

# Exploring the Evolution of Physics Cognition in Video Generation: A Survey

Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang,  
Zhengrong Zuo, Nong Sang, Siteng Huang, and Donglin Wang

**Huazhong University of Science and Technology, Westlake University, Shandong University,  
Tsinghua University, Zhejiang University**

**Abstract**—Recent advancements in video generation have witnessed significant progress, especially with the rapid advancement of diffusion models. Despite this, their deficiencies in physical cognition have gradually received widespread attention - generated content often violates the fundamental laws of physics, falling into the dilemma of “visual realism but physical absurdity”. Researchers began to increasingly recognize the importance of physical fidelity in video generation and attempted to integrate heuristic physical cognition such as motion representations and physical knowledge into generative systems to simulate real-world dynamic scenarios. Considering the lack of a systematic overview in this field, this survey aims to provide a comprehensive summary of architecture designs and their applications to fill this gap. Specifically, we discuss and organize the evolutionary process of physical cognition in video generation from a cognitive science perspective, while proposing a three-tier taxonomy: 1) basic schema perception for generation, 2) passive cognition of physical knowledge for generation, and 3) active cognition for world simulation, encompassing state-of-the-art methods, classical paradigms, and benchmarks. Subsequently, we emphasize the inherent key challenges in this domain and delineate potential pathways for future research, contributing to advancing the frontiers of discussion in both academia and industry. Through structured review and interdisciplinary analysis, this survey aims to provide directional guidance for developing interpretable, controllable, and physically consistent video generation paradigms, thereby propelling generative models from the stage of “visual mimicry” towards a new phase of “human-like physical comprehension”. A comprehensive list of papers studied in this survey is available at [here](#).

**Index Terms**—Video Generation, Physics Cognition, World Models.

## 1 INTRODUCTION

### 1.1 Overview

RECENT years have witnessed groundbreaking advancements in video generation tasks [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47]. These generative video models, typically trained on vast amounts of real-world video data, demonstrate remarkable capabilities in producing temporally and spatially coherent video sequences based on multimodal conditional signals (e.g., text [48], [49], [50], [51], [52], [53], images [54], [55], [56], [57], [58], or videos [59], [60], [61], [62]). These existing techniques such as Sora [1], Kling [63], and HunyuanVideo [64] have demonstrated realistic visual quality, temporal continuity, and powerful capability of prompt following, and have

also achieved great success in many downstream tasks, including video customization [30], [32], [62], [65], video editing [37], [66], [67], [68], and video super-resolution [69], [70], etc. More importantly, video generation is increasingly being applied to domains such as gaming [2], [71], [72], robotics [73], [74], autonomous driving [75], [76], [77], and scientific research [78] through techniques like instruction tuning [79], contextual learning [80], planning [81], and reinforcement learning (RL) [82], playing a crucial role in the development of Artificial General Intelligence (AGI). As noted by Yang et al. [83], video generation models, much like language models, are progressively evolving into autonomous agents, planners, environment simulators, and computational engines. Ultimately, video generation models have the potential to serve as artificial brains capable of reasoning and acting within the physical world.

Despite remarkable success, studies [84], [85], [86] have shown that these models often exhibit significant deficiencies in physical cognition when dealing with complex dynamic scenes. For instance, generated results frequently violate fundamental physical laws (e.g., Newtonian dynamics, momentum conservation, and energy conservation) in scenarios involving rigid body collisions, fluid dynamics, or elastic deformations, resulting in “visually realistic yet physically absurd” content, see in Fig. 1. These contradictions underscore the bottlenecks in video generation models’ capacity for physical cognition modeling, which may have significant negative impacts on AI applications such as robotics and autonomous driving.

- M. Lin is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. Email: [minghui.lin@hust.edu.cn](mailto:minghui.lin@hust.edu.cn)
- X. Wang, Z. Zuo, and S. Nong are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.
- Y. Wang, F. Dai, P. Ding, D. Wang are with the School of Engineering, Westlake University, Hangzhou, China.
- S. Wang is with the School of Control Science and Engineering, Shandong University, Jinan, China.
- C. Wang is with Tsinghua University, Beijing, China.
- S. Huang is with Zhejiang University, Hangzhou, China. Email: [siteng.huang@gmail.com](mailto:siteng.huang@gmail.com)
- Corresponding Author: Siteng Huang and Donglin Wang

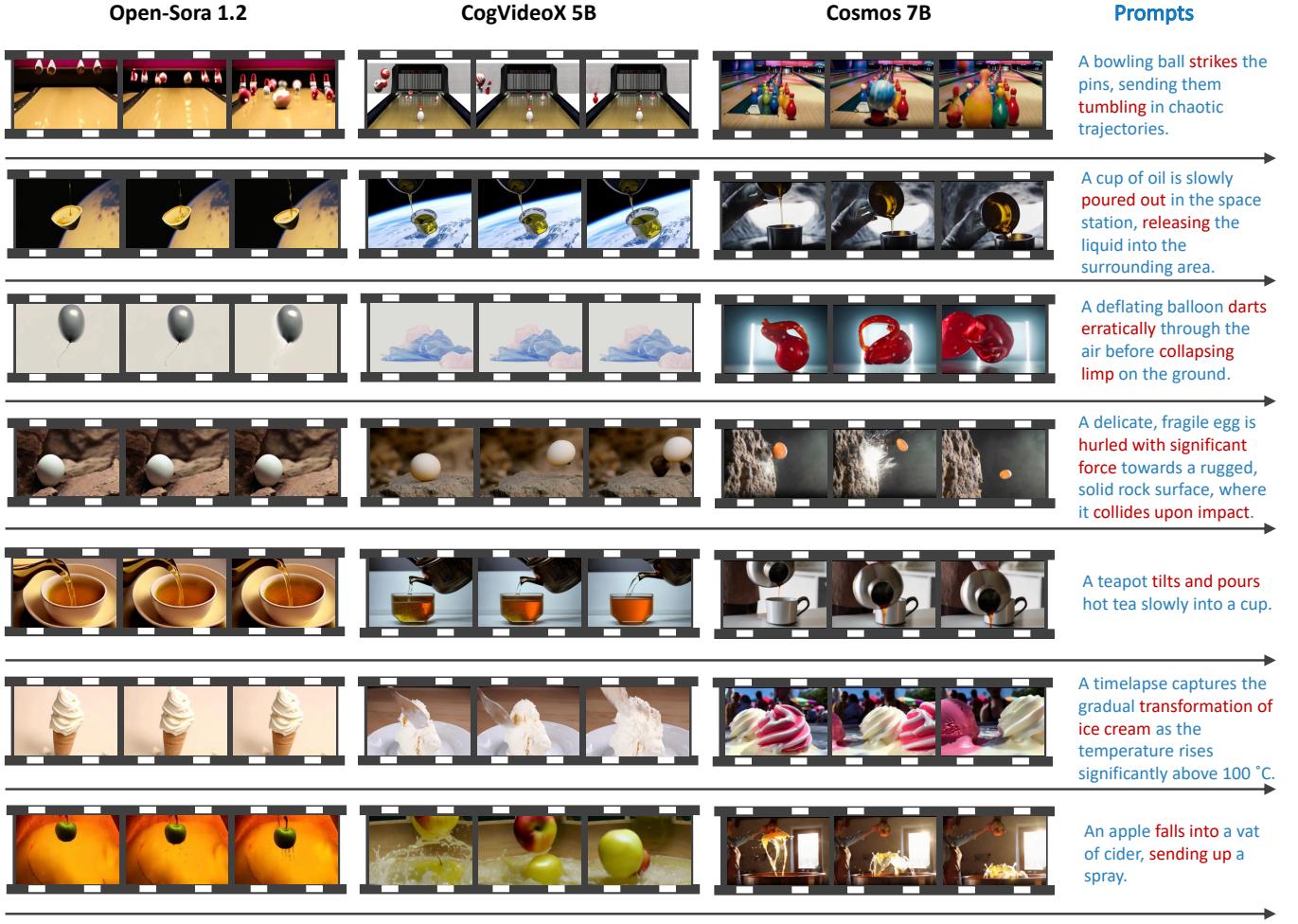


Fig. 1. Video cases generated by three typical state-of-the-art generative video models [1], [48], [87]. We can observe that these advanced models still struggle to produce satisfying videos that strictly conform to physical laws.

Therefore, research on physical cognition in video generation has begun to receive widespread attention in both academia and industry [88], [89], [90]. Recent advancements involve the systematic integration of diverse forms of physical information into generative architectures, such as motion-driven video generation and the integration of physical simulators with 3D representation-based rendering for interactive dynamics. With the rapid progress in physical cognition in video generation, efforts to track and compare the latest research on this topic have become extremely important and meaningful. However, existing surveys on this topic are limited in the general AI-Generated Content (AIGC) field [91], [92] or paying less attention on video generation [93]. To this end, this survey aims to fill this gap and sort out the precise and comprehensive development of physical cognition in video generation for readers.

To enhance the interpretability of physical cognition in video generation and strengthen its human-like capability as an artificial brain for reasoning and acting in the physical world [83], we draw inspiration from human physical cognitive mechanisms to systematically categorize physical cognition in video generation. Through this approach, we aim to provide cognition-driven solution guidance for addressing the persistent “physical embedding bottleneck” in

video generation.

Overall, this survey aims to establish an evolutionary pathway for modeling physical cognition in video generation from a cognitive science perspective. By providing a comprehensive and structured review of existing methods, we seek to offer guidance for developing explainable, controllable, predictable, and physically consistent video generation paradigms.

## 1.2 Taxonomy

The evolution of cognitive systems in human development exhibits distinct stages, forming the fundamental mechanism through which individuals understand and explore the physical world. Based on Piaget’s theory of cognitive development [94], we adapt and refine the characterization of an individual’s cognition of the physical world, proposing that it evolves in a spiral manner through three stages: “**Intuitive perception-Symbolic learning-Interaction**” (Fig. 2). In the initial stage (exemplified by infancy), individuals develop an intuitive sense of physical reality (e.g., object permanence) through primitive sensory schemas; however, this perception remains chaotic despite the omnipresence of physical principles. As cognitive development progresses to

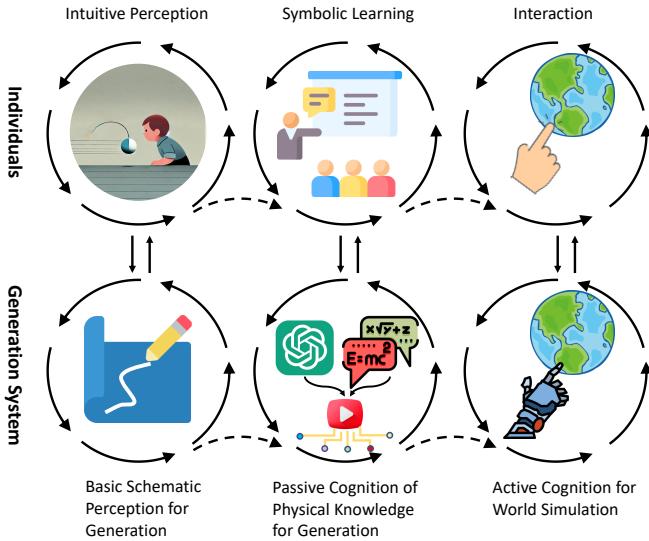


Fig. 2. Cognitive evolution processes of individuals and generation system.

the next stage, individuals begin to acquire physical knowledge passively through observation and symbolic learning (e.g., witnessing falling apples or memorizing Newtonian laws). At a more advanced stage of cognition, humans develop the ability to actively reason about and predict physical phenomena, continuously refining their cognitive models through active interaction with the environment. Contemporary video generation systems exhibit profound mapping with this evolutionary trajectory of human physics cognition. Video generation methods encompass different approaches, including fundamental schematic perception, passive learning of physical knowledge, and active interaction with the environment.

This survey establishes an evolutionary framework for physical cognition modeling in generation systems through the lens of individual cognitive development, as shown in Fig. 5. We systematically categorize state-of-the-art research into three pivotal areas: **Basic Schematic Perception for Generation**: Relies on unidirectional stimulation of low-fidelity visual patterns, leading to intuitive responses (e.g., object localization without contextual awareness); **Passive Cognition of Physical Knowledge for Generation**: Grounding generation through pre-stored static physical knowledge from physics simulators or Large Language Models (LLMs); **Active Cognition for World Simulation**: Emphasizing active interaction with the environment to achieve more physically faithful future predictions. Specifically, this survey provides a systematic analysis of the following aspects:

- **Basic Schematic Perception for Generation (Sec. 4):** We discuss how video and motion-based generative models integrate fundamental motion patterns to enhance motion consistency in dynamic scenes. Additionally, we explore relighting techniques and zero-shot self-guided generation approaches;
- **Passive Cognition of Physical Knowledge for Generation (Sec. 5):** We systematically review various mechanisms for embedding physical knowledge into generative models, demonstrating how such cognitive grounding improves physical interpretability

and physical consistency in generated content;

- **Active Cognition for World Simulation (Sec. 6):** We investigate generative models that predict the future through active interaction with the environment, illustrating how this approach effectively bridges the gap between video generators and real-world physical dynamics.

Finally, we discuss existing physical evaluation benchmarks and highlight unresolved challenges, such as constructing large foundational physics models, improving the physical fidelity of world simulators, incorporating multi-sensor data, enhancing the efficiency of physical simulation, addressing data scarcity and the Sim2Real gap, advancing physical quality assessment, and other related issues.

### 1.3 Structure

The overall category structure of existing works discussed in this survey is illustrated in Fig. 5, and is organized as follows: In Sec. 1, we introduce the importance of physical fidelity in video generation and provide an overview of the classification criteria. In Sec. 3, we introduce fundamental background knowledge that encompasses physical commonsense, mainstream generative models, and physics simulators, laying the groundwork for subsequent discussions. From Sec. 4 to Sec. 6, we detail the evolutionary progression of physical cognition in video generation: In Sec. 4, we focus on Basic Schematic Perception for Generation, discussing open-loop video generation methods based on fundamental representation signals such as video-based and motion-based generation; In Sec. 5, we explore Passive Cognition-Based Generation, emphasizing approaches that incorporate symbolic knowledge embeddings; In Sec. 6, we investigate diverse environment-interactive mechanisms, including multi-modal data-driven methods, spatial awareness, and external feedback mechanisms. In Sec. 7, we outline existing benchmarks and evaluation metrics used for assessing the physical plausibility of generated videos. In Sec. 8, we discuss current challenges and explore potential future directions in the field. Finally, in Sec. 9, we summarize the key findings and contributions of this survey.

## 2 SURVEY SCOPE AND COMPARISON

**Survey Scope.** This survey focuses on the investigation of video generation methods with physical fidelity, including the generation of 2D videos, dynamic 3D, and 4D. To emphasize the physical significance in video generation, we exclude methods of unconditional video generation [95] and long video generation [96] in the general sense. Additionally, pure visual transformation methods that do not involve any individual motion, physical priors, dynamic modeling, or real-world constraints, such as art style transfer [97] and video super-resolution [69], are also outside the scope of this survey. Our focus is on video generation techniques that adhere to physical laws, such as kinematics, dynamics, or optical properties, to ensure visually realistic and physically plausible results.

**Comparison with Other Surveys.** Unlike existing surveys on physics AI, which adopt the classification of “explicit simulation and implicit learning” [93] or focus solely

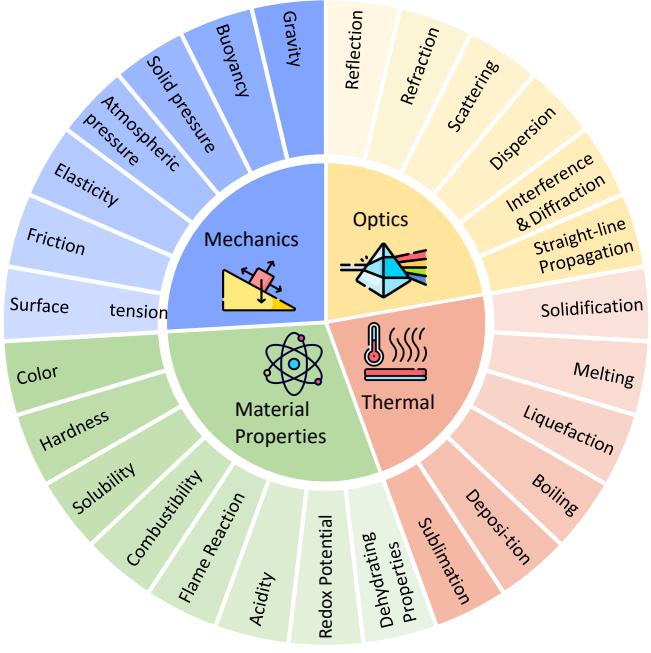


Fig. 3. The taxonomy of PhysGenBench [86] benchmark, including 4 physical commonsense and 27 physical laws.

on 3D/4D generation [91], this survey introduces an innovative classification paradigm inspired by the human cognitive perspective. By drawing on the evolutionary trajectory of human physical cognition, we systematically reconstruct the developmental pathway of physical knowledge in video generation techniques. This classification not only provides design guidance for physics-augmented generative models but also offers multi-pathway analysis for embedding cognitive science into AGI development [98], [99].

### 3 BACKGROUND

This section first introduces the classification of physical commonsense in daily life and provides illustrative examples (see Sec. 3.1). Subsequently, we provide a detailed introduction to several mainstream generative models in Sec. 3.2. Finally, in Sec. 3.3, we explore the physics simulators, physics engines, and related platforms discussed in this survey. These aspects lay the foundation for understanding physics cognition-based generation approaches.

#### 3.1 Physical Commonsense

Physical commonsense refers to the basic intuitive understanding of physical objects and behaviors encountered in daily life, while physical laws are universal scientific principles used to describe consistent behaviors in nature. Physical phenomena, on the other hand, are observable events or processes caused by the interaction of physical laws [86]. The function of a physically plausible generative model is to generate physically accurate phenomena based on cues that describe physical laws. The most universal physical commonsense in the world can be divided into four main domains: mechanics, optics, thermodynamics, and material properties. For example, in Fig. 3, the PhyGenBench [86]

benchmark covers these four key physical domains and includes 27 physical laws (e.g., gravity and reflection).

- **Mechanics:** The study of the motion of objects and the interactions of forces with objects. It includes branches such as statics, dynamics, and fluid mechanics. For example, the friction between a car's tires and the ground allows the car to start and stop.
- **Optics:** The study of light propagation, refraction, reflection, and interaction with objects. For example, mirrors utilize the principle of reflection to reflect light and form images.
- **Thermal:** The study of energy conversion, heat transfer, and temperature changes in materials. It focuses on how heat flows between objects and interacts with other forms of energy (such as mechanical or chemical energy). For example, when a pot is heated, heat is conducted through the metal to the upper part of the pot.
- **Material Properties:** Refers to the characteristics of a substance's behavior and reactions under different environmental conditions. For example, salt dissolves in water, while oil does not.

### 3.2 Generative Models

In this subsection, we introduce several mainstream 2D and 3D generative models, including GANs [100], Diffusion Models [101], NeRF [102], Gaussian Splatting [103], with a schematic illustration provided in Fig. 4.

#### 3.2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [100] consist of two integral components: the Generator and the Discriminator. The Generator  $G(z)$  takes random noise  $z$  as input and attempts to produce samples resembling the true data distribution. Conversely, the Discriminator  $D(x) \in [0, 1]$  evaluates a given sample  $x$  and outputs the probability of the sample being real. These components are trained in a minimax game manner, where the Generator aims to maximize the Discriminator's error, while the Discriminator strives to accurately differentiate between real and generated samples. The objective function, which quantifies their adversarial training process, can be expressed as:

$$\min_G \max_D \mathcal{L}(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

This objective encapsulates the dual aims of both components, facilitating the synthesis of high-quality data through adversarial learning.

