

VLIPP: Towards Physically Plausible Video Generation with Vision and Language Informed Physical Prior

Xindi Yang^{1*}, Baolu Li^{2*}, Yiming Zhang², Zhenfei Yin^{4,5†}, Lei Bai^{3†}, Liqian Ma⁶, Zhiyong Wang⁵, Jianfei Cai¹, Tien-Tsin Wong¹, Huchuan Lu², Xu Jia^{2†}

¹Monash University ²Dalian University of Technology ³Shanghai Artificial Intelligence Laboratory

⁴Oxford University ⁵The University of Sydney ⁶ZMO AI

Abstract

Video diffusion models (VDMs) have advanced significantly in recent years, enabling the generation of highly realistic videos and drawing the attention of the community in their potential as world simulators. However, despite their capabilities, VDMs often fail to produce physically plausible videos due to an inherent lack of understanding of physics, resulting in incorrect dynamics and event sequences. To address this limitation, we propose a novel two-stage image-to-video generation framework that explicitly incorporates physics with vision and language informed physical prior. In the first stage, we employ a Vision Language Model (VLM) as a coarse-grained motion planner, integrating chain-of-thought and physics-aware reasoning to predict a rough motion trajectories/changes that approximate real-world physical dynamics while ensuring the inter-frame consistency. In the second stage, we use the predicted motion trajectories/changes to guide the video generation of a VDM. As the predicted motion trajectories/changes are rough, noise is added during inference to provide freedom to the VDM in generating motion with more fine details. Extensive experimental results demonstrate that our framework can produce physically plausible motion, and comparative evaluations highlight the notable superiority of our approach over existing methods. More video results are available on our Project Page: https://madaoer.github.io/projects/physically_plausible_video_generation/.

1. Introduction

Video diffusion models (VDMs) trained on large-scale video datasets have made remarkable progress in terms of realism, demonstrating significant potential for various content creation applications. Despite the absence of explicit geometric modeling, the generated videos still exhibit co-

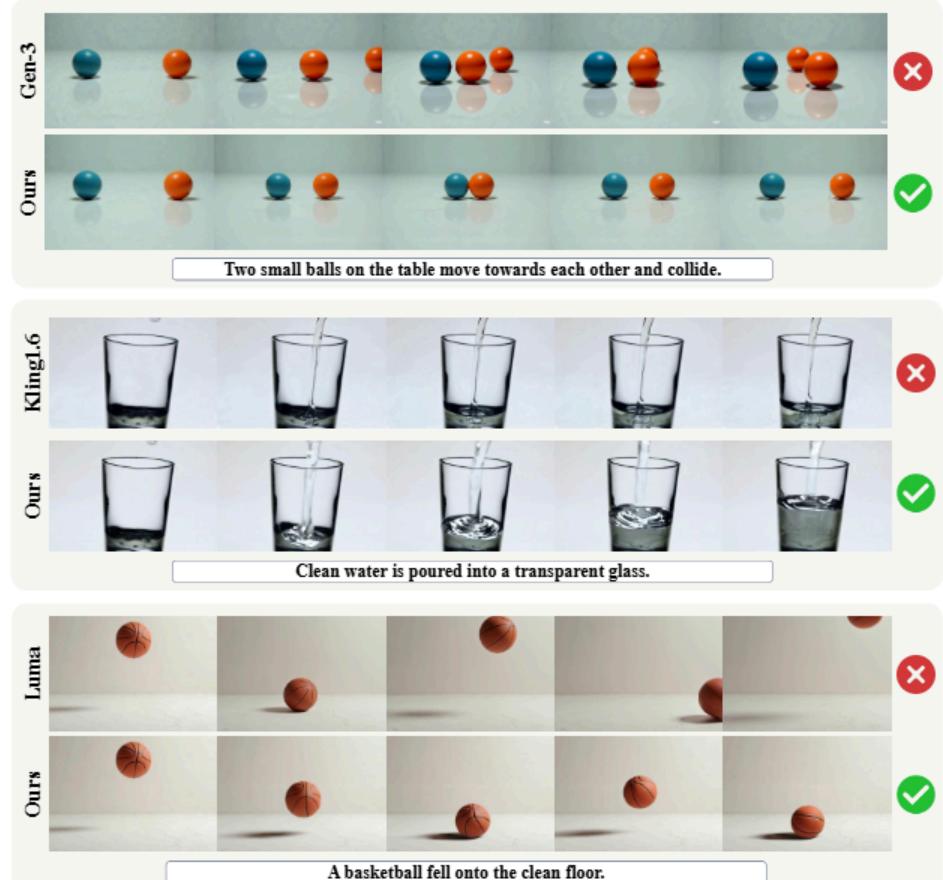


Figure 1. Existing commercial closed-source VDMs fail to generate physically plausible motion, whereas our video generation framework is able to achieve this by incorporating external physical prior knowledge.

herent spatial relationships among objects, rich textured details, and realistic lighting effects, including reflections and shadows. Such qualities often make the generated videos nearly indistinguishable from real-world footages. This drives the research community to explore the potential of VDMs as world models. However, they still struggle with understanding the physical laws of the real world and generating videos that adhere to these principles.

Although existing VDMs can produce visually realistic videos, they fail to mimic the real-dynamic physical motions. As shown in Fig 1, even the current commercial closed-source VDMs struggle with the task of generating videos that conform to physical laws. PhyT2V [58] refines text prompts by incorporating detailed descriptions of phys-

* Equal contribution † Corresponding author

VLIPP：基于视觉与语言信息物理先验的物理合理视频生成方法

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, Xu Jia

¹ 莫纳什大学 大连理工大学 上海人工智能实验室

⁴ 牛津大学 新南威尔士大学 ZMO AI

摘要

视频扩散模型（VDMs）近年来取得了显著进展，能够生成高度逼真的视频，并因其作为世界模拟器的潜力吸引了社区的关注。然而，尽管具备这些能力，VDMs 由于缺乏对物理学的内在理解，往往无法生成物理上合理的视频，导致动力学和事件序列不正确。为解决这一局限性，我们提出了一种新颖的两阶段图像到视频生成框架，明确地将物理与受视觉和语言信息指导的物理先验相结合。在第一阶段，我们采用视觉语言模型（VLM）作为粗粒度运动规划器，整合思维链和物理感知推理，预测近似真实世界物理动力学的粗略运动轨迹/变化，同时确保帧间一致性。在第二阶段，我们使用预测的运动轨迹/变化来指导 VDM 的视频生成。由于预测的运动轨迹/变化较为粗略，在推理过程中添加噪声，为 VDM 在生成具有更精细节的运动提供自由度。大量的实验结果表明，我们的框架能够生成符合物理规律的动画，而比较评估则突显了我们的方法相对于现有方法的显著优势。更多视频结果可以在我们的项目页面查看：https://madaoer.github.io/projects/physically_plausible_video_generation/。

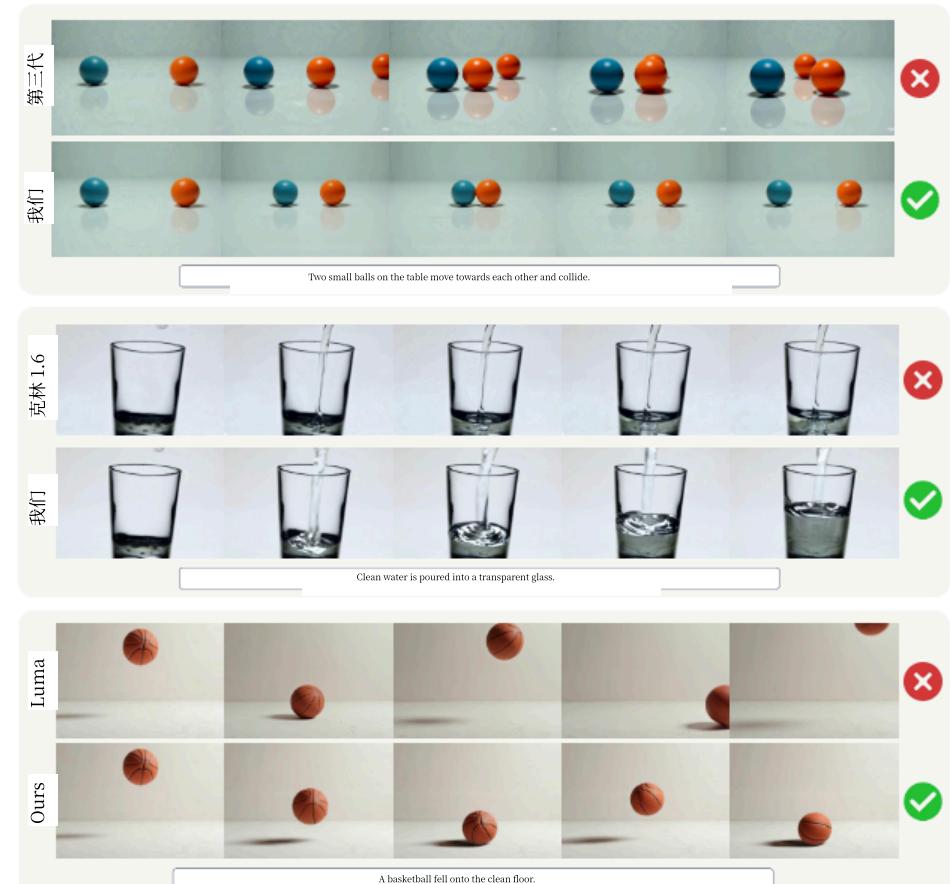


图 1. 现有的商业闭源 VDM 无法生成符合物理规律的运动，而我们的视频生成框架通过结合外部物理先验知识能够实现这一点。

1. 引言

在大型视频数据集上训练的视频扩散模型（VDMs）在逼真度方面取得了显著进步，展示了其在各种内容创作应用中的巨大潜力。尽管缺乏显式的几何建模，生成的视频仍然表现出高度一致性和连贯性。

物体之间固有的空间关系、丰富的纹理细节以及逼真的光照效果，包括反射和阴影。这些品质通常使得生成的视频几乎无法与真实世界影像区分开来。这促使研究界探索 VDM 作为世界模型的潜力。然而，它们仍然难以理解现实世界的物理规律，并生成遵循这些原则的视频。

尽管现有的视觉动态模型（VDMs）能够生成视觉上逼真的视频，但它们无法模拟真实的动态物理运动。如图 1 所示，即使是当前的商业闭源 VDMs 在生成符合物理定律的视频的任务上也存在困难。PhyT2V [58] 通过结合详细的物理描述来优化文本提示。

* 同等贡献 通讯作者

ical processes to guide VDMs in generating physically plausible videos. However, despite being pretrained on internet-scale real-world video–text pairs, VDMs do not inherently understand physical laws. This limitation arises from the gap and ambiguity between text descriptions and the actual motion in the video[33]. Moreover, VDMs tend to overfit the training data rather than developing a general understanding of physical laws[21]. Inspired by the success of graphics-based physical rendering, some methods have guided VDMs to generate physically plausible videos using simulations from graphics engines[18, 29, 32, 56, 64]; however, these approaches rely on the physical effects that graphics engines can simulate and incur high computational costs.

The gap and ambiguity between text and real-world motion makes it difficult to enable physically plausible video generation through detailed text descriptions alone. Moreover, it is challenging to gather scalable physical data for training due to the abstractness and diversity of physical phenomena. Consequently, a viable approach could be to model abstract physical laws as conditions for diffusion models. However, it is less practical to explicitly model the physics equation for every kind of motion. Instead, we resort to current large foundation models for their ability to “understand” basic physics[6] and reason about physical phenomena based on the knowledge they extract. For example, given two colliding balls, the Large Language Model (LLM) can *approximately* predict the paths of the balls after collision. Inspired by this observation, we propose a novel video generation framework that employs a Vision Language Model (VLM) to predict the path/change during a physics event, described by a given image and a text prompt.

In this paper, we propose VLIPP, a two-stage approach to incorporate physics as conditions into VDM, enabling the generation of physically plausible motion. In the first stage, the VLM serves as a *coarse-level motion planner*, while a VDM serves as a *fine-level motion synthesizer*. The idea of stage one is to utilize the chain-of-thought and the physics-aware reasoning of VLM planning to ensure that coarse-level motion trajectories approximately follow real-world physics dynamics. In stage two, we can generate fine-level motion using an image-to-video diffusion model conditioned by the approximated path/change planned by VLM from stage one. Note that the approximated path/changes are not in the level to tell the speed or acceleration of the motion. We choose an existing image-to-video model [7] to accept our coarse-level path/change, by injecting noise to the motion path during both the training and inference phases. Notably, during the VLM planning stage, generating entire physically plausible motion trajectories is not required. Instead, we leverage the generative priors of VDM to produce fine-level physically plausible videos based on coarse-level motion trajectories provided by the VLM. So

that the detail-level motion such as speed, acceleration, and vibration are left to the VDM to synthesize.

We evaluate our physically plausible video generation framework with two major video physics benchmarks and achieved satisfactory results. Furthermore, we discuss and analyze multiple insightful design choices in our video generation framework, such as employing a motion planner tailored for different physics categories, and enhancing the robustness of diffusion model to noisy trajectories. Our contributions are summarized as follows:

1. We introduce a novel image-to-video generation framework for generating physically plausible videos by leveraging the VLM and VDM priors, significantly outperforming the contemporary competitors.
2. We propose a novel chain-of-thought and physics-aware reasoning approach in VLM, along with random noise injection in the latent space during video generation, which effectively improves both the generation quality and physical plausibility.
3. We conduct a comprehensive experiments and user studies to demonstrate the effectiveness and generalization of our framework in physically plausible videos generation.

2. Related Work

2.1. Physically Plausible Visual Content Generation

Generating physically plausible videos offers substantial value to real-world applications such as scientific simulations[43], robotics [4, 59], and autonomous driving [10, 48]. Traditional graphics pipelines rely on simulation systems to model physical phenomena [36, 42]. Inspired by these approaches, recent studies [18, 29] have performed dynamic simulations in image space based on physical engines. Furthermore, some methods [56, 64] incorporate physical priors into 3D representations to enable the synthesis of physically plausible motions. However, these rule-based or solver-based simulators face limitations in expressiveness, efficiency, generalizability, and parameter tuning. Furthermore, these simulators require significant expertise, rendering them inaccessible and unfriendly for users.

In addition, Some studies have explored VDMs for generating physically plausible videos. Li *et al.* [24] models natural oscillations and swaying in frequency-domain. A downstream rendering module then animates static images based on the generated motion information. PhysDiff [62] introduces physical simulator as constrain into the diffusion process by projecting denoised motion of a diffusion step into a physically plausible motion. These methods mainly focus only on specific types of physical motion and do not establish a generalizable approach for generating physically plausible videos.

基于物理过程的指导 VDMs 生成物理上合理的视频。然而，尽管 VDMs 在互联网规模的现实世界视频-文本对上进行了预训练，但它们并不天生理解物理定律。这种局限性源于文本描述与视频实际运动之间的差距和模糊性[33]。此外，VDMs 倾向于过拟合训练数据，而不是发展对物理定律的普遍理解[21]。受基于图形的物理渲染成功的启发，一些方法指导 VDMs 使用图形引擎的模拟来生成物理上合理的视频[18, 29, 32, 56, 64]；然而，这些方法依赖于图形引擎可以模拟的物理效应，并导致高昂的计算成本。

文本与现实世界运动的差距和模糊性使得仅通过详细的文本描述难以实现物理上合理的视频生成。此外，由于物理现象的抽象性和多样性，收集可扩展的物理数据进行训练是一项挑战。因此，一个可行的方法可能是将抽象的物理定律建模为扩散模型的条件。然而，为每种运动显式地建模物理方程并不实用。相反，我们借助当前的大型基础模型，利用它们“理解”基本物理[6]以及基于提取的知识推理物理现象的能力。例如，对于两个相撞的球，大型语言模型（LLM）可以近似预测碰撞后球的运动路径。受此观察的启发，我们提出了一种新型视频生成框架，该框架采用视觉语言模型（VLM）来预测由给定图像和文本提示描述的物理事件中的路径/变化。

在这篇论文中，我们提出了 VLIPP，一种分两阶段的方法，将物理条件融入 VDM，以生成符合物理的运动。在第一阶段，VLM 作为粗级运动规划器，而 VDM 作为细级运动合成器。第一阶段的思想是利用 VLM 规划的思维链和物理感知推理，确保粗级运动轨迹大致遵循现实世界的物理动力学。在第二阶段，我们使用由 VLM 在第一阶段规划的近似路径/变化条件下的图像到视频扩散模型来生成细级运动。请注意，近似路径/变化并未说明运动的速度或加速度。我们选择一个现有的图像到视频模型[7]来接受我们的粗级路径/变化，通过在训练和推理阶段向运动路径注入噪声。值得注意的是，在 VLM 规划阶段，并不需要生成整个符合物理的运动轨迹。相反，我们利用 VDM 的生成先验，基于 VLM 提供的粗级运动轨迹生成精细级物理上合理的视频。

因此，速度、加速度和振动等细节级运动由 VDM 合成。我们使用两个主要的视频物理基准评估了我们的物理上合理的视频生成框架，并取得了令人满意的结果。此外，我们还讨论并分析了我们的视频生成框架中的多个富有见地的设计选择，例如采用针对不同物理类别的运动规划器，以及增强扩散模型对噪声轨迹的鲁棒性。我们的贡献总结如下：

1. 我们介绍了一种利用 VLM 和 VDM 先验生成物理上合理的视频的新颖图像到视频生成框架，显著优于当代竞争对手。
2. 我们在 VLM 中提出了一种新颖的思维链和物理感知推理方法，并在视频生成期间对潜在空间注入随机噪声，有效提高了生成质量和物理合理性。
3. 我们进行了全面的实验和用户研究，以证明我们的框架在物理上合理的视频生成中的有效性和泛化能力。

2. 相关工作

2.1. 物理上合理的视觉内容生成

生成物理上合理的视频为科学模拟[43]、机器人[4, 59]和自动驾驶[10, 48]等现实应用提供了巨大价值。传统图形管线依赖模拟系统来建模物理现象[36, 42]。受这些方法的启发，近期研究[18, 29]基于物理引擎在图像空间中执行动态模拟。此外，一些方法[56, 64]将物理先验知识融入三维表示中，以实现物理上合理的运动合成。然而，这些基于规则或求解器的模拟器在表达力、效率、泛化能力和参数调整方面存在局限性。此外，这些模拟器需要大量专业知识，使其对用户来说难以使用且不友好。

