TelePhysics: Physics-Grounded Multi-Object Scene Generation from a Single Image in Real Time

Xin Zhang[1,2], **Yabo Chen**[2], **Yijie Fang**[2], **Wanying Qu**[1], **Haibin Huang**[2], **Chi Zhang**[2], **Feng Xu**[✉1], **Xuelong Li**[✉2]

[1]Fudan University, [2]Institute of Artificial Intelligence, China Telecom (TeleAI)
[*]Equal Contributions

Recent generative video models achieve impressive visual quality but remain constrained by limited physical consistency and controllability. Existing video generation methods provide minimal physical control, and single-image-to-3D conversion approaches often suffer from object interpenetration. Furthermore, physics-based scene-level 3D generation methods exhibit spatial misalignment, stylized artifacts, and inconsistencies with the input data, restricting their use in realistic interactive video synthesis. We propose TelePhysics, a training-free framework that converts a single image into a physically consistent and controllable video through holistic scene-level 3D reconstruction. By representing the full scene geometry in a unified spatial coordinate system, TelePhysics resolves object penetration and alignment ambiguity. This formulation enables accurate multi-object interactions and mechanics-based control, such as applying forces and simulating rigid-body dynamics, while preserving the visual fidelity of the input. Unlike diffusion-based or autoregressive video priors that incur high inference latency, our approach allows instantaneous physics simulation and real-time rendering after initialization. Experimental results demonstrate that TelePhysics substantially outperforms prior methods in physical fidelity, spatial coherence, and controllability. The open-source code is available at https://github.com/xinzhang007/TelePhysics.

TeleAI

1 Introduction

Recent advances in generative video modeling have achieved remarkable visual realism, largely driven by diffusion and autoregressive architectures. However, these predominantly appearance-driven methods lack explicit mechanical reasoning. Consequently, current video generators are physically uncontrollable and fail to reliably support object-level interactions or physics-grounded manipulation. Although some studies incorporate physics through auxiliary priors or intermediate signals Gillman *et al.* (2025); Li *et al.* (2025b); Xie *et al.* (2025); Wang *et al.* (2025); Liu *et al.* (2026); Zhang *et al.* (2025c); Yang *et al.* (2025); Satish *et al.* (2026), they typically rely on manually specified parameters and soft constraints. As a result, they offer limited fine-grained temporal control and yield short, simplistic motions that miss long-horizon, complex physical interactions.

To improve physical reasoning, recent work has explored 3D reconstruction and physics-based scene generation from visual inputs. Yet, single-image-to-3D conversion methods Zhang *et al.* (2024a); Lai *et al.* (2025); Xiang *et al.* (2025a); Xu *et al.* (2024); Wu *et al.* (2025); Chen *et al.* (2024b,c); Li *et al.* (2025a) commonly reconstruct objects independently. Despite segmentation-aware formulations Liu *et al.* (2024a); Chen *et al.* (2024a), this lack of global spatial constraints causes severe interpenetration. Scene-level 3D generation methods jointly model object layouts and physical plausibility Paschalidou *et al.* (2021); Tang *et al.* (2024); Yang *et al.* (2024c,a); Zhou *et al.* (2024); Chen *et al.* (2025b); Lin *et al.* (2025a), often using scalable explicit representations like Gaussian splatting Zhou *et al.* (2024); Yang *et al.* (2024a); Go *et al.* (2025); Lee *et al.* (2024). Nevertheless, aligning generated scenes with the input image remains challenging.

Figure 1 TelePhysics is a unified, training-free framework designed to facilitate holistic 3D scene generation and physically grounded video synthesis from a single input image. The figure showcases interactions among multiple objects across diverse scenes. Additional results are provided in Fig. 11

Spatial discrepancies in object pose, scale, and orientation are frequent, and enforcing physical constraints can conflict with image evidence under occlusion and limited viewpoints Yang *et al.* (2024c); Yao *et al.* (2025); Zheng *et al.* (2025).

We present TelePhysics, a training-free unified framework for holistic 3D scene generation and physics simulation from a single image. Unlike prior methods that rely on learned priors or fragmented object-centric pipelines, TelePhysics jointly models all scene elements within a shared 3D representation under consistent spatial and physical constraints without requiring additional training. It integrates scene-level geometry, physics-driven simulation, and photorealistic rendering, decoupling physical reasoning from iterative generative inference. After initialization, TelePhysics supports immediate physics simulation and real-time rendering. This transforms a single image into a dynamic, interactive 3D environment that obeys physical laws, preserves visual fidelity, and supports long-horizon multi-object interactions.