#### 3.2.2 Diffusion Models

Inspired by non-equilibrium thermodynamics, diffusion model [101] gradually adds random noise to the data distribution by defining a Markov chain of diffusion steps, and then learns the inverse denoising process to generate new data. Its core idea is to gradually destroy the data and learn to recover, contrasting with the direct adversarial training of GANs.

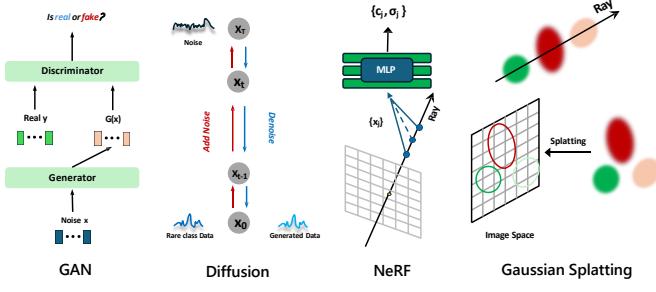


Fig. 4. Introduction to mainstream generative models: GANs [100], Diffusion Models [101], NeRF [102], Gaussian Splatting [103].

The forward process (diffusion process) incrementally adds Gaussian noise to the data  $x_0$  via a Markov chain until the data is transformed into pure noise  $x_T$ .

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right) \quad (2)$$

where,  $\beta_t$  represents the noise level at step  $t$ , controlling the rate of noise addition.

The reverse process (denoising process) trains a neural network  $\epsilon_\theta$  to predict the noise introduced at each step, thereby generating data through iterative denoising:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (3)$$

The mean  $\mu_\theta(x_t, t)$  and covariance  $\Sigma_\theta(x_t, t)$  are parameterized by a learnable neural network.

The loss function is defined as the mean squared error (MSE) for noise prediction:

$$\mathcal{L} = \mathbb{E}_{t, x_0, \epsilon} \left[ \|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right] \quad (4)$$

where  $\epsilon$  represents the noise added during the forward process, and  $\epsilon_\theta$  is the neural network's prediction of this noise.

### 3.2.3 Neural Radiance Fields

Neural Radiance Fields (NeRF) [102] is a deep learning-based approach for novel view synthesis. It leverages sparse multi-view input data to train a neural network that implicitly models the volumetric density and color information of a 3D scene, enabling the synthesis of high-quality novel views. Specifically, given a spatial position  $\mathbf{x} \in \mathbb{R}^3$  and a viewing direction  $\mathbf{d} \in \mathbb{R}^3$  (typically represented as a unit vector), NeRF employs a continuous mapping function  $F$  to approximate the scene's volumetric density  $\sigma$  and color  $\mathbf{c}$ . This function is parameterized by a neural network and is formally defined as:

$$F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma) \quad (5)$$

where  $\mathbf{x} = (x, y, z)$  represents the coordinates of a point in 3D space. The viewing direction is given by  $\mathbf{d} = (\theta, \phi)$ , which is typically expressed in spherical coordinates. The color at the point is denoted as  $\mathbf{c} = (r, g, b)$ , while  $\sigma$  represents the volumetric density, determining the extent to which light is absorbed at that point. The neural network parameters are denoted by  $\Theta$ . During the image rendering stage, NeRF employs the classical volume rendering technique, integrating color along camera rays to synthesize high-quality novel view images.

### 3.2.4 Gaussian Splatting

The core idea of Gaussian splatting [103] is to represent a scene as a set of anisotropic 3D Gaussian kernels  $G = \{G_p : (\mathbf{x}_p, \alpha_p, \mathbf{A}_p, \mathbf{c}_p)\}_{p \in \mathcal{P}}$ , where each kernel is defined by its center position  $\mathbf{x}_p$ , opacity  $\alpha_p$ , covariance matrices  $\mathbf{A}_p$ , and color function  $\mathbf{c}_p$ . During the rendering phase, each Gaussian kernel is projected onto the viewpoint (i.e., splatted) and weighted accumulation is performed based on opacity and depth. The final color of the  $i$ -th pixel in the synthesized image from a new viewpoint is computed as follows:

$$\mathbf{c}_i = \sum_k G_k(i) \alpha_k \mathbf{c}_k(\mathbf{r}_i) \prod_{j=1}^{k-1} (1 - G_j(\mathbf{i}) \alpha_j). \quad (6)$$

where,  $G_k(i)$  represents the projection weight of the  $k$ -th Gaussian kernel, and  $\mathbf{r}_i$  is the viewpoint direction of the camera.

For 4D generation [71], [132], [151], which aims to synthesize time-varying 3D scenes, simply incorporating the time variable into NeRF/3DGS [102], [103] allows for the reconstruction and generation of dynamically evolving scenarios. Models based on NeRF/3DGS can effectively capture and synthesize realistic 4D dynamics, making them powerful tools for generating time-varying 3D scenes, such as dynamic human motions, fluid simulations, and physics-driven environments.

## 3.3 Physics Simulation

In physics modeling, physics simulators serve as one of the critical components, translating physical laws into computable forms through numerical methods to enable precise simulation of object and environmental dynamics in complex scenarios. Meanwhile, to meet diverse application demands, simulation engines and platforms integrating multiple physics simulators have emerged. These engines and platforms unify various physics simulation functionalities through standardized interfaces and modular designs, offering more flexible and efficient solutions. In the following sections, we provide a detailed discussion of the mainstream physics simulation methods (see Sec. 3.3.1), along with the leading physics engines and platforms (see Sec. 3.3.2).

### 3.3.1 Physics Simulation Methods

Currently, mainstream physical simulation approaches include grid-based Lagrangian-Eulerian hybrid methods (e.g., the Material Point Method, MPM [132], [157], widely used for simulating elastomers, fluids, and granular materials) and position-based methods (e.g., Position-Based Dynamics, PBD [161], XPBD [162], commonly applied in real-time simulations of deformable objects and cloth).

**Position-Based Dynamics, PBD** [161]: PBD [161] adjusts particle positions using constraint conditions defined as cost functions, which are optimized to satisfy the desired constraints. In a typical dynamic system, each particle (or vertex) is assigned a mass, position, and velocity. The simulation process involves predicting new positions based on current velocities and external forces, iteratively correcting these positions to enforce the constraints, and then updating velocities accordingly.

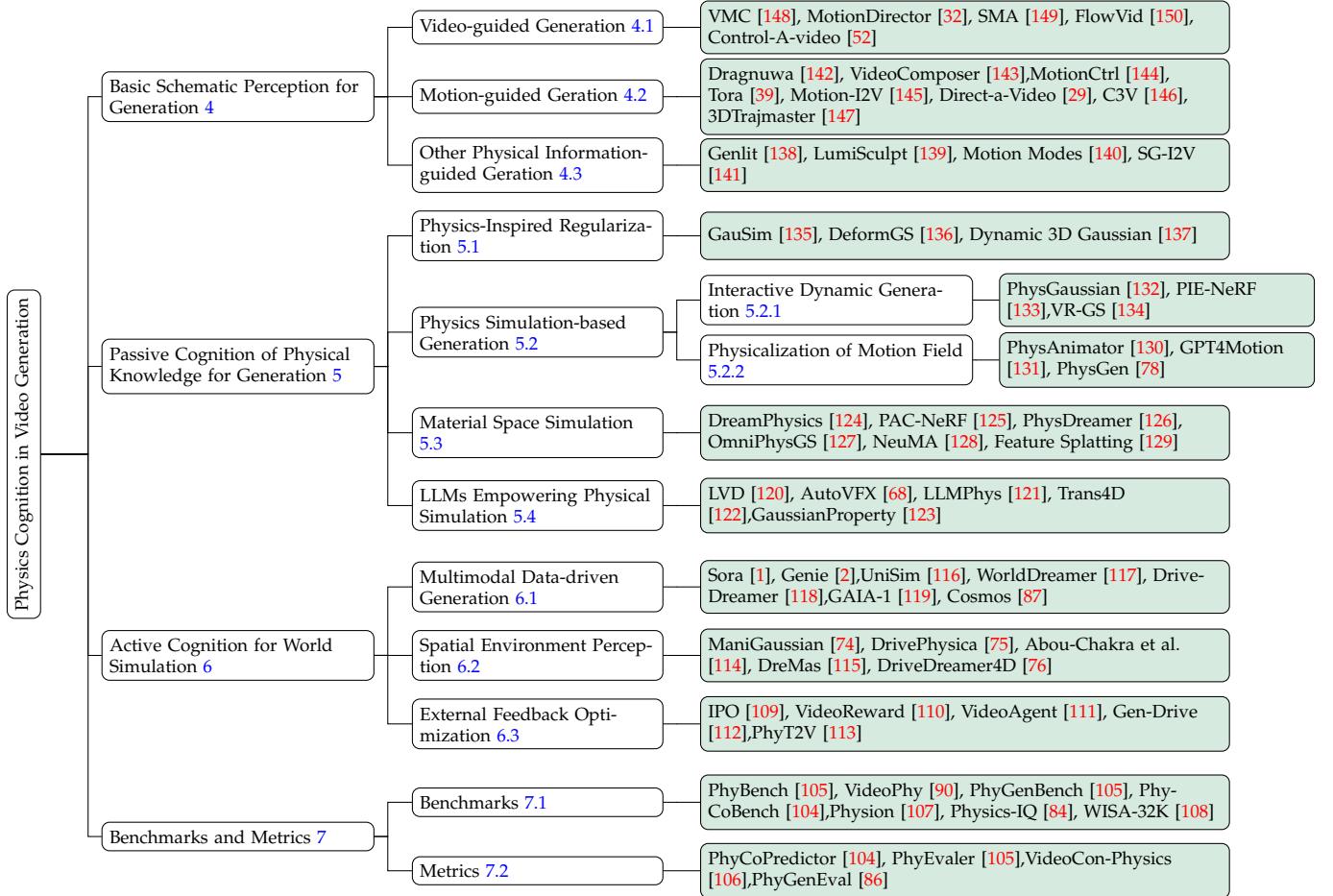


Fig. 5. Overview of the evolution of physical cognition in video generation. Please note that the typical methods listed here cover only a subset of the relevant literature and do not represent all existing studies.

**Extended Position-Based Dynamics, XPBD [162]:** XPBD is an extension of the PBD [161] framework, designed to address several inherent limitations of PBD, such as time step dependency and imprecise control over constraint stiffness. The core idea of XPBD is to model constraints as elastic potential energy and solve constraint forces using implicit integration methods. This approach maintains the computational efficiency of PBD while delivering more accurate and physically plausible simulations.

**Material Point Method, MPM [132], [157]:** The MPM combines the Eulerian and Lagrangian approaches to solve PDEs by discretizing space into a grid-particle framework, thereby simulating the motion and deformation of materials at macroscopic scales. Specifically, in MPM, an object is represented as a collection of discrete particles, each carrying its own material properties (e.g., mass, density, and velocity). The simulation alternates between transferring particle data to a fixed grid (P2G) for computing forces and updating grid-based quantities, and then transferring the updated information back to the particles (G2P) to revise their positions and velocities. This iterative process allows MPM to accurately simulate complex deformations and interactions in a physically consistent manner.

Notably, compared to PBD [161] and XPBD [162], MPM [157] is a differentiable simulator. Its mathematical differentiability enables a deep integration between physical sys-

tems and data-driven methods (e.g., neural networks). Such frameworks can generate high-quality videos that strictly adhere to physical laws, offering both visual fidelity and dynamic realism.

### 3.3.2 Physics Engines and Platforms

In this subsection, we will introduce several mainstream physics engines and platforms that are covered in this survey, as shown in Tab. 1.

**Bullet Physics:** Bullet [152] is a robust and open-source physics engine widely utilized in game development, film special effects, and robotics simulation. It supports rigid body dynamics, soft body dynamics, collision detection, and constraint solving, boasting high performance and cross-platform compatibility.

**PyBullet:** PyBullet [153] is a Python interface encapsulation of the Bullet [152], designed to offer a user-friendly programming experience. It not only supports complex physical simulations but also integrates graphic rendering functionalities, enabling the easy creation of robotic simulation environments and reinforcement learning scenarios.

**Blender:** Blender [131] is a comprehensive open-source 3D creation software that integrates modeling, animation, rendering, and video editing features. Its built-in physics engine supports simulations of rigid bodies, soft bodies, fluids, fabrics, and particle systems. Although it may not

Physics Engines and Platforms	Programming Language	GPU Acceleration	Supported Physics Types	Open Source	Typical Application Scenarios
Bullet [152]	C++	Partial support	Rigid, soft body, collision detection, constraint solving	✓	Game development, film special effects, and robotics simulation
PyBullet [153]	Python	Partial support	Rigid, soft body, collision detection, constraint solving	✓	Robot simulation, reinforcement learning
Blender Physics [131]	Python	Partial support	Rigid body, soft body, fluids, fabrics and particle systems	✓	3D animation and film special effects
Isaac Gym [154]	Python	Fully support	Rigid body and joint drive	✗	Large-scale parallel simulation, complex robot motion control
NVIDIA PhysX [155]	C++	Fully support	Rigid body, soft body, cloth, fluids and particle system	✗	Game development, virtual reality, and industrial simulation
Taichi [156]	Python/C++	Fully support	Fluids, elastic bodies, particle systems	✓	Support for custom physical models (e.g. MPM [157])
NVIDIA Omniverse [158]	Python/C++/USD [159]	Fully support	Multi-physics engine integration	✗	Cross-software collaboration, digital twins
Genesis [160]	Python	Fully support	Coupling various physical models	✓	Robot realistic data generation

TABLE 1  
Comparison of features of physics engines and platforms.

match the performance of specialized physics engines, it excels in artistic creation and visual effect production.

**Isaac Gym:** Isaac Gym [154] is a high-performance physics simulation platform developed by NVIDIA, specifically designed for reinforcement learning and robotics simulation. Utilizing GPU acceleration technology, it can run thousands of simulation environments in parallel on a single GPU, significantly enhancing training efficiency.

**NVIDIA PhysX:** NVIDIA PhysX [155] has emerged as a predominant solution in gaming and real-time simulation domains by achieving real-time computation of large-scale physical interactions through GPU optimization, effectively balancing computational accuracy with system performance.

**Taichi:** Taichi [156] is an open-source library for computer graphics and physical simulations, focusing on high-performance computing and programmability. It employs Just-In-Time (JIT) compilation technology to convert Python code into highly efficient low-level code, supporting both CPU and GPU operations.

**NVIDIA Omniverse:** NVIDIA Omniverse [158] is a real-time 3D design collaboration and physics simulation platform built on the Universal Scene Description (USD) [159] framework. Its core features include cross-software collaboration, physics-level precision simulation (digital twins), and a generative AI-driven toolchain, aimed at advancing the development of generative physical AI.

**Genesis:** Genesis [160] is an open-source generative physics engine designed for robotics, embodied intelligence, and physical AI. Built in Python, it integrates multiple physics solvers [157], [161], [162] into a unified framework, enabling fast simulations. A key feature is its differentiable simulators, which support reinforcement learning by embedding gradient information. Genesis also supports language-driven generative simulations, allowing the creation of high-fidelity 4D dynamic worlds.

## 4 BASIC SCHEMATIC PERCEPTION FOR GENERATION

Although traditional video generation models demonstrate remarkable generative capabilities, their mechanisms for physical dynamic control remain significantly constrained. These models typically rely on simple signal conditioning (e.g., text [48], [49], [50], [51] or images [54], [55], [56]), which inadequately specifies fine-grained dynamic details. To address this limitation, recent approaches abstract physical visual patterns into controllable basic schemas, such as optical flow fields [150], trajectories [142], [144], motion bounding boxes [65], or motion video sequences [32], [148]. These basic schemas encode the temporal evolution of object entities or pixel positions as motion fields, which are then injected into the model's latent space, ultimately achieving motion-consistent video generation. This section's paradigm pipeline is shown in Fig. 6.

### 4.1 Video-guided Generation

The core objective of Video-to-Video (V2V) generation lies in extracting and transferring motion attributes (e.g., motion direction, velocity distribution, and temporal pose evolution) from reference videos while reconstructing the visual environment of target scenes. As shown in Fig. 6(a), current methodologies achieve this through a dual-prior driven framework: reference videos provide motion pattern priors (such as rigid-body motion continuity and temporal coherence of human actions), while text/image prompts supply content semantic priors (e.g., scene layouts, object materials).

**Unwrapping Motion and Appearance.** To address the complex entanglement between motion and appearance, some studies focus on separating and refining motion features. For example, VMC (Video Motion Customization) [148] aligns the ground truth residuals between consecutive frames with their denoising predictions, enabling the distillation of motion information and fine-tuning the keyframe generation module to improve temporal consistency. At a higher level, MotionDirector [32] adopts a dual-path architecture to decouple appearance and motion features from

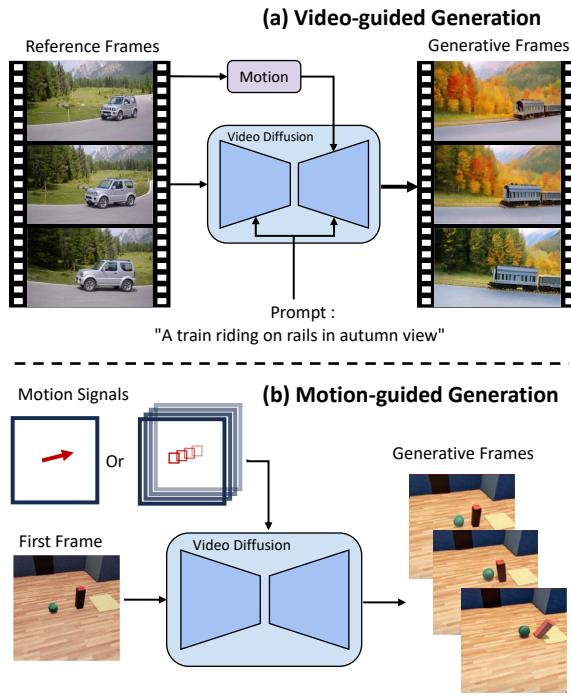


Fig. 6. Basic motion signal-guided generation pipeline. (a) is a video-guided generation pipeline in Sec. 4.1, while (b) is a motion-guided generation pipeline in Sec. 4.2.

reference videos. This method not only reproduces the same motion across different appearances but also supports the blending of motion and appearance features from different videos, enhancing the diversity and controllability of generation. Spectral Motion Alignment (SMA) [149] addresses the issues of contextual information deficiency and local distortions that may arise from motion vectors obtained through traditional inter-frame residuals by coordinating global and local frequency domain regularization. Sun et al. [163] disentangles appearance and motion by distinguishing frame-level and spatiotemporal representations, achieving motion control by minimizing appearance information.

**Extracting Motion Signals.** Additionally, some studies the concept of explicit motion correlation between reference and generated videos. These approaches typically extract conditional signals from the reference video and embed them into the generation framework for learning. For instance, FlowVid [150] addresses the limitations of conventional optical flow constraints by introducing depth maps as spatial conditions to assist motion control, significantly improving spatial consistency. Control-A-Video [52] combines motion priors from the reference video's motion flow and inter-frame residuals to warp latent noise, thereby controlling motion consistency across generated video frames. Li et al. [164] focuses on natural motion by extracting motion spectral volume [165] from real-world videos and modeling them as oscillatory patterns. These patterns are used to guide the generation of seamless looping videos or interactive dynamic simulations. Wang et al. [166] integrates the strengths of explicit and implicit methods by abstracting the motion information from reference videos into Q, K, and V embeddings within temporal attention. It uses an optical

flow-like inter-frame difference approach to eliminate static appearance biases, capturing global and local motion patterns across the temporal dimension and further ensuring both temporal and motion consistency.