此外，一些研究探索了 VDMs 在生成物理上合理的视频方面的应用。Li 等人[24]在频域中建模了自然的振荡和摇摆。然后，下游渲染模块根据生成的运动信息对静态图像进行动画处理。PhysDiff[62]通过将去噪运动投影到物理上合理的运动中，将物理模拟器作为约束引入到扩散过程中。这些方法主要只关注特定类型的物理运动，并没有建立一种通用的生成物理上合理视频的方法。

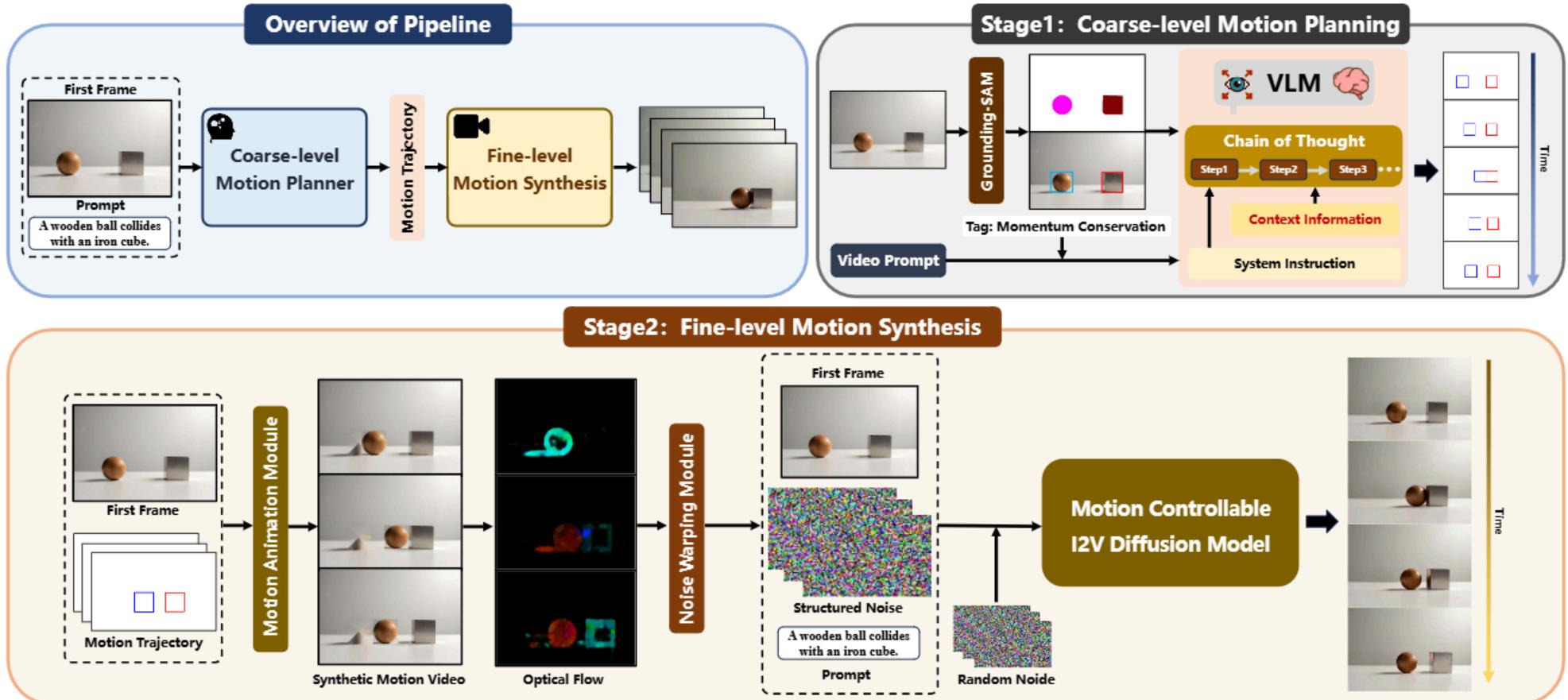


Figure 2. The illustration of our physically plausible image-to-video generation pipeline. Our pipeline consists of two stages. In the first stage, the VLM generates a coarse-grained, physically plausible motion trajectory based on the provided input conditions. In the second stage, We simulate a synthetic video using the predicted trajectory to provide the motion condition. We then extract the optical flow from this video and convert it into structured noise. These conditions are fed into a motion controllable image-to-video diffusion model, and ultimately generates a physically plausible video.

2.2. Motion Controllable Video Generation

Existing studies commonly provide one of the following three types of motion control: bounding box control [19, 23, 30, 46, 52], point trajectory control [34, 37, 40, 50, 54] and camera control [2, 15, 47, 61]. Bounding box control provides object motion guidance by generating a sequence of bounding boxes that track the object’s position over time. Point-trajectory control offers motion cues through point-based trajectories, enabling drag-style manipulation. Camera motion control guides video generation using explicit 3D camera parameters, ensuring consistency and realistic viewpoint changes. However, these approaches prioritize motion control but often overlook physical plausibility. To address the limitation, we propose a novel framework for physically plausible video generation that incorporates physics as conditions into video diffusion models.

2.3. Generation based on VLMs Planning

VLMs have exhibited robust capabilities in visual understanding and planning [27, 39, 65]. Their strong performance in domains such as robot path planning and video understanding shows their ability in understanding the real physical world. Prior work has successfully leveraged LLMs to guide the layout of images or videos, yielding promising results [25, 53]. VideoDirectorGPT [26] leverages LLMs for fine-grained scene-by-scene planning, explicitly controlling spatial layout to generate temporally consistent long videos. Pandora [55] utilizes LLMs for real-time control through free-text action commands, achieving

domain generality, video consistency, and controllability. However, these efforts have yet to address interactions with real-world physical phenomena, such as collision, fall, and melting.

Moreover, the absence of visual information can cause severe hallucination issues in language models for spatial planning tasks, leading to problems like overlapping object boundaries, disproportionate scaling, and incorrect planning [20, 57]. In this paper, we propose utilizing VLMs as coarse-level motion planners within the image space and incorporate physics-aware reasoning and Chain of Thought (CoT) [51] into the inference process.

3. Method

Task Fomulation. In this paper, our goal is to enable an image-to-video diffusion model to generate physically plausible videos. Since VDMs rely more on memory and case-based imitation and struggle to understand general physical rules[21], the key challenge is how to incorporate physical laws into the models. To achieve this, we need to identify a method to incorporate physical principles into the video diffusion framework. Given an image $I \in R^{H \times W \times C}$ (H is height, W is width and C is the number of channels) and a text description d of possible events based on image I , our framework should infer a physics-compliant guidance as the input condition and synthesize a video that adheres to both physical laws and real-world dynamics.

Overall Pipeline. Overall pipeline of VLIPP is illustrated

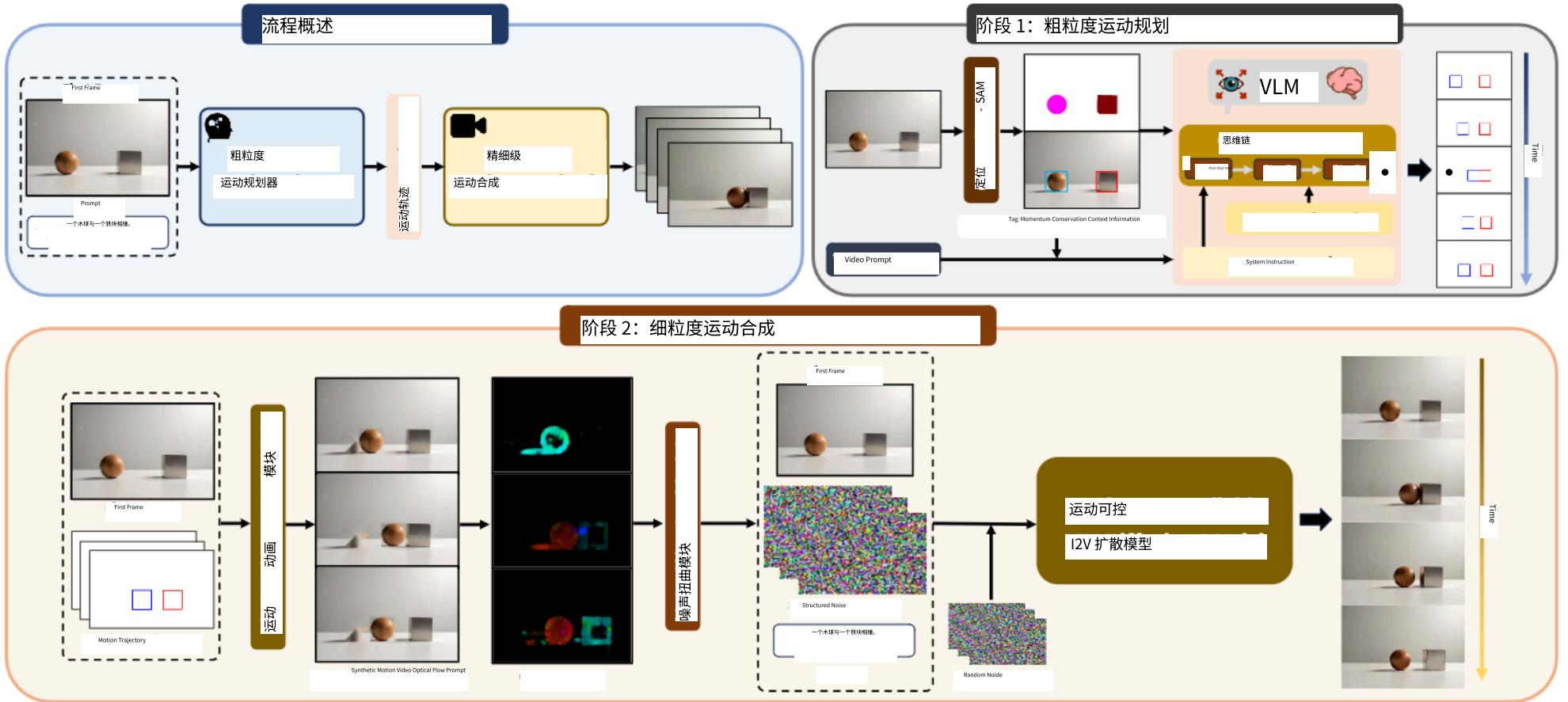


图 2. 我们物理上合理的图像到视频生成流程的示意图。我们的流程包含两个阶段。在第一阶段，视觉语言模型根据提供的输入条件生成粗粒度的物理上合理的运动轨迹。在第二阶段，我们使用预测的轨迹模拟合成视频以提供运动条件。然后我们从该视频中提取光流并将其转换为结构化噪声。这些条件被输入到运动可控的图像到视频扩散模型中，最终生成物理上合理的视频。

2.2. 运动可控视频生成

现有研究通常提供以下三种类型的运动控制：边界框控制[19, 23, 30, 46, 52]、点轨迹控制[34, 37, 40, 50, 54]和相机控制[2, 15, 47, 61]。边界框控制通过生成一系列跟踪物体随时间位置变化的边界框来提供物体运动指导。点轨迹控制通过基于点的轨迹提供运动提示，实现拖拽式操作。相机运动控制使用显式的3D相机参数来指导视频生成，确保视角变化的一致性和真实性。然而，这些方法优先考虑运动控制，但往往忽略了物理合理性。为解决这一局限性，我们提出了一种基于物理合理性的视频生成新框架，将物理作为条件融入视频扩散模型。

实现领域泛化、视频一致性和可控性。然而，这些努力尚未解决与真实世界物理现象的交互，例如碰撞、跌倒和融化。此外，缺乏视觉信息会导致空间规划任务中的语言模型出现严重的幻觉问题，例如物体边界重叠、比例失调和规划错误[20, 57]。在本文中，我们提出将VLMs作为图像空间中的粗粒度运动规划器，并将物理感知推理和思维链(CoT)[51]融入推理过程。

2.3. 基于 VLMs 的规划生成

VLMs 在视觉理解和规划方面展现出强大的能力 [27, 39, 65]。它们在机器人路径规划、视频理解等领域的出色表现显示了其理解真实物理世界的能力。先前工作已成功利用LLMs指导图像或视频的布局，并取得了令人满意的结果 [25, 53]。VideoDirectorGPT [26] 利用LLMs进行细粒度的逐场景规划，明确控制空间布局以生成时间上连贯的长视频。Pandora [55] 利用LLMs通过自由文本动作指令实现实时控制

3. 方法

任务公式化。在本文中，我们的目标是使图像到视频扩散模型能够生成符合物理规律的视频。由于VDMs更依赖于记忆和基于案例的模仿，难以理解一般的物理规则[21]，因此关键挑战是如何将物理定律融入模型。为实现这一目标，我们需要找到一种将物理原理融入视频扩散框架的方法。给定一个图像 $I \in R$ (H 是高度， W 是宽度， C 是通道数)以及基于图像 I 的可能事件的文本描述 d ，我们的框架应推断出符合物理的指导作为输入条件，并合成一个既符合物理定律又遵循真实世界动态的视频。

整体流程。VLIPP的整体流程如图所示。

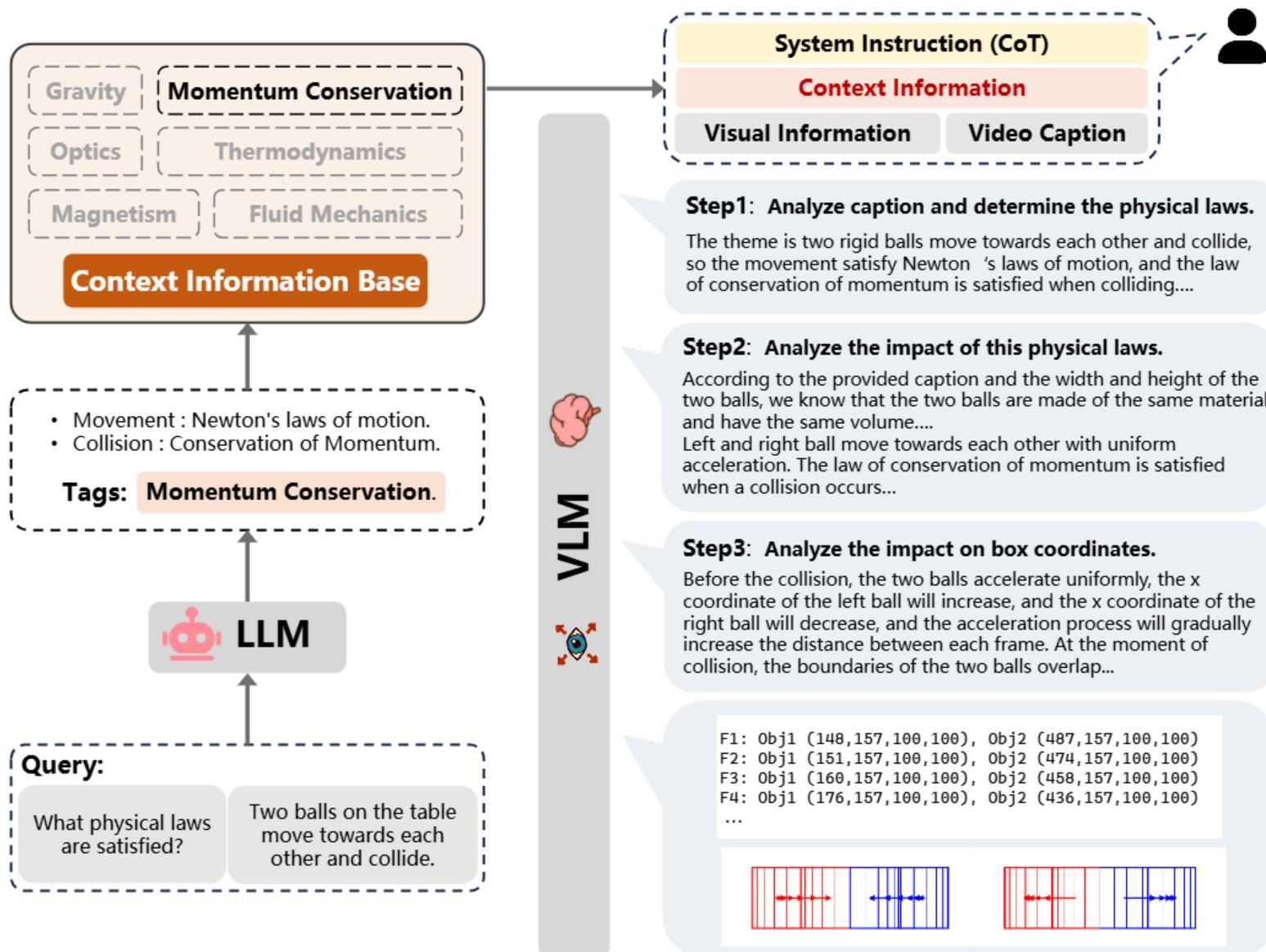


Figure 3. The illustration of chain-of-thought reasoning in the VLM for generating a coarse-grained motion trajectory. First off, the VLM determines the corresponding physical laws and its context for the given scene. Then, the VLM performs step-by-step reasoning to predict the physically plausible motions of objects in image space, leveraging physical context and chain-of-thought prompting. Finally, the VLM predicts bounding boxes according to real-world physics.

in Figure 2. In the first stage, the VLM conducts semantic analysis and physical attribute analysis on the given image I and a description d to obtain the bounding boxes of the objects in the scene, denoted as b_1, b_2, \dots, b_n , along with the applicable physical laws l . Next, the VLM infers possible future physical scenarios in the current scene to derive the coarse-level motion trajectories of p_1, p_2, \dots, p_t in the image space. Finally, we utilize an image-to-video diffusion model to synthesize the detailed dynamics in the video.

3.1. VLM as a Coarse-Level Motion Planner

Our motivation is to incorporate physical laws as constraints into a video diffusion model to enhance the physical plausibility of the generated videos. To achieve this, we must identify a method to inject physical laws into the video diffusion model. Given a video description and the first frame, the task at this stage is to generate coarse-level motion trajectory aligned with physical laws.