We extensively evaluate TelePhysics on scenes containing single and multiple objects. In single-object settings, it reconstructs geometry from one image and produces stable, physically plausible dynamics under diverse forces, substantially reducing artifacts like drifting and penetration over long-horizon simulations. In challenging multi-object scenes, TelePhysics maintains global spatial coherence and models complex interactions including contact, support, collision, and stacking. Under occlusion and limited viewpoints, the unified representation aligns the pose, scale, and relative layout of objects, significantly mitigating interpenetration. Quantitative and qualitative results show that TelePhysics outperforms prior methods in physical realism, temporal consistency, and visual fidelity. Our main contributions are summarized as follows.

- We introduce TelePhysics, a training-free, unified framework that converts a single image into a physically consistent, photorealistic, and interactive 3D scene, enabling instantaneous physics simulation

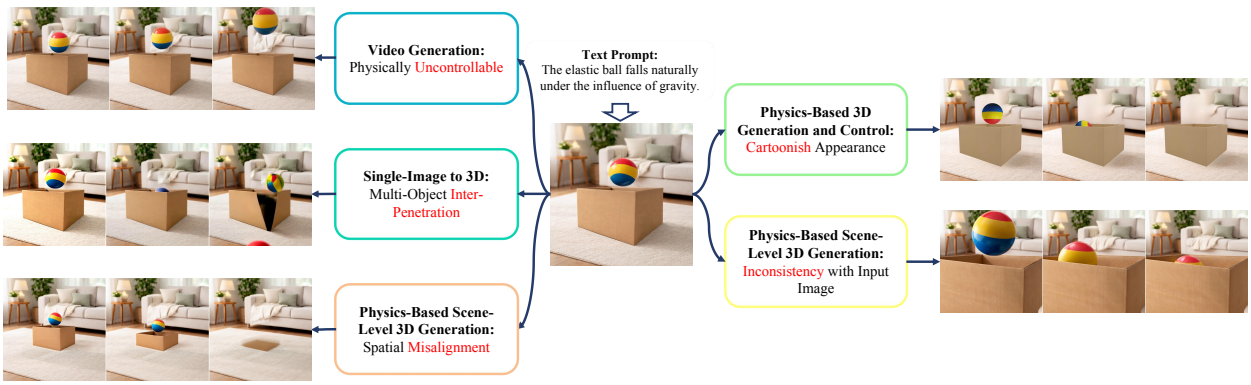

Figure 2 Primary issues of physics-grounded scene generation. Given a single reference image and a text prompt describing a physical event (e.g., a ball falling under gravity), existing paradigms show distinct limitations. Pure video generation models often yield physically uncontrollable motion that defies natural laws. Single-image-to-3D conversion and scene-level generation approaches frequently suffer from geometric artifacts like object interpenetration and spatial misalignment. Meanwhile, physics-based control methods often degrade visual fidelity, causing a cartoon-like appearance or content inconsistent with the original input image.

and real-time rendering.

- We propose a Scene-Aware Pose Alignment mechanism that transforms egocentric mesh representations into a unified world coordinate system and anchors them to a canonical ground plane, enforcing coherent and physically plausible scene configurations.
- We introduce a coarse-to-fine camera pose optimization strategy for perspective alignment, ensuring accurate photometric and geometric consistency between reconstructed objects and the input image.
- Comprehensive evaluation of single- and multi-object scenarios, as well as video generation tasks, demonstrates superior physical realism, temporal stability, and image-faithful reconstruction compared to appearance-driven and object-centric baselines.

2 Related Works

2.1 Physics-Aware Generative Models

Recent progress in video generation Ho *et al.* (2022); OpenAI (2025); Google DeepMind (2025); Wan *et al.* (2025); Kong *et al.* (2024); Yang *et al.* (2024b); Huang *et al.* (2025b); Xiang *et al.* (2025b) increasingly incorporates physics-inspired signals to improve realism, temporal coherence, and dynamic consistency. For example, An *et al.* An *et al.* (2026) explore the various perspectives and approaches in integrating AI with physics, providing a comprehensive overview of methods that incorporate physical principles into generative models to enhance dynamic realism. A prevalent strategy enforces physics via explicit simulators: methods such as PhysGaussian Xie *et al.* (2024) and its extensions Huang *et al.* (2025a); Lin *et al.* (2025b); Liu *et al.* (2025b); Mittal *et al.* (2025); Zhang *et al.* (2024b) reconstruct 3D geometry from multi-view inputs, simulate scene or object dynamics, and render the outcomes into videos. While effective when high-quality reconstructions are available, these pipelines depend heavily on accurate geometry and are typically confined to object-centric or single-scene settings, limiting scalability to complex environments and fine-grained control over physical properties.