**Causal Reasoning Outside of Motion.** In scenarios involving complex interactive dynamics, InterDyn [167] introduces hand control signals in the form of masks, which are processed through a ControlNet-like [168] branch network to drive action generation. This approach enables the prediction of complex interactions involving counterfactual future scenarios and force propagation dynamics. Unlike traditional methods, InterDyn leverages the implicit physical priors in pre-trained video generation models, allowing for continuous causal dynamic reasoning beyond the scope of the control signals. Although it does not require explicit reconstruction or physical simulation, it still relies on pre-defined motion signals to guide subsequent predictions.

## 4.2 Motion-guided Generation

Video-to-video (V2V) generation methods are primarily limited to reproducing existing motion patterns. To achieve greater flexibility in adapting to the generation of diverse motion patterns, the research community has increasingly focused on motion-guided generation paradigms—introducing explicit motion control signals as conditional constraints to steer models toward generating controllable dynamic content. Motion-guided video generation (see Fig. 6(b)) typically adopts either a single-stage [22], [29], [65], [144] or two-stage paradigm [39], [57], [142], [145], [169], [170], [171], [172]. The two-stage paradigm involves: 1) extracting low-dimensional motion representations (e.g., optical flow fields, motion vectors, or point trajectories) from mask bounding boxes or arrow-guided trajectories; and 2) augmenting pre-trained video diffusion models with motion encoders to embed these motion features into the latent space of the diffusion process. In contrast, the single-stage framework omits the explicit motion extraction step, instead directly mapping simplified motion control signals to the generative model through end-to-end learning (equivalent to Stage 2 in two-stage frameworks). To enable diffusion models to effectively respond to controllable signals, various approaches have been proposed. Some integrate conditional signals and latent representations through attention mechanisms [29], [39], [57], [145], [170], while others [22], [65], [142], [144], [169], [171], [172] employ additional structures similar to ControlNet [168], encoding external conditions into the latent space via multi-scale skip connections.

Despite progress in generating controllable motion-driven videos under these paradigms, challenges remain, including motion inconsistencies, ambiguities in camera transitions and object movements, and difficulties in modeling 3D scene dynamics. In the following sections, we introduce improvements to motion-guided video generation across three critical dimensions: motion consistency enhancement, camera-object motion control, and 3D spatial motion generation.

**Motion Consistency Enhancement.** Motion consistency directly impacts the physical plausibility of generated videos. Ensuring motion consistency allows an object's velocity, direction, and acceleration to remain coherent

across multiple frames, avoiding interruptions or unrealistic changes in motion. This enhances the video's visual coherence and viewing experience. Current research focuses on designing motion signals to drive the generation of physically plausible motion videos.

Initially, Dragnuwa [142] achieved controllable video generation across semantic, spatial, and temporal dimensions by integrating three fundamental control signals: text, images, and trajectories. Through an end-to-end pipeline, it sampled dense optical flow directly from video streams and fused it with text and image features at multiple scales, adaptively training to generate trajectory-consistent videos. VideoComposer [143] presented a compositional video generation framework and leveraged motion vectors as a temporal condition for motion transfer. Animate Anyone [173] and UniAnimate [4] took a character image as a reference and used a pose sequence to control the desired movement of the character. MotionCtrl [144] introduced independent modules for controlling camera and object movements, enabling 2D point-driven object motion and 3D trajectory-driven camera control. MOFA-Video [171] and Motion Dreamer [174] used a sparse-to-dense motion generation network to extract explicit motion cues and incorporated various motion fields as control conditions to fine-tune the diffusion model. Motion prompting [22] expands user requests into point trajectory motion prompts, encoding both sparse and dense motion in spatial and temporal dimensions. Motion-I2V [145] employed explicit motion modeling to generate videos of moving objects while preserving their appearance. The method introduced a diffusion-based motion field predictor and a motion-guided I2V generator, leveraging a temporal attention module to connect the two stages and propagate motion trajectories consistently across synthesized frames. Dreamvideo-2 [65] introduced reference attention to learn subject appearance features, while bounding box mask sequences provided motion control signals. To achieve a balance between subject representation and motion control, the method masked reference attention to enhance subject identity representation. Tora [39] built upon Sora [1] and further emphasized the importance of motion control by incorporating dynamic trajectory components to guide scalable video generation.

Earlier trajectory-controlled motion methods relied solely on single-point controls to generate pixel-level motion, which struggled to produce realistic physical entity movements. These limitations often resulted in undesired global motion or appearance distortions. To address these challenges, DragAnything [169] defined sequences of entity-centric points and Gaussian maps to guide motion at the entity level. TrackGo [170] combined masks and trajectory tracking to obtain detailed multi-point trajectories and introduced a TrackAdapter to manipulate dual-branch attention maps for motion intervention. Additionally, the application of spatiotemporal attention embedding may be misled by irrelevant blocks. Flatten [66] utilizes optical flow to connect patches across different frames, guiding accurate attention and promoting information exchange between patches along the same trajectory.

When multiple trajectories in dynamic motion come into close proximity (e.g., collisions or overlaps), sparse optical flow may merge and erroneously swap trajectories,

as observed in Tora [39]. To mitigate this issue, InTraGen [175] introduced multi-modal interaction encoding to ensure stable generation of multi-object interaction videos. The method designed a target ID map to provide interactive priors for sparse dynamic information, enabling the model to accurately identify both static and dynamic objects under trajectory-controlled conditions.

**Camera-object Motion Control.** Given the limitations of earlier methods in handling only a single type of motion signal, AnimateAnything [57] integrates multiple conditions (e.g., motion annotations, camera motion) to control video generation. By combining explicit and implicit signal injection, this method leverages a Flow Generation Model (FGM) to iteratively denoise and transform various dynamic signals into a unified dense optical flow, which is ultimately used to guide video generation. While both Motion-I2V [145] and AnimateAnything adopt a two-stage architecture to incorporate control signals, AnimateAnything stands out by enabling the integration of multiple conditions and providing balanced guidance for video generation. Similarly, Image Conductor [172] focuses on generating camera transitions and object motion but takes a different approach than AnimateAnything. Specifically, it decouples camera and object motion through distinct LoRA weights, allowing precise control over either motion type. On the other hand, Direct-a-Video [29] employs a self-supervised training strategy to flexibly decouple camera and object motion without pre-training. This is achieved by cropping static-camera video sequences to simulate moving-camera perspectives and embedding them into temporal cross-attention. Additionally, it emphasizes the spatial attention of text tokens corresponding to motion bounding boxes, enabling independent or joint control over camera and object motion.

**3D Spatial Motion Generation.** The motion trajectories of objects in the real physical world are inherently three-dimensional dynamic processes, involving complex spatial interactions such as depth displacement, multi-view occlusion, rotational and orbital motions. However, existing motion control methods are predominantly confined to two-dimensional planar trajectory guidance, struggling to accurately characterize physical dynamics scene within 3D spatial. To obtain well-defined 3D trajectories, LeviTor [176] introduces a novel approach that combines depth information with K-means clustering of multiple points to control 3D object trajectories in video synthesis, thereby guiding the model to generate physically plausible videos in 3D space. Chen et al. [177] extends 2D conditional optical flow to a 3D point cloud motion field, achieving cross-dimensional feature fusion by concatenating video frames and 3D point vectors and feeding them into a diffusion model for training, thereby enhancing the dynamics of video generation at a higher level. C3V [146] relies on textual priors processed through LLMs to obtain object trajectories in 3D space, optimizing spatial layout and temporal dynamics in video generation. This method decomposes the input text into sub-prompts, which are processed by expert models to generate 3D representations. It then incrementally estimates object trajectory coordinates in the scene and refines dynamic objects using Score Distillation Sampling (SDS) [178] to extract 2D diffusion priors, achieving coarse-to-fine text-driven 3D video generation. TC4D [179] decomposes the deformation

field into global and local components, modeling global motion as rigid body transformations along trajectories. It introduces video score distillation sampling to activate the internal motion priors of pretrained generative models, optimizing the local deformation field. This approach achieves significant progress in generating large-scale scenes.

Previous approaches either focus on obtaining 3D trajectories or directly guide video generation using trajectories in 3D space, while overlooking the issue of the lack of realistic 3D trajectory data during training. 3DTrajMaster [147] addresses this gap by introducing the first 360°-Motion Dataset specifically designed for 3D trajectory-guided video generation. Furthermore, 3DTrajMaster develops a plug-and-play 3D trajectory embedding framework, which incorporates domain adapters and an annealed sampling strategy to enable controllable video generation with multi-entity 6DoF (six degrees of freedom).

#### 4.3 Other Physical Information-guided Generation

**Video Relighting.** In addition to motion realism, lighting consistency is another indispensable aspect of physical plausibility in high-fidelity video generation. Previous models for lighting modeling primarily focused on relighting for single images, either through explicit inverse rendering [67], [180] of geometry and materials or by employing neural methods [181], [182], [183] for end-to-end learning. GenLit [138] uniquely explores the potential of video diffusion models to comprehend the physical world, particularly lighting. It reformulates the single-image relighting task as a video generation task, enabling controllable light manipulation in video generation through external light source signals. LumiSculpt [139] shifts the focus of the relighting task to text-to-video generation, decoupling lighting representation from appearance and enabling the diverse generation of videos with controllable lighting. Lux Post Facto [184] encodes High Dynamic Range (HDR) maps into “lighting embeddings” for refined lighting control, integrating them as conditional inputs into a diffusion model. By training delighting and relighting models on a hybrid dataset with lighting-rich and motion-rich, this approach achieves more precise and nuanced video relighting.

**Zero-Shot Self-guidance.** Previous methods [52], [142], [147] that leverage the generative capabilities of pre-trained models often rely heavily on conditional guidance to produce controllable motion videos, overlooking the latent motion priors embedded within pre-trained models. Motion Modes [140] aims to uncover the motion generation potential of pre-trained diffusion models by guiding energy-based iterative sampling during inference to construct a motion set, achieving diverse motion generation without additional training or conditional control. Similarly, SG-I2V [141] optimizes noise latent variables to enforce feature similarity within bounding box trajectories, enabling self-guided trajectory control.

**Discussion.** Overall, basic schematic perception-based generation methods highlight the effectiveness of control signals in controllable video generation tasks. By embedding various forms of schematic representations, these approaches address key challenges such as motion consistency, camera-object control, 3D spatial motion, and relighting,

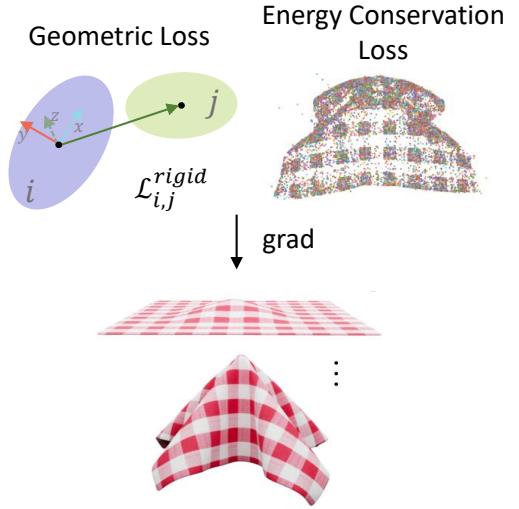


Fig. 7. Physics-inspired regularization pipeline.

However, they face several critical limitations: 1) **Lack of Physical Consistency:** Generation methods guided solely by simple visual patterns often fail to capture fundamental physical principles, resulting in physically implausible video outputs; 2) **Limited Generalization to Complex Scenes:** These models struggle to adapt to diverse environments involving complex object interactions, such as collisions, deformations, and external forces. To enhance physical plausibility and improve video realism, future research could explore the integration of physical knowledge into generative frameworks or leverage real-world physical interactions to enrich the modeling of dynamic environments.

### 5 PASSIVE COGNITION OF PHYSICAL KNOWLEDGE FOR GENERATION

In addition to the aforementioned schema-guided intuitive video generation methods, the generative system can also leverage physical rules and strictly follow symbolic logic for deductive reasoning. The sources of symbolic knowledge include physical loss function constraints(in Sec. 5.1), knowledge embedding from physics simulators(in Sec. 5.2), and world physics knowledge from LLMs (in Sec. 5.4). Explicit physics knowledge embedding methods demonstrate unique advantages in complex interactive scenarios: by directly embedding physical knowledge, the generated videos not only ensure visual coherence but also achieve physical verifiability at both the microscopic particle motion and macroscopic energy transfer levels, thus laying a theoretical foundation for building reliable world simulators.

#### 5.1 Physics-Inspired Regularization

Physics-informed regularization adds constraint terms related to physical systems within the loss function, enabling the model to learn data features while adhering to known physical laws, thereby enhancing its generalization ability and physical consistency, as shown in Fig. 7.

**Regularization of Energy Conservation.** To enable the model to more accurately reflect long-term complex deformations in the real world, GauSim [135] combines continuum mechanics with neural networks, effectively capturing dynamic laws through hierarchical simulations from coarse to fine. Additionally, explicit constraints for mass conservation and momentum conservation are introduced to ensure physical interpretability. DeformGS [136] focuses on large deformations, shadows, and occlusion in highly deformable objects, shaping the 3DGS deformation field by optimizing deformation functions, while introducing local isometric loss and momentum conservation regularization terms to constrain the relative distance and trajectory smoothness of adjacent Gaussians.

**Regularization of Geometric.** In dynamic scene modeling tasks, Dynamic 3D Gaussians [137] relies on regularization to enforce underlying physical models, integrating explicit modeling with a physical metric space. Dynamic 3D Gaussians [137] are designed for complex motion scenarios, incorporating local-rigidity loss (to enforce rigid body transformations), local-rotational similarity loss (to maintain local rotational consistency), and long-term isometric loss to achieve dense 6-DOF tracking. Dynamic 3D Gaussians and DeformGS provide high-precision 3D tracking solutions for deformable object manipulation and complex motion scenarios. Rai et al. [185] significantly enhances the temporal consistency and rigidity of shape preservation in animations by incorporating Length-Area (LA) regularization and shape-preserving As-Rigid-As-Possible (ARAP) loss. While incorporating 3D point clouds into the diffusion framework, Chen et al. [177] jointly optimizes the noise distribution and local rigidity preservation of the point cloud structure, thereby enhancing the geometric fidelity and physical coherence of generated content.

While these methods generally demonstrate good performance under predefined settings, they cannot guarantee an accurate capture of unlearned physical phenomena, such as complex interactions, which may lead to physically implausible results. Therefore, further exploration is needed for more explicit physical video generation.

## 5.2 Physics Simulation-based Generation

Physics simulation-based video generation leverages explicit physical rules and models, embedding domain-specific knowledge (e.g., dynamics, fluid mechanics, and material science) to simulate motion and interactions within a scene. In Sec. 5.2.1, we investigate interactive dynamic generation methods that employ external forces as inputs for static 3D scenes. These methods model the interactions between scenes and forces using various physical simulators, infer changing trends that comply with the physics laws, and generate corresponding dynamic scenes. In Sec. 5.2.2, we introduce approaches for the physicalization of the motion field. In Sec. 4.2, we have discussed motion-guided generation based on manually defined motion signals, which inherently exhibit high uncertainty. To address this, the methods described in Sec. 5.2.2 leverage physics simulators to generate precise motion fields, effectively guiding the diffusion model for more physically accurate video synthesis.

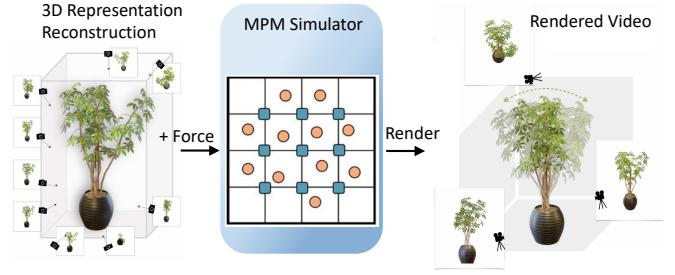


Fig. 8. Generation pipeline based on physical simulation, cf. Physgaussian [132].

### 5.2.1 Interactive dynamic generation

Interactive dynamic generation utilizes external forces as inputs to static 3D scenes, employing physics simulators to simulate the interactions between the scene and the applied forces, infer the evolution trends that comply with physical laws, and generate corresponding dynamic scenes.

**Dynamic Generation Based on GS.** To capture realistic motion patterns and causal relationships for creating more lifelike interactive dynamics, PhysGaussian [132] was the first to integrate 3D Gaussian Splatting (3DGS) with the Material Point Method (MPM) into a unified simulation-rendering pipeline, as shown in Fig. 8. This pioneering approach treats 3D Gaussian kernels as discrete particles, allowing the deformation of Gaussian kernels during continuous media transformation to seamlessly integrate physical simulation with visual rendering. Compared to PhysGaussian [132], VR-GS [134] addresses the real-time inefficiencies of MPM by adopting XPBD [162]. Additionally, it introduces a two-stage embedding strategy to resolve sharp artifact issues: Gaussian kernels are first embedded into local tetrahedra independently, and the tetrahedral vertices are then embedded into a global grid, allowing the gaussian kernels to adapt smoothly to the mesh. Gaussian Splashing (GSP) [186] innovatively integrates Lagrangian fluid-solid interactions within 3DGS scenarios through a unified PBD framework. Notably, GSP decouples the simulation and rendering processes at the solid level through separate instantiation of particles and Gaussian kernels. Furthermore, it enhances fluid simulation and rendering by optimizing Gaussian kernels with the integration of diffuse reflection, normals, specular highlights, and surface roughness. Phys-motion [187] and Phy124 [89] focus on generating physics-consistent dynamic videos from a single image. Sync4D [188] maps the motion from the reference video onto a skeleton with skinning weights, and integrates MPM-based physical simulation and displacement loss to optimize the velocity field.

**Dynamic Generation Based on NeRF.** In addition to 3DGS-based physical simulation, PIE-NeRF [133] and Video2Game [71] integrate NeRF with physics simulation. PIE-NeRF demonstrates the feasibility of integrating classical Lagrangian dynamics with a mesh-free NeRF approach, introducing Quadratic Generalized Moving Least Square (Q-GMLS) and Voronoi partitioning to handle nonlinear deformations and intensive computations. Video2Game [71] employs NeRF to construct 3D worlds for games, further

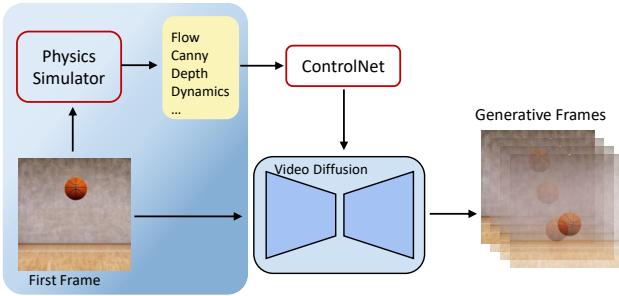


Fig. 9. Physics simulator generates motion signals.

refining mesh models to enhance compatibility with game engines.

**Neural Networks for Solving PDEs.** Beyond incorporating physics simulators to constrain and guide physical video generation, some works (e.g., ElastoGen [151]) embed explicit physical priors into neural models. During training, these methods optimize for the numerical solutions of partial differential equations (PDEs), enabling collaborative learning between physics and neural networks.