Scene Understanding. In the real world, most physical phenomena arise from interactions among objects and their motion trajectories. We first initiate the process by identifying and locating objects within a scene. Inspired by the recent studies [1, 9, 28] in VLMs for scene understand-

ing, we employ GPT-4o [38] to recognize all objects that could be involved in physical phenomena as described in the text description d . These objects are subsequently detected and segmented using Grounded-SAM2 [41], yielding their bounding boxes. By leveraging the pretrained knowledge and common-sense reasoning capabilities of foundation models, we effectively determine the relevant objects in the scene.

Physical-Aware Recognition. To perform more effective reasoning in predicting the motion, it is necessary to determine what specific physical principle to apply in the given context. We utilize the pretrained prior of the LLM to determine the physical laws applicable to the current scene. Following the configuration in the physical benchmark [3, 31, 33], we currently classify common physical phenomena in videos into six categories: gravity, momentum conservation, optics, thermodynamics, magnetism, and fluid mechanics. Note that such list can be easily extended within our framework. Given a video description d , the LLM infers the physical law l that governs the current scene. We provide the specific physical context information for VLM to enhance its understanding of physical laws [11]. Detailed context design is presented in the Appendix.

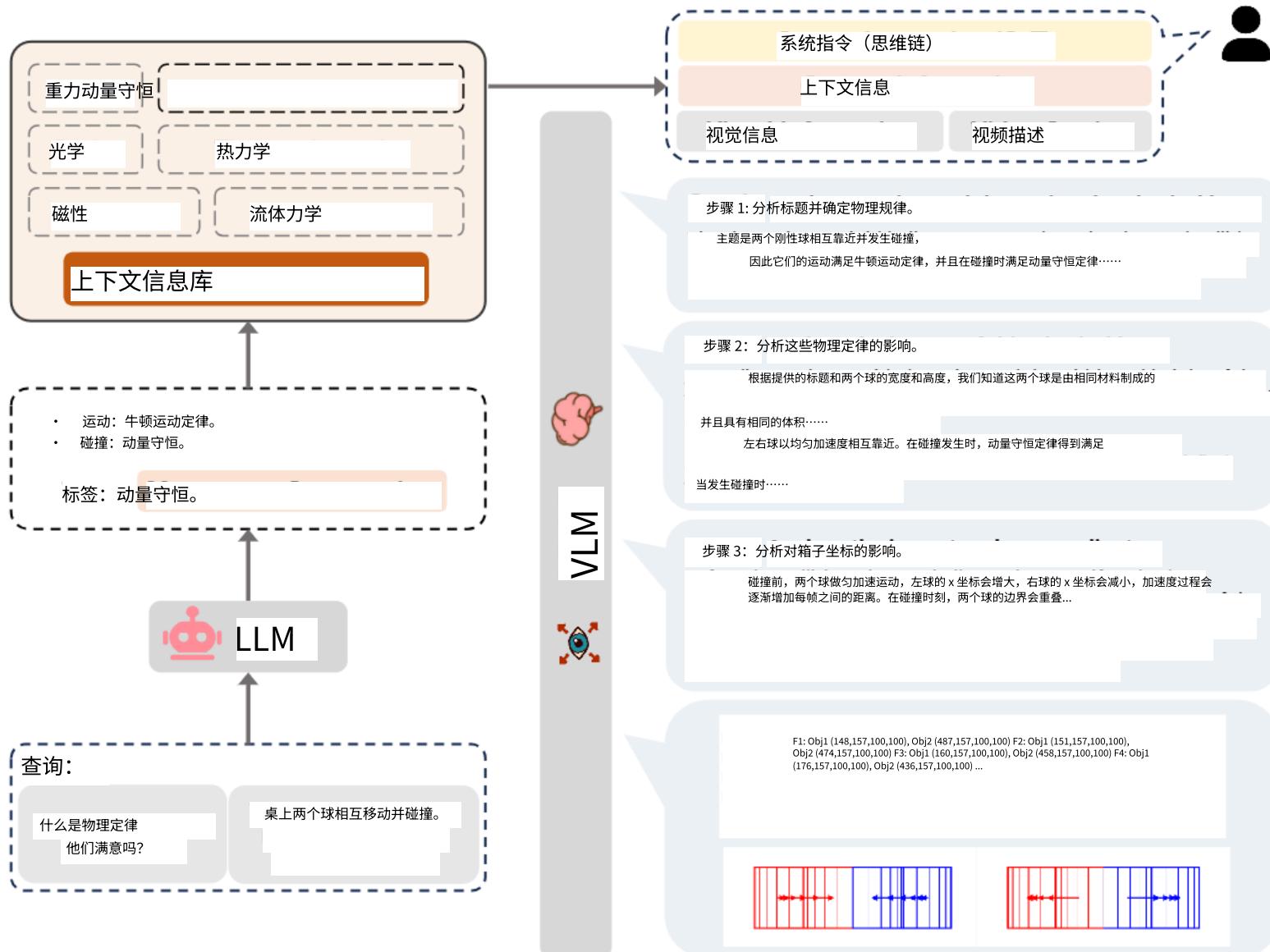


图 3. VLM 生成粗粒度运动轨迹的推理链示意图。首先，VLM 确定给定场景对应的物理规律及其上下文。然后，VLM 进行逐步推理，利用物理上下文和推理链提示，预测图像空间中物体的物理上合理的运动。最后，VLM 根据现实世界物理预测边界框。

在图 2 中。在第一阶段，VLM 对给定的图像 I 和描述 d 进行语义分析和物理属性分析，以获取场景中对象的边界框 b_1, b_2, \dots, b_n ，以及适用的物理定律 l 。接下来，VLM 推断当前场景中可能的未来物理场景，以导出图像空间中 p_1, p_2, \dots, p_n 的粗粒度运动轨迹。最后，我们利用图像到视频扩散模型合成视频中的详细动态。

3.1. VLM 作为粗粒度运动规划器

我们的动机是将物理定律作为约束纳入视频扩散模型，以增强生成视频的物理合理性。为实现这一目标，我们必须找到一种将物理定律注入视频扩散模型的方法。给定视频描述和第一帧，本阶段的任务是为生成与物理定律一致的粗粒度运动轨迹。

场景理解。 在现实世界中，大多数物理现象源于物体之间的相互作用及其运动轨迹。我们首先通过识别和定位场景中的物体来启动该过程。受近期研究[1, 9]的启发

[28] 在视觉语言模型 (VLMs) 中进行场景理解时，我们采用 GPT-4o [38] 识别文本描述 d 中可能参与物理现象的所有对象。随后，使用 Grounded-SAM2 [41] 对这些对象进行检测和分割，并获取它们的边界框。通过利用基础模型的前置知识和常识推理能力，我们有效地确定了场景中的相关对象。**物理感知识别。** 为了在预测运动时进行更有效的推理，有必要确定在给定情境中应应用哪些具体的物理原理。我们利用 LLM 的预训练先验来确定当前场景适用的物理定律。遵循物理基准[3, 31, 33]中的配置，我们目前将视频中常见的物理现象分为六类：重力、动量守恒、光学、热力学、磁学和流体力学。请注意，该列表可以在我们的框架内轻松扩展。给定视频描述 d ，LLM 推断出支配当前场景的物理定律 l 。我们为 VLM 提供具体的物理上下文信息，以增强其对物理定律的理解[11]。

详细的上下文设计在附录中呈现。

Chain of Thought Reasoning in VLM. Given the physical law l , an image I and a video description d for the scene, we prompt the VLM to predict the future bounding box positions of objects within the image-space. We choose to predict in the image space for two primary reasons. Firstly, motion in image space aligns more with our subsequent video synthesizer. Secondly, image space dynamics can effectively represent a wide range of real-world motions [29].

At a given time t , the predicted position of i -th object bounding box b_i^t is denoted as $[x_t^i, y_t^i, w_t^i, h_t^i]$, where (x_t^i, y_t^i) represents its top-left coordinate; w_t^i & h_t^i denote its width and height, respectively. Governed by the physical law, the four values of the bounding box may change over time. The VLM reasons the bounding box positions of N future frames for every object o_i based on the condition. To help VLM better understand physical laws, we adapt a chain-of-thought [51] into its reasoning, to significantly enhance its reasoning capabilities. As shown in Figure 3, we formulate our analysis of physical phenomena in videos as step-by-step reasoning: beginning with broad conceptual ideas and progressing to a detailed and practical examination:

1. Given the physical law l and context information, the VLM analysis video caption and detail the physical law.
2. The VLM analyzes the potential interactions and movement of each object within the scene;
3. The VLM predicts the detailed changes in position and shape of the bounding box corresponding to each object over time.

Through the structured planning process, the VLM plans coarse-level motion trajectories for the objects, approximating real-world physics dynamics. In particular, our VLM infers the changes of object bounding boxes for next 12 frames, constrained by the token length limitation. To be compatible with the generation process of the chosen VDM in the next stage, these inferred 12 frames are further linearly interpolated to produce a total of 49 frames.

3.2. VDM Serves as a Fine-Level Motion Synthesizer

In the previous stage, the motion trajectory planned by the VLM is neither precise nor fully compliant with physical laws. On the other hand, while VDM may not be able to produce realistic global motion trajectories, it is able to generate sound motion in finer scale. In this stage, our key insight is that the VDM can refine the coarse-level motion to produce physically plausible motion that aligns with real-world dynamics with its powerful generative prior.

Motion Animation. To incorporate physical laws into the video diffusion model, we use the inferred coarse motion trajectory to guide the generation process of the diffusion model. Optical flow provides a unified representation of motion, and recent studies [7, 12, 13] have demonstrated

its effectiveness in guiding diffusion models. Accordingly, we leverage the coarse-level motion trajectory to animate a synthetic motion sequence and derive the corresponding optical flow. Specifically, for each object o_i , we extract its bounding box from the first frame and move it to the bounding box location b_i specified by the motion trajectory. To animate the change of shape (e.g., due to compression or expansion), we resize object o_i according to the difference between o_i and o_{i+1} during inpainting. The synthetic motion video is generated as follows:

$$\hat{V}(t) = \text{Animation}(B, rs(o_0^0, b_0^0) \dots rs(o_t^0, b_t^0)) \quad (1)$$

where $\hat{V}(t)$ denotes the corresponding frame of inpainted video at timestep t , B denotes the inpainted background with the foreground object removed, o_i^0 denotes the i -th object at timestep 0, b_i^t represents the i -th bounding box at timestep t , and rs denotes resize function.

Structured Noise from Synthetic Video. Optical flow is an effective representation for guiding VDMs [12, 13]. Follow prior work [7, 8], we employ RAFT[44] to extract optical flow from the synthetic video and formulate it as structural noise, which retains Gaussian properties. Given the synthetic video $\hat{V}(t) \in \mathbb{R}^{F \times C \times H \times W}$, we calculate its per-frame optical flow to get a structured noise tensor $Q \in \mathbb{R}^{F \times C \times H \times W}$. The structured noise enables the VDM to generate videos that exhibit motion patterns closely aligned with those in the optical flow, thereby improving the realism of the output.

Noise Injection in Video Synthesis. We adopt Go-with-the-Flow [7] as our video synthesis model, a fine-tuned CogVideoX [60], which is designed to accept structured noise Q as input and synthesize videos that adhere to the implicit optical flow. The vanilla Go-with-the-Flow tends to tightly follow the provided structured noise Q . However, our Q is derived from a coarse-level motion trajectory and may not be sufficiently accurate to follow the physical laws of the real world. To address this limitation, we inject noise during the inference phase to give more flexibility to the VDM to generate detail-level motion changes as

$$Q_i = \frac{(1 - \gamma)Q_i + \zeta\gamma}{\sqrt{(1 - \gamma)^2 + \gamma^2}} \quad (2)$$

where Q_i is structured noise at i -th frames, $\zeta \in \mathbb{R}^{C \times H \times W}$ is Gaussian noise and $\gamma \in [0, 1]$. We set $\gamma = 0.4$ for even frame index and $\gamma = 0.6$ for odd frame index.

With this approach, the VDM is able to generate motion deviate from the coarse-level motion trajectory whenever necessary for producing high-quality fine-level motion.

4. Empirical Analysis and Discussion

In this section, we conduct extensive experiments to demonstrate the effectiveness of our video generation framework

VLM 中的思维链推理。给定物理定律 \mathcal{L} 、图像 I 以及场景的视频描述 d ，我们提示 VLM 预测图像空间中物体未来的边界框位置。我们选择在图像空间中进行预测，主要有两个原因。首先，图像空间中的运动与后续的视频合成器更加一致。其次，图像空间动力学能够有效地表示广泛的真实世界运动[29]。

在某一时刻 t ，第 i 个对象边界框的预测位置表示为 $[x, y, w, h]$ ，其中 (x, y) 代表其左上角坐标； w 和 h 分别表示其宽度和高度。受物理定律的支配，边界框的四个值可能会随时间变化。VLM 基于条件推理每个对象在未来 N 帧中的边界框位置。为了帮助 VLM 更好地理解物理定律，我们将其推理中引入了思维链[51]，显著增强了其推理能力。如图 3 所示，我们将视频中物理现象的分析表述为逐步推理：从宏观概念开始，逐步深入到详细和实用的检查：

1. 基于物理定律 \mathcal{L} 和上下文信息，VLM 分析视频标题并详细说明物理定律。
2. VLM 分析场景中每个对象的潜在交互和运动；
3. VLM 预测每个对象对应的边界框随时间变化的位置和形状的详细变化。

通过结构化规划过程，VLM 为对象规划粗粒度运动轨迹，近似真实世界的物理动力学。特别是，我们的 VLM 推断对象边界框在接下来 12 帧内的变化，受限于标记长度。为了与下一阶段所选 VDM 的生成过程兼容，这些推断的 12 帧进一步线性插值，产生总共 49 帧。

13]已经证明它在指导扩散模型方面的有效性。因此，我们利用粗略级别的运动轨迹来对合成运动序列进行动画处理，并导出相应的光流。具体来说，对于每个对象 o ，我们从第一帧中提取其边界框，并将其移动到运动轨迹指定的边界框位置 b 。为了动画化形状的变化（例如，由于压缩或膨胀），我们根据 o 和 b 在修复过程中的差异来调整对象 o 的大小。合成运动视频的生成如下：

$$\hat{V}(t) = \text{Animation}(B, rs(o, b) \dots rs(o, b)) \quad (1)$$

其中 $\hat{V}(t)$ 表示在时间步 t 处修复视频的相应帧， B 表示移除前景对象后的修复背景， o 表示在时间步 0 处的第 i 个对象， b 表示在时间步 t 处的第 i 个边界框， rs 表示调整大小函数。

合成视频中的结构化噪声。光流是一种有效的表示方法，用于指导视频生成模型 (VDM) [12, 13]。遵循先前工作[7, 8]，我们采用 RAFT[44]从合成视频中提取光流，并将其形式化为结构化噪声，该噪声保留高斯特性。给定合成视频 $\hat{V}(t) \in \mathbb{R}$ ，我们计算其逐帧光流，以获得结构化噪声张量 $Q \in \mathbb{R}$ 。结构化噪声使 VDM 能够生成运动模式与光流高度一致的视频，从而提高输出的逼真度。

视频合成中的噪声注入。我们采用 Go-with-the-Flow[7]作为我们的视频合成模型，这是一个经过微调的 CogVideoX[60]，其设计为接受结构化噪声 Q 作为输入，并合成遵循隐式光流的视频。普通的 Go-with-the-Flow 倾向于紧密跟随提供的结构化噪声 Q 。然而，我们的 Q 是从粗粒度的运动轨迹中推导出来的，可能不足以遵循现实世界的物理规律。为了解决这一局限性，我们在推理阶段注入噪声，以赋予 VDM 更多灵活性，使其能够生成细节级别的运动变化。

3.2. VDM 作为细粒度运动合成器

在上一阶段，VLM 规划的运动轨迹既不精确也不完全符合物理定律。另一方面，虽然 VDM 可能无法生成逼真的全局运动轨迹，但它能够以更精细的尺度生成合理的运动。在本阶段，我们的关键洞察是，VDM 可以通过其强大的生成先验，将粗粒度运动细化为符合真实世界动力学且物理上合理的运动。

运动动画。为了将物理定律融入视频扩散模型，我们使用推断出的粗略运动轨迹来指导扩散模型的生成过程。光流提供了一个统一的运动表示，最近的研究[7, 12]

$$Q = p \frac{(1 - \gamma)Q + \zeta\gamma}{(1 - \gamma) + \gamma} \quad (2)$$

其中 Q 为第 i 帧的结构化噪声， $\zeta \in \mathbb{R}$ 为高斯噪声， $\gamma \in [0, 1]$ 。我们设置偶数帧索引时 $\gamma = 0.4$ ，奇数帧索引时 $\gamma = 0.6$ 。

采用这种方法，VDM 能够在必要时从粗级运动轨迹生成运动偏差，以产生高质量的细级运动。

4. 实验分析与讨论

在本节中，我们进行了广泛的实验，以证明我们的视频生成框架的有效性。

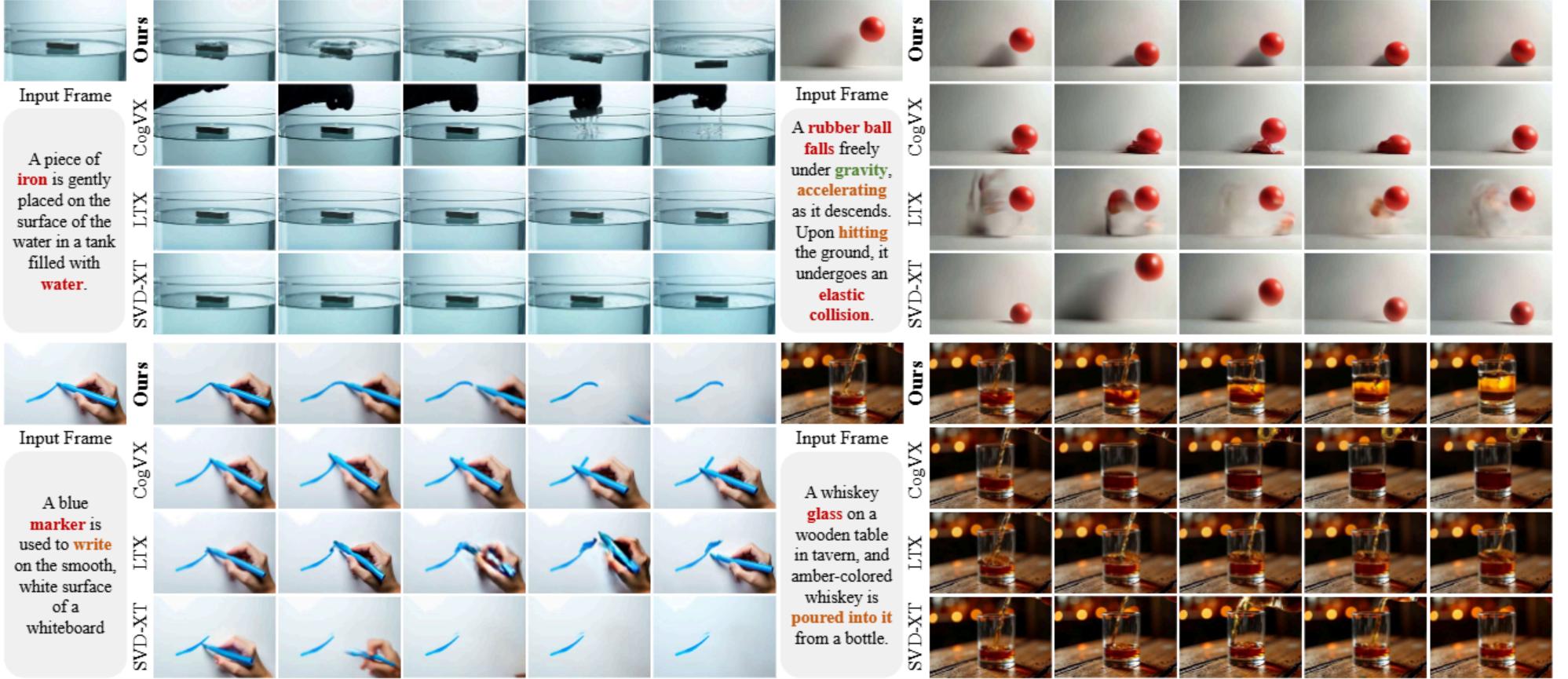


Figure 4. Visual comparisons of physically plausible video generation results from our framework, CogVideoX-I2V-5B [60], LTX-Video-I2V [14] and SVD-XT [5].

