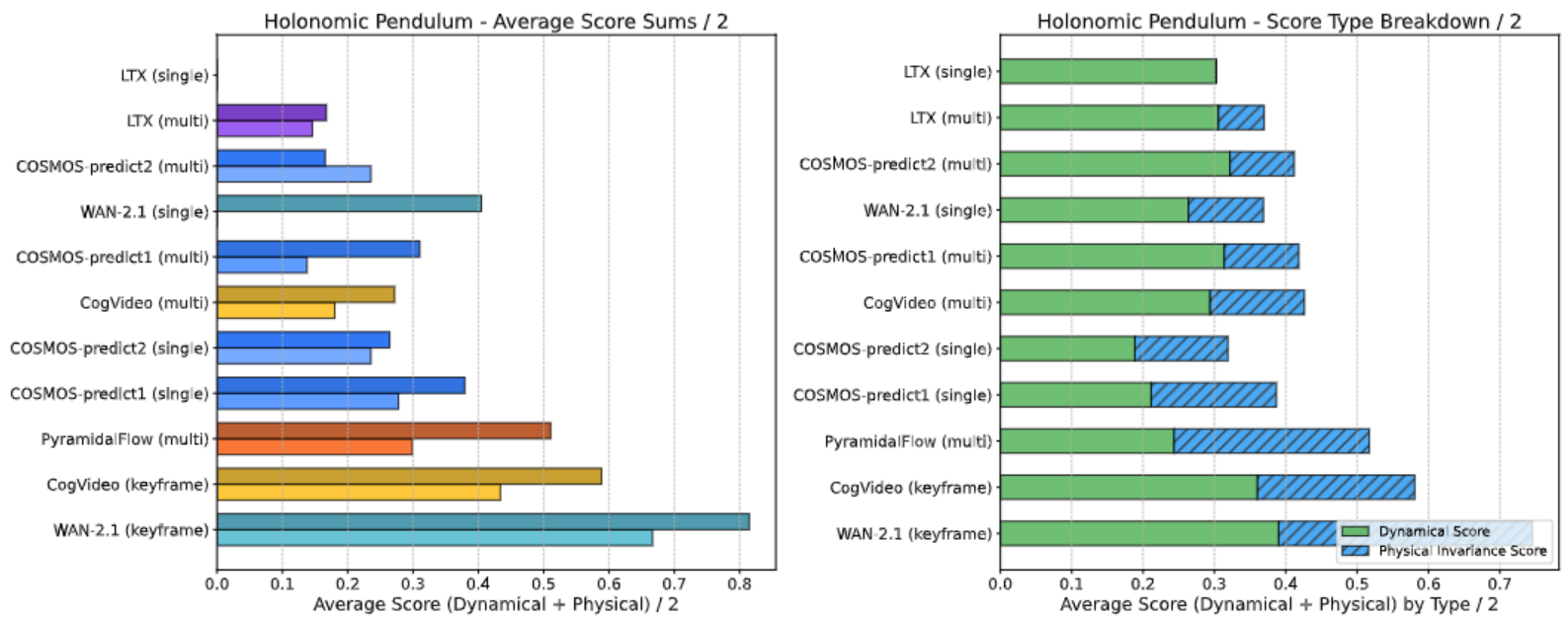


Figure 24: The scores for the holonomic pendulum experiment. On the left, darker colors denote *enhanced* textual prompt.