### 5.2.2 Physicalization of Motion Field

While physics simulation-based methods (see Sec. 5.2.1) can faithfully reconstruct dynamic scenes that adhere to physical laws, they are limited in scope and cannot adapt to the diversity of real-world scenarios. In contrast, the motion-guided generation paradigm (see Sec. 4.2) offers a certain degree of flexibility but relies on manually defined motion signals, often leading to inaccuracies due to weak geometric understanding and missing physical constraints. To this end, the academic community proposes a new framework (see Fig. 9) – before the generation stage of the video diffusion model, first construct motion signals that conform to physical laws (such as optical flow fields, depth maps, dynamic assets, etc.), and encode these physical constraint conditions into the generation process of the diffusion model, thereby achieving precise control of video synthesis.

To optimize the generation of motion signals, MotionCraft [189] introduces a physics-informed zero-shot video generation approach that achieves dual-consistency motion mapping between latent and pixel spaces by strategically warping noise latent vectors through physically simulated optical flow. Similarly, PhysAnimator [130] innovatively integrates physical simulation with data-driven generative models to enable the generation of deformable animation sequences. This approach solves the kinematic equations in 2D space to evolve dynamics and generate conditional optical flow fields. The optical flow fields are then used to generate deformable sketch sequences as control signals, with sampled keyframes guiding the coherent animation generation. GPT4Motion [131] utilizes GPT-4 to generate Blender scripts that simulate continuous motion sequences, producing depth and edge maps. These outputs serve as conditional inputs for ControlNet, guiding a diffusion model to render realistic videos frame by frame. Similarly, PhysGen [78] employs rigid-body physics and large language model-inferred parameters to simulate realistic dynamic interactions, driving a video diffusion module for

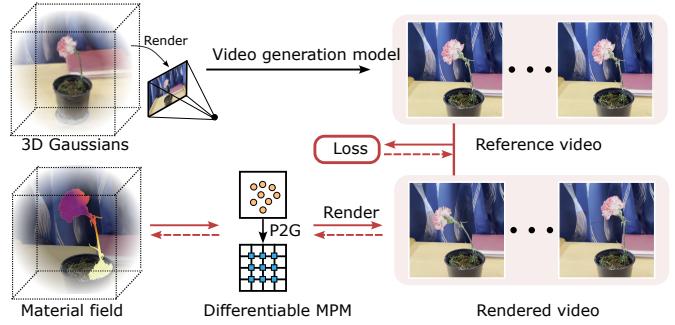


Fig. 10. Architecture of PhysDreamer [126]. Optimize the material field and velocity field by minimizing the discrepancy between the rendered video and the reference video.

rendering and refinement. Phymotion [187] uses physically plausible foreground dynamics generated by MPM-simulated 3DGS as intermediate coarse 3DGS dynamics, and then constructs a diffusion pipeline employing DDIM+ inversion to obtain latent noise codes. During the sampling stage, it mixes coarse and enhanced sampled videos to achieve high-quality results.

Moreover, using such above simulations as physical priors to guide diffusion models in video generation may introduce discrepancies between simulated and real-world conditions (Sim2Real Gap). To address this issue, SimGen [190] proposes a cascaded generation framework, where a simulator first generates simulated conditions (e.g., depth and segmentation) combined with text prompts, which are then fed into a lightweight diffusion model to transform them into realistic conditions. These refined conditions subsequently guide the diffusion model in generating realistic driving scenes.

## 5.3 Material Space Simulation

Interactive dynamic generation (see Sec. 5.2.1), based on rigorously defined mathematical equations and physical simulators, generates high-precision interactive dynamics [71], [132], [133] with physical laws through numerical simulations. However, the interpretability of this approach heavily relies on manually defined physical parameters (e.g., Young's modulus, friction coefficients) [126], [191]. These parameterized representations are constrained by parameter sensitivity and diversity in real-world scenarios, making it challenging to configure complex material fields in scenes. To address these limitations, current research utilizes various physical priors to guide the adaptive modeling of material fields. By autonomously learning and inferring material fields, these methods [124], [125], [126], [128], [191], [193] enable more accurate and adaptable physical simulations.

**Score Distillation Sampling.** To enable the automatic estimation of material properties, DreamFusion [178] introduced Score Distillation Sampling (SDS), which provides a novel approach to bridging the gap between 2D and 3D representations. SDS leverages the score function of diffusion models to optimize the parameters of 3D models (e.g., NeRF or other neural implicit representations) via backpropagation. By minimizing a noise reconstruction loss in an adversarial manner, SDS ensures that the 2D-rendered images generated by the 3D representation align with the

	Methods	Input Type	Physics Simulator	Material Field			Representation	Materials Types
				Manual Parameter init.	Learnable	LLM inference		
Interactive Dynamic Generation	PhysGaussian [132]	Multi-view	MPM	✓			3DGS	varieties materials
	Phy124 [89]	Single image	MPM	✓			3DGS	elastoplasticity
	VR-GS [134]	Multi-view	XPBD	✓			3DGS	elastoplasticity
	Gaussian Splashing [186]	Multi-view	PBD	✓			3DGS	solids and fluids
	PIE-NeRF [133]	Multi-view	Q-GMLS/Taichi	✓			NeRF	hyperelastic
	Video2Game [71]	Dynamic Video	Cannon.js/Blender/Unreal	✓	✓		NeRF	Rigid-body
Physicalization of Motion Field	ElastoGen [151]	3D model	NeuralIMTL	✓			NeRF/3DGS	hyperelastic
	PhysGen [78]	Single image	Pymunk			✓	2D	rigid-body
	MotionCraft [189]	Text	$\phi$ -Flow	✓			2D	rigid-body and Fluids
	PhysMotion [187]	Single image	MPM	✓			3DGS & 2D	varieties materials
	GPT4Motion [131]	Text	Blender			✓	2D	varieties materials
Material Space Simulation	PhysAnimator [130]	Single anime illustration	Taichi [156]	✓			2D	deformable body
	PhysDreamer [126]	3D model	MPM	✓			3DGS	hyperelastic
	PAC-NeRF [125]	Dynamic Video	MPM	✓			NeRF	varieties materials
	Physics3D [191]	3D model	MPM	✓			3DGS	elastoplastic and viscoelastic
	DreamPhysics [124]	3D model & Text & Image	MPM	✓			3DGS	elastoplastic
	Liu et al. [192]	Multi-view & Text	MPM	✓	✓		3DGS	varieties materials
	NeuMA [128]	Multi-view	MPM	✓			3DGS	varieties materials
	OmniPhysGS [127]	3D model & Text	MPM	✓			3DGS	varieties materials
	Feature Splatting [129]	Multi-view & Text	GS-Taichi-MPM			✓	3DGS	varieties materials
	Phys4DGen [193]	Singe image	MPM			✓	3DGS	varieties materials
Diffusion-based Methods	Sim Anything [194]	Multi-view	MLS-MPM			✓	3DGS	varieties materials
	GaussianProperty [123]	Multi-view	MPM			✓	3DGS	varieties materials

TABLE 2  
Summary of physics simulation-based generation methods.

distribution of the diffusion model, typically conditioned on textual guidance for controlled generation.

$$\nabla_{\theta} \mathcal{L}_{\text{SDS}} = \mathbb{E}_{\epsilon, t} [w(t) \cdot (\epsilon_{\theta}(\mathbf{x}_t) - \epsilon) \cdot \nabla_{\theta} \mathbf{x}_0] \quad (7)$$

where  $\epsilon_{\theta}$  represents the noise predictor of the diffusion model,  $\mathbf{x}_t$  is the noisy sample, and  $\mathbf{x}_0$  is the current rendered image.

Through SDS, researchers can leverage existing 2D diffusion models to optimize differentiable material fields directly. Physics3D [191] extends the physical parameters in MPM to capture both the elastic and viscous properties of materials. DreamPhysics [124], based on the physically modeling-friendly KAN (Kolmogorov Arnold Networks) [195] representation, introduces Motion Distillation Sampling (MDS) to emphasize motion information in videos, ensuring temporal consistency. MDS builds on SDS by reducing biases caused by color and placing greater focus on motion. However, applying multiple iterations of SDS using a diffusion model results in a significant increase in computational overhead. Constrained by the computationally intensive requirements of SDS, Liu et al. [192] proposes a resource-efficient optical flow loss as an alternative to optimize material properties.

**Learning from the Reference Video.** PAC-NeRF [125] introduces a hybrid particle and grid-based NeRF representation and estimates the unknown geometry and physical parameters of dynamic objects in a supervised manner. PhysDreamer [126] (see Fig. 10) distills physics priors via aligning a reference video from video generation models, modeling and optimizing physical material fields (e.g., Young’s modulus and Poisson’s ratio) while incorporating MPM for material dynamics simulation. Different from the previous method of treating 3D particles equally, Fluid-Nexus [196] divides 3D particles into physical particles and visual particles, and uses a differentiable physical simulator to reconstruct and predict physical particles and synthesize new perspective fluid videos to optimize the visual appearance of 3D fluids.

**Learning Constitutive Models.** Compared to approaches that model individual materials, methods capable of handling a wide range of materials and complex objects are evidently more aligned with the real world. OmniPhysGS [127] extends gaussian kernels into learnable constitutive Gaussian kernels and predicts material properties using expert-designed constitutive models, enabling flexible adaptation to diverse materials without the need for manual

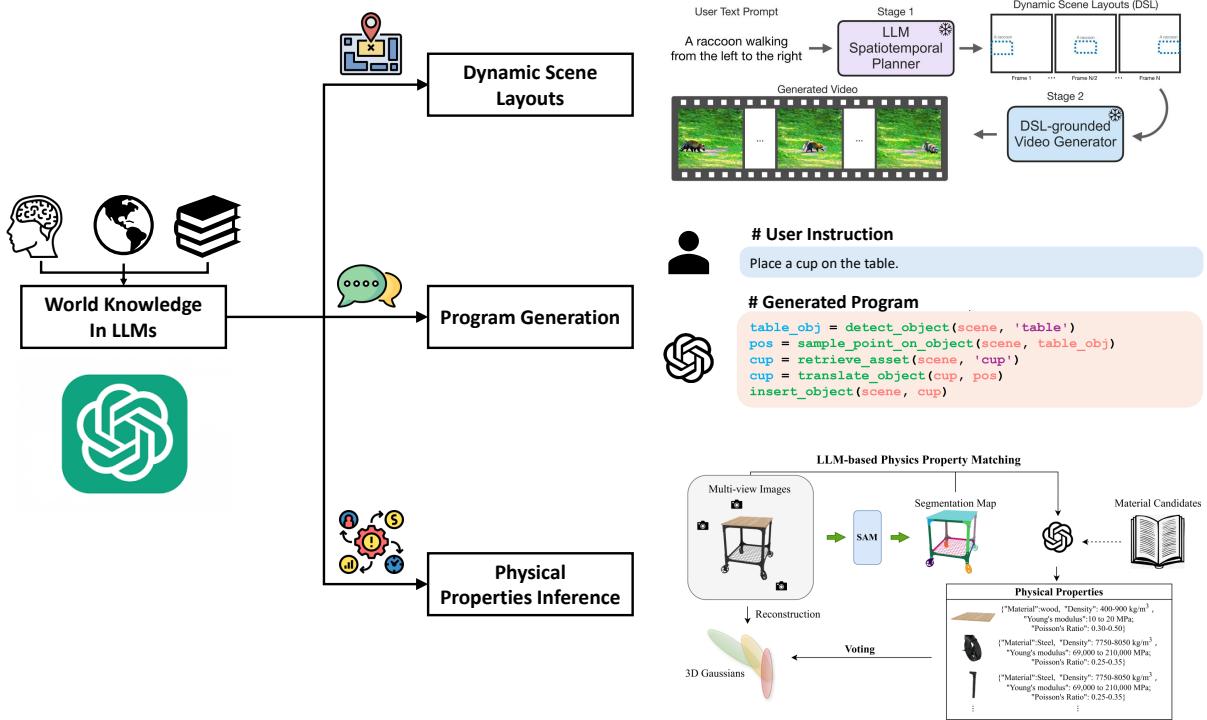


Fig. 11. World physics knowledge in LLMs is used for scene layouts, program generation, and properties inference.

configuration. On the other hand, Physics3D [191] adopts inelastic and dissipative terms in its constitutive model, which is limited to modeling a few specific physical quantities. NCLaw [197] innovatively supplants classical elasto-plastic constitutive modeling with a neural network-based approach, strategically embedding data-driven constitutive laws within a differentiable, partial differential equation (PDE)-governed simulation framework. The key idea behind NeuMA [128] is to learn the discrepancy between expert physical models and real-world scene dynamics, and perform personalized residual correction. Specifically, it uses the NCLaw [197] encoded with universal material priors, combined with the efficient, adaptive learning of elasto-plastic properties via the Lora adapter. Additionally, unlike PhysGaussian, which directly discretizes Gaussian kernels into particles, NeuMA samples physical particles to drive the differentiable rendering of Gaussian kernels.

**LLMs Inference Physics Properties.** Additionally, the natural physical knowledge and reasoning abilities embedded in LLMs open up new possibilities for material reasoning in 4D dynamic generation. These models can capture complex physical phenomena and material properties from images or even videos. We will provide a detailed introduction to the methods involving LLMs for reasoning about physics properties in Sec. 5.4, such as Feature Splatting [129], Phys4DGen [193], Sim Anything [194], GaussianProperty [123].

#### 5.4 LLMs Empowering Physical Simulation

Although the methods discussed above provide explicit strategies for spatiotemporal [171], [176], [179] layout modeling or material field distillation [126], [127], [191], they

often rely on complex priors, which can reduce generalizability. In contrast, LLMs demonstrate remarkable capability in memorizing world knowledge due to their vast model capacity and extensive training datasets [198]. This enables LLMs to recall relevant physical knowledge to provide contextually appropriate responses across diverse scenarios [199], [200]. Consequently, recent studies have sought to leverage LLMs to enable more efficient spatiotemporal reasoning and automated material property estimation, thereby generating high-quality videos, as shown in Fig. 11.

**Dynamic Scene Layouts.** LLM-Grounded Video Diffusion (LVD) [120] capitalizes on the remarkable natural language understanding and reasoning capabilities of LLMs to generate physically plausible spatiotemporal scene layouts, eliminating the need for additional motion priors by using cross-attention alignment. C3V [146] provides coarse 3D object trajectories estimated by LLMs and employs video diffusion models for fine-grained supervision. Trans4D [122] uses MLLMs to generate scene descriptions that include physical properties and dynamic spatiotemporal information for initializing 4D scene modeling. It then employs a geometric transition network to predict intermediate scenes of complex object interactions from coarse to fine levels.

**Program Generation.** Although LLMs are generally recognized for their capability to understand physical common sense, they face significant limitations in reasoning about dynamic real-world interactions. To overcome this, Kubrick [201] focuses on generating videos with physical correctness, precise camera control, and temporal consistency in an end-to-end manner. Built on an RAG framework, it uses LLM/VLM agents along with 3D engines (e.g., Blender) to design an LLM Director, LLM Programmer, and LLM Reviewer, enabling iterative refinement of synthetic videos.

Similarly, both AutoVFX [68] and GPT4Motion [131] integrate the physics engine Blender with the large language model GPT-4 [202]. AutoVFX enables users to create executable programs through natural language instructions, driving dynamic editing and rendering of 3D modeling scenes. GPT4Motion, on the other hand, embeds components simulated by the physics engine (such as edge maps and depth maps) as conditional signals into the Stable Diffusion model to generate video frames with physically plausible motion. LLMPHys [121] combines LLMs with physics engines for collaborative reasoning of dynamic scene changes. Specifically, LLMs infer physical parameters and generate simulation programs, iteratively refining reasoning accuracy through feedback. The inferred parameters are then used in simulation tasks to produce complete dynamic sequences.

**Physical Properties Generation.** LLMs and other large pretrained models also play a pivotal role in material-level reasoning for dynamic generation. When dealing with complex interactions in 4D scenes, Feature Splatting [129] manipulates object appearance and assigns material properties through natural language guidance. Then, the method extends MPM using GS-Taichi-MPM by integrating Gaussian-specific features (e.g., isotropic opacity and covariance) to address issues such as collision collapse and artifacts. Similar to PhysDreamer [126], Phys4DGen [193] focuses on material-centric modeling in constructing 4D dynamic spaces. While PhysDreamer [126] models physical material fields via neural representations, Phys4DGen leverages large pre-trained models to segment and infer material properties. PhysGen [78] leverages GPT-4o [202] to automatically assign physical parameters to physically accurate motion fields, enhancing video generation with diffusion models. GaussianProperty [123] utilizes GPT-4V to estimate material properties, applying a voting strategy to project physical attributes onto 3D Gaussian representations for realistic dynamic simulations. It also leverages these material properties for Robot grasp prediction, ensuring objects remain intact by avoiding excessive deformation during manipulation. Similarly, Sim Anything [194] relies on MLLMs to predict the overall material properties of objects and reformulates local material property variations as probabilistic distribution estimates, aiming to capture a comprehensive material distribution for realistic physical simulation.

**Discussion.** Overall, passive physical cognition-based generation relies on pre-stored physical knowledge, such as physics simulators, symbolic representations, or LLMs, to enhance the physical plausibility of generated videos. While this approach improves physical interpretability and consistency, it faces several fundamental limitations: 1) Limited Adaptability to Unseen Scenarios: These models passively retrieve and apply predefined physical knowledge, restricting their ability to generalize beyond observed physical phenomena or adapt to novel interactions and environmental conditions. 2) When relying on LLMs or other symbolic representations, there is often a gap between theoretical physical principles and their practical applicability in real-world video generation tasks. 3) Integrating high-fidelity physics simulators often incurs significant computational costs, making real-time video generation a challenging task. To address these limitations, future research could explore

the following directions: (i) incorporating active interaction with the environment to allow models to dynamically refine physical world understanding, enhancing both generalization and adaptability; and (ii) developing more efficient and differentiable physics simulators that can be seamlessly integrated with world models to improve both computational efficiency and physical fidelity in video generation.

## 6 ACTIVE COGNITION FOR WORLD SIMULATION

In the development of cognitive architectures, physical symbol-driven generative systems (e.g., physics engine-based simulations) operate under strict rules but often fail to capture the diverse physical phenomena present in complex real-world scenarios, limiting their predictive power in open-ended environments (see Sec. 5). In contrast, a key advantage of world models lies in their ability to infer potential scenarios beyond the training data distribution, thereby enabling diverse world simulations for embodied agents. As envisioned by LeCun, the ultimate goal of world models is to reason and plan for unknown patterns, understanding and predicting the evolution of the world in a human-like manner [203]. All these analyses highlight the necessity for world models to incorporate an imagination mechanism grounded in physical commonsense. Consequently, a critical challenge for developers is how to integrate real-world physical cognition into the design and learning of world models, enabling a profound understanding of the physical world's dynamics and ensuring that future predictions are physically faithful. To achieve this, world models may dynamically update through active environmental interaction, bridging the gap between simulation and real-world complexity.

### 6.1 Multimodal Data-driven Generation

Currently, most world models learn statistical representations from large-scale datasets and are gradually showing their potential as foundational physical simulators. World models need to integrate multimodal inputs, such as visual, linguistic, and action information, in order to construct a comprehensive representation of the environment and enable planning in the latent space (similar to how humans perceive the world through a combination of sight, sound, and touch).