Model	Mechanics(\uparrow)	Optics(\uparrow)	Thermal(\uparrow)	Material(\uparrow)	Average(\uparrow)
CogvideoX-T2V-5B	0.43	0.55	0.40	0.42	0.45
LTX-Video-T2V	0.35	0.45	0.36	0.38	0.39
OpenSora	0.43	0.50	0.44	0.37	0.44
PhyT2V	0.49	0.61	0.49	0.47	0.52
LLM-Grounding Video Diffusion	0.32	0.41	0.26	0.24	0.31
CogvideoX-I2V-5B	0.48	0.69	0.43	0.41	0.52
SVD-XT	0.46	0.68	0.48	0.41	0.52
LTX-Video-I2V	0.47	0.65	0.46	0.37	0.50
SG-I2V	0.52	0.69	0.51	0.39	0.54
Ours	0.55	0.71	0.60	0.53	0.60

Table 1. Quantitative results of VDMs on PhyGenBench.

compared to existing methods. We evaluate our approach on two established benchmarks for physically plausible video generation. Our framework consistently achieves superior performance across all benchmarks.

4.1. Implementation Details

We propose a two-stage physically plausible image-to-video generation framework. In the first stage, we utilize ChatGPT-4o as the coarse-level motion planner. In the second stage, we utilize an open-source I2V model, Go-with-the-Flow [7], as a fine-level motion synthesizer. Unless otherwise specified, in all experiments, we generate each video with a resolution of 720×480 and 49 frames.

4.2. Benchmarks and Models

Traditional metrics in the visual domain, such as the Peak Signal-to-Noise Ratio (PSNR)[17], the Structural Similarity Index (SSIM)[49], the Learned Perceptual Image

Patch Similarity (LPIPS)[63], the Fréchet Inception Distance (FID)[16] and the Fréchet Video Distance (FVD)[45], do not account for the physical realism of the generated videos [31, 33]. Recent studies have begun to address this limitation by developing benchmarks and metrics that evaluate physical realism. In this work, we adopt two benchmarks, described below.

PhyGenBench [31] categorizes physical properties into four domains: mechanics, optics, thermal, and material. It includes 27 physical phenomena, each governed by real world physical laws, reflected in 160 carefully designed text prompts. As PhyGenbench provides only text prompts, we adapt it to our image-to-video setting by generating a corresponding first frame for each prompt with FLUX[22]. We adhere to the predefined benchmark evaluation protocol, i.e., employing GPT-4o to assess the physical realism of the generated videos.



图 4. 我们框架 CogVideoX-I2V-5B [60]、LTX-VideoI2V [14]和 SVD-XT [5]生成的物理上合理的视频结果的视觉比较。

模型	力学(↑)	光学(↑)	热学(↑)	材料学(↑)	平均(↑)
CogvideoX-T2V-5B	0.43	0.55	0.40	0.42	0.45
LTX-Video-T2V	0.35	0.45	0.36	0.38	0.39
OpenSora	0.43	0.50	0.44	0.37	0.44
Diffusion	0.32	0.41	0.26	0.24	0.31
CogvideoX-I2V-5B	0.48	0.69	0.43	0.41	0.52
SVD-XT	0.46	0.68	0.48	0.41	0.52
			LTX-Video-I2V	0.47	0.65
				0.46	0.37
				0.50	SG-I2V
				0.52	0.69
				0.51	
					0.39
					0.54
我们	0.55	0.71	0.60	0.53	0.60

表 1. PhyGenBench 上 VDMs 的定量结果。

与现有方法相比。我们在两个已建立的物理合理视频生成基准上评估了我们的方法。我们的框架在所有基准上始终实现了优越的性能。

4.1. 实现细节

我们提出了一种两阶段物理上合理的图像到视频生成框架。在第一阶段，我们使用 ChatGPT-4o 作为粗粒度运动规划器。在第二阶段，我们使用开源 I2V 模型 Go-with-the-Flow [7]作为细粒度运动合成器。除非另有说明，在所有实验中，我们以 720×480 分辨率和 49 帧生成每个视频。

学习感知图像块相似性 (LPIPS) [63]、弗雷歇特 inception 距离 (FID) [16]和弗雷歇特视频距离 (FVD) [45]，均未考虑生成视频的物理真实性[31, 33]。近期研究开始通过开发评估物理真实性的基准和指标来解决这个问题。在本工作中，我们采用了两个基准，如下所述。

PhyGenBench [31] 将物理属性分为四个领域：力学、光学、热学和材料学。它包含 27 种物理现象，每种现象都受现实世界物理定律的支配，体现在 160 个精心设计的文本提示中。由于 PhyGenbench 仅提供文本提示，我们通过使用 FLUX[22]为每个提示生成相应的第一帧图像，将其应用于我们的图像到视频设置。我们遵循预定义的基准评估协议，即使用 GPT-4o 来评估生成视频的物理真实性。

4.2. 基准和模型

视觉领域的传统指标，如峰值信噪比 (PSNR) [17]、结构相似性指数 (SSIM) [49]

Model	S.M.(↑)	F.D.(↑)	Optics(↑)	Magnetism(↑)	Thermodynamics(↑)	Average(↑)
Cogvideo-I2V-5B	30.4	29.8	16.7	13.3	8.5	27.1
SVD-XT	21.9	20.5	6.8	8.4	17.1	19.1
LTX-Video-I2V	30.2	29.8	15.9	13.2	8.4	26.8
SG-I2V	34.6	31.2	15.9	13.1	8.4	29.7
Ours	42.3	34.1	16.9	13.4	8.8	34.6

Table 2. Quantitative results of physically plausible video generation on Physics-IQ Benchmark. S.M. refers to Solid Mechanics, and F.D. refers to Fluid Dynamics.

Physics-IQ [33] comprises 396 real-world videos spanning 66 distinct physical scenarios. For each scenario, videos are recorded from three different perspectives and filmed twice under identical conditions to eliminate randomness. This benchmark evaluates real-world physical phenomena, including collisions, object continuity, occlusion, object permanence, and fluid dynamics. This benchmark assesses physical realism from semantic and temporal perspectives, using semantic metrics and visual metrics to compare generated videos against the real-world reference videos.

Compared Models. In the context of text-to-video generation, we compare our framework with CogVideoX-T2V-5B [60], LTX-Video-T2V [14], and OpenSora [66]. Moreover, we evaluate our framework against PhyT2V [58], which enhances physical realism by iteratively refining the prompt. For the image-to-video generation scenario, our framework is evaluated alongside CogVideoX-I2V-5B [60], SVD-XT [5], and LTX-Video-I2V [14]. Additionally, we conducted experiments in the motion-controllable setting. In this setting, we leverage the motion trajectory predicted by the VLM as a condition to guide VDM generation. We benchmark our approach against image-to-video motion controllable model, SG-I2V [35] and text-to-video motion controllable model LLM-grounded Video Diffusion Models [25]. The experimental details are presented in the Appendix.

4.3. Quantitative Evaluation

We begin with an empirical study on PhyGenBench and Physics-IQ, comparing our framework against widely adopted open-source models in the research community. Based on different physical properties, we categorize the benchmark samples accordingly. Additionally, we classify VDMs into text-to-video (T2V) diffusion models and image-to-video (I2V) diffusion models based on the input conditions.

In Table 1, we present our experimental results on PhyGenBench, evaluating different video generation models following its evaluation protocol. The results show that our framework achieves state-of-the-art performance across four different physical phenomena. **Our framework outperforms the best T2V method by an average of 15.3% and the best I2V method by 11.1%.** Specifically, our

framework demonstrates significant advantages in the Mechanics, Thermal, and Material domains, outperforming the best I2V method by 5.7%, 17.6%, and 35.8%, respectively. These advantages are particularly evident in these three types of physical phenomena, which involve more substantial changes in motion, volume, or shape. Our framework is better equipped to understand and reason about bounding box sequences to represent these changes effectively.

Similarly, for the Physics-IQ benchmark, we evaluate the performance of different video generation models following its evaluation protocol. **Our framework achieves the best results across four different physical phenomena, with improvements of 22.2% in Solid Mechanics and 9.2% in Fluid Dynamics compared to the second-best models.** These significant improvements demonstrate the effectiveness of our framework in generating physically plausible videos.

4.4. Qualitative Evaluation

Figures 4 and 5 demonstrate a qualitative comparison between our video generation framework and baseline methods. Among all evaluated approaches, our framework consistently produces videos with the highest degree of physical realism. In the ball falling sample in Figure 4, while CogVideoX shows a bouncing effect, artifacts are present in the video; LTX-Video and SVD-XT exhibit motions that do not adhere to the laws of physics. In Figure 5, we analyze two examples from Physics-IQ. In the pouring water example, the baseline methods fail to show the simultaneous decrease in water level of the glass beverage dispenser and the increase in water level of the glass below; in the ball collision example, none of the baseline methods correctly depict the collision of balls. More videos are provided in the supplement.

4.5. Ablation Study

We perform an ablation study to evaluate the contributions of key components in our framework. We design four variants to analyze the effectiveness of different components in our framework.

1. **Ours w/o VLM Planner:** To assess the overall functionality of our framework, we replace the structured noise input of the VDM with random noise to evaluate the ef-

S.M.(↑) F.D.(↑) 光学(↑) 磁学(↑) 热力学(↑) 平均(↑)						
Cogvideo-I2V-5B	30.4	29.8	16.7	13.3	8.5	27.1
SVD-XT	21.9	20.5	6.8	8.4	17.1	19.1
LTX-Video-I2V	30.2	29.8	15.9	13.2	8.4	26.8
SG-I2V	34.6	31.2	15.9	13.1	8.4	29.7
Ours	42.3	34.1	16.9	13.4		
					8.8	34.6

表 2. Physics-IQ 基准上物理合理视频生成的定量结果。S.M.指固体力学，F.D. 指的是流体动力学。

Physics-IQ [33] 包含 396 个真实世界视频，涵盖 66 种不同的物理场景。对于每个场景，从三个不同的视角进行录像，并在相同条件下拍摄两次以消除随机性。该基准评估真实世界的物理现象，包括碰撞、物体连续性、遮挡、物体持久性和流体动力学。该基准从语义和时间角度评估物理真实性，使用语义指标和视觉指标将生成的视频与真实世界参考视频进行比较。

对比模型。在文本到视频生成的背景下，我们将我们的框架与 CogVideoX-T2V5B [60]、LTX-Video-T2V [14] 和 OpenSora [66] 进行比较。此外，我们通过迭代优化提示来增强物理真实性的 PhyT2V [58] 评估我们的框架。对于图像到视频生成的场景，我们的框架与 CogVideoX-I2V-5B [60]、SVD-XT [5] 和 LTX-Video-I2V [14] 一起进行评估。此外，我们在可控运动设置中进行了实验。在此设置中，我们利用 VLM 预测的运动轨迹作为条件来指导 VDM 生成。我们将我们的方法与图像到视频可控运动模型 SG-I2V [35] 和文本到视频可控运动模型 LLM-grounded Video Diffusion Models [25] 进行基准测试。实验细节在附录中呈现。

我们的框架在力学、热学和材料学领域展现出显著优势，分别比最优 I2V 方法高出 5.7%、17.6% 和 35.8%。这些优势在涉及运动、体积或形状发生较大变化的这三种物理现象中尤为明显。我们的框架更擅长理解和推理边界框序列，以有效表示这些变化。

同样地，在 Physics-IQ 基准测试中，我们遵循其评估协议，评估了不同视频生成模型的性能。我们的框架在四种不同的物理现象中均取得了最佳结果，与次优模型相比，在固体力学中提升了 22.2%，在流体动力学中提升了 9.2%。

这些显著改进证明了我们的框架在生成物理上合理的视频方面的有效性。

4.4. 定性评估

图 4 和图 5 展示了我们的视频生成框架与基线方法之间的定性比较。在所有评估方法中，我们的框架始终生成物理真实性最高的视频。在图 4 中的球体下落示例中，虽然 CogVideoX 显示了弹跳效果，但视频中存在伪影；LTX-Video 和 SVD-XT 表现出不符合物理定律的运动。在图 5 中，我们分析了 Physics-IQ 中的两个示例。在倒水示例中，基线方法未能显示玻璃饮料分配器的水位同时下降和下方玻璃的水位同时上升；在球体碰撞示例中，没有任何基线方法正确描绘了球体的碰撞。补充材料中提供了更多视频。

4.3. 定量评估

我们从 PhyGenBench 和 Physics-IQ 的实证研究开始，将我们的框架与科研界广泛采用的开放源码模型进行比较。根据不同的物理属性，我们将基准样本相应分类。此外，根据输入条件，我们将 VDMs 分为文本到视频 (T2V) 扩散模型和图像到视频 (I2V) 扩散模型。

在表 1 中，我们展示了遵循 PhyGenBench 评估协议的实验结果，评估了不同的视频生成模型。结果表明，我们的框架在四种不同的物理现象中均实现了最先进的性能。我们的框架平均比最佳 T2V 方法高 15.3%，比最佳 I2V 方法高 11.1%。具体来说

4.5. 消融研究

我们进行消融研究以评估我们框架中关键组件的贡献。我们设计了四个变体来分析我们框架中不同组件的有效性。

1. 去除 VLM 规划器：为了评估我们框架的整体功能，我们将 VDM 的结构化噪声输入替换为随机噪声，以评估其效果。

Model	S.M.(\uparrow)	F.D.(\uparrow)	Optics(\uparrow)	Magnetism(\uparrow)	Thermodynamics(\uparrow)	Average(\uparrow)
Ours	42.3	34.1	16.9	13.4	8.8	34.9
w/o VLM Planner.	16.3	20.8	13.4	5.8	5.6	16.2
w/o C.I	26.3	28.1	16.9	11.2	8.4	24.3
w/o CoT	21.4	26.9	16.1	8.6	6.9	21.0
w/o C.C	18.7	22.4	14.9	7.2	6.1	18.1

Table 3. Ablation study on VLM, in-context learning and COT. S.M. refers to Solid Mechanics, and F.D. refers to Fluid Dynamics.

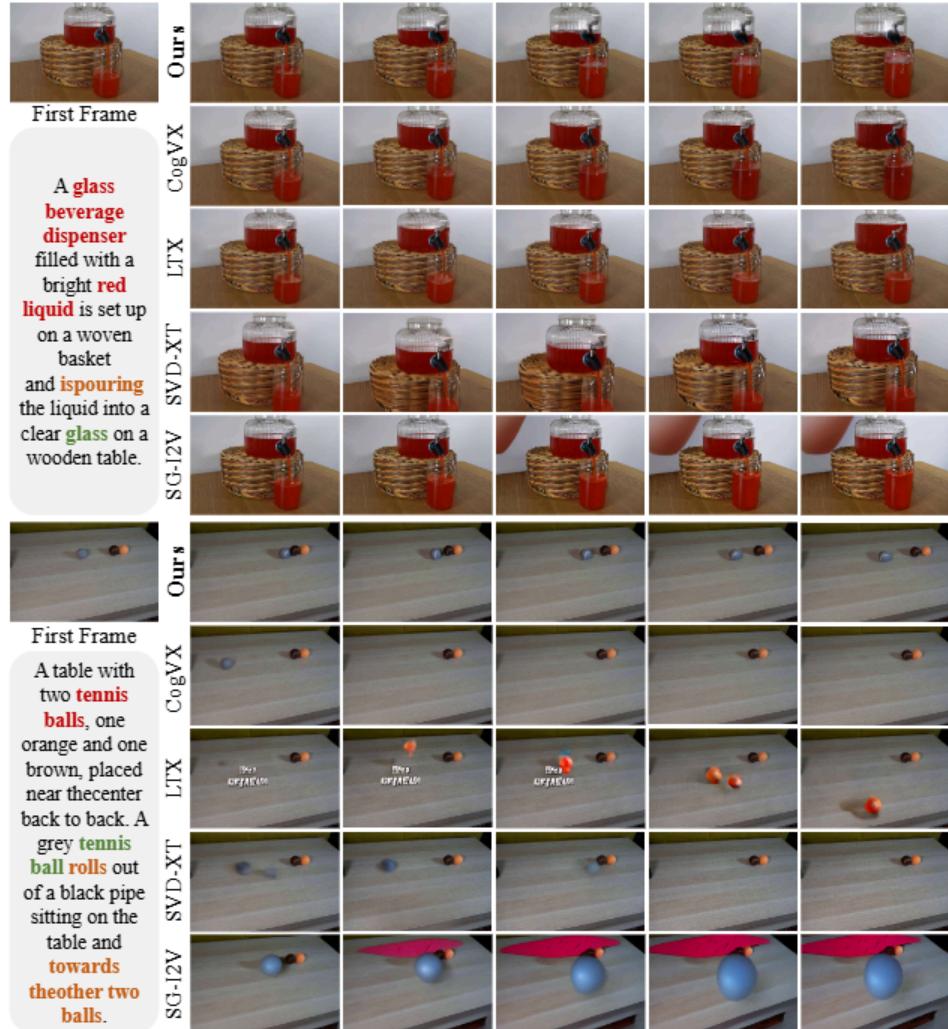


Figure 5. Visual comparisons of physically plausible video generation results from our framework, CogVideoX-I2V-5B, LTX-Video-I2V, SVD-XT and SG-I2V [35] in the Physics-IQ dataset.

fectiveness of the VLM planner.