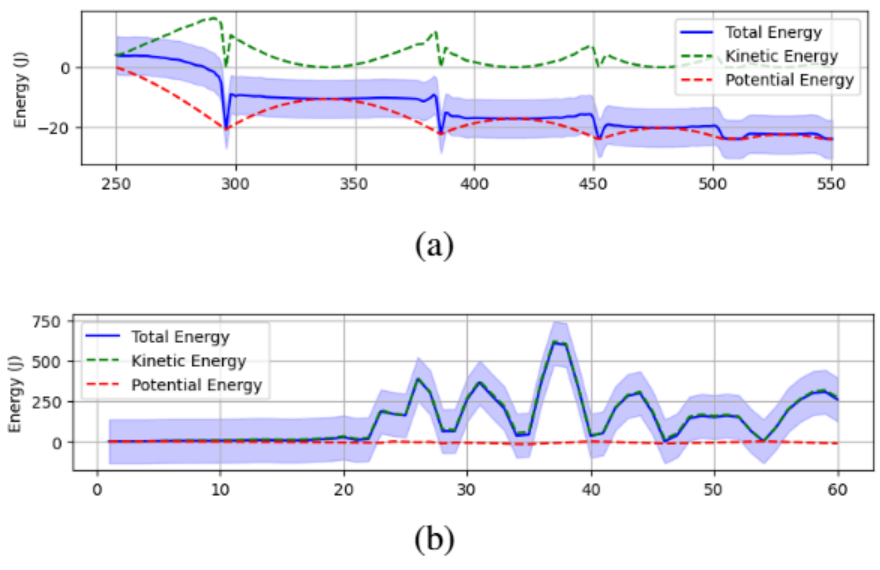


Figure 25: Energy analysis of real-world and generated falling + bouncing ball videos: (a) Real-world video energy conservation (b) CogVideoX plain single frame generated video energy conservation