Building on OpenAI's remarkable success in LLMs [202], [204], [205], Sora [1] aims to achieve AGI (Artificial General Intelligence) in the vision domain and emergent intelligence at the physical level by significantly increasing the scale of training data and parameters. Sora stands out with its ability to generate high-quality, controllable, and realistic videos up to one minute long. Its introduction marks a critical step in the evolution from generative models to world models, inspiring further advancements in this field. However, Sora faces significant challenges in accurately simulating physical processes and interactions, and merely scaling data has yet to fulfill the envisioned higher-level intelligence of world models [206].

**Interactive.** Genie [2], a contemporary of Sora, focuses on creating interactive environments and controllable virtual worlds, offering users diverse interaction experiences.

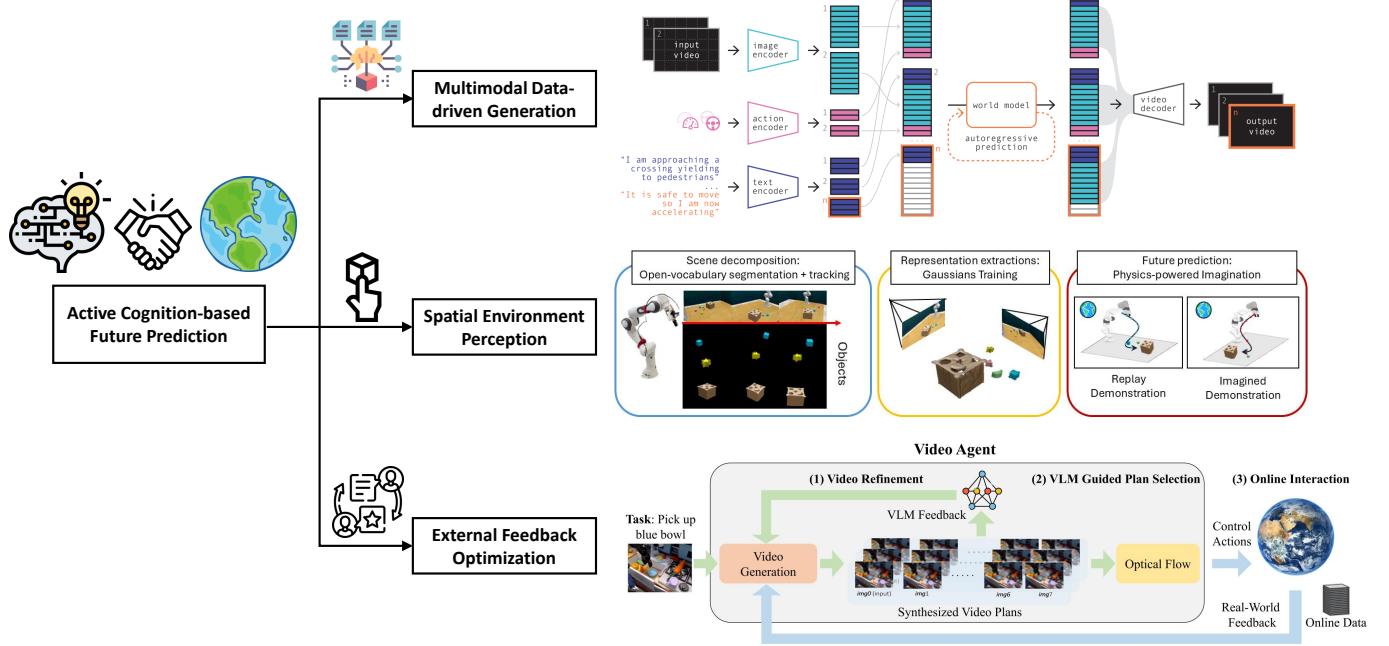


Fig. 12. The model actively interacts with the environment to achieve future prediction through multimodal data-driven generation, spatial environment perception, and external feedback optimization.

By combining implicit action models, video tokenizers, and dynamic models, Genie extracts latent actions, compresses video frames into discrete tokens, and predicts subsequent video tokens through the dynamics model. Genie 2 [207] extends this capability into 3D scenarios, generating dynamic virtual scenes from textual prompts or images. Leveraging Imagen 3 [208], an advanced text-to-image model, and incorporating Classifier-Free Guidance (CFG), Genie 2 achieves flexible and precise scene generation, simulating physical effects such as fluid dynamics, smoke, lighting, gravity, and more. UniSim [116] recognizes “interacting with the environment” as a key trend and introduces a universal world-generation simulator conditioned on actions. It unifies various action modalities (e.g., language instructions, robot controls, and camera movements) into a shared action space, which serves as conditional prompts to guide a diffusion model in generating subsequent frames. UniSim provides a unified environment for training both high-level vision-language policies and low-level reinforcement learning strategies, bridging the gap between simulated and real-world environments. To enhance generative models’ causal control and intervention capabilities in the physical world, WorldDreamer [117] proposes a novel approach using vision-text-action triplet datasets for supervised training. Its U-ViT-based [209] STPT (Spatial Temporal Patchwise Transformer) integrates dynamic masked strategies, embedding visual and prompt information to generate results in approximately 10 steps, 3 to 20 times faster than diffusion models.

The application of generative world models in autonomous driving has become a widely recognized research direction. In particular, autonomous driving [118], [119], [210] requires safe and reliable decision-making in unstruc-

tured and complex scenarios. Pioneering work, such as GAIA-1 [119], encodes past images, text, and actions to predict the next image token, emphasizing a deep understanding of driving data. Building on this, DriveDreamer [118] introduces a two-stage generation strategy: in the first stage, it integrates various driving elements to comprehensively understand underlying traffic structures and construct accurate scene maps; in the second stage, it predicts future video frames to enable controllable generation of interactive driving scenarios. Unlike GAIA-1, which focuses on scene video generation, DriveDreamer places greater emphasis on the decision-making aspects of generating driving behaviors and scenes.

**Digitalized Physical World.** Despite these advances, Kang et al. [85] remain skeptical about the ability of vision-based world models to abstract physical laws. Studies involving diverse physical scene datasets evaluated models’ capabilities under in-distribution, out-of-distribution, and compositional generalization settings. Even with aggressive scaling, models struggled to reproduce correct physical phenomena, merely memorizing and imitating observed physics dynamics. Similarly, Motamed et al. [84] proposes to evaluate various general video generation models on the Physics-IQ dataset, and the conclusions obtained are similar to the view of Kang et al. [85]. These views indicate that vision-based representations alone are insufficient for precise physical modeling. Cosmos [87] seeks to digitize the physical world with the goal of creating “Physical AI”, leveraging generalist generative models and physical simulators to design controlled scenarios. These scenarios are employed to fine-tune the model’s understanding of physical attributes such as gravity, collisions, torque, and inertia. Cosmos enables generate high-quality, 3D-consistent videos

while advancing physical AI tasks. The World Simulation Assistant (WISA) [108] constructs a dataset of 32,000 videos with captions to fine-tune T2V models. It decomposes physics-based captions into textual descriptions, qualitative categories, and quantitative attributes for multidimensional interpretation. By integrating a Mixture-of-Physical-Experts Attention (MoPA) mechanism and a Physics Classifier, WISA enables independent understanding of physical properties. This structured approach effectively embeds physical principles into the model, enhancing its ability to generate physics-consistent videos. PISA [211] focuses on physics-based generation in free-fall scenarios, exploring a two-stage post-training strategy using Open-Sora [212] as the foundation model, incorporating physics-supervised fine-tuning and object reward optimization. Evaluation results demonstrate that fine-tuning an open-source model on a small dataset with this approach enables it to acquire new capabilities for generating more physically accurate videos.

## 6.2 Spatial Environment Perception

Traditional world models typically rely on semantic representations and lack scene-level spatiotemporal dynamic modeling, which may result in poor performance in interactive tasks. Especially in embodied environments, robots need to execute precise spatiotemporal actions, and relying solely on semantic abstractions may lead to physically infeasible decisions. Spatiotemporal perception-driven world models address the limitations of multimodal data-driven approaches in the spatial perception dimension, physical consistency, and the bottleneck of diverse data imagination, by leveraging structured spatiotemporal representations.

**Spatial Perception.** Spatial perception, based on geometric priors, enables 3D scene reconstruction, overcoming the inherent planar bias of traditional 2D video generation. ManiGaussian [74] extends this approach by predicting robotic actions from geometric, semantic, and dynamic perspectives, propagating Gaussian particles over time to capture spatiotemporal scene dynamics. It supports language-controlled agents by dynamically propagating semantic features within a Gaussian distribution. Additionally, it constructs a Gaussian world model to parameterize the dynamic Gaussian splatting framework, leveraging real-world scenes to supervise the learning of Gaussian deformation fields, thereby predicting future scenarios. Realistic driving video generation must also adhere to fundamental physical laws, such as absolute and relative motion and spatial relationships. The key insight of DrivePhysica [75] is aligning ego-vehicle coordinates with global world coordinates to achieve complementary perspectives while capturing the relative motion of surrounding objects to generate instance flows. By embedding 3D bounding box coordinates, DrivePhysica incorporates depth information to preserve correct spatial relationships. By integrating critical physical principles, DrivePhysica generates high-quality, multi-view driving videos.

**Physics Consistency.** Physics consistency constraint modeling explicitly encodes the physical laws in space, providing a corrigible framework of physical constraints for feedback regulation in the perception, planning, and control modules, ensuring that the model's decisions align with

the action logic of the real world. Abou-Chakra et al. [114] models robotic environments as 3D Gaussian distributions representing visual states and introduces position-based dynamics (PBD) [161] simulation. This enables the model to predict future states and perform real-time corrections under strict physical constraints. By embedding explicit physical priors (particles) into the 3D Gaussian representation, the model utilizes visual feedback to adjust Gaussian distributions, thereby refining particle positions. This approach allows robots to robustly understand physical laws and synchronize physical simulation with visual feedback, advancing perception, planning, and control algorithms.

**Diverse Data Imagination.** Most of the aforementioned methods require a large volume of real-world samples; however, in practice, embodied data is often scarce. Developing a world model capable of comprehensive learning under limited sample conditions has become a significant challenge. DreMa [115] introduces an innovative and valuable compositional manipulative world model designed to create diverse realistic environments (accommodating the motion dynamics and physical properties) for robots. By leveraging generative simulations and physical simulators, DreMa generates physically plausible, novel, and imagined dynamic demonstrations (with specific equivariant transformations applied). This enables robots to achieve significant improvements in imitation learning with limited data. By extending the applicability of world models beyond training domains, DreMa narrows the gap with real-world environments. Moreover, reconstructing the complexity of dynamic interactive driving scenes remains a significant challenge. Considering the natural advantages of generative world models in producing diverse and controllable high-fidelity 2D videos, DriveDreamer4D [76] innovatively leverages world model priors to advance autonomous driving 4D scene reconstruction. Specifically, it adjusts original trajectories to generate new trajectories, guiding the world model to produce diverse dynamic data (e.g., lane changes, acceleration, and deceleration). Then, it integrates real and synthetic data to optimize the performance of the 4D scene generation model.

## 6.3 External Feedback Optimization

External Feedback-based video generation dynamically adjusts the model's generation of the world by incorporating external environmental knowledge or real-time environmental signals. The core idea is to leverage feedback information from sources external to the model's own training data to ensure that the generated content aligns with specific domain physical laws, semantic logic, or human preferences, thereby achieving an active cognitive closed-loop optimization of “generation-environment observation-correction update.” Tab. 3 presents the implementation methods and characteristics of different external supervision feedback approaches.

**Training the Reward Model through Human Preference Annotations.** To address the physical illusions in dynamic interactive scenes of video generation, the Iterative Preference Optimization (IPO) [109] framework trains a reward model that automatically evaluates the physical plausibility of generated videos and then iteratively optimizes the generation model based on these physics-based

Methods	Reward Model	Feedback Type	Optimization Technique	Optimization Direction
IPO [109]	Human-Annotated Training VLM	Pair-wise & Point-wise feedback	Diffusion-DPO [213] & Diffusion-KTO [214]	Subject consistency, motion smoothness and aesthetic quality
VideoReward [110]	Human-Annotated Training VLM	Pair-wise feedback	Flow-DPO& Flow-RWR&Flow-NRG	Visual quality, motion quality, and text alignment
Furuta et al. [82]	Gemini-1.5-Pro [215] & Metric	Pair-wise & Point-wise feedback	RWR] [216] & DPO [217]	Overall coherence, physical accuracy, task completion, and the existence of inconsistencies
VideoAgent [111]	GPT4-turbo [202] & Online Execution Feedback	Binary value {0, 1}	Consistency models [218] & Online finetuning	Trajectory smoothness, physical stability and achieving the goal
Gen-Drive [112]	GPT-4o [202]	Pair-wise feedback	DDPO [219]	Complex traffic environment, scene consistency and interactive dynamics
PhyT2V [113]	GPT-4o [202]	Mismatch between video semantics and prompts	LLM global step-back reasoning	Adherence physical rules

TABLE 3  
Overview of the characteristics of various methods in external feedback optimization.

preference feedbacks. VideoReward [110] proposed a large-scale human preference dataset consisting of 182k annotations, covering human evaluations of video generation in terms of visual quality, motion quality, and text alignment. This dataset is used to train the VLM reward model, where multidimensional reward tokens are decoupled to maintain contextual independence across evaluation levels. Finally, through exploring alignment algorithms-two training-time strategies (Flow-DPO [213] and Flow-RWR [220]) and one inference-time technique (reward-guided)-video alignment is achieved.

**Replacing Human Feedback with AI Models.** However, the cost of manual annotation is high. To solve this problem, some methods replace manual annotation with AI agents to obtain human preference feedback. Furuta et al. [82] combines RL fine-tuning strategies to accept external feedback and iteratively optimize object motion to match the real world. This approach explores feedback mechanisms based on metrics from complex motion scenarios, as well as AI-driven feedback from VLMs, which serve as a substitute for human preferences. The feedback system is designed to evaluate generated videos in terms of overall coherence, physical accuracy, and task completion. Experimental results show that AI feedback serves as the best proxy for human preferences. Similarly, VideoAgent [111] explores two feedback mechanisms for iteratively refining generative models to mitigate physical hallucinations. The first mechanism leverages VLM to assess video quality, guiding iterative refinement and filtering generated videos to produce action candidates for online execution. The second mechanism involves interactive feedback from real-world environments, where generated videos are translated into action control policies, executed in the physical world, and used to collect additional environmental data for further improvement. This approach introduces an offline-online feedback paradigm that enables video refinement without requiring large-scale data. In the autonomous driving interaction scene prediction, Gen-Drive [112] transforms the traditional prediction-planning paradigm into a generation-

evaluation prediction paradigm, achieving feedback fine-tuning of the generative model by introducing a reward model that selects human preferences. PhyT2V [113] employs LLMs to achieve a three-step iterative optimization process: 1) parsing textual prompts to extract objects and underlying physical rules; 2) back-inferencing the semantics of the generated video and evaluating its alignment with the original prompt, which serves as a reward signal for the LLM; 3) leveraging the derived rewards and physical constraints to reconstruct and refine the textual prompt. This closed-loop framework surpasses the limitations of traditional text-to-video (T2V) techniques by incorporating a self-optimization mechanism, thereby enhancing the physical plausibility and accuracy of generated content.

**Discussion.** Overall, actively cognitive-driven world simulation aims to achieve the simulation and prediction of the physical world by enabling models to interact with the environment, thereby facilitating a more comprehensive understanding of physical dynamics. This approach significantly improves generation efficiency, counterfactual prediction accuracy, and generalization ability. However, it heavily relies on large-scale datasets and lacks dedicated foundation models for physical understanding. To address these challenges, future research could overcome the dependence on large-scale datasets by incorporating more diverse training data, as suggested in [85]. Additionally, enhancing the integration and synergy of multi-sensor data could enable a more comprehensive perception of the surrounding environment, as proposed in [221]. Another promising direction is the development of large-scale physics foundation models to further enhance the model's ability to understand and reason about physical phenomena.

## 7 BENCHMARKS AND METRICS

Existing video evaluation benchmarks [224] and metrics (e.g., PSNR [225], SSIM [226], FVD [227]) primarily focus on assessing pixel-level similarity or visual and semantic quality, yet fail to effectively measure whether the video content adheres to physical laws (e.g., gravity, collision,

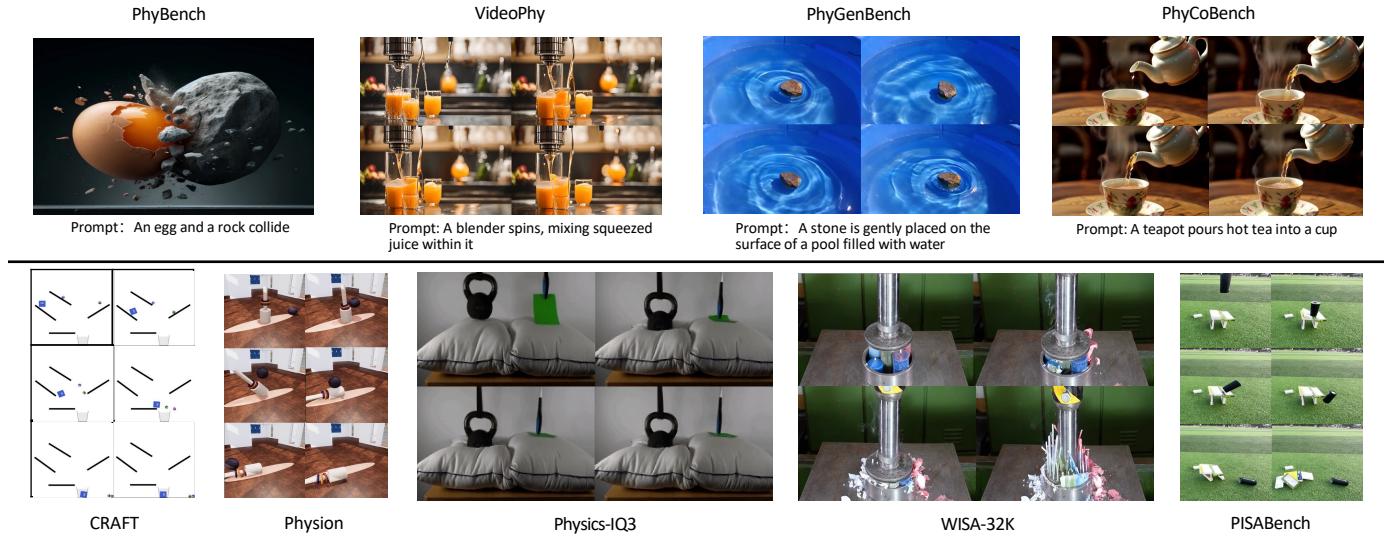


Fig. 13. Physics benchmark dataset examples.

Dataset	Physics Categories					Prompts	Videos	Scenarios
	Mechanics	Optics	Thermal	Material properties	Magnetism			
Caption Level	PhyBench [105]	✓	✓	✓	✓	700	-	31
	VideoPhy [90]	✓			✓	688	-	3
	PhyGenBench [86]	✓	✓	✓	✓	160	-	27
	PhyCoBench [104]	✓			✓	120	-	7
Video Level	CRAFT [222]	✓			✓	-	58K	20
	Physion [107]	✓			✓	-	17K	8
	Physion++ [223]	✓			✓	-	9.5K	9
	LLMPH [121]	✓			✓	-	100	
	Physics-IQ [84]	✓	✓	✓	✓	✓	-	396
Caption & Video	WISA-32K [108]	✓	✓	✓	✓	32000	32000	17
	PisaBench [211]	✓				361	361	1

TABLE 4  
Comprehensive benchmark dataset based on physical rules.

fluid dynamics). Due to the lack of evaluation for physical plausibility, models may over-optimize visual aesthetics (e.g., texture, color, spatiotemporal consistency) while neglecting the modeling of physical principles, resulting in generated content that “appears realistic but is physically implausible.” Although general-purpose large-scale video datasets enable the training of versatile models, the scarcity of videos with physical properties often leads to suboptimal performance in generating physically plausible videos [84]. In this section, we introduce the benchmarks and metrics used to evaluate the physical fidelity of video generation models.