2. **Ours w/o CI:** Keeping the overall structure unchanged, we remove the in-context information from the VLM.
3. **Ours w/o CoT:** Similarly, while keeping other components unchanged, we remove the CoT reasoning process from the VLM.
4. **Ours w/o CC:** Lastly, we remove both the in-context information and the CoT reasoning process from the VLM while maintaining all other components.

Table 3 presents a quantitative comparison between our full method and these variants. Among all variants, **Ours w/o VLM Planner** shows the most significant performance drop, as removing the planner completely eliminates our ability in understanding the physical laws, leading to nearly random results. Notably, **Ours w/o CoT** exhibits a more pronounced decline compared to **Ours w/o CI**, indicating that the reasoning process in CoT enhances the understanding of physics. While in-context information contributes to the physical reasoning ability of VLM, compared to CoT it

is less effective in preventing errors caused by VLM hallucination.

4.6. User Study

To complement our above evaluations, we conduct a user study to assess the subjective human perception of the generated videos. We follow the gold standard experimental approaches from psychophysics, a 2AFC paradigm, which means two-alternative-forced-choice [33]. In our case, participants completed a questionnaire in which they were presented with pairs of videos and asked to select the one that better aligned with their expectations of physical realism. Responses from 50 participants are summarized in Table 4. The result indicates a strong preference for videos generated by our framework over those from competitors. A detailed analysis of these findings follows in the subsequent discussion.

Model	P.P.(\uparrow)	V.R.(\uparrow)
CogVideoX-I2V-5B	34%	40%
LTX-Video	22%	18%
Ours	52%	48%

Table 4. User study statistics of the preference rate for Physical Plausibility (P.P.) & Visual Realism (V.R.).

4.7. Limitations

Although our framework can generate physically plausible videos, its performance remains constrained by the base model. Firstly, we cannot model physical events that cannot be represented by image space bounding box trajectories. For example, phenomena that involve intrinsic state changes of objects such as solid fragmentation and gas solidification. Moreover, our pipeline lacks 3D spatial perception. It is unable to understand the spatial relationships within the scene. Finally, the optical flow of small objects is prone to noise interference. This will cause our framework to generate ambiguous content. With the recent progress in video generation model, we anticipate that our framework will be further improved in generating videos under more challenging physical conditions.

S.M.(↑) F.D.(↑) 光学(↑) 磁学(↑) 热力学(↑) 平均(↑)

我们的 42.3 34.1 16.9 13.4 8.8 34.9

不含 VLM 规划器。 16.3 20.8 13.4 5.8 5.6 16.2

不含 C.I 26.3 28.1 16.9 11.2 8.4 24.3 不含 CoT 21.4 26.9 16.1 8.6 6.9 21.0 不含 C.C 18.7 22.4 14.9 7.2 6.1 18.1 表 3. 关于 VLM、情境学习和 CoT 的消融研究。S.M. 指的是固体力学，F.D. 指的是流体动力学。

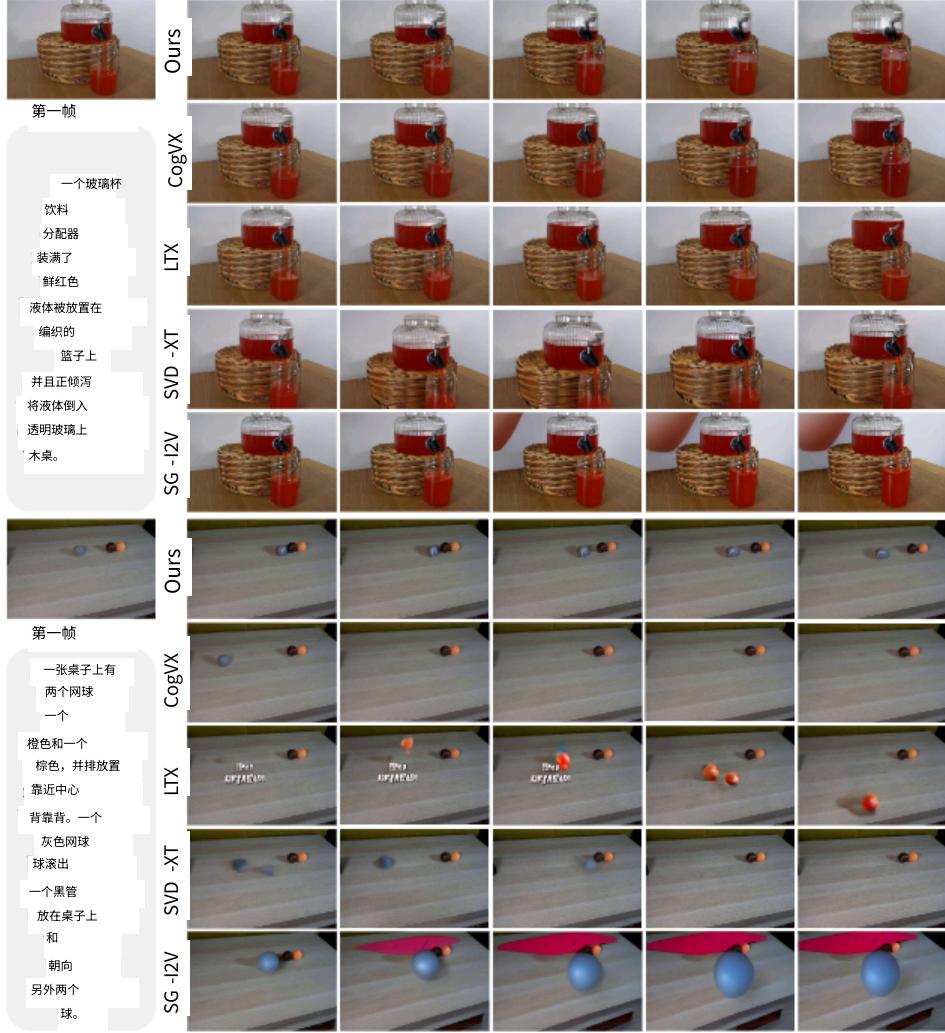


图 5. 在 Physics-IQ 数据集上，我们的框架 CogVideoX-I2V-5B、LTXVideo-I2V、SVD-XT 和 SG-I2V 生成的物理上合理的视频结果的视觉比较[35]。

VLM 规划器的有效性。

2. 我们的无 CI 版本：保持整体结构不变，我们从 VLM 中移除了情境信息。
3. 我们的无 CoT 版本：类似地，在保持其他组件不变的情况下，我们从 VLM 中移除了 CoT 推理过程。
4. 我们的无 CC 版本：最后，我们移除了 VLM 规划器中的情境信息和 CoT 推理过程，同时保持所有其他组件。

表 3 展示了我们的完整方法与这些变体之间的定量比较。在所有变体中，我们的无 VLM 规划器版本表现出最显著的性能下降，因为移除规划器完全消除了我们理解物理定律的能力，导致结果几乎随机。值得注意的是，我们的无 CoT 版本相比无 CI 版本表现出更明显的下降，表明 CoT 中的推理过程增强了物理理解能力。而情境信息则有助于 VLM 的物理推理能力。

与 CoT 相比，它在预防由 VLM 幻觉引起的错误方面效果较差。

4.6. 用户研究

为了补充我们之前的评估，我们进行了一项用户研究，以评估生成视频的主观人类感知。我们遵循心理物理学中的黄金标准实验方法，采用 2AFC 范式，即双选择强制选择[33]。在我们的案例中，参与者完成了一份问卷，其中他们被展示成对的视频，并被要求选择更符合他们对物理现实期望的那个。50 名参与者的回答总结在表 4 中。结果表明，人们强烈倾向于我们框架生成的视频，而不是竞争对手的视频。对这些发现的详细分析将在随后的讨论中进行。

模型 P.P.(↑) V.R.(↑)

CogVideoX-I2V-5B	34%	40%
LTX-Video	22%	18%

我们 52% 48%

表 4. 用户研究统计数据：对物理合理性 (P.P.) 与视觉真实性 (V.R.) 的偏好率

4.7. 局限性

尽管我们的框架能够生成物理上合理的视频，但其性能仍然受限于基础模型。首先，我们无法模拟无法用图像空间边界框轨迹表示的物理事件。例如，涉及物体内在状态变化的物理现象，如固体碎裂和气体凝固。此外，我们的流程缺乏三维空间感知能力，无法理解场景内的空间关系。最后，小物体的光流容易受到噪声干扰，这将导致我们的框架生成模糊内容。随着视频生成模型的最新进展，我们预计我们的框架将在更具有挑战性的物理条件下生成视频方面得到进一步改进。

5. Conclusion

Recently, VDMs have achieved great empirical success and are receiving considerable attention in computer vision and computer graphics. However, due to the lack of understanding of physical laws, VDMs are unable to generate physically plausible videos. In this paper, we introduce VLIPP, a novel two-stage physically plausible video generation framework that incorporates physical laws into video diffusion models through vision and language informed physical prior. Our experimental results demonstrate the effectiveness of our method compared to existing approaches.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4
- [2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 3
- [3] 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. 4
- [4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024. 2
- [5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6, 7, 1
- [6] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. 2
- [7] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. *arXiv preprint arXiv:2501.08331*, 2025. 2, 5, 6
- [8] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *The Twelfth International Conference on Learning Representations*. OpenReview, 2024. 5
- [9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 4
- [10] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [11] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 4
- [12] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. *arXiv preprint arXiv:2401.18085*, 2024. 5
- [13] Daniel Geng, Charles Herrmann, Junhua Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 5
- [14] 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. 6, 7, 1
- [15] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6
- [17] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6
- [18] Hao-Yu Hsu, Zhi-Hao Lin, Albert Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically realistic video editing from natural language instructions. *arXiv preprint arXiv:2411.02394*, 2024. 2
- [19] Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. Fine-grained controllable video generation via object appearance and context. *arXiv preprint arXiv:2312.02919*, 2023. 3
- [20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 3
- [21] 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. 2, 3
- [22] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 6, 1
- [23] Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Guo Zhou, Hua Yao, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned

5. 结论

最近，视觉动态模型（VDMs）在计算机视觉和计算机图形学领域取得了巨大的实证成功，并受到了广泛关注。然而，由于缺乏对物理定律的理解，VDMs 无法生成物理上合理的视频。在本文中，我们介绍了 VLIPP，这是一个新颖的两阶段物理上合理的视频生成框架，它通过视觉和语言引导的物理先验将物理定律融入视频扩散模型。我们的实验结果表明，与现有方法相比，我们的方法具有有效性。

参考文献

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, 等。
Gpt-4 技术报告。arXiv 预印本 arXiv:2303.08774, 2023. 4 [2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: 分析和改进视频扩散中的三维相机控制
前人。arXiv 预印本 arXiv:2411.18673, 2024. 3
[3] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, KaiWei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520, 2024. 4
[4] Amir Bar, 高岳 Zhou, Danny Tran, Trevor Darrell 和 Yann LeCun. 导航世界模型. arXiv 预印本 arXiv:2412.03572, 2024. 2
[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, 等. 稳定视频扩散：将潜在视频扩散模型扩展到大规模 datasets. arXiv preprint arXiv:2311.15127, 2023 年 6 月 7 日, 第 1 页
[6] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, 等. 通用人工智能的火花：gpt-4 的早期实验, 2023. 2 [7] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, 等. 随遇而安：实时扭曲控制的运动可控视频扩散模型
噪声。arXiv 预印本 arXiv:2501.08331, 2025 年 2 月 5 日, 第 6 页
[8] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. 如何扭曲你的噪声：用于扩散模型的时序相关噪声先验。在第十二届国际
国际学习表征会议. OpenReview, 2024. 5 [9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source
套房。中国科学信息科学, 67(12):220101, 2024. 4
- [10] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas、Noah Snavely 和 Gordon Wetzstein。街景：使用自回归视频扩散生成大规模一致街景。在 ACM SIGGRAPH 2024 会议论文集中, 第 1-11 页, 2024 年。2
[11] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu 等人。情境学习的综述。arXiv 预印本 arXiv:2301.00234, 2022. 4 [12] Daniel Geng 和 Andrew Owens。运动引导：
基于扩散的图像编辑与可微分的运动
定时器。arXiv 预印本 arXiv:2401.18085, 2024 年 5 月。
[13] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, 等人。运动提示：通过运动轨迹控制视频生成。arXiv 预印本 arXiv:2412.02700, 2024 年 5 月。
[14] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon 等。Ltx-video：实时视频潜在扩散。arXiv 预印本 arXiv:2501.00103, 2024 年 6 月 7 日, 第 1 页。
[15] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li 和 Ceyuan Yang。Cameractrl：实现文本到视频生成的相机控制。arXiv 预印本 arXiv:2404.02101, 2024 年 3 月
[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, 和 Sepp Hochreiter。通过双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡。神经信息处理系统进展, 第 30 卷, 2017 年 6 月。
[17] Alain Hore 和 Djemel Ziou。图像质量指标：PSNR vs. ssim. 在 2010 年第 20 届国际模式识别, 第 2366-2369 页。IEEE, 2010 年。6 [18] 许浩宇, 林志豪, 赵伯, 夏洪池, 和
王神龙. Autovfx: 从自然语言指令进行物理真实视频编辑.
arXiv 预印本 arXiv:2411.02394, 2024. 2 [19] 黄欣平, 苏宇川, 孙德清, 姜路,
徐慧嘉、朱雨坤和杨明轩。基于物体外观和上下文的细粒度可控视频生成。arXiv 预印本 arXiv:2312.02919, 2023. 3
[20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, 张寅峰, 王浩天, 陈强龙, 彭伟华, 冯晓程, 秦冰, 等. 大型语言模型中的幻觉现象综述：原理、分类、挑战与开放性问题.
ACM 计算机学会信息系统汇刊, 43(2):1-55, 2025. 3
[21] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, 王凯欣、黄高、冯嘉石. 视频生成与世界模型的距离：从物理规律的角度看. arXiv 预印本 arXiv:2411.02385, 2024. 2, 3
[22] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024 年 6 月 1 日
[23] Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Guo Zhou, Hua Yao, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned

- video generation via diffusion models. *arXiv preprint arXiv:2312.00651*, 2023. 3
- [24] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024. 2
- [25] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 3, 7, 1
- [26] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 3
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3
- [28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 4
- [29] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenglong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024. 2, 5
- [30] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3
- [31] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quan-feng 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. 4, 6, 1
- [32] Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2
- [33] 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. 2, 4, 6, 7, 8, 1
- [34] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505, 2024. 3
- [35] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 7, 8, 1
- [36] Andrew Nealen, Matthias Müller, Richard Keiser, Eddy Boxerman, and Mark Carlson. Physically based deformable models in computer graphics. In *Computer graphics forum*, pages 809–836. Wiley Online Library, 2006. 2
- [37] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*, pages 111–128. Springer, 2024. 3
- [38] Openai. <https://openai.com/index/hello-gpt-4o/>, 2024. 4
- [39] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 3
- [40] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024. 3
- [41] 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. 4
- [42] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4): 1–10, 2013. 2
- [43] Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: Biomedical generalist video generation model. *arXiv preprint arXiv:2407.08944*, 2024. 2
- [44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5
- [45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6
- [46] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 3
- [47] Xi Wang, Robin Courant, Marc Christie, and Vicky Kalogeiton. Akira: Augmentation kit on rays for optical video generation. *arXiv preprint arXiv:2412.14158*, 2024. 3
- [48] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 2
- [49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [50] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3
- [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3, 5