In order to reduce the dependency on multi-view data, recent works aim to synthesize physical dynamics from a single image by coupling generative models with physics simulators, grounding monocular image-to-video or image-to-3D generation in rigid-body physics Liu *et al.* (2024b); Chen *et al.* (2025a); Tan *et al.* (2024); Zhang *et al.* (2025a); Liu *et al.* (2025a). Zhang *et al.* Zhang *et al.* (2025b) propose a method that incorporates variational positive-incentive noise, showing how this noise can benefit models by enhancing

their ability to simulate physical dynamics and interact with noisy real-world data. Similarly, Li Li (2022) explores the role of positive-incentive noise in improving model robustness, demonstrating how this noise can encourage models to explore a wider range of physical behaviors, facilitating better performance in scenarios with limited data.

Further progress in the field introduces the idea of using physics as auxiliary guidance rather than enforcing strict constraints. This method injects physical priors or intermediate signals into the generative process to encourage dynamic consistency while allowing for more flexibility Gillman *et al.* (2025); Li *et al.* (2025b); Xie *et al.* (2025); Wang *et al.* (2025). For instance, Liang *et al.* Liang *et al.* (2025) examine how reinforcement learning (RL) can be integrated with visual generative models to enhance the simulation of complex physical environments. Similarly, Hou *et al.* Hou *et al.* (2024) survey advances in 3D pre-training techniques, highlighting their potential in improving downstream tasks, such as those in physics-aware generative models. These methods often rely on manually chosen parameters and provide limited fine-grained temporal control, yielding short, simplified motions that struggle with long-horizon, multi-object interactions. Overall, physics-aware generative models significantly improve over purely appearance-driven methods, yet challenges remain in controllability, scalability, and long-term physical fidelity.

2.2 Single-Image and Multi-Object 3D Reconstruction

Single-image 3D reconstruction has advanced substantially in recovering geometry and pose from monocular inputs Lai *et al.* (2025); Xiang *et al.* (2025a); Xu *et al.* (2024); Wu *et al.* (2025); Chen *et al.* (2024b,c). Fueled by powerful 2D diffusion priors, these methods have progressed from primitive shapes to high-fidelity, textured assets. Extending them to multi-object settings, however, remains difficult due to depth ambiguity and complex occlusions. Most approaches reconstruct instances independently without explicitly modeling inter-object relations or collision constraints, often yielding physically implausible outcomes such as inter-penetration or inconsistent global scaling.

Segmentation-aware pipelines address part of this challenge by adopting a compositional formulation: decoupling object instances in 2D before lifting enables independent 3D asset reconstruction while preserving occlusion ordering and supporting downstream editing Liu *et al.* (2024a); Chen *et al.* (2024a). Leveraging instance-level masks allows these frameworks to handle clutter and maintain geometric integrity at the instance level. Nonetheless, without global scene constraints and explicit physical reasoning, multi-object compositions can still suffer from spatial misalignment and penetration.

2.3 Scene-Level 3D Generation and Image Consistency

Scene-level 3D generation aims to jointly synthesize multiple objects with coherent spatial arrangements, and recent methods increasingly incorporate layout and physical priors to improve plausibility and reduce artifacts such as interpenetration or floating objects Paschalidou *et al.* (2021); Tang *et al.* (2024); Yang *et al.* (2024c,a); Zhou *et al.* (2024); Chen *et al.* (2025b); Lin *et al.* (2025a). Meanwhile, advances in explicit 3D representations (e.g., Gaussian splatting) further enable scalable scene synthesis and rendering for complex environments Zhou *et al.* (2024); Yang *et al.* (2024a); Go *et al.* (2025); Lee *et al.* (2024).

Nevertheless, aligning generated 3D scenes with input images remains an open problem, especially for image-conditioned scene-level generation from sparse or monocular observations. Due to the ill-posed nature of 2D-to-3D inference, existing approaches often suffer from spatial misalignment in object translation, scale, and orientation, and may introduce inconsistencies between the reconstructed geometry and the reference image under occlusion and limited viewpoint cues Yao *et al.* (2025); Zheng *et al.* (2025). Moreover, when physical optimization (e.g., collision-free or stable layouts) is imposed, it can conflict with image evidence, further exacerbating the tension between physical plausibility and visual fidelity, thereby limiting the reliability of current methods for image-conditioned scene-level 3D generation Yang *et al.* (2024c); Yao *et al.* (2025).