## 7.1 Benchmarks

Existing physics benchmarks can be categorized into caption-level, video-level and caption&video level benchmarks. Fig. 13 presents examples of physics benchmarks. In addition, Tab. 4 provides a comparative analysis of various benchmarks.

**Benchmarks for Caption Level.** Meng et al. [105] introduced PhyBench, a comprehensive dataset designed to evaluate the understanding of physical commonsense in text-to-image (T2I) models. The dataset encompasses four major categories: mechanics, optics, thermodynamics, and material properties, comprising 31 physical scenarios and 700 textual prompts. Each scenario is enriched with fine-grained physical principles derived from textbook knowledge and GPT-4o. Bansal et al. [90] proposed the VideoPhy dataset, which consists of 688 textual prompts (with an average caption length of 8.5 words), covering three types of physical dynamic interactions: rigid body-rigid body, rigid body-fluid, and fluid-fluid. The dataset provides a diverse range of visual concepts and action descriptions. Compared to VideoPhy [90], which only describes physical phenomena in text, VideoPhy-2 [228] provides explicit annotations of physical rules. It selects 197 real-world actions related to physical commonsense(such as object interactions, sports, and physical activities) and uses LLMs to generate 3940 text

prompts showcasing specific physical interactions. These prompts guide video generation, from which candidate physical rules are inferred. PhyGenBench [86] is a benchmark specifically designed to evaluate the physical commonsense understanding of generative models. It comprises 160 textual prompts, covering 27 representative physical laws across four domains, and providing comprehensive coverage of physical phenomena. Its scope extends beyond VideoPhy [90], which focuses solely on simple interactions between rigid bodies and fluids. PhyCoBench [104] includes seven observable physical phenomena (e.g., Newton's laws, conservation principles, collisions) and 120 corresponding benchmark evaluation prompts.

**Benchmarks for Video Level.** In addition to evaluating the physical video generation capabilities of models in text-to-video (T2V) tasks, some studies have also benchmarked their ability to understand and reason about physical dynamics. In tasks involving the prediction of conditional continuation frames (I2V/V2V), models are required to demonstrate a profound understanding of physical laws, going beyond the mere reproduction of scenes from pre-trained memory, thereby showcasing strong generalization capabilities. CRAFT [222] is designed to evaluate models' understanding of physical forces and causal relationships between objects. It utilizes a 2D physics simulator to generate virtual scenes with 20 distinct layouts, each producing a 10-second video clip. Physion [107] evaluates models' comprehension of eight physical phenomena, covering rigid-soft body interactions and complex multi-component interactions. Evaluation metrics include overall accuracy, the correlation between model outputs and human responses (Pearson correlation coefficient), and Cohen's kappa index. Building upon Physion, Physion++ [223] extends the benchmark by incorporating additional mechanical properties, such as mass, friction, elasticity, and deformability, offering a more comprehensive evaluation framework. LLMPH [121] employs a black-box optimization approach, integrating the physical knowledge of LLMs with the simulation capabilities of physics engines to form feedback loops. Additionally, the created TraySim dataset comprises 100 simulated scenarios, focusing on the task of predicting stable poses of object instances during complex interactions. The method proposed by Kang et al. [85] generates videos governed by classical mechanics laws, such as uniform motion, elastic collisions, and parabolic motion, aiming to investigate whether video generation models can discover physical laws through learning from video data. It also evaluates their performance in in-distribution (ID), out-of-distribution (OOD), and compositional generalization scenarios. The aforementioned benchmarks [85], [107], [121], [222], [223] for physical video evaluation are all synthesized using physics simulators, highlighting the need for real-world videos capturing diverse and complex physical phenomena to address this limitation. The Physics-IQ [84] dataset encompasses five domains: solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism, comprising a total of 396 high-quality videos (66 scenes  $\times$  3 perspectives  $\times$  2 recordings). Each video lasts 8 seconds and covers 66 distinct physical scenarios. Physics-IQ provides the research community with comprehensive and high-quality real-world physical interactions, and is expected to

significantly advance the development of video generation models in terms of physical realism.

**Benchmarks for Caption&Video.** The WISA-32K dataset [108] manually collects 32,000 video samples covering three physical categories (dynamics, thermodynamics, and optics), with detailed physical annotations generated using GPT-4o mini. These annotations are categorized into textual physical descriptions, qualitative physical categories, and quantitative physical properties for subsequent model fine-tuning and training inputs. For example, a textual description such as "A large-scale explosion generates massive smoke and dust" has the qualitative physical categories of "gas motion, explosion phenomena, etc." and the quantitative physical properties of "Density: debris: 1 to  $2.5 \text{ g/cm}^3$ ". PisaBench [211] focuses on simple drop tasks to evaluate the ability of generative models to produce accurate physical phenomena. This benchmark consists of 361 real-world free-fall videos, capturing physical properties such as gravity and dynamic collisions, along with manually annotated captions. Additionally, SAM2 [229] is used to generate segmentation masks for all objects in the videos.

In summary, these datasets collectively provide comprehensive support for research on physical plausibility in video generation, spanning tasks from text-to-video generation to frame sequence prediction, thereby driving profound advancements in the field. However, nearly all benchmarks have reached a similar conclusion: current video generation models still fall short of fully capturing physical laws [84], [85], [105], [228]. As a result, existing generative models remain far from becoming true world simulators.

## 7.2 Metrics

Apart from the high-cost human evaluation, existing automated methods for assessing physical fidelity can be broadly categorized into quantitative score-based approaches [84], [104], [211] and automated evaluations leveraging VLMs [86], [105], [106]. Quantitative metrics provide explicit numerical computing assessments of a model's adherence to physical principles, while VLM-based evaluations enable a more flexible assessment by incorporating high-level reasoning and contextual understanding. In this section, we discuss two major paradigms of physical consistency evaluation: Quantitative Score-Based Evaluation and VLM-based Automatic Evaluation.

**Quantitative Score-Based Evaluation.** PhyCoPredictor [104] is a tool for automatically evaluating the physical consistency of generative models. This approach first constructs a flow-guided generative model and trains it across diverse motion scenarios. Performance evaluation is then conducted by comparing the optical flow and videos generated by the model under assessment with those produced by the reference model. Motamed et al. [84] aim to quantify the discrepancy between generated and real videos from multiple perspectives, including spatial IoU, spatiotemporal IoU, weighted spatial IoU, and Mean Squared Error (MSE). These four metrics are integrated into a single score, the Physical IQ score, which comprehensively tracks the model's capability in physical understanding and generation. PISA [211] introduces three spatial metrics to evaluate state-of-the-art I2V models in a fundamental physical scenario-free fall.

This approach assesses the accuracy of trajectories, shape fidelity, and object persistence by computing the Trajectory L2, Chamfer Distance (CD), and Intersection Over Union (IoU) between generated and real videos, respectively.

**VLM-based Automatic Evaluation.** Meng et al. [105] proposed PhyEvaler, an automated evaluation framework based on GPT-4o [202], which generates images from input prompts and assesses them based on scene accuracy and physical correctness. This precise design of physical text prompts can also be extended to video generation tasks. VideoCon-Physics [106] leverages the VIDEOCON [106] model and is fine-tuned using human feedback on Semantic Adherence (SA) and Physical Commonsense (PC) to achieve more accurate assessment. SA evaluates whether the video accurately depicts the entities, actions, and relationships described in the textual prompt (e.g., a red sphere rolling and colliding with a blue cube). PC assesses whether the video adheres to fundamental physical commonsense (e.g., after the collision, the sphere and the cube move in opposite directions, following the principle of momentum conservation). The accompanying PhyGenEval [86] framework, similar to VideoCon-Physics, also evaluates from two dimensions: SA and PC. Leveraging GPT-4o, it performs key physical frame detection, physical sequence verification, and overall naturalness analysis, enabling a hierarchical assessment of physical commonsense consistency. Furthermore, based on the SA and PC scores, VideoPhy-2-Autoeval [228] introduces physics rule classification (evaluating whether the generated videos comply with or violate specific physical laws), thereby improving the accuracy of video generation assessment. In the construction of the automated evaluator, VideoPhy-2-Autoeval integrates human assessment knowledge to fine-tune the aforementioned VideoCon-Physics evaluation model [106], enabling automated evaluation of the physical consistency of generated videos.

## 8 PROSPECTS AND CHALLENGES

As video generation advances toward more realistic world simulators and world models, the development of generative systems with physical cognition capabilities presents significant potential. However, several key challenges remain to be addressed.

**Building Large Physics Foundation Models.** While current general-purpose LLMs have demonstrated cognitive breakthroughs in multi-modal understanding and reasoning, their ability to conduct in-depth research in scientific physics remains limited. Therefore, developing and evaluating dedicated large physics models (LPMs) represents a highly promising direction. Barman et al. [230] provides a potential roadmap for designing physics-specific LLMs. By integrating physics-specific knowledge into these models, the exploration of LLMs' problem-solving and innovation capabilities in the field of physics can be facilitated. Leveraging LPMs will significantly advance the development of generative models and has the potential to give rise to the next generation of world models with physical reasoning and evolutionary capabilities.

**Advancing Physical Fidelity in World Simulators.** The development of world simulators aims to reproduce, pre-

dict, or reason about complex real-world phenomena. This requires precise modeling of physical environments, object interactions, and dynamic behaviors to ensure both physical consistency and interpretability in simulations. Future works can focus on: 1) integrating LLMs or physics engines with generative models [231], allowing the system to extract and embed physical knowledge from LLMs and physics engines into the video generation process, thereby enhancing physical realism. 2) incorporating reinforcement learning-based fine-tuning techniques and human feedback [6], [110], where physical rule feedback is introduced during training to dynamically correct the generative system, improving its adherence to physical laws and generalization capabilities.

**Incorporating Multi-Sensor Data.** Physical perception extends beyond vision and text, encompassing various forms of physical data from mechanics, thermodynamics, and material science (e.g., vibration, temperature, and tactile feedback). By integrating a wider range of sensor-derived physical data into models, this approach enables multi-level physical information embedding and allows for more comprehensive interactions with the environment. The combination of these enriched representations enhances the model's ability to understand and describe complex real-world physical phenomena.

**Data Scarcity and the Sim2Real Gap.** Although video generation systems based on physical cognition have made significant progress, they often rely on large-scale, high-quality physical scene data (either synthetic or real) to capture underlying physical patterns. At the same time, embodied agent learning (such as robotics and autonomous driving) also relies on high-precision, large-scale physical data to enable reasoning, planning, situational learning, and other AI applications. However, collecting such data is typically time-consuming and labor-intensive. Moreover, the diversity of physical phenomena necessitates a corresponding diversity in training data. Therefore, a key challenge moving forward is developing efficient methods for synthesizing large-scale, physically faithful and diverse video datasets for model training and evaluation. A promising solution is leveraging high-fidelity physics simulators, such as Cosmos [87], to generate large-scale synthetic data, which can then be used to enhance world model training. However, it is important to acknowledge that a gap still exists between simulated environments and the real world, and bridging this gap remains an open challenge.

**Efficiency of Physical Simulation.** Physical simulators are used to solve physical states and predict target interaction states, but this approach often involves frequent numerical computations during the simulation process, resulting in significant computational overhead. This makes real-time simulation particularly challenging. Potential solutions include: 1) GPU acceleration for parallel simulation. For example, Genesis [160], which integrates various physical solvers, combines GPU parallel acceleration to achieve unprecedented simulation speeds. 2) Designing efficient model layers, such as replacing the full attention layer in Transformer with a linear attention layer [11]. 3) Distillation acceleration, distilling teacher model with complex arithmetic into computationally efficient student model [9].

**Physical Quality Assessment.** Existing methods [23], [64], [212] usually use some pixel/visual numerical metrics

to evaluate the performance of the model, such as FID, SSIM, PSNR and FVD. However, these numerical metrics usually cannot fully demonstrate the advantages and disadvantages of the model, and are often inconsistent with human preferences. Therefore, recent methods have started to introduce human evaluation or benchmarks from multiple perspectives [224] to comprehensively and accurately evaluate the model. Nevertheless, existing methods either cannot evaluate the physical quality or require a lot of manpower. A feasible idea is to develop an automatic physical quality evaluation approach in combination with cutting-edge multimodal large language models (MLLM) to fully mine their powerful abilities of physical understanding, such as PhyEvalu [105]. At the same time, a powerful physical evaluator can replace human preferences as a reward model to optimize the generative model (such as Videoagent [111]), making the generated content more consistent with physical laws.

## 9 CONCLUSION

In this survey, we provide a comprehensive overview of the latest advancements in physics cognition-based video generation. We categorize existing research based on the evolutionary progression of physical cognition-ranging from schematic perception to passive and to active cognition-and offer an in-depth discussion of each category. Furthermore, we summarize the available datasets and commonly used evaluation metrics. Despite the rapid progress in this field, significant challenges remain, warranting further exploration. Looking ahead, with the advancement of AGI, physics-faithful video generation models are expected to play a crucial role as world simulators, becoming an indispensable component in the pathway toward AGI realization.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No.623B2039 and U22B2053), STI 2030-Major Projects (2022ZD0208800), NSFC General Program (Grant No. 62176215).

## REFERENCES

- [1] OpenAI, "Video generation models as world simulators," 2024. [Online]. Available: <https://openai.com/research/video-generation-models-as-world-simulators>
- [2] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. Bechtle, F. Behbahani, S. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. de Freitas, S. Singh, and T. Rocktäschel, "Genie: Generative interactive environments," 2024. [Online]. Available: <https://arxiv.org/abs/2402.15391>
- [3] Runway, "Introducing Gen-3 Alpha: A new frontier for video generation," 2024.
- [4] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang, "UniAnimate: Taming unified video diffusion models for consistent human image animation," *Science China Information Sciences*, 2025.
- [5] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, "Video-P2P: Video editing with cross-attention control," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [6] H. Yuan, S. Zhang, X. Wang, Y. Wei, T. Feng, Y. Pan, Y. Zhang, Z. Liu, S. Albanie, and D. Ni, "InstructVideo: Instructing video diffusion models with human feedback," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [7] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, and J. Zhou, "I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models," *arXiv preprint arXiv:2311.04145*, 2023.
- [8] Z. Qing, S. Zhang, J. Wang, X. Wang, Y. Wei, Y. Zhang, C. Gao, and N. Sang, "Hierarchical spatio-temporal decoupling for text-to-video generation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [9] X. Wang, S. Zhang, H. Zhang, Y. Liu, Y. Zhang, C. Gao, and N. Sang, "VideoLCM: Video latent consistency model," *arXiv preprint arXiv:2312.09109*, 2023.
- [10] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "ModelScope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.
- [11] H. Wang, C.-Y. Ma, Y.-C. Liu, J. Hou, T. Xu, J. Wang, F. Juefei-Xu, Y. Luo, P. Zhang, T. Hou *et al.*, "LinGen: Towards high-resolution minute-length text-to-video generation with linear computational complexity," *arXiv preprint arXiv:2412.09856*, 2024.
- [12] Y. Hou, L. Zheng, and P. Torr, "Learning camera movement control from real-world drone videos," *arXiv preprint arXiv:2412.09620*, 2024.
- [13] Z. Lin, W. Liu, C. Chen, J. Lu, W. Hu, T.-J. Fu, J. Allardice, Z. Lai, L. Song, B. Zhang *et al.*, "STIV: Scalable text and image conditioned video generation," *arXiv preprint arXiv:2412.07730*, 2024.
- [14] S. Chen, C. Ge, Y. Zhang, Y. Zhang, F. Zhu, H. Yang, H. Hao, H. Wu, Z. Lai, Y. Hu *et al.*, "Goku: Flow based video generative foundation models," *arXiv preprint arXiv:2502.04896*, 2025.
- [15] X. Wang, C. Gao, Y. Wang, and N. Sang, "Replace anyone in videos," *arXiv preprint arXiv:2409.19911*, 2024.
- [16] S. Liu, T. Wang, J.-H. Wang, Q. Liu, Z. Zhang, J.-Y. Lee, Y. Li, B. Yu, Z. Lin, S. Y. Kim *et al.*, "Generative video propagation," *arXiv preprint arXiv:2412.19761*, 2024.
- [17] Y. Wu, Z. Zhang, Y. Li, Y. Xu, A. Kag, Y. Sui, H. Coskun, K. Ma, A. Lebedev, J. Hu *et al.*, "SnapGen-V: Generating a five-second video within five seconds on a mobile device," *arXiv preprint arXiv:2412.10494*, 2024.
- [18] T. Yin, Q. Zhang, R. Zhang, W. T. Freeman, F. Durand, E. Shechtman, and X. Huang, "From slow bidirectional to fast causal video generators," *arXiv preprint arXiv:2412.07772*, 2024.
- [19] M. Zheng, Y. Xu, H. Huang, X. Ma, Y. Liu, W. Shu, Y. Pang, F. Tang, Q. Chen, H. Yang *et al.*, "VideoGen-of-Thought: A collaborative framework for multi-shot video generation," *arXiv preprint arXiv:2412.02259*, 2024.
- [20] A. Melnik, M. Ljubljana, C. Lu, Q. Yan, W. Ren, and H. Ritter, "Video diffusion models: A survey," *arXiv preprint arXiv:2405.03150*, 2024.
- [21] Y. HaCohen, N. Chirrut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon *et al.*, "LTX-Video: Realtime video latent diffusion," *arXiv preprint arXiv:2501.00103*, 2024.
- [22] D. Geng, C. Herrmann, J. Hur, F. Cole, S. Zhang, T. Pfaff, T. Lopez-Guevara, C. Doersch, Y. Aytar, M. Rubinstein *et al.*, "Motion prompting: Controlling video generation with motion trajectories," *arXiv preprint arXiv:2412.02700*, 2024.
- [23] B. Lin, Y. Ge, X. Cheng, Z. Li, B. Zhu, S. Wang, X. He, Y. Ye, S. Yuan, L. Chen *et al.*, "Open-Sora plan: Open-source large video generation model," *arXiv preprint arXiv:2412.00131*, 2024.
- [24] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang *et al.*, "Movie Gen: A cast of media foundation models," *arXiv preprint arXiv:2410.13720*, 2024.
- [25] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding, "VDT: general-purpose video diffusion transformers via mask modeling," in *Proceedings of the International Conference on Learning Representations*, 2024.
- [26] X. Wang, S. Zhang, H. Yuan, Z. Qing, B. Gong, Y. Zhang, Y. Shen, C. Gao, and N. Sang, "A recipe for scaling up text-to-video generation with text-free videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6572–6582.