- 通过扩散模型生成视频。arXiv 预印本 arXiv:2312.00651, 2023. 3
- Trackdiffusion: Tracklet-conditioned [24] 郑琦, 理查德·塔克, 诺亚·斯奈利和霍林斯基。生成图像动力学. 在 IEEE/CVF 计算机视觉与模式识别会议论文集中, 第 24142-24153 页, 2024. 2
- [25] 龙连, 白凤石, 亚当·亚拉, 特雷弗·达雷尔和 Boyi Li. 基于 LLM 的视频扩散模型. arXiv 预印本 arXiv:2309.17444, 2023. 3, 7, 1
- [26] Han Lin, Abhay Zala, Jaemin Cho, 和 Mohit Bansal. Videodirectorgpt: 通过 LLM 引导规划实现一致的多场景视频生成. arXiv 预印本 arXiv:2309.15091, 2023. 3
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 使用视觉指令微调改进的基线, 2023 年 3 月 [28] 刘浩天, 李春元, 李宇恒, 李博, 韩元 张胜申, 李永宰. Llavanext: 改进的推理、OCR 和世界知识, 2024. 4
- [29] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shen-王龙. Physgen: 基于刚体物理的图像到视频生成. 在欧洲计算机视觉会议, 第 360-378 页. 斯普林格出版社, 2024. 2, 5
- [30] 万多·库尔特·马, 约翰·P·刘易斯, 和 W·巴斯蒂安·克莱因。开拓者: 基于扩散的视频生成的轨迹控制。在 SIGGRAPH Asia 2024 会议论文集中, 第 1-11 页, 2024 年 3 月。
- [31] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quan-feng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, 和 Ping Luo。迈向世界模拟器: 构建基于物理常识的视频生成基准。arXiv 预印本 arXiv:2410.05363, 2024 年 4 月, 6 月, 1 日。
- [32] Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, 和 Enrico Magli。Motioncraft: 基于物理的零样本视频生成。在神经信息处理系统大会第 38 届年会上。
- 2
- [33] 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. 2, 4, 6, 7, 8, 1
- [34] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang 张英山, 和 张建. Revideo: 具备运动和内容控制的视频重制. 神经信息处理系统进展, 37:18481–18505, 2024. 3
- [35] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, David B Lindell。Sg-i2v: 图像到视频生成中的自引导轨迹控制。arXiv 预印本 arXiv:2411.04989, 2024 年 7 月 8 日, 第 1 页。
- [36] Andrew Nealen, Matthias Müller, Richard Keiser, Eddy Boxerman, 以及 Mark Carlson。计算机图形中的基于物理的可变形模型。在《计算机图形论坛》中, 第 809–836 页。Wiley 在线图书馆, 2006 年。第 2 页。
- [37] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, 以及 Yinqiang Zheng。Mofa-video: 通过冻结图像到视频扩散模型中的生成运动场自适应实现可控图像动画。在《欧洲计算机视觉会议》中, 第 111–128 页。Springer, 2024 年。
- 3
- [38] Openai. <https://openai.com/index/hello-gpt-4o/>, 2024 年 4 月
- [39] 陈斌, 巴鲁汉丁·亚曼, 托马索·内斯蒂, 阿比鲁普·雷。Vlp: 视觉语言规划用于自动驾驶。在 IEEE/CVF 计算机视觉与模式识别会议论文集中, 第 14760-14769 页, 2024 年。3
- [40] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. arXiv preprint arXiv:2406.16863, 2024. 3 [41] 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. 4
- [42] 亚历克谢·斯托马京, 克雷格·施罗德, 劳伦斯·蔡, 约瑟夫·Teran 和 Andrew Selle. 一种用于雪模拟的材料点方法. ACM Transactions on Graphics (TOG), 32(4): 1–10, 2013. 2
- [43] Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: 生物医学通用视频生成模型。arXiv preprint arXiv:2407.08944, 2024. 2
- [44] Zachary Teed 和 Jia Deng. Raft: 循环全对字段 用于光流的变换。在计算机视觉-ECCV 2020: 第 16 届欧洲会议, 英国格拉斯哥, 2020 年 8 月 23 日至 28 日, 会议录, 第二部分 16, 第 402–419 页。Springer, 2020. 5
- [45] 托马斯·翁特廷纳, 斯约尔德·范·斯滕基斯特, 卡罗尔·库拉奇, 拉斐尔·马利尼埃, 马辛·米哈尔斯基和西尔万·盖利。Fvd: A new metric for video generation. 2019. 6 [46] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guo-强伟、袁丽平和李航。Boximator: 为视频合成生成丰富且可控的运动。arXiv 预印本 arXiv:2402.01566, 2024 年 3 月。
- [47] 西王, 罗宾·库兰, 马克·克里斯蒂和维姬·卡拉-geiton. Akira: 光学视频用射线上的增强套件 生成。arXiv 预印本 arXiv:2412.14158, 2024. 3
- [48] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jia-Gang Zhu, 和 Jiwen Lu. Drivedreamer: 迈向自动驾驶的真实世界驾驶世界模型。在《欧洲计算机视觉会议》, 第 55-72 页。Springer, 2024 年。2
- [49] 周王, Alan C Bovik, Hamid R Sheikh, 和 Eero P Simoncelli. 图像质量评估: 从错误可见性到结构相似性。IEEE 图像处理汇刊, 13(4):600–612, 2004. 6
- [50] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: 一个用于视频生成的统一且灵活的运动控制器。ACM SIGGRAPH 2024 会议论文集, 第 1-11 页, 2024 年 3 月。
- [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 思维链提示在大语言模型中激发推理。神经信息处理系统进展, 第 35 卷第 24824-24837 页, 2022 年 3 月, 第 5 期。

- [52] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*, 2024. 3
- [53] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336, 2024. 3
- [54] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 3
- [55] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 3
- [56] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 2
- [57] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024. 3
- [58] Qiya Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. *arXiv preprint arXiv:2412.00596*, 2024. 1, 7
- [59] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 2
- [60] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenqi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 5, 6, 7, 1
- [61] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3
- [62] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023. 2
- [63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6
- [64] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406. Springer, 2024. 2
- [65] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 3
- [66] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 7

- [5] [52] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: 基于运动的定制化文本到视频生成版本。arXiv 预印本 arXiv:2406.17758, 2024 年 3 月
- [53] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, 和 Trevor Darrell。自纠正 LLM 控制的扩散模型。在 IEEE/CVF 计算机视觉与模式识别会议论文集中, 第 6327–6336 页, 2024 年 3 月
- [54] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, 和 Di Zhang. Draganything: 使用实体表示的任何物体的运动控制。在欧几里得计算机视觉会议中, 第 331–348 页。Springer, 2024 年 3 月
- [55] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting 宁宇恒, 赵泽宇, 冯天华, 郝时博, 石垚明, 等. Pandora: 面向通用世界模型的自然语言动作与视频状态。arXiv 预印本 arXiv:2406.09455, 2024. 3
- [56] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, 陈帆夫江. Physgaussian: 物理集成 3D 高斯用于生成动力学。在 IEEE/CVF 计算机视觉与模式识别会议论文集中, 第 4389-4398 页, 2024 年. 2
- [57] 许子伟、Sanjay Jain 和 Mohan Kankanhalli. Hallucination 一个国家是不可避免的：大型语言的天生局限性 模型。arXiv 预印本 arXiv:2401.11817, 2024 年 3 月
- [58] Qiya Xue, Xiangyu Yin, Boyuan Yang, and Wei 高. Phyt2v: 基于 LLM 的迭代自改进物理约束文本到视频生成。arXiv 预印本 arXiv:2412.00596, 2024. 1, 7 [59] 杨梦娇, 杜一伦, 卡米尔 · 加塞米普尔, 乔纳森
- Tompson、Dale Schuurmans 和 Pieter Abbeel. 学习交互式真实世界模拟器。arXiv 预印本 arXiv:2310.06114, 1(2):6, 2023. 2
- [60] 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. arXiv preprint arXiv:2408.06072, 2024. 5, 6, 7, 1
- [61] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, 黄志鹏, 高祥军, 王天信, 山英, 田永红. Viewcrafter: 针对高保真新视角合成的视频扩散模型控制方法。arXiv 预印本 arXiv:2409.02048, 2024. 3 [62] 叶远, 宋嘉明, 优马尔 · 伊克巴尔, 阿拉斯 · 瓦赫达特, 简
- Kautz. Physdiff: 基于物理指导的人类运动扩散模型。在 IEEE/CVF 国际计算机视觉会议论文集中, 第 16010-16021 页, 2023 年。2
- [63] 张 Richard, Phillip Isola, Alexei A Efros, Eli Shecht-一位男士, 以及 Oliver Wang。深度特征作为感知度量的不合理有效性。在 IEEE 计算机视觉与模式识别会议论文集中, 第 586-595 页, 2018 年。6
- [64] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y 冯长喜、郑诺、吴嘉俊和威廉 · T · 弗里曼
- Physdreamer: 通过视频生成实现基于物理的 3D 物体交互。发表于《欧洲计算机视觉会议》, 第 388-406 页。Springer 出版社, 2024 年。2
- [65] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llavanext: A strong zero-shot video understanding model, 2024. 3
- [66] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, 李胜贵, 刘红鑫, 周宇坤, 李天一, 游阳。Open-sora: 为所有人实现高效视频制作民主化。arXiv 预印本 arXiv:2412.20404, 2024 年。7

VLIPP: Towards Physically Plausible Video Generation with Vision and Language Informed Physical Prior

Supplementary Material

A. Coarse-Level Motion Planning Details

In this section, we present the experimental setting and details for reproducing the results. The main principle of our experimental setting is to fairly compare different Video Diffusion Models(VDMs) in generating physically plausible videos. Our adapt well-known open source model to serve as Compared Models. We discuss these models in details.

1. **CogVideoX[60]**: CogVideoX is capable of performing both text-to-video generation and image-to-video generation. It provides two model variants, featuring 2 billion and 5 billion parameters, respectively. In our experiments, we configured CogVideoX to generate 49 frames with a resolution of 720x480.
2. **LTX-Video[14]**: LTX-Video is also capable of performing both text-to-video generation and image-to-video generation. In our experiments, we compared two versions of LTX-Video with corresponding methods. It can generate videos with 49 frames with a resolution of 768x512.
3. **SVD-XT[5]**: SVD-XT is capable of performing image-to-video generation. In our experiments, we configured SVD-XT to generate 25 frames with a resolution of 1024x576.
4. **SG-I2V[35]**: SG-I2V is a motion trajectory-guided image-to-video generation model. It is capable of generating bounding box-controllable videos with 14 frames with a resolution of 1024x576.
5. **LLM-grounded Video Diffusion Models[25]**: LLM-grounded Video Diffusion Models are capable of predicting future frame bounding boxes based on input prompts and injecting the box information in a training-free manner. In our experiments, we configured LLM-grounded to generate 24 frames with a resolution of 576x320.

We additionally present the Reasoning Template utilized during the stage 1 Coarse-Level Motion Planning process, as shown in Fig 6 and Fig 7. This includes system instructions to ensure the proper functioning of the chain of thought and provides the VLM with context information to guarantee the accuracy of predictions.

B. Experiment Details

In this section, we present the experimental details of our benchmark, PhyGenBench[31] and Physics-IQ[33].

PhyGenBench comprises 160 prompts, spanning four

domains of physical knowledge: Mechanics (40), Optics (50), Thermal (40), and Material (20), along with 27 types of physical laws. It also includes 165 objects and 42 actions. The evaluation focuses on two aspects: semantic alignment and physical commonsense alignment. The degree of semantic alignment is assessed by extracting objects and actions from the prompts using a Vision-Language Model (VLM), determining whether the objects appear, and evaluating based on the presence of objects and the occurrence of actions. The degree of physical commonsense alignment is determined through a three-step process: detecting whether the physical phenomena occur and whether the order of occurrence is correct; and finally conducting an overall naturalness evaluation.

Physic-IQ categorizes real-world physical laws into Solid Mechanics, Fluid Dynamics, Optics, Magnetism, and Thermodynamics, encompassing 114, 45, 24, 6, and 9 videos, respectively. The evaluation approach is twofold, focusing on physical comprehension and visual authenticity. Physical comprehension is determined by identifying the timing, location, and frequency of actions, ultimately calculating the mean squared error between corresponding pixels in the generated and real frames to derive a physical comprehension score. Visual authenticity is evaluated using a Vision-Language Model (VLM), employing the gold standard experimental method from psychophysics. The VLM receives pairs of real and generated videos of the same scene in random order and is tasked with identifying the real scene, a design intended to reflect visual authenticity.

During the experimental phase of this paper, we utilized the prompts provided by the PhyGenBench dataset to infer the initial frame’s prompts using an LLM, which were then generated by FLUX[22]. To ensure fairness in comparison, all I2V models were supplied with the same initial frame image. Given that different models produce videos with varying numbers of frames, a uniform sampling ratio was applied during the testing phase to extract key frames consistently across all models.

C. More Qualitative Results

In this section, we further demonstrate examples of the proposed framework across various scenarios using prompts from the PhyGenBench dataset. Fig 8, 9, and 10 show video generation results driven by physical conditions like fluid dynamics and thermodynamics, while Fig 11 highlights effects in light reflection scenarios.

VLIPP：基于视觉与语言信息物理先验的物理合理视频生成方法

补充材料

A. 粗粒度运动规划细节

在本节中，我们介绍了实验设置和重现结果的详细信息。我们实验设置的主要原则是公平比较不同的视频扩散模型 (VDMs) 在生成物理上合理的视频方面的表现。我们将著名的开源模型适配作为对比模型。我们详细讨论了这些模型。

1. CogVideoX[60]: CogVideoX 能够执行文本到视频生成和图像到视频生成。它提供两个模型变体，分别具有 20 亿和 50 亿参数。在我们的实验中，我们将 CogVideoX 配置为生成 49 帧，分辨率为 720×480 。
2. LTX-Video[14]: LTX-Video 也能执行文本到视频生成和图像到视频生成。在我们的实验中，我们比较了 LTX-Video 的两个版本与对应方法。它可以生成分辨率为 768×512 、包含 49 帧的视频。
3. SVD-XT[5]: SVD-XT 能够执行图像到视频生成。在我们的实验中，我们配置 SVD-XT 生成分辨率为 1024×576 、包含 25 帧的视频。
4. SG-I2V[35]: SG-I2V 是一种基于运动轨迹引导的图像到视频生成模型。它能生成边框框可控、分辨率为 1024×576 、包含 14 帧的视频。
5. LLM-grounded Video Diffusion Models[25]: LLM-grounded Video Diffusion Models 能够根据输入提示预测未来帧的边框框，并以无需训练的方式注入框信息。在我们的实验中，我们配置 LLM-grounded 生成分辨率为 576×320 、包含 24 帧的视频。

我们此外展示了在阶段 1 粗粒度运动规划过程中使用的推理模板，如图 6 和图 7 所示。这包括系统指令以确保思维链的适当功能，并为视觉语言模型提供上下文信息以保证预测的准确性。

B. 实验细节

在本节中，我们介绍了我们的基准测试 PhyGenBench[31]和 Physics-IQ[33]的实验细节。

PhyGenBench 包含 160 个提示，涵盖四个

物理知识领域：力学 (40)、光学 (50)、热学 (40) 和材料 (20)，以及 27 种物理定律。还包括 165 个物体和 42 个动作。评估主要关注两个方面：语义对齐和物理常识对齐。语义对齐程度通过使用视觉语言模型 (VLM) 从提示中提取物体和动作，判断物体是否出现，并根据物体是否存在以及动作是否发生进行评估。物理常识对齐程度通过三步过程确定：检测物理现象是否发生以及发生顺序是否正确；最后进行整体自然度评估。

Physic-IQ 将现实世界中的物理定律分为固体力学、流体动力学、光学、磁学和热力学，分别包含 114、45、24、6 和 9 个视频。评估方法分为两个方面，即物理理解和视觉真实性。物理理解通过识别动作的时间、位置和频率来确定，最终通过计算生成帧和真实帧中对应像素的均方误差来得出物理理解分数。视觉真实性使用视觉语言模型 (VLM) 进行评估，采用心理物理学中的金标准实验方法。VLM 随机接收同一场景的真实视频和生成视频的对，并要求识别真实场景，这种设计旨在反映视觉真实性。

在本文的实验阶段，我们利用 PhyGenBench 数据集提供的提示，通过 LLM 推断初始帧的提示，然后由 FLUX[22]生成。为确保比较的公平性，所有 I2V 模型都接收了相同的初始帧图像。考虑到不同模型生成的视频帧数不同，在测试阶段应用了统一的采样率，以在所有模型中一致地提取关键帧。

C. 更多定性结果

在本节中，我们使用 PhyGenBench 数据集的提示，在多种场景中进一步展示了所提出的框架的示例。图 8、9 和 10 展示了由流体动力学和热力学等物理条件驱动的视频生成结果，而图 11 突出了光反射场景中的效果。

Physics-Aware Reasoning Template (Gravity/Momentum)

System Instruction

You are an expert in real-world physical motion. Your task is to predict video bounding boxes. You will receive a reference image, its segmentation map, the initial bounding box where motion starts, and a textual description of the video content. Your task is to predict the bounding box coordinates for the next 13 frames. When predicting, ensure that different bounding boxes do not overlap or exceed the frame boundaries.

Each frame should be represented as: `[{`id`: unique object identifier, `name`: object name, `box`: [box top-left x-coordinate, box top-left y-coordinate, box width, box height]}, ...]`. Your predicted frames must cover the entire described scene. Critical moments in the movement must be included in the predicted sequence, such as the moment of collision. You need to consider the material and properties of the moving objects in the video. You only need to focus on the objects provided to you in the first frame, and no need to pay attention to anything else. Your predictions should infer which physical laws apply, such as Gravity, Conservation of Momentum, Fluid Mechanics, Thermodynamics, Magnetic Force, etc. For example, under the action of gravity, the ball will accelerate and rebound after colliding with the ground. When two balls move towards each other, the material, volume, density, etc. of each ball must be considered, and the collision process satisfies the law of conservation of momentum. Assume objects move and interact based on real-world physics. The camera follows perspective geometry. Even if an object disappears and reappears, its bounding box should maintain the same ID across frames. If necessary, you may make reasonable assumptions. Before generating bounding box coordinates for each frame, provide step-by-step reasoning.

Context Information

 **User:** Caption: Generate a high-quality video of a wooden ball dropping from a height of about 1 meter onto a smooth, solid table. Capture the physics of the ball's descent, including its acceleration due to gravity, the moment of impact with the table, and the subsequent bounce. Show the ball compressing slightly upon impact, then rebounding, losing a bit of height with each bounce until it comes to rest. Include subtle details like the sound of the impact, the slight vibration of the table, and the ball's rotation as it moves.

initial_boxes: [{`id`: 0, `name`: 'ball', `box`: [303.0, 82.2, 108, 108]}]

 **Assistant:** Reasoning:

Step1-Analysis Caption and determine the physical laws to be satisfied: As can be seen from the caption, the topic of this study is about the free fall of a small ball, so this topic needs to satisfy Newton's laws of motion, focusing on the impact of gravity.

Step2-Analyze the impact of this physical law: In accordance with Newton's laws of motion, gravity will cause the small ball to perform free fall motion, vertically accelerate and hit the table, and rebound after the collision. Due to energy loss, the height of each rebound will decrease, which is in line with the law of conservation of energy and momentum.

Step3-Analyze the impact on box coordinates: The horizontal motion of the ball is negligible, so the x coordinate remains basically unchanged. Due to the vertical acceleration, the y coordinate of the small ball will gradually increase and the interpolation of the y coordinate between each two frames will also gradually increase. After colliding with the table, the small ball rebounds, the y coordinate decreases, and the small ball decelerates vertically upward. Subsequently, the small ball continues to accelerate vertically and collides with the table for the second time. After the third collision, the small ball is stationary on the table, and the y coordinate remains unchanged. Since the ball is a rigid body, it will be slightly compressed at the moment of collision with the desktop. Therefore, the width in the box coordinates remains unchanged, the height decreases slightly, and then quickly returns to normal. The width and height will not change for the rest of the time.