2.4 Physics-based Simulation and Rendering

Physics-based simulation serves as the cornerstone for generating dynamically consistent motion, with established solvers spanning rigid body dynamics to continuum mechanics-based approaches such as the Material

Point Method (MPM) Sulsky *et al.* (1994a); Ram *et al.* (2015); Klár *et al.* (2016); Jiang *et al.* (2016); Hu *et al.* (2018); Xie *et al.* (2024) and Position-Based Dynamics (PBD) Müller *et al.* (2007a) providing robust mathematical frameworks for modeling object interactions and deformations; by explicitly enforcing physical laws, these methods enable stable and interpretable handling of collisions, contacts, and material responses across diverse scenarios Baraff (1997). Most simulation frameworks rely on simplified geometric or particle-based representations that lack real-world visual complexity, causing direct rendering of simulation outputs to appear synthetic and fail to bridge the gap between abstract physical states and photorealistic video. Recent approaches address this by integrating large-scale video diffusion priors into the rendering pipeline, replacing conventional graphics engines with pre-trained video generation models as conditional renderers Ho *et al.* (2022); Blattmann *et al.* (2023). Conditioned on coarse physical constraints such as geometry, motion fields, or contact events, these methods synthesize realistic textures, lighting, and materials absent from the simulation, effectively transforming low-fidelity physical proxies into photorealistic video Li *et al.* (2024); Wu *et al.* (2023). While this improves appearance, it does not by itself guarantee physical controllability or long-horizon consistency, underscoring the need for frameworks that tightly couple scene-level geometry, physics, and rendering.

3 Methods

Existing pipelines for physics-grounded scene generation struggle with several interconnected issues (Fig. 2). The inherent difficulty in controlling generative video models complicates the enforcement of target motions and physical constraints. When lifting multiple objects into 3D, current methods often produce interpenetrating shapes and spatial misalignments, creating layouts that are physically implausible and globally incoherent. Furthermore, synthesized assets typically lack geometric and photometric realism, adopting a cartoon-like appearance. These problems are compounded by a tendency for generated outputs to drift from the input image, degrading both structural consistency and visual fidelity.

We address these challenges with TelePhysics, a training-free framework that unifies 3D scene generation and physics-based video synthesis from a single image. The pipeline (Fig. 3) operates through four main stages. The Scene Perception module (Section 3.1) initially decomposes the input into a background and interactive 3D primitives, translating pixel-level data into a manipulable 3D structure. The Scene Alignment stage (Section 3.2) then establishes a shared world coordinate frame, assigning consistent poses, scales, and orientations to all entities. This aligns the abstract geometry with the original image perspective to create a physically actionable scene state.

To connect semantic information with physical properties, a Vision-Language Model (VLM) evaluates the material parameters of these aligned entities. The Physics Simulation Engine (Section 3.3) processes these assets to resolve multibody dynamics and contact constraints, yielding physically plausible trajectories. To bypass the computational cost of traditional graphics rendering while maintaining visual quality, WonderTrace (Section 3.4) refines the simulated frames into photorealistic, temporally coherent sequences using priors from modern video models. Together, these processes convert a static image into an interactive, physically consistent 3D environment that remains faithful to the source and supports long-horizon control.

3.1 Scene Perception

The reconstruction of complex scenes from a single image is an inherently ill-posed problem. Many existing single-view 3D methods lack holistic spatial reasoning, which often results in artifacts such as inter-object penetration and fragmented geometry. To address these limitations, we decompose the input image into instance-level foreground geometry and a globally consistent background. Foreground objects are segmented and reconstructed independently as separate 3D meshes within local coordinate frames, a strategy that prevents geometric interference and inter-penetration. These meshes are subsequently integrated through geometric constraints and coordinate alignment. In parallel, the background is reconstructed using a two-stage completion strategy: inpainting is employed to fill occluded regions, while generative outpainting extends the scene beyond the visible field of view(Fig. 3a).

Instance-level Mesh Reconstruction. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ and a set of semantic prompts P, we first employ the Segment Anything Model 3 (SAM 3) Carion *et al.* (2025) to generate high-fidelity