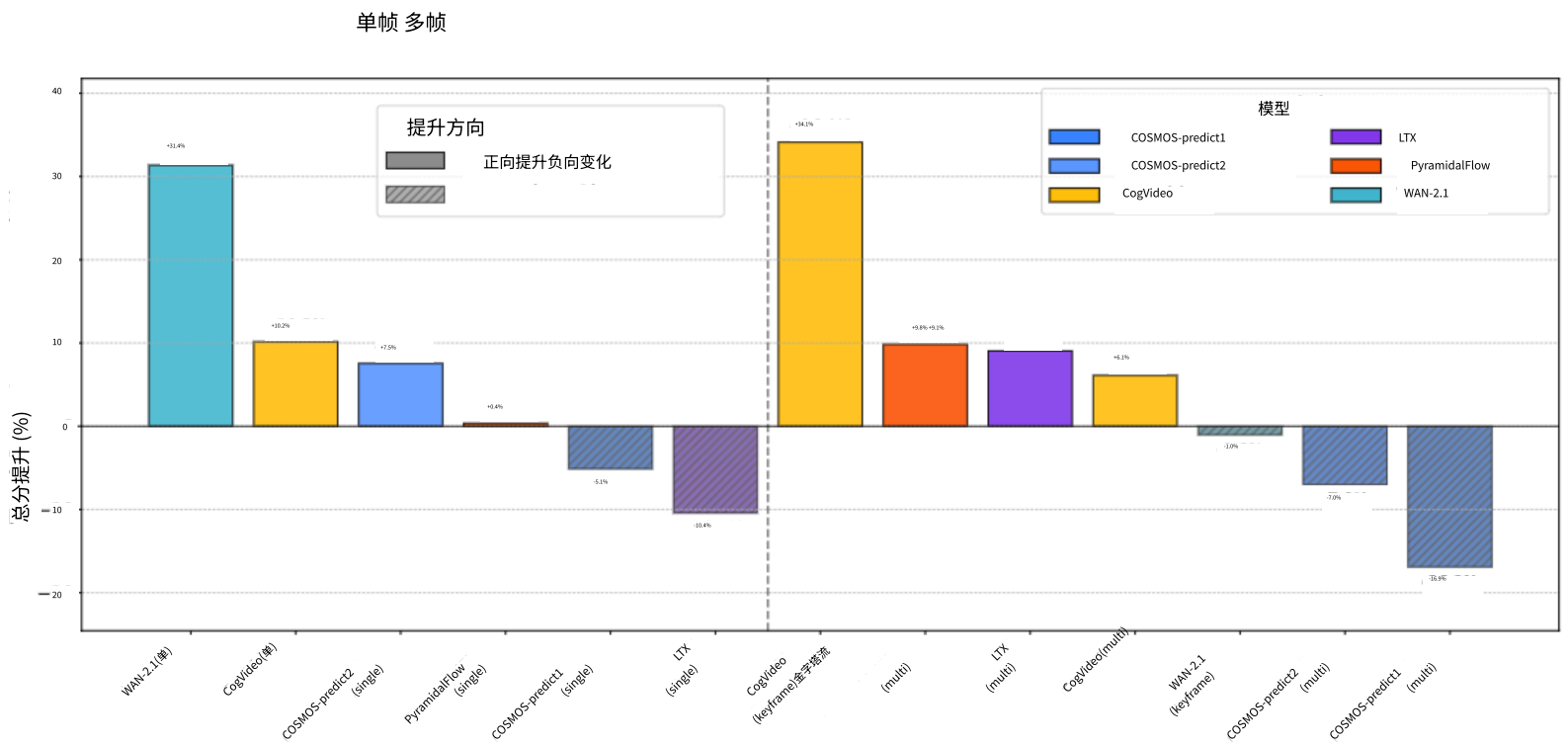


图 23：使用增强提示与普通提示的分数相对变化。

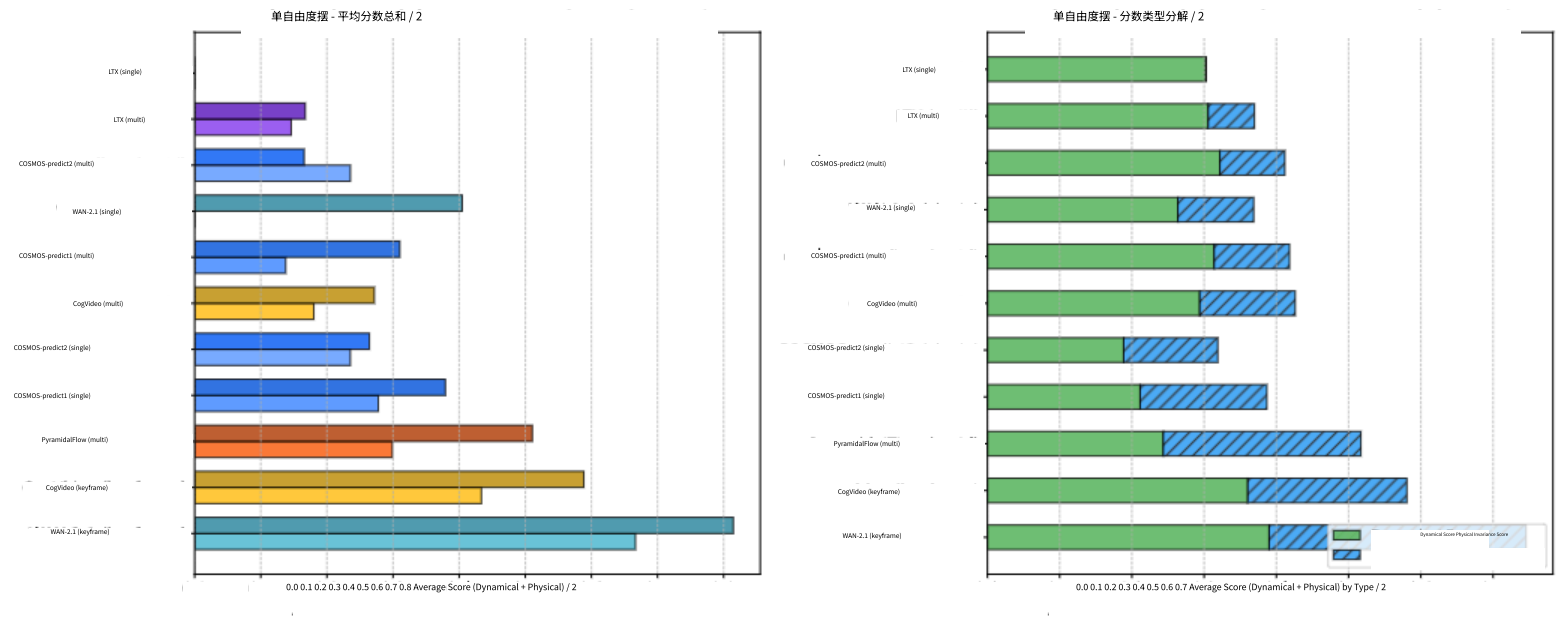


图 24：单自由度摆实验的分数。左侧，较深的颜色表示增强文本提示。

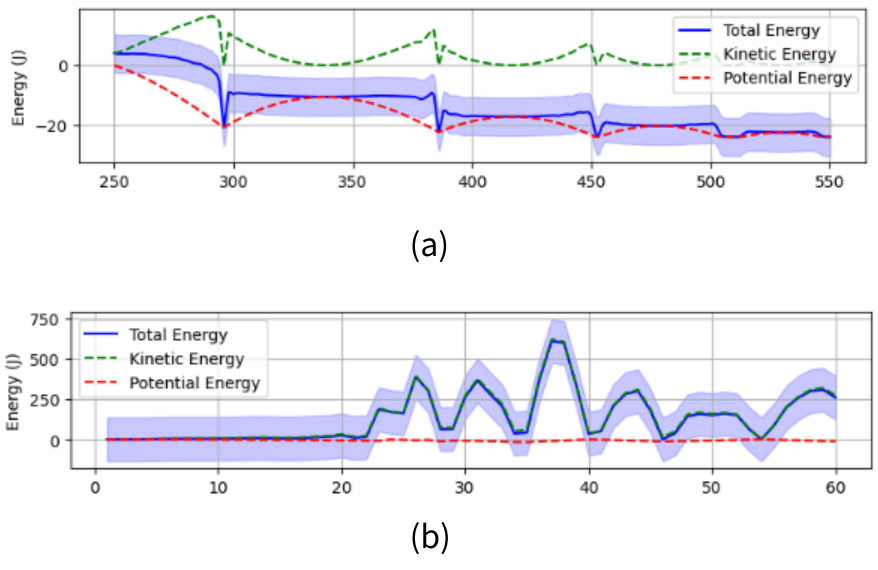


图 25：真实世界和生成下落+弹跳球视频的能量分析：(a) 真实世界视频能量守恒 (b) CogVideoX 普通单帧生成视频能量守恒

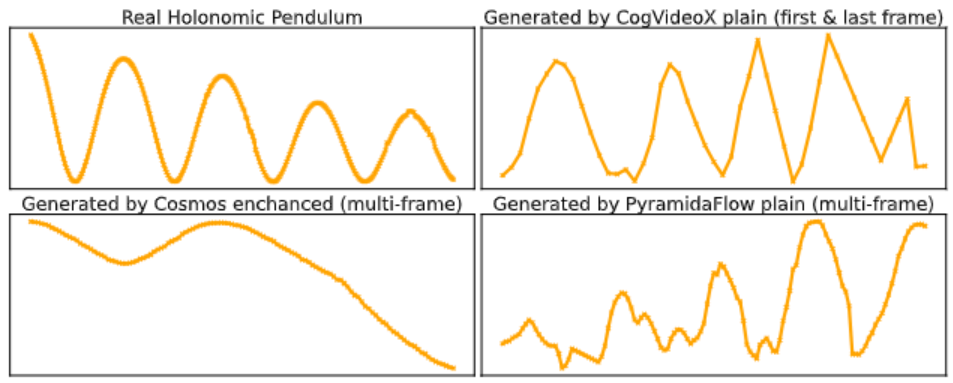


Figure 26: Real (top left) and generated trajectories for the holonomic pendulum.