Predict the coordinates for the next 13 frames:

```
Frame 1: [{"id": 0, "name": "wooden ball", "box": [303, 82, 108, 108]}]
Frame 2: [{"id": 0, "name": "wooden ball", "box": [303, 137, 108, 108]}]
Frame 3: [{"id": 0, "name": "wooden ball", "box": [303, 184, 108, 108]}]
Frame 4: [{"id": 0, "name": "wooden ball", "box": [303, 231, 108, 108]}]
Frame 5: [{"id": 0, "name": "wooden ball", "box": [303, 255, 108, 108]}]
Frame 6: [{"id": 0, "name": "wooden ball", "box": [303, 284, 108, 104]}]
Frame 7: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 108]}]
Frame 8: [{"id": 0, "name": "wooden ball", "box": [303, 298, 108, 108]}]
Frame 9: [{"id": 0, "name": "wooden ball", "box": [303, 305, 108, 108]}]
Frame 10: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 104]}]
Frame 11: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 108]}]
Frame 12: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 108]}]
Frame 13: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 104]}]
```

Figure 6. Physics-Aware Reasoning Template for Rigid Body Motion

物理感知推理模板（重力）

系统指令

您是现实世界物理运动的专家。您的任务是预测视频边界框。您将收到一张参考图像、其分割图、运动开始的初始边界框以及视频内容的文字描述。您的任务是预测接下来 13 帧的边界框坐标。预测时，确保不同的边界框不重叠或超出帧边界。

每个帧应表示为：`[{'id':唯一对象标识符, 'name':对象名称, 'box':[框左上角 x 坐标, 框左上角 y 坐标, 框宽度, 框高度]}, ...]`。你的预测帧必须覆盖整个描述的场景。运动中的关键时刻必须包含在预测序列中，例如碰撞的瞬间。你需要考虑视频中运动对象的材料和属性。你只需要关注第一帧中提供的对象，无需关注其他任何事物。你的预测应推断适用的物理定律，例如重力、动量守恒、流体力学、热力学、磁力等。例如，在重力的作用下，球体撞击地面后会加速并反弹。当两个球体相互靠近时，必须考虑每个球体的材料、体积、密度等，碰撞过程满足动量守恒定律。假设对象根据现实世界物理学进行运动和相互作用。摄像机遵循透视几何。即使物体消失后重新出现，其边界框应该在所有帧中保持相同的 ID。如有必要，你可以为每一帧生成合理的assumptions.Before边界框坐标，并提供逐步推理过程。

上下文信息

标题: 生成一个高质量的木球从约 1 米的高度落到一个光滑、坚固的桌子上的视频。捕捉球的下降过程中的物理学，包括由于重力产生的加速度、球与桌子碰撞的瞬间以及随后的弹跳。展示球在撞击时轻微压缩，然后反弹，每次弹跳时高度略有下降，直到静止。包括细微的细节，如撞击声、桌子轻微的振动以及球运动时的旋转。初始框：[{'id': 0, 'name': 'ball', 'box': [303.0, 82.2, 108, 108]}]

步骤 1-分析标题并确定需要满足的物理定律：从标题可以看出，本研究的主题是关于小球的自由落体，因此这个主题需要满足牛顿运动定律，重点关注重力的作用。

步骤 2-分析该物理定律的影响：根据牛顿运动定律，重力会使小球进行自由落体运动，垂直加速并撞击桌面，然后在碰撞后反弹。由于能量损失，每次反弹的高度都会降低，这符合能量守恒和动量守恒定律。

步骤 3-分析对盒子坐标的影响：球的水平运动可以忽略不计，所以 x 坐标基本保持不变。由于垂直加速度，小球 y 坐标会逐渐增加，并且每两帧之间 y 坐标的插值也会逐渐增大。小球撞击桌面后反弹，y 坐标减小，并且小球垂直向上减速。随后，小球继续垂直加速，并第二次撞击桌面。第三次撞击后，小球静止在桌面上，y 坐标保持不变。由于球是刚体，撞击桌面时会发生轻微压缩。因此，在盒子坐标中，宽度保持不变，高度略微减小，然后迅速恢复正常。其余时间内，宽度和高度都不会变化。

预测接下来 13 帧的坐标：

```
帧数 1: [{'id': 0, 'name': 'wooden ball', 'box': [303, 82, 108, 108]}] 帧数 2: [{'id': 0, 'name': 'wooden ball', 'box': [303, 137, 108, 108]}] 帧数 3: [{'id': 0, 'name': 'wooden ball', 'box': [303, 184, 108, 108]}]

帧数 4: [{'id': 0, 'name': 'wooden ball', 'box': [303, 313, 108, 108]}] 帧数 5: [{'id': 0, 'name': 'wooden ball', 'box': [303, 255, 108, 108]}] 帧数 6: [{'id': 0, 'name': 'wooden ball', 'box': [303, 284, 108, 108]}] 帧数 7: [{'id': 0, 'name': 'wooden ball', 'box': [303, 313, 108, 108]}] 帧数 8: [{'id': 0, 'name': 'wooden ball', 'box': [303, 298, 108, 108]}] 帧数 9: [{'id': 0, 'name': 'wooden ball', 'box': [303, 305, 108, 108]}] 帧数 10: [{'id': 0, 'name': 'wooden ball', 'box': [303, 313, 108, 104]}] 帧数 11: [{'id': 0, 'name': 'wooden ball', 'box': [303, 313, 108, 108]}]

帧数 12: [{'id': 0, 'name': 'wooden ball', 'box': [303, 313, 108, 108]}] 帧数 13: [{'id': 0, 'name': 'wooden ball', 'box': [303, 313, 108, 104]}]
```

图 6. 物理感知推理模板用于刚体运动

Physics-Aware Recognition Template (Fluid)

System Instruction

You are an expert in real-world physical motion. Your task is to predict video bounding boxes. You will receive a reference image, its segmentation map, the initial bounding box where motion starts, and a textual description of the video content. Your task is to predict the bounding box coordinates for the next 13 frames. When predicting, ensure that different bounding boxes do not overlap or exceed the frame boundaries.

Each frame should be represented as: `[{'id': unique object identifier, 'name': object name, 'box': [box top-left x-coordinate, box top-left y-coordinate, box width, box height]}, ...]'. Your predicted frames must cover the entire described scene. Critical moments in the movement must be included in the predicted sequence, such as the moment of collision. You need to consider the material and properties of the moving objects in the video. Your predictions should infer which physical laws apply, such as Gravity, Conservation of Momentum, Fluid Mechanics, Thermodynamics, Magnetic Force, etc. For example, under the influence of fluid mechanics, when pouring water into a cup, the water level in the cup should rise and ripples should occur. The top boundary of a liquid should move when it's being poured, while the bottom edge of the liquid (for instance, in a container) should stay fixed. When an object (such as a liquid) is in a container and is being poured, the bottom boundary of the liquid typically stays fixed. Assume objects move and interact based on real-world physics. The camera follows perspective geometry. Even if an object disappears and reappears, its bounding box should maintain the same ID across frames. If necessary, you may make reasonable assumptions. Before generating bounding box coordinates for each frame, provide step-by-step reasoning.

Context Information

 **User:** Caption: The video shows a close-up of a clear glass being filled with tea. The tea is being poured from above, and we can see the stream of tea hitting the bottom of the glass and causing ripples and splashes. The background is a plain white surface, which contrasts with the transparency of the glass and the clarity of the tea. The glass is cylindrical in shape and appears to be of a standard size for a drinking glass.

initial_boxes: [{'id': 0, 'name': 'tea', 'box': [241.8, 335.3, 165, 88]}]

 **Assistant:** Reasoning:

Step1-Analysis Caption and determine the physical laws to be satisfied: As can be seen from the caption, the theme of this session is to pour tea into a glass teacup. Therefore, the tea is affected by gravity during its fall, and the state of the tea in the cup satisfies fluid mechanics.

Step2-Analyze the impact of this physical law: Tea falls into the cup under the influence of gravity. According to fluid mechanics, the water level in the cup will gradually rise, and the water surface will ripple due to the pouring of water.

Step3-Analyze the impact on box coordinates: When tea is poured into the cup, under the restriction of the cup, the x coordinate of the tea box will not change much. As the water level rises, the y coordinate will gradually decrease. As the water level rises, the overall volume of the tea gradually increases, the height of the box will also increase, and the width will change with the boundary of the cup.

Predict the coordinates for the next 13 frames:

```
Frame 1: [ {'id': 0, 'name': 'tea', 'box': [241.8, 335.3, 165, 88]} ]
Frame 2: [ {'id': 0, 'name': 'tea', 'box': [241.2, 330.0, 165, 94]} ]
Frame 3: [ {'id': 0, 'name': 'tea', 'box': [241.0, 322.2, 165, 102]} ]
Frame 4: [ {'id': 0, 'name': 'tea', 'box': [241.3, 317.6, 165, 105]} ]
Frame 5: [ {'id': 0, 'name': 'tea', 'box': [240.6, 313.7, 166, 111]} ]
Frame 6: [ {'id': 0, 'name': 'tea', 'box': [239.1, 305.6, 167, 118]} ]
Frame 7: [ {'id': 0, 'name': 'tea', 'box': [237.7, 301.1, 170, 123]} ]
Frame 8: [ {'id': 0, 'name': 'tea', 'box': [236.8, 295.3, 172, 129]} ]
Frame 9: [ {'id': 0, 'name': 'tea', 'box': [235.7, 290.2, 172, 134]} ]
Frame 10: [ {'id': 0, 'name': 'tea', 'box': [235.1, 283.3, 173, 140]} ]
Frame 11: [ {'id': 0, 'name': 'tea', 'box': [234.0, 278.3, 175, 145]} ]
Frame 12: [ {'id': 0, 'name': 'tea', 'box': [233.1, 272.0, 177, 152]} ]
Frame 13: [ {'id': 0, 'name': 'tea', 'box': [231.7, 268.0, 179, 156]} ]
```

Figure 7. Physics-Aware Reasoning Template for fluid dynamics and thermodynamics.

物理感知识别模板（流体）

系统指令

您是现实世界物理运动的专家。您的任务是预测视频边界框。您将收到一张参考图像、其分割图、运动开始的初始边界框以及视频内容的文字描述。您的任务是预测接下来 13 帧的边界框坐标。预测时，确保不同的边界框不重叠或超出帧边界。

每个帧应表示为：`[{'id':唯一对象标识符, 'name':对象名称, 'box':[框左上角 x 坐标, 框左上角 y 坐标, 框宽度, 框高度]}, ...]`。你的预测帧必须覆盖整个描述的场景。运动中的关键时刻必须包含在预测序列中，例如碰撞的瞬间。你需要考虑视频中运动对象的材料和属性。你的预测应推断适用的物理定律，例如重力、动量守恒、流体力学、热力学、磁力等。例如，在流体力学的影响下，当将水倒入杯中时，杯中的水位应上升并产生涟漪。当液体被倒入时，液体的上边界应移动，而液体底部边缘（例如在容器中）应保持固定。当对象（例如液体）在容器中并被倒入时，液体的底部边界通常保持固定。假设对象根据现实世界物理学移动和相互作用。摄像机遵循透视几何。即使物体消失后重新出现，其边界框应该在所有帧中保持相同的 ID。如有必要，你可以为每一帧生成合理的assumptions.Before边界框坐标，并提供逐步推理过程。

上下文信息

 **视频** 展示了一个透明的玻璃杯被茶水注满的特写镜头。茶水从上方倾倒，我们可以看到茶水流击打杯底，引起涟漪和飞溅。背景是一个纯白色的平面，与玻璃的透明性和茶水的清澈形成对比。这个玻璃杯呈圆柱形，看起来是标准饮用杯的大小。

initial_boxes: [{"id": 0, "name": "tea", "box": [241.8, 335.3, 165, 88]}]

 **步骤 1-分析标题并确定需要满足的物理定律：**从标题可以看出，本次课的主题是向玻璃茶杯倒茶。因此，茶水在落下时受到重力的影响，杯中茶水的状态满足流体力学。

步骤 2-分析该物理定律的影响：茶水在重力影响下落入杯中。根据流体力学，杯中水位会逐渐上升，水面因倒水而泛起涟漪。

步骤 3-分析茶盒坐标的影响：当茶被倒入杯中时，在杯子的限制下，茶盒的 x 坐标不会变化很大。随着水位上升，y 坐标会逐渐减小。随着水位上升，茶的整体体积逐渐增加，盒子的高度也会增加，宽度会随着杯子的边界变化。

预测接下来 13 帧的坐标：

```
帧 1: [{"id": 0, "name": "tea", "box": [241.8, 335.3, 165, 88]}] 帧 2: [{"id": 0, "name": "tea", "box": [241.2, 330.0, 165, 94]}] 帧 3: [{"id": 0, "name": "tea", "box": [241.0, 322.2, 165, 102]}] 帧 4: [{"id": 0, "name": "tea", "box": [241.3, 317.6, 165, 105]}] 帧 5: [{"id": 0, "name": "tea", "box": [240.6, 313.7, 166, 111]}] 帧 6: [{"id": 0, "name": "tea", "box": [239.1, 305.6, 167, 118]}] 帧 7: [{"id": 0, "name": "tea", "box": [237.7, 301.1, 170, 123]}] 帧 8: [{"id": 0, "name": "tea", "box": [236.8, 295.3, 172, 129]}] 帧 9: [{"id": 0, "name": "tea", "box": [235.7, 290.2, 172, 134]}] 帧 10: [{"id": 0, "name": "tea", "box": [235.1, 283.3, 173, 140]}] 帧 11: [{"id": 0, "name": "tea", "box": [234.0, 278.3, 175, 145]}] 帧 12: [{"id": 0, "name": "tea", "box": [233.1, 272.0, 177, 152]}] 帧 13: [{"id": 0, "name": "tea", "box": [231.7, 268.0, 179, 156]}]
```

图 7. 物理感知推理模板，用于流体动力学和热力学。



Figure 8. More examples of generated videos related to fluid dynamics and thermodynamics.