- [27] X. Chen, Z. Liu, M. Chen, Y. Feng, Y. Liu, Y. Shen, and H. Zhao, "LivePhoto: Real image animation with text-guided motion control," in *Proceedings of the European Conference on Computer Vision*. Springer, 2024, pp. 475–491.
- [28] H. Qiu, M. Xia, Y. Zhang, Y. He, X. Wang, Y. Shan, and Z. Liu, "FreeNoise: Tuning-free longer video diffusion via noise rescheduling," in *Proceedings of the International Conference on Learning Representations*, 2024.
- [29] S. Yang, L. Hou, H. Huang, C. Ma, P. Wan, D. Zhang, X. Chen, and J. Liao, "Direct-a-video: Customized video generation with user-directed camera movement and object motion," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.
- [30] H. Jeong, G. Y. Park, and J. C. Ye, "VMC: Video motion customization using temporal attention adaption for text-to-video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9212–9221.
- [31] Z. Xing, Q. Dai, H. Hu, Z. Wu, and Y.-G. Jiang, "SimDA: Simple diffusion adapter for efficient video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7827–7839.
- [32] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, J.-W. Liu, W. Wu, J. Keppo, and M. Z. Shou, "MotionDirector: Motion customization of text-to-video diffusion models," in *Proceedings of the European Conference on Computer Vision*. Springer, 2024, pp. 273–290.
- [33] Y. Zeng, G. Wei, J. Zheng, J. Zou, Y. Wei, Y. Zhang, and H. Li, "Make pixels dance: High-dynamic video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8850–8860.
- [34] J. Xing, M. Xia, Y. Zhang, H. Chen, W. Yu, H. Liu, G. Liu, X. Wang, Y. Shan, and T.-T. Wong, "DynamicalCrafter: Animating open-domain images with video diffusion priors," in *Proceedings of the European Conference on Computer Vision*. Springer, 2024, pp. 399–417.
- [35] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu, "VACE: All-in-one video creation and editing," *arXiv preprint arXiv:2503.07598*, 2025.
- [36] G. Zheng, T. Li, R. Jiang, Y. Lu, T. Wu, and X. Li, "CamI2V: Camera-controlled image-to-video diffusion model," *arXiv preprint arXiv:2410.15957*, 2024.
- [37] W. Sun, R.-C. Tu, J. Liao, and D. Tao, "Diffusion model-based video editing: A survey," *arXiv preprint arXiv:2407.07111*, 2024.
- [38] Y. Deng, R. Wang, Y. Zhang, Y.-W. Tai, and C.-K. Tang, "DragVideo: Interactive drag-style video editing," in *Proceedings of the European Conference on Computer Vision*. Springer, 2024, pp. 183–199.
- [39] Z. Zhang, J. Liao, M. Li, Z. Dai, B. Qiu, S. Zhu, L. Qin, and W. Wang, "Tora: Trajectory-oriented diffusion transformer for video generation," *arXiv preprint arXiv:2407.21705*, 2024.
- [40] S. Bahmani, I. Skorokhodov, A. Siarohin, W. Menapace, G. Qian, M. Vasilkovsky, H.-Y. Lee, C. Wang, J. Zou, A. Tagliasacchi *et al.*, "VD3D: Taming large video diffusion transformers for 3D camera control," *arXiv preprint arXiv:2407.12781*, 2024.
- [41] D. J. Zhang, D. Li, H. Le, M. Z. Shou, C. Xiong, and D. Sahoo, "Moonshot: Towards controllable video generation and editing with multimodal conditions," *arXiv preprint arXiv:2401.01827*, 2024.
- [42] E. Soleimani and G. Khodabandeh, "A survey of emerging approaches and advances in video generation," 2024.
- [43] R. Sun, Y. Zhang, T. Shah, J. Sun, S. Zhang, W. Li, H. Duan, B. Wei, and R. Ranjan, "From Sora what we can see: A survey of text-to-video generation," *arXiv preprint arXiv:2405.10674*, 2024.
- [44] J. Liu, J. Zhu, L. Gao, H. T. Shen, and J. Song, "AICL: Action in-context learning for video diffusion model," *arXiv preprint arXiv:2403.11535*, 2024.
- [45] Y. He, S. Yang, H. Chen, X. Cun, M. Xia, Y. Zhang, X. Wang, R. He, Q. Chen, and Y. Shan, "ScaleCrafter: Tuning-free higher-resolution visual generation with diffusion models," in *Proceedings of the International Conference on Learning Representations*, 2023.
- [46] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang, X. Wang, C. Weng, Y. Shan *et al.*, "Animate-a-story: Storytelling with retrieval-augmented video generation," *arXiv preprint arXiv:2307.06940*, 2023.
- [47] WanTeam, "Wan: Open and advanced large-scale video generative models," 2025.
- [48] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "CogVideoX: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.
- [49] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.
- [50] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning," in *Proceedings of the International Conference on Learning Representations*, 2024.
- [51] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7623–7633.
- [52] W. Chen, Y. Ji, J. Wu, H. Wu, P. Xie, J. Li, X. Xia, X. Xiao, and L. Lin, "Control-A-Video: Controllable text-to-video diffusion models with motion prior and reward feedback learning," *arXiv preprint arXiv:2305.13840*, 2023.
- [53] H. Ni, B. Egger, S. Lohit, A. Cherian, Y. Wang, T. Koike-Akino, S. X. Huang, and T. K. Marks, "TI2V-Zero: Zero-shot image conditioning for text-to-video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9015–9025.
- [54] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "ST-Adapter: Parameter-efficient image-to-video transfer learning," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, pp. 26462–26477, 2022.
- [55] Y. Hu, C. Luo, and Z. Chen, "Make it move: controllable image-to-video generation with text descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18219–18228.
- [56] W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, W. Huang, and W. Chen, "ConsistI2V: Enhancing visual consistency for image-to-video generation," *arXiv preprint arXiv:2402.04324*, 2024.
- [57] G. Lei, C. Wang, H. Li, R. Zhang, Y. Wang, and W. Xu, "AnimateAnything: Consistent and controllable animation for video generation," *arXiv preprint arXiv:2411.10836*, 2024.
- [58] X. Guo, M. Zheng, L. Hou, Y. Gao, Y. Deng, P. Wan, D. Zhang, Y. Liu, W. Hu, Z. Zha *et al.*, "I2V-Adapter: A general image-to-video adapter for diffusion models," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.
- [59] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, G. Schindler, R. Hornung, V. Birodkar, J. Yan, M. Chiu, K. Somandepalli, H. Akbari, Y. Alon, Y. Cheng, J. V. Dillon, A. Gupta, M. Hahn, A. Hauth, D. Hendon, A. Martinez, D. Minnen, M. Sirotenko, K. Sohn, X. Yang, H. Adam, M. Yang, I. Essa, H. Wang, D. A. Ross, B. Seybold, and L. Jiang, "VideoPoet: A large language model for zero-shot video generation," in *Proceedings of the International Conference on Machine Learning*, 2024.
- [60] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, G. Liu, A. Raj *et al.*, "Lumiere: A space-time diffusion model for video generation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.
- [61] B. Wu, C.-Y. Chuang, X. Wang, Y. Jia, K. Krishnakumar, T. Xiao, F. Liang, L. Yu, and P. Vajda, "Fairy: Fast parallelized instruction-guided video-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8261–8270.
- [62] Y. Wei, S. Zhang, H. Yuan, B. Gong, L. Tang, X. Wang, H. Qiu, H. Li, S. Tan, Y. Zhang *et al.*, "DreamRelation: Relation-centric video customization," *arXiv preprint arXiv:2503.07602*, 2025.
- [63] "Kling." [Online]. Available: <https://kling.kuaishou.com/>
- [64] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, "HunyuanVideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.
- [65] Y. Wei, S. Zhang, H. Yuan, X. Wang, H. Qiu, R. Zhao, Y. Feng, F. Liu, Z. Huang, J. Ye *et al.*, "DreamVideo-2: Zero-shot subject-driven video customization with precise motion control," *arXiv preprint arXiv:2410.13830*, 2024.
- [66] Y. Cong, M. Xu, C. Simon, S. Chen, J. Ren, Y. Xie, B. Rosenhahn, T. Xiang, and S. He, "FLATTEN: optical flow-guided attention for consistent text-to-video editing," in *Proceedings of the International Conference on Learning Representations*, 2024.

- [67] Z. Li, J. Shi, S. Bi, R. Zhu, K. Sunkavalli, M. Hašan, Z. Xu, R. Ramamoorthi, and M. Chandraker, "Physically-based editing of indoor scene lighting from a single image," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 555–572.
- [68] H.-Y. Hsu, Z.-H. Lin, A. Zhai, H. Xia, and S. Wang, "AutoVFX: Physically realistic video editing from natural language instructions," *arXiv preprint arXiv:2411.02394*, 2024.
- [69] S. Zhou, P. Yang, J. Wang, Y. Luo, and C. C. Loy, "Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2535–2545.
- [70] Q. Jiang, Q. L. Wang, L. H. Chi, X. H. Chen, Q. Y. Zhang, R. Zhou, Z. Q. Deng, J. S. Deng, B. B. Tang, S. H. Lv *et al.*, "TempDiff: Enhancing temporal-awareness in latent diffusion for real-world video super-resolution," in *Computer Graphics Forum*, vol. 43, no. 7. Wiley Online Library, 2024, p. e15211.
- [71] H. Xia, Z.-H. Lin, W.-C. Ma, and S. Wang, "Video2Game: Real-time interactive realistic and browser-compatible environment from a single video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4578–4588.
- [72] Z. Ren, Y. Wei, X. Guo, Y. Zhao, B. Kang, J. Feng, and X. Jin, "VideoWorld: Exploring knowledge learning from unlabeled videos," *arXiv preprint arXiv:2501.09781*, 2025.
- [73] H. Zhao, X. Liu, M. Xu, Y. Hao, W. Chen, and X. Han, "TASTE-Rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation," 2025. [Online]. Available: <https://arxiv.org/abs/2503.11423>
- [74] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Mani-Gaussian: Dynamic gaussian splatting for multi-task robotic manipulation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2025, pp. 349–366.
- [75] Z. Yang, X. Guo, C. Ding, C. Wang, and W. Wu, "Physical informed driving world model," 2024. [Online]. Available: <https://arxiv.org/abs/2412.08410>
- [76] G. Zhao, C. Ni, X. Wang, Z. Zhu, X. Zhang, Y. Wang, G. Huang, X. Chen, B. Wang, Y. Zhang, W. Mei, and X. Wang, "DriveDreamer4D: World models are effective data machines for 4D driving scene representation," 2024. [Online]. Available: <https://arxiv.org/abs/2410.13571>
- [77] A. Fu, Y. Zhou, T. Zhou, Y. Yang, B. Gao, Q. Li, G. Wu, and L. Shao, "Exploring the interplay between video generation and world models in autonomous driving: A survey," 2024. [Online]. Available: <https://arxiv.org/abs/2411.02914>
- [78] S. Liu, Z. Ren, S. Gupta, and S. Wang, "PhysGen: Rigid-body physics-grounded image-to-video generation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2025, pp. 360–378.
- [79] B. Lai, X. Dai, L. Chen, G. Pang, J. M. Rehg, and M. Liu, "LEGO: learning egocentric action frame generation via visual instruction tuning," in *Proceedings of the European Conference on Computer Vision*. Springer, 2024, pp. 135–155.
- [80] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang *et al.*, "Make-your-video: Customized video generation using textual and structural guidance," *IEEE Transactions on Visualization and Computer Graphics*, 2024.
- [81] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3D scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16750–16761.
- [82] H. Furuta, H. Zen, D. Schuurmans, A. Faust, Y. Matsuo, P. Liang, and S. Yang, "Improving dynamic object interactions in text-to-video generation with ai feedback," *arXiv preprint arXiv:2412.02617*, 2024.
- [83] S. Yang, J. Walker, J. Parker-Holder, Y. Du, J. Bruce, A. Barreto, P. Abbeel, and D. Schuurmans, "Video as the new language for real-world decision making," *arXiv preprint arXiv:2402.17139*, 2024.
- [84] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos, "Do generative video models learn physical principles from watching videos?" *arXiv preprint arXiv:2501.09038*, 2025.
- [85] B. Kang, Y. Yue, R. Lu, Z. Lin, Y. Zhao, K. Wang, G. Huang, and J. Feng, "How far is video generation from world model: A physical law perspective," 2024. [Online]. Available: <https://arxiv.org/abs/2411.02385>
- [86] F. Meng, J. Liao, X. Tan, W. Shao, Q. Lu, K. Zhang, Y. Cheng, D. Li, Y. Qiao, and P. Luo, "Towards world simulator: Crafting physical commonsense-based benchmark for video generation," *arXiv preprint arXiv:2410.05363*, 2024.
- [87] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding *et al.*, "Cosmos world foundation model platform for physical AI," *arXiv preprint arXiv:2501.03575*, 2025.
- [88] Z. Yang, X. Guo, C. Ding, C. Wang, and W. Wu, "Physical informed driving world model," *arXiv preprint arXiv:2412.08410*, 2024.
- [89] J. Lin, Z. Wang, Y. Hou, Y. Tang, and M. Jiang, "Phy124: Fast physics-driven 4D content generation from a single image," *arXiv preprint arXiv:2409.07179*, 2024.
- [90] H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, C. Jiang, Y. Sun, K.-W. Chang, and A. Grover, "VideoPhy: Evaluating physical commonsense for video generation," 2024. [Online]. Available: <https://arxiv.org/abs/2406.03520>
- [91] S. Meng, Y. Luo, and P. Liu, "Grounding creativity in physics: A brief survey of physical priors in AIGC," *arXiv preprint arXiv:2502.07007*, 2025.
- [92] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," *arXiv preprint arXiv:2303.04226*, 2023.
- [93] D. Liu, J. Zhang, A.-D. Dinh, E. Park, S. Zhang, and C. Xu, "Generative physical AI in vision: A survey," *arXiv preprint arXiv:2501.10928*, 2025.
- [94] P. Barrouillet, "Theories of cognitive development: From Piaget to today," pp. 1–12, 2015.
- [95] S. Gupta, A. Keshari, and S. Das, "RV-GAN: Recurrent gan for unconditional video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2024–2033.
- [96] C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, and L. Bai, "A survey on long video generation: Challenges, methods, and prospects," *arXiv preprint arXiv:2403.16407*, 2024.
- [97] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "VToonify: Controllable high-resolution portrait video style transfer," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 1–15, 2022.
- [98] P. Wang, K. Liu, and Q. Dougherty, "Conceptions of artificial intelligence and singularity," *Information*, vol. 9, no. 4, p. 79, 2018.
- [99] P. Voss and M. Jovanovic, "Why we don't have AGI yet," *arXiv preprint arXiv:2308.03598*, 2023.
- [100] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [101] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [102] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [103] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>
- [104] Y. Chen, X. Zhu, T. Li, H. Chen, and C. Shen, "A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction," 2025. [Online]. Available: <https://arxiv.org/abs/2502.05503>
- [105] F. Meng, W. Shao, L. Luo, Y. Wang, Y. Chen, Q. Lu, Y. Yang, T. Yang, K. Zhang, Y. Qiao *et al.*, "PhyBench: A physical commonsense benchmark for evaluating text-to-image models," *arXiv preprint arXiv:2406.11802*, 2024.
- [106] H. Bansal, Y. Bitton, I. Szekely, K.-W. Chang, and A. Grover, "VideoCon: Robust video-language alignment via contrast captions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13927–13937.
- [107] D. M. Bear, E. Wang, D. Mrowca, F. J. Binder, H.-Y. F. Tung, R. Pramod, C. Holdaway, S. Tao, K. Smith, F.-Y. Sun *et al.*, "Physion: Evaluating physical prediction from vision in humans and machines," *arXiv preprint arXiv:2106.08261*, 2021.
- [108] J. Wang, A. Ma, K. Cao, J. Zheng, Z. Zhang, J. Feng, S. Liu, Y. Ma, B. Cheng, D. Leng, Y. Yin, and X. Liang, "WISA: World simulator assistant for physics-aware text-to-video generation," 2025. [Online]. Available: <https://arxiv.org/abs/2503.08153>