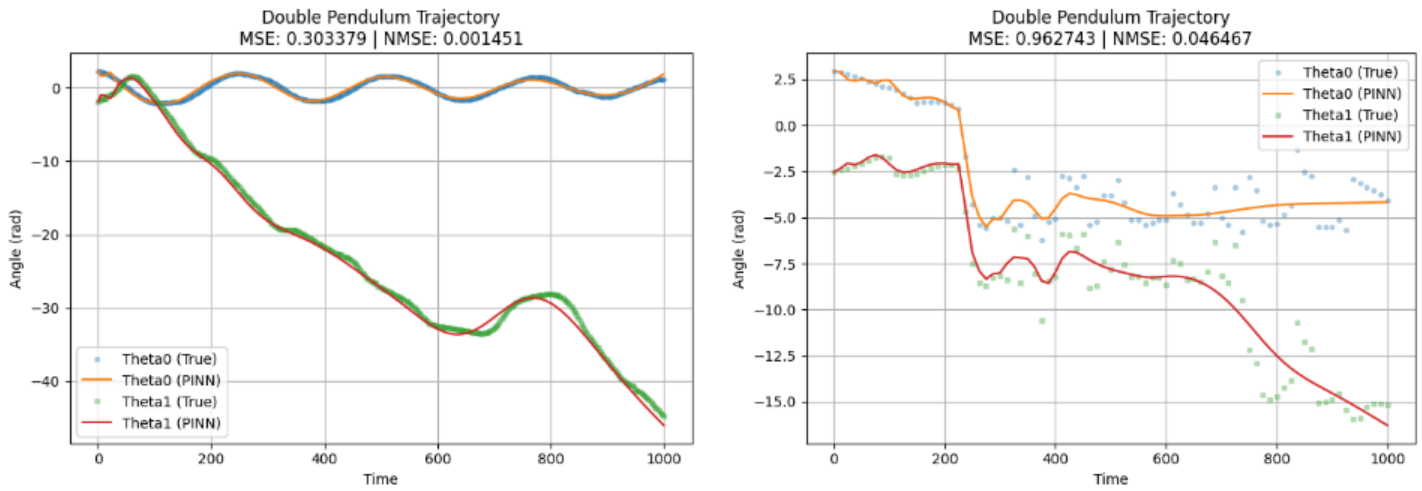


Figure 27: Real (left) and generated trajectory (right) for the double pendulum and corresponding fitting curve with PINN. While NMSE for generated trajectory is small 0.05, it is still 50 times worse that PINN with the same parameters fitted to real-world trajectory.