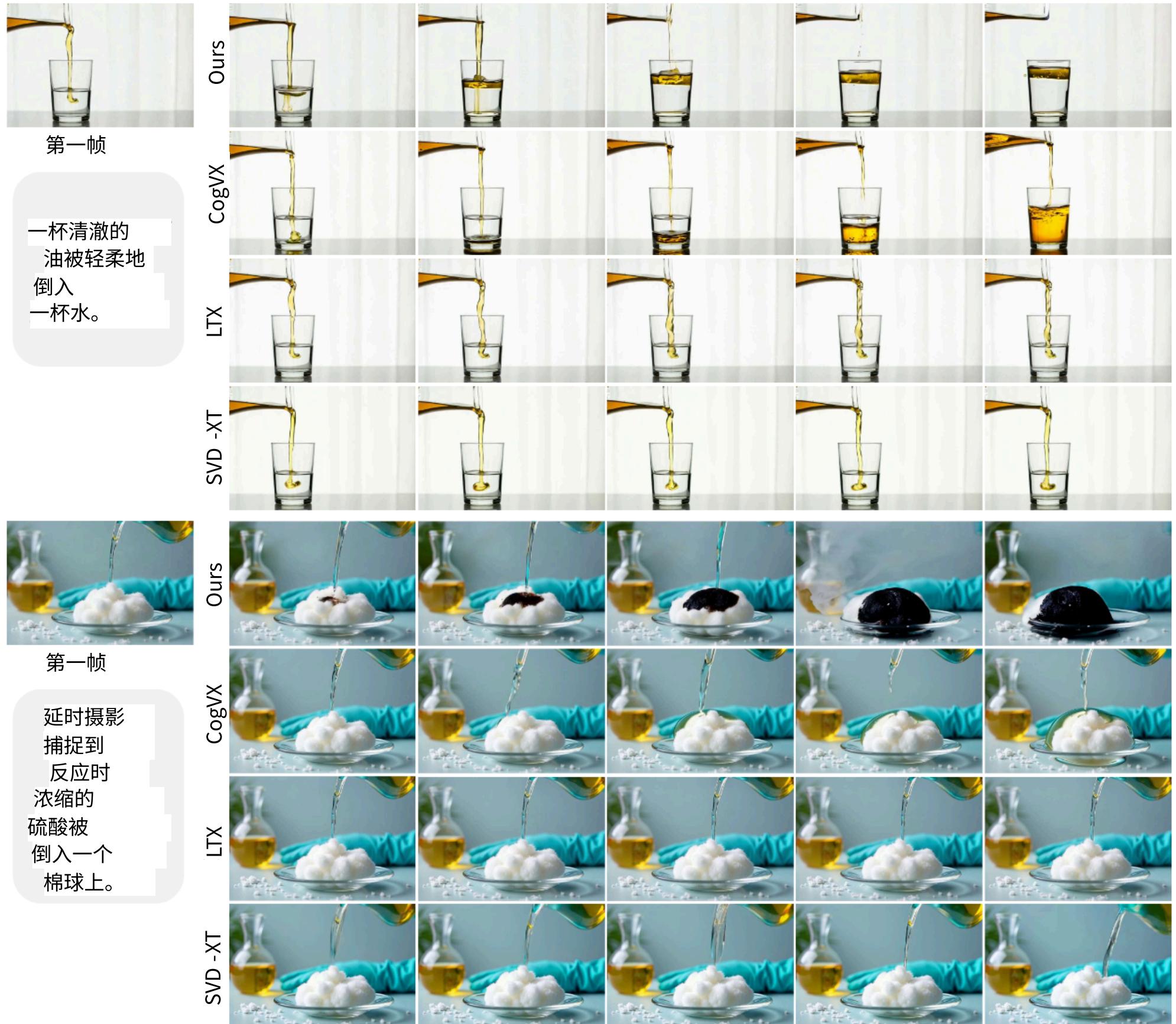


图 8. 更多与流体动力学和热力学相关的生成视频示例。

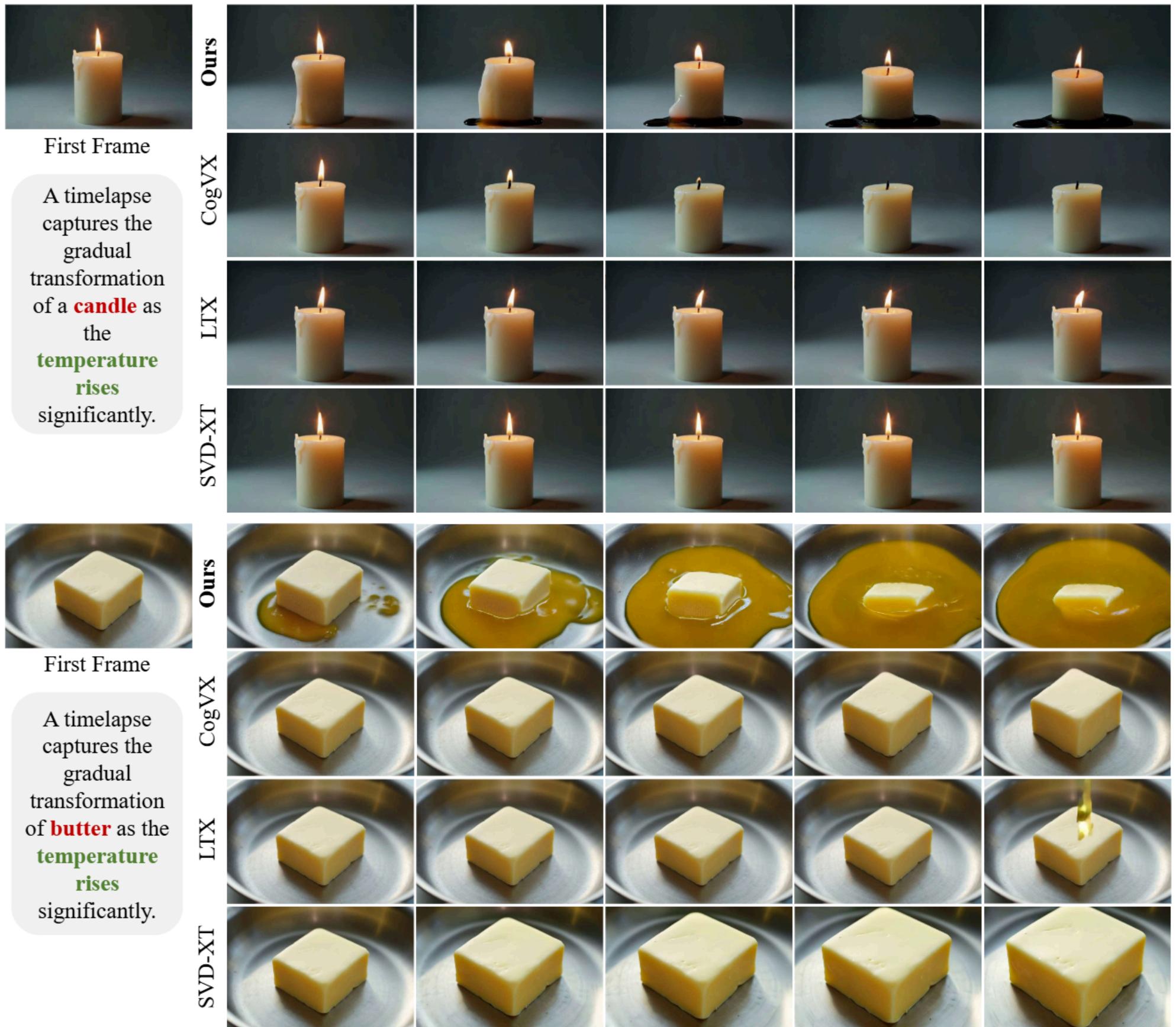


Figure 9. More examples of generated videos related to thermodynamics.



图 9。更多与热力学相关的生成视频示例。

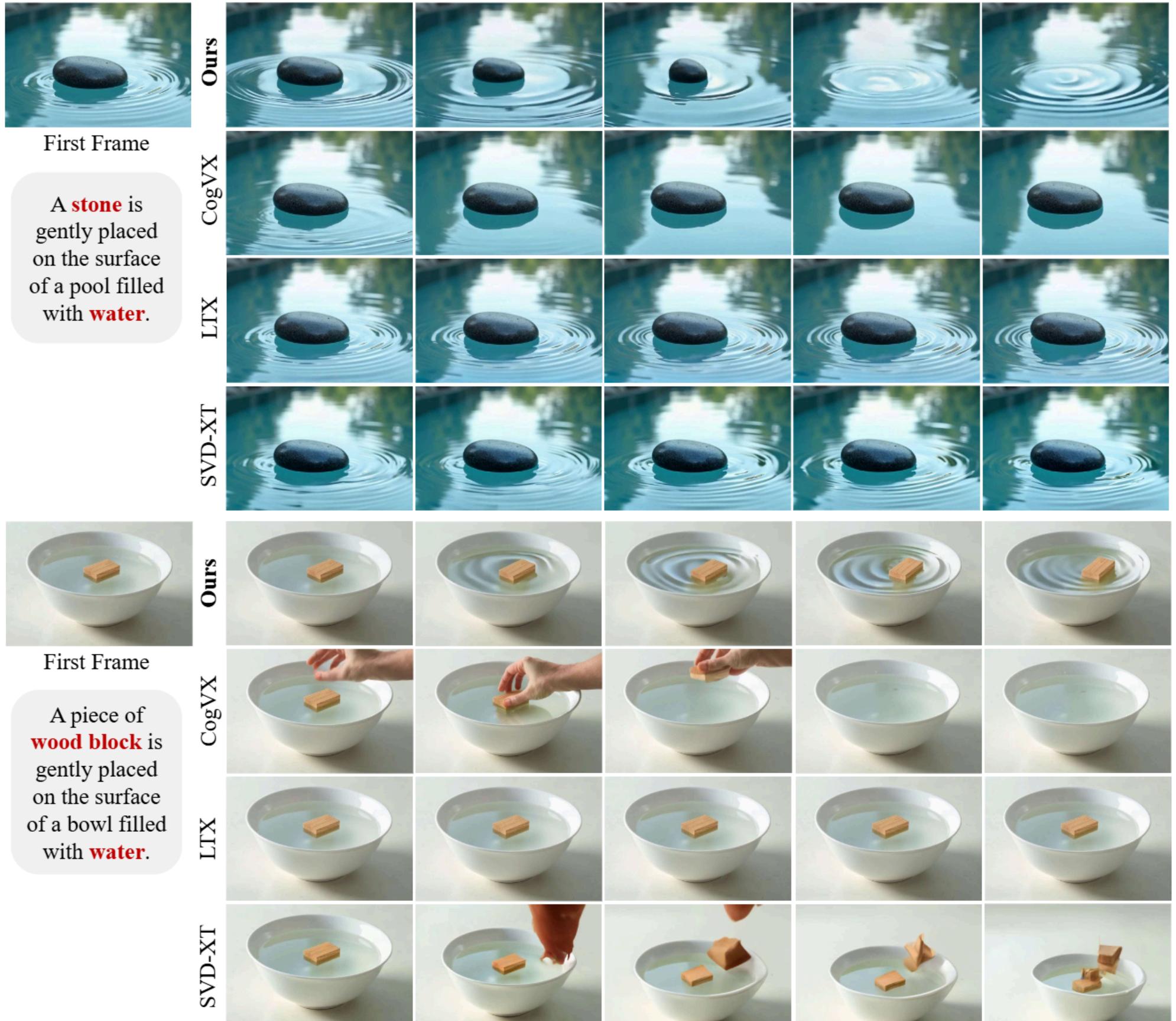


Figure 10. More examples of generated videos related to fluid dynamics.

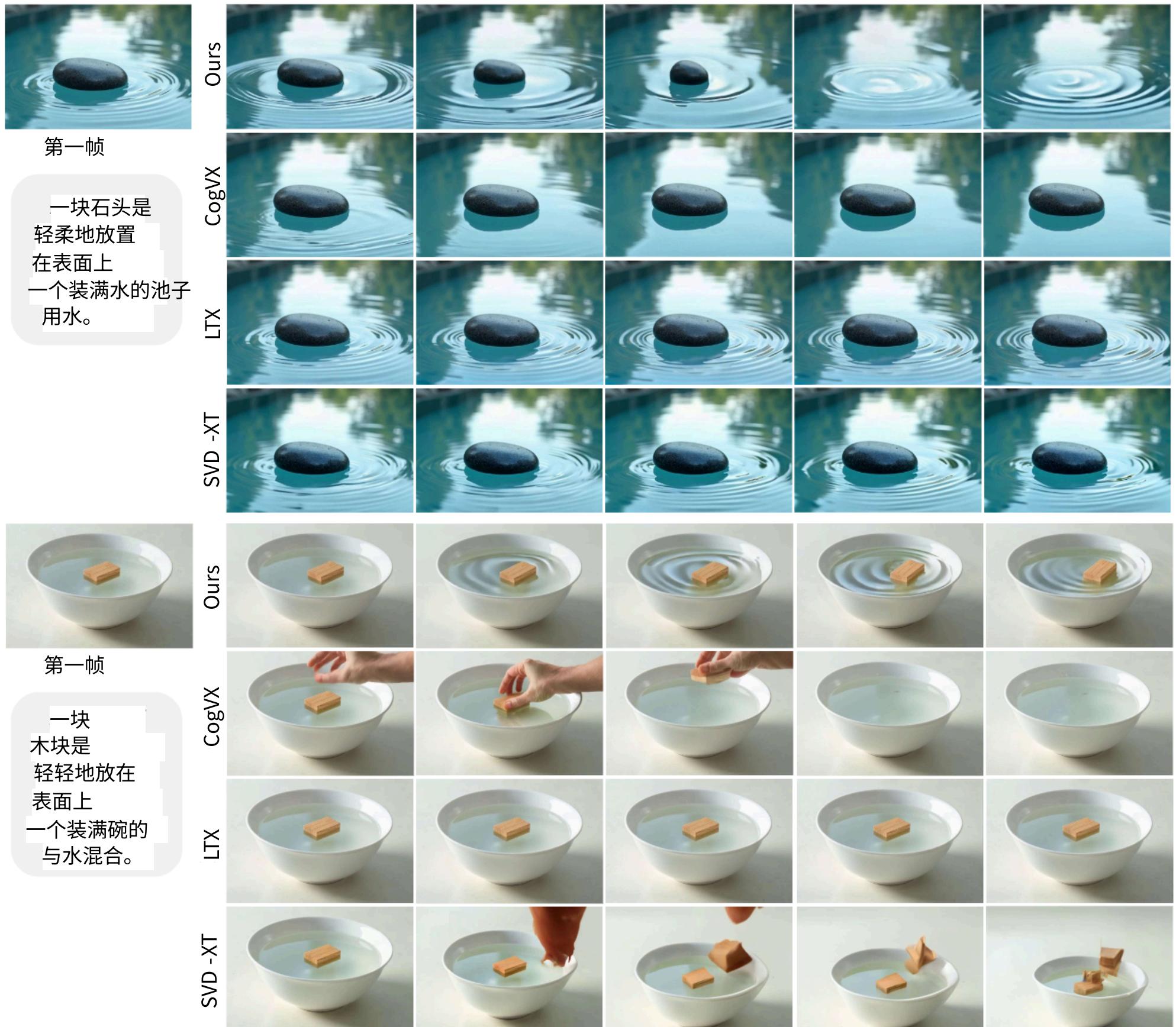


图 10。更多与流体动力学相关的生成视频示例。

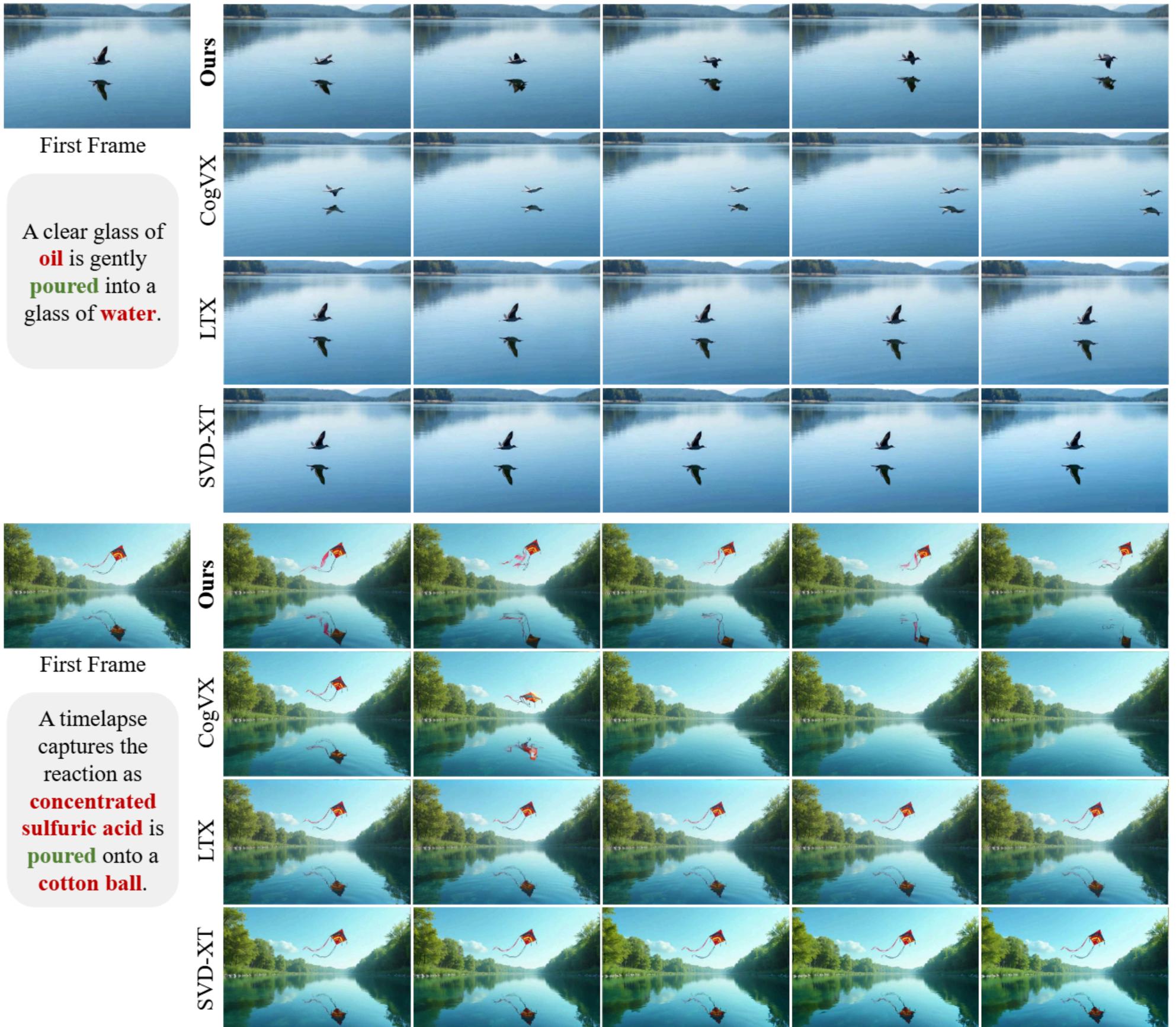


Figure 11. More examples of generated videos related to optics.

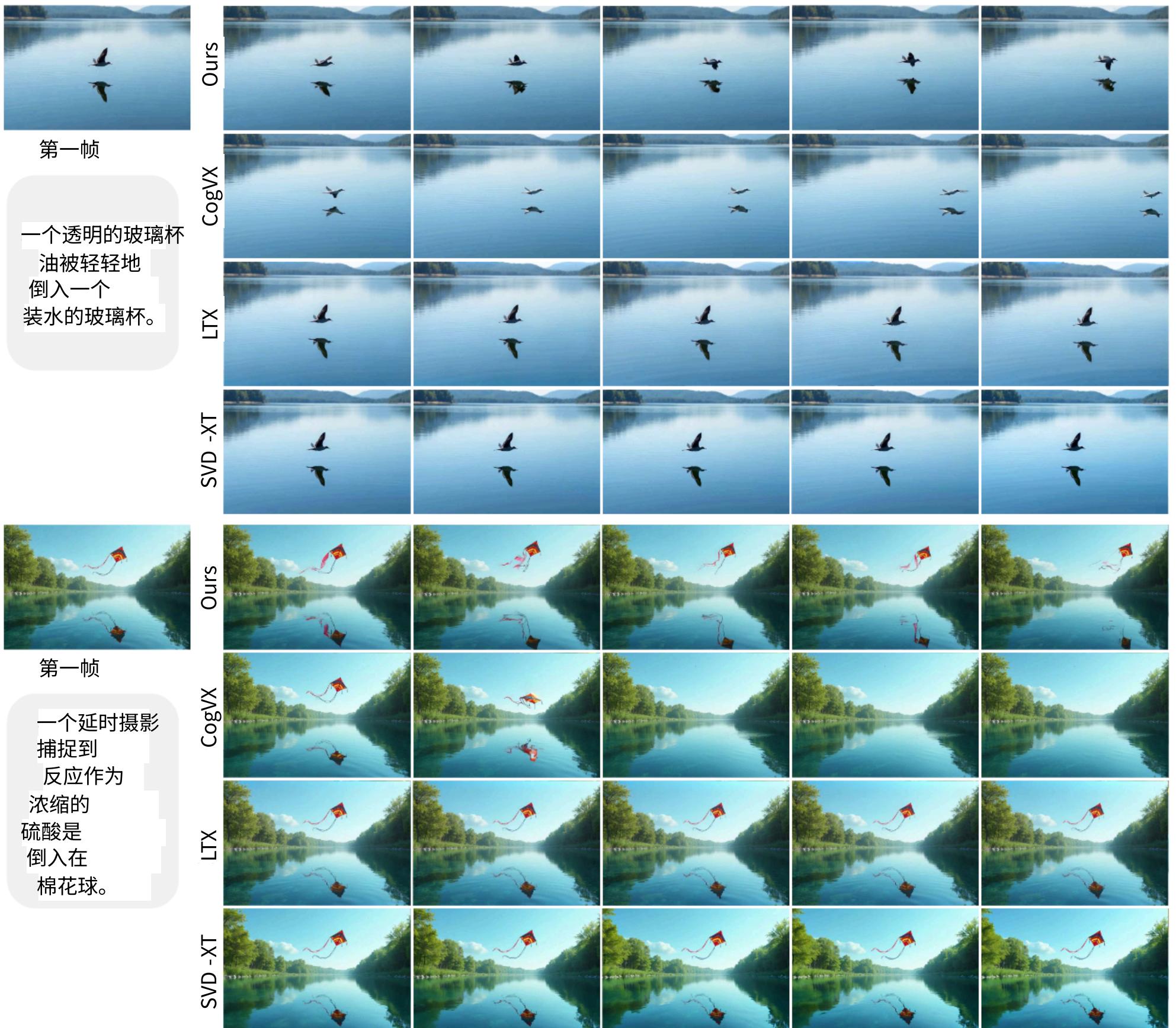


图 11。更多与光学相关的生成视频示例。