- [109] X. Yang, Z. Tan, X. Nie, and H. Li, "IPO: Iterative preference optimization for text-to-video generation," *arXiv preprint arXiv:2502.02088*, 2025.
- [110] J. Liu, G. Liu, J. Liang, Z. Yuan, X. Liu, M. Zheng, X. Wu, Q. Wang, W. Qin, M. Xia *et al.*, "Improving video generation with human feedback," *arXiv preprint arXiv:2501.13918*, 2025.
- [111] A. Soni, S. Venkataraman, A. Chandra, S. Fischmeister, P. Liang, B. Dai, and S. Yang, "VideoAgent: Self-improving video generation," *arXiv preprint arXiv:2410.10076*, 2024.
- [112] Z. Huang, X. Weng, M. Igl, Y. Chen, Y. Cao, B. Ivanovic, M. Pavone, and C. Lv, "Gen-Drive: Enhancing diffusion generative driving policies with reward modeling and reinforcement learning fine-tuning," *arXiv preprint arXiv:2410.05582*, 2024.
- [113] Q. Xue, X. Yin, B. Yang, and W. Gao, "PhyT2V: Llm-guided iterative self-refinement for physics-grounded text-to-video generation," *arXiv preprint arXiv:2412.00596*, 2024.
- [114] J. Abou-Chakra, K. Rana, F. Dayoub, and N. Suenderhauf, "Physically embodied gaussian splatting: A visually learnt and physically grounded 3d representation for robotics," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: <https://openreview.net/forum?id=AEq0onGrN2>
- [115] L. Barcellona, A. Zadaianchuk, D. Allegro, S. Papa, S. Ghidoni, and E. Gavves, "Dream to manipulate: Compositional world models empowering robot imitation learning with imagination," *arXiv preprint arXiv:2412.14957*, 2024.
- [116] S. Yang, Y. Du, S. K. S. Ghasemipour, J. Tompson, L. P. Kaelbling, D. Schuurmans, and P. Abbeel, "Learning interactive real-world simulators," in *Proceedings of the International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=sFyTZEqmUY>
- [117] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu, "World-Dreamer: Towards general world models for video generation via predicting masked tokens," *arXiv preprint arXiv:2401.09985*, 2024.
- [118] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "DriveDreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.
- [119] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "GAIA-1: A generative world model for autonomous driving," 2023. [Online]. Available: <https://arxiv.org/abs/2309.17080>
- [120] L. Lian, B. Shi, A. Yala, T. Darrell, and B. Li, "LLM-grounded video diffusion models," in *Proceedings of the International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=exKHiougU>
- [121] A. Cherian, R. Corcodel, S. Jain, and D. Romeres, "Llmphy: Complex physical reasoning using large language models and world models," *arXiv preprint arXiv:2411.08027*, 2024.
- [122] B. Zeng, L. Yang, S. Li, J. Liu, Z. Zhang, J. Tian, K. Zhu, Y. Guo, F.-Y. Wang, M. Xu *et al.*, "Trans4D: Realistic geometry-aware transition for compositional text-to-4D synthesis," *arXiv preprint arXiv:2410.07155*, 2024.
- [123] X. Xu, W. Ge, D. Qiu, Z. Chen, D. Yan, Z. Liu, H. Zhao, H. Zhao, S. Zhang, J. Liang *et al.*, "GaussianProperty: Integrating physical properties to 3D gaussians with lmms," *arXiv preprint arXiv:2412.11258*, 2024.
- [124] T. Huang, H. Zhang, Y. Zeng, Z. Zhang, H. Li, W. Zuo, and R. W. Lau, "DreamPhysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors," *arXiv preprint arXiv:2406.01476*, 2024.
- [125] X. Li, Y.-L. Qiao, P. Y. Chen, K. M. Jatavallabhula, M. Lin, C. Jiang, and C. Gan, "PAC-NeRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification," *arXiv preprint arXiv:2303.05512*, 2023.
- [126] T. Zhang, H.-X. Yu, R. Wu, B. Y. Feng, C. Zheng, N. Snavely, J. Wu, and W. T. Freeman, "PhysDreamer: Physics-based interaction with 3d objects via video generation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2025, pp. 388–406.
- [127] Y. Lin, C. Lin, J. Xu, and Y. MU, "OmniPhysGS: 3D constitutive gaussians for general physics-based dynamics generation," in *Proceedings of the International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=9HZtP6l5lv>
- [128] J. Cao, S. Guan, Y. Ge, W. Li, X. Yang, and C. Ma, "NeuMA: Neural material adaptor for visual grounding of intrinsic dynamics," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024.
- [129] R.-Z. Qiu, G. Yang, W. Zeng, and X. Wang, "Feature splatting: Language-driven physics-based scene synthesis and editing," 2024. [Online]. Available: <https://arxiv.org/abs/2404.01223>
- [130] T. Xie, Y. Zhao, Y. Jiang, and C. Jiang, "PhysAnimator: Physics-guided generative cartoon animation," 2025. [Online]. Available: <https://arxiv.org/abs/2501.16550>
- [131] J. Lv, Y. Huang, M. Yan, J. Huang, J. Liu, Y. Liu, Y. Wen, X. Chen, and S. Chen, "GPT4Motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1430–1440.
- [132] T. Xie, Z. Zong, Y. Qiu, X. Li, Y. Feng, Y. Yang, and C. Jiang, "PhysGaussian: Physics-integrated 3d gaussians for generative dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4389–4398.
- [133] Y. Feng, Y. Shang, X. Li, T. Shao, C. Jiang, and Y. Yang, "PIE-NeRF: Physics-based interactive elastodynamics with nerf," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4450–4461.
- [134] Y. Jiang, C. Yu, T. Xie, X. Li, Y. Feng, H. Wang, M. Li, H. Lau, F. Gao, Y. Yang *et al.*, "VR-GS: A physical dynamics-aware interactive gaussian splatting system in virtual reality," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–1.
- [135] Y. Shao, M. Huang, C. C. Loy, and B. Dai, "GauSim: Registering elastic objects into digital world by gaussian simulator," *arXiv preprint arXiv:2412.17804*, 2024.
- [136] B. P. Duisterhof, Z. Mandi, Y. Yao, J.-W. Liu, J. Seidenschwarz, M. Z. Shou, D. Ramanan, S. Song, S. Birchfield, B. Wen, and J. Ichnowski, "DeformGS: Scene flow in highly deformable scenes for deformable object manipulation," 2024. [Online]. Available: <https://arxiv.org/abs/2312.00583>
- [137] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, "Dynamic 3D gaussians: Tracking by persistent dynamic view synthesis," in *2024 International Conference on 3D Vision*, 2024, pp. 800–809.
- [138] S. Bharadwaj, H. Feng, V. Abrevaya, and M. J. Black, "GenLit: Reformulating single-image relighting as video generation," *arXiv preprint arXiv:2412.11224*, 2024.
- [139] Y. Zhang, D. Zheng, B. Gong, J. Chen, M. Yang, W. Dong, and C. Xu, "LumiSculpt: A consistency lighting control network for video generation," 2024. [Online]. Available: <https://arxiv.org/abs/2410.22979>
- [140] K. Pandey, M. Gadelha, Y. Hold-Geoffroy, K. Singh, N. J. Mitra, and P. Guerrero, "Motion modes: What could happen next?" *arXiv preprint arXiv:2412.00148*, 2024.
- [141] K. Namekata, S. Bahmani, Z. Wu, Y. Kant, I. Gilitschenski, and D. B. Lindell, "SG-i2V: Self-guided trajectory control in image-to-video generation," *arXiv preprint arXiv:2411.04989*, 2024.
- [142] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, "Drag-NUWA: Fine-grained control in video generation by integrating text, image, and trajectory," *arXiv preprint arXiv:2308.08089*, 2023.
- [143] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "VideoComposer: Compositional video synthesis with motion controllability," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 36, pp. 7594–7611, 2023.
- [144] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan, "MotionCtrl: A unified and flexible motion controller for video generation," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.
- [145] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin *et al.*, "Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.
- [146] H. Zhu, T. He, A. Tang, J. Guo, Z. Chen, and J. Bian, "Compositional 3D-aware video generation with LLM director," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024.
- [147] X. Fu, X. Liu, X. Wang, S. Peng, M. Xia, X. Shi, Z. Yuan, P. Wan, D. Zhang, and D. Lin, "3DTrajMaster: Mastering 3D trajectory for multi-entity motion in video generation," *arXiv preprint arXiv:2412.07759*, 2024.
- [148] H. Jeong, G. Y. Park, and J. C. Ye, "VMC: Video motion customization using temporal attention adaption for text-to-video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9212–9221.

- [149] G. Y. Park, H. Jeong, S. W. Lee, and J. C. Ye, "Spectral motion alignment for video motion transfer using diffusion models," *arXiv preprint arXiv:2403.15249*, 2024.
- [150] F. Liang, B. Wu, J. Wang, L. Yu, K. Li, Y. Zhao, I. Misra, J.-B. Huang, P. Zhang, P. Vajda *et al.*, "FlowVid: Taming imperfect optical flows for consistent video-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8207–8216.
- [151] Y. Feng, Y. Shang, X. Feng, L. Lan, S. Zhe, T. Shao, H. Wu, K. Zhou, H. Su, C. Jiang *et al.*, "ElastoGen: 4D generative elastodynamics," *arXiv preprint arXiv:2405.15056*, 2024.
- [152] E. Coumans, "Bullet physics simulation," in *ACM SIGGRAPH 2015 Courses*, 2015, p. 1. [Online]. Available: <https://github.com/bulletphysics/bullet3?tab=readme-ov-file>
- [153] E. Coumans and Y. Bai, "PyBullet, a python module for physics simulation for games, robotics and machine learning," 2016–2021. [Online]. Available: <http://pybullet.org>
- [154] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.
- [155] NVIDIA, "PhysX," 2024. [Online]. Available: <https://github.com/NVIDIA-Omniverse/PhysX>
- [156] Y. Hu, L. Anderson, T.-M. Li, Q. Sun, N. Carr, J. Ragan-Kelley, and F. Durand, "DiffTaichi: Differentiable programming for physical simulation," in *Proceedings of the International Conference on Learning Representations*, 2020.
- [157] C. Jiang, C. Schroeder, J. Teran, A. Stomakhin, and A. Selle, "The material point method for simulating continuum materials," in *ACM SIGGRAPH 2016 courses*, 2016, pp. 1–52.
- [158] NVIDIA, "Omniverse," 2021. [Online]. Available: <https://developer.nvidia.com/omniverse>
- [159] ——, "OpenUSD," 2016. [Online]. Available: <https://developer.nvidia.com/usd#section-getting-started>
- [160] G. Authors, "Genesis: A universal and generative physics engine for robotics and beyond," December 2024. [Online]. Available: <https://github.com/Genesis-Embodied-AI/Genesis>
- [161] M. Müller, B. Heidelberger, M. Hennix, and J. Ratcliff, "Position based dynamics," *Journal of Visual Communication and Image Representation*, vol. 18, no. 2, pp. 109–118, 2007.
- [162] M. Macklin, M. Müller, and N. Chentanez, "XPBD: position-based simulation of compliant constrained dynamics," in *Proceedings of the 9th International Conference on Motion in Games*, 2016, pp. 49–54.
- [163] Y. Sun, H. Zhou, L. Yuan, J. J. Sun, Y. Li, X. Jia, H. Adam, B. Hariharan, L. Zhao, and T. Liu, "Video creation by demonstration," *arXiv preprint arXiv:2412.09551*, 2024.
- [164] Z. Li, R. Tucker, N. Snavely, and A. Holynski, "Generative image dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [165] M. A. Davis, "Visual vibration analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2016.
- [166] L. Wang, Z. Mai, G. Shen, Y. Liang, X. Tao, P. Wan, D. Zhang, Y. Li, and Y. Chen, "Motion inversion for video customization," *arXiv preprint arXiv:2403.20193*, 2024.
- [167] R. Akkerman, H. Feng, M. J. Black, D. Tzionas, and V. F. Abrevaya, "InterDyn: Controllable interactive dynamics with video diffusion models," *arXiv preprint arXiv:2412.11785*, 2024.
- [168] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
- [169] W. Wu, Z. Li, Y. Gu, R. Zhao, Y. He, D. J. Zhang, M. Z. Shou, Y. Li, T. Gao, and D. Zhang, "DragAnything: Motion control for anything using entity representation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2025, pp. 331–348.
- [170] H. Zhou, C. Wang, R. Nie, J. Lin, D. Yu, Q. Yu, and C. Wang, "TrackGo: A flexible and efficient method for controllable video generation," *arXiv preprint arXiv:2408.11475*, 2024.
- [171] M. Niu, X. Cun, X. Wang, Y. Zhang, Y. Shan, and Y. Zheng, "MOFA-Video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model," in *Proceedings of the European Conference on Computer Vision*. Springer, 2025, pp. 111–128.
- [172] Y. Li, X. Wang, Z. Zhang, Z. Wang, Z. Yuan, L. Xie, Y. Zou, and Y. Shan, "Image conductor: Precision control for interactive video synthesis," *arXiv preprint arXiv:2406.15339*, 2024.
- [173] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.
- [174] T. Xu, Z. Chen, L. Wu, H. Lu, Y. Chen, L. Jiang, B. Liu, and Y. Chen, "Motion dreamer: Realizing physically coherent video generation through scene-aware motion reasoning," 2024. [Online]. Available: <https://arxiv.org/abs/2412.00547>
- [175] Z. Liu, A. Yanev, A. Mahmood, I. Nikolov, S. Motamed, W.-S. Zheng, X. Wang, L. Van Gool, and D. P. Paudel, "InTraGen: Trajectory-controlled video generation for object interactions," *arXiv preprint arXiv:2411.16804*, 2024.
- [176] H. Wang, H. Ouyang, Q. Wang, W. Wang, K. L. Cheng, Q. Chen, Y. Shen, and L. Wang, "LeviTori: 3D trajectory oriented image-to-video synthesis," *arXiv preprint arXiv:2412.15214*, 2024.
- [177] Y. Chen, J. Cao, A. Kag, V. Goel, S. Korolev, C. Jiang, S. Tulyakov, and J. Ren, "Towards physical understanding in video generation: A 3d point regularization approach," 2025. [Online]. Available: <https://arxiv.org/abs/2502.03639>
- [178] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
- [179] S. Bahmani, X. Liu, W. Yifan, I. Skorokhodov, V. Rong, Z. Liu, X. Liu, J. J. Park, S. Tulyakov, G. Wetzstein *et al.*, "TC4D: Trajectory-conditioned text-to-4D generation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2025, pp. 53–72.
- [180] Z. Wang, J. Philion, S. Fidler, and J. Kautz, "Learning indoor inverse rendering with 3d spatially-varying lighting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 538–12 547.
- [181] P. Kocsis, J. Philip, K. Sunkavalli, M. Nießner, and Y. Hold-Geoffroy, "LightIt!: Illumination modeling and control for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9359–9369.
- [182] C. Zeng, Y. Dong, P. Peers, Y. Kong, H. Wu, and X. Tong, "DiLightNet: Fine-grained lighting control for diffusion-based image generation," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.
- [183] M. Ren, W. Xiong, J. S. Yoon, Z. Shu, J. Zhang, H. Jung, G. Gerig, and H. Zhang, "Relightful harmonization: Lighting-aware portrait background replacement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6452–6462.
- [184] Y. Mei, M. He, L. Ma, J. Philip, W. Xian, D. M. George, X. Yu, G. Dedic, A. L. Taşel, N. Yu *et al.*, "Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset," *arXiv preprint arXiv:2503.14485*, 2025.
- [185] G. Rai and O. Sharma, "Enhancing sketch animation: Text-to-video diffusion models with temporal consistency and rigidity constraints," *arXiv preprint arXiv:2411.19381*, 2024.
- [186] Y. Feng, X. Feng, Y. Shang, Y. Jiang, C. Yu, Z. Zong, T. Shao, H. Wu, K. Zhou, C. Jiang *et al.*, "Gaussian splashing: Dynamic fluid synthesis with gaussian splatting," *arXiv preprint arXiv:2401.15318*, 2024.
- [187] X. Tan, Y. Jiang, X. Li, Z. Zong, T. Xie, Y. Yang, and C. Jiang, "PhysMotion: Physics-grounded dynamics from a single image," *arXiv preprint arXiv:2411.17189*, 2024.
- [188] Z. Fu, J. Wei, W. Shen, C. Song, X. Yang, F. Liu, X. Yang, and G. Lin, "Sync4D: Video guided controllable dynamics for physics-based 4D generation," *arXiv preprint arXiv:2405.16849*, 2024.
- [189] L. S. Aira, A. Montanaro, E. Aiello, D. Valsesia, and E. Magli, "MotionCraft: Physics-based zero-shot video generation," *arXiv preprint arXiv:2405.13557*, 2024.
- [190] Y. Zhou, M. Simon, Z. M. Peng, S. Mo, H. Zhu, M. Guo, and B. Zhou, "SimGen: Simulator-conditioned driving scene generation." *Proceedings of the Advances in Neural Information Processing Systems*, vol. 37, pp. 48 838–48 874, 2024.
- [191] F. Liu, H. Wang, S. Yao, S. Zhang, J. Zhou, and Y. Duan, "Physics3D: Learning physical properties of 3D gaussians via video diffusion," 2024. [Online]. Available: <https://arxiv.org/abs/2406.04338>
- [192] Z. Liu, W. Ye, Y. Luximon, P. Wan, and D. Zhang, "Unleashing the potential of multi-modal foundation models and video diffusion for 4D dynamic physical scene simulation," *arXiv preprint arXiv:2411.14423*, 2024.

- [193] J. Lin, Z. Wang, S. Jiang, Y. Hou, and M. Jiang, "Phys4DGen: A physics-driven framework for controllable and efficient 4D content generation from a single image," *arXiv preprint arXiv:2411.16800*, 2024.
- [194] H. Zhao, H. Wang, X. Zhao, H. Wang, Z. Wu, C. Long, and H. Zou, "Automated 3D physical simulation of open-world scene with gaussian splatting," *arXiv preprint arXiv:2411.12789*, 2024.
- [195] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "KAN: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.
- [196] Y. Gao, H.-X. Yu, B. Zhu, and J. Wu, "FluidNexus: 3D fluid reconstruction and prediction from a single video," *arXiv preprint arXiv:2503.04720*, 2025.
- [197] P. Ma, P. Y. Chen, B. Deng, J. B. Tenenbaum, T. Du, C. Gan, and W. Matusik, "Learning neural constitutive laws from motion observations for generalizable pde dynamics," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2023, pp. 23 279–23 300.
- [198] B. Min, H. Ross, E. Sulem, A. P. B. Veysel, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.
- [199] M. Yu, L. Liu, J. Wu, T. T. Chung, S. Zhang, J. Li, D.-Y. Yeung, and J. Zhou, "The stochastic parrot on LLM's shoulder: A summative assessment of physical concept understanding," *arXiv preprint arXiv:2502.08946*, 2025.
- [200] M. Mitchell and D. C. Krakauer, "The debate over understanding in AI's large language models," *Proceedings of the National Academy of Sciences*, vol. 120, no. 13, p. e2215907120, 2023.
- [201] L. He, Y. Song, H. Huang, D. Aliaga, and X. Zhou, "Kubrick: Multimodal agent collaborations for synthetic video generation," 2024. [Online]. Available: <https://arxiv.org/abs/2408.10453>
- [202] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [203] A. Dawid and Y. LeCun, "Introduction to latent variable energy-based models: A path towards autonomous machine intelligence," *arXiv preprint arXiv:2306.02572*, 2023.
- [204] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [205] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
- [206] Z. Zhu, X. Wang, W. Zhao, C. Min, N. Deng, M. Dou, Y. Wang, B. Shi, K. Wang, C. Zhang *et al.*, "Is sora a world simulator? a comprehensive survey on general world models and beyond," *arXiv preprint arXiv:2405.03520*, 2024.
- [207] G. DeepMind, "Genie 2: A large-scale foundation world model," 2024.
- [208] J. Baldridge, J. Bauer, M. Bhutani, N. Brichtova, A. Bunner, K. Chan, Y. Chen, S. Dieleman, Y. Du, Z. Eaton-Rosen *et al.*, *arXiv preprint arXiv:2408.07009*, 2024.
- [209] E. Hoogeboom, J. Heek, and T. Salimans, "simple diffusion: End-to-end diffusion for high resolution images," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2023, pp. 13 213–13 232.
- [210] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024. [Online]. Available: <https://openreview.net/forum?id=Tw9nfNyOMy>
- [211] C. Li, O. Michel, X. Pan, S. Liu, M. Roberts, and S. Xie, "PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop," 2025. [Online]. Available: <https://arxiv.org/abs/2503.09595>
- [212] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-Sora: Democratizing efficient video production for all," *arXiv preprint arXiv:2412.20404*, 2024.
- [213] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik, "Diffusion model alignment using direct preference optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8228–8238.
- [214] S. Li, K. Kallidromitis, A. Gokul, Y. Kato, and K. Kozuka, "Aligning diffusion models by optimizing human utility," *arXiv preprint arXiv:2404.04465*, 2024.
- [215] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
- [216] J. Peters and S. Schaal, "Reinforcement learning by reward-weighted regression for operational space control," in *Proceedings of the International Conference on Machine Learning*, 2007, pp. 745–750.
- [217] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 36, pp. 53 728–53 741, 2023.
- [218] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2023, pp. 32 211–32 252.
- [219] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," *arXiv preprint arXiv:2305.13301*, 2023.
- [220] X. B. Peng, A. Kumar, G. Zhang, and S. Levine, "Advantage-weighted regression: Simple and scalable off-policy reinforcement learning," *arXiv preprint arXiv:1910.00177*, 2019.
- [221] T. Feng, C. Jin, J. Liu, K. Zhu, H. Tu, Z. Cheng, G. Lin, and J. You, "How far are we from agi: Are llms all we need?" *Transactions on Machine Learning Research*, 2024.
- [222] T. Ates, M. S. Atesoglu, C. Yigit, I. Kesenci, M. Kobas, E. Erdem, A. Erdem, T. Goksun, and D. Yuret, "CRAFT: A benchmark for causal reasoning about forces and interactions," 2022. [Online]. Available: <https://arxiv.org/abs/2012.04293>
- [223] H.-Y. Tung, M. Ding, Z. Chen, D. Bear, C. Gan, J. Tenenbaum, D. Yamins, J. Fan, and K. Smith, "Physion++: Evaluating physical scene understanding that requires online inference of different physical properties," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [224] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit *et al.*, "VBench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818.
- [225] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.
- [226] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [227] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.
- [228] H. Bansal, C. Peng, Y. Bitton, R. Goldenberg, A. Grover, and K.-W. Chang, "Videophy-2: A challenging action-centric physical commonsense evaluation in video generation," *arXiv preprint arXiv:2503.06800*, 2025.
- [229] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.
- [230] K. G. Barman, S. Caron, E. Sullivan, H. W. de Regt, R. R. de Austra, M. Boon, M. Färber, S. Fröse, F. Hasibi, A. Ipp, R. Kapoor, G. Kasieczka, D. Kostić, M. Krämer, T. Golling, L. G. Lopez, J. Marco, S. Otten, P. Pawłowski, P. Vischia, E. Weber, and C. Weniger, "Large physics models: Towards a collaborative approach with large language models and foundation models," 2025. [Online]. Available: <https://arxiv.org/abs/2501.05382>
- [231] K. Song, T. Hou, Z. He, H. Ma, J. Wang, A. Sinha, S. Tsai, Y. Luo, X. Dai, L. Chen *et al.*, "DirectorLLM for human-centric video generation," *arXiv preprint arXiv:2412.14484*, 2024.