Reproducability statement. For reproducibility, we release the full code on GitHub and all benchmark data on Hugging Face.

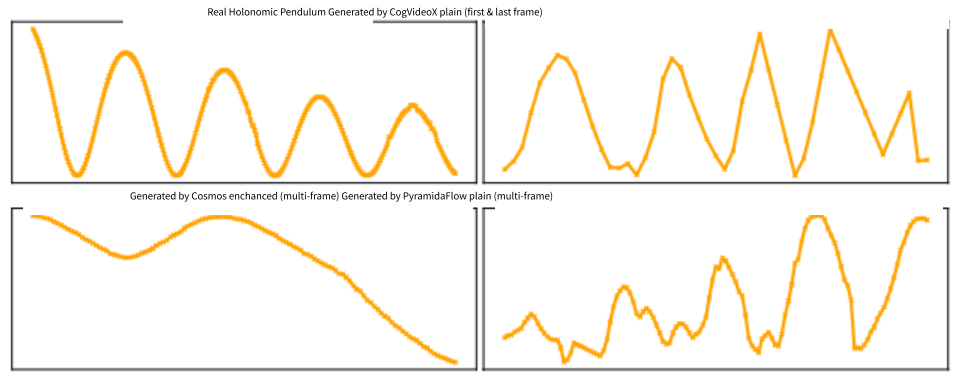


图 26：完整摆的实轨迹（左上）和生成轨迹

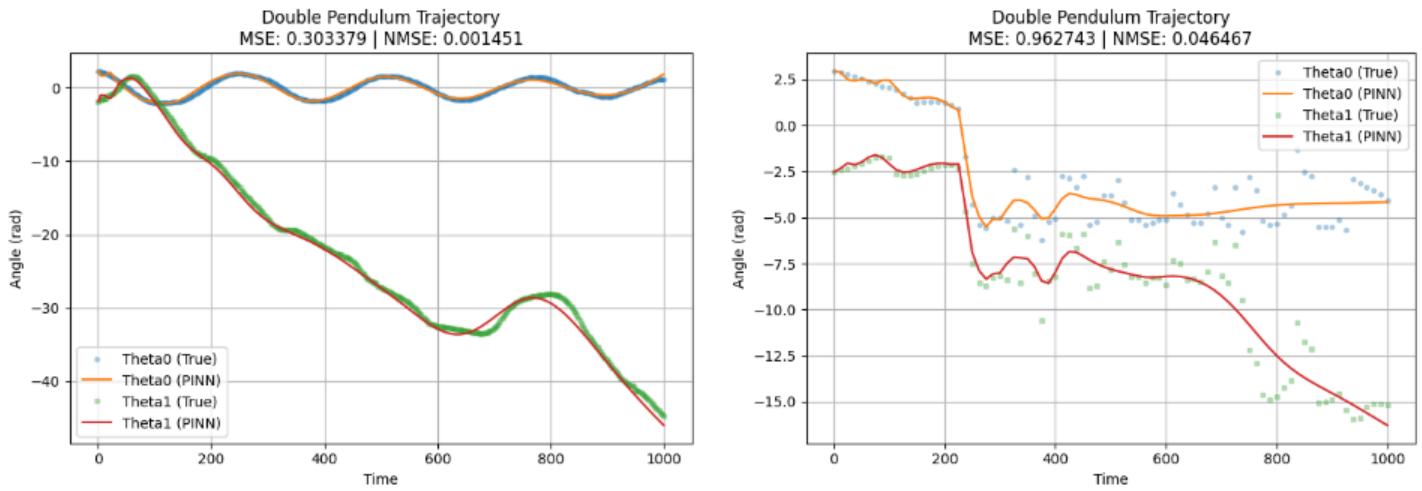


图 27：双摆的真实轨迹（左）和生成轨迹（右）以及相应的 PINN 拟合曲线。虽然生成轨迹的 NMSE 很小（0.05），但与使用相同参数拟合真实轨迹的 PINN 相比，它仍然差 50 倍。

可复现性声明。为了确保可复现性，我们在 GitHub 上发布了完整代码，并在 Hugging Face 上发布了所有基准数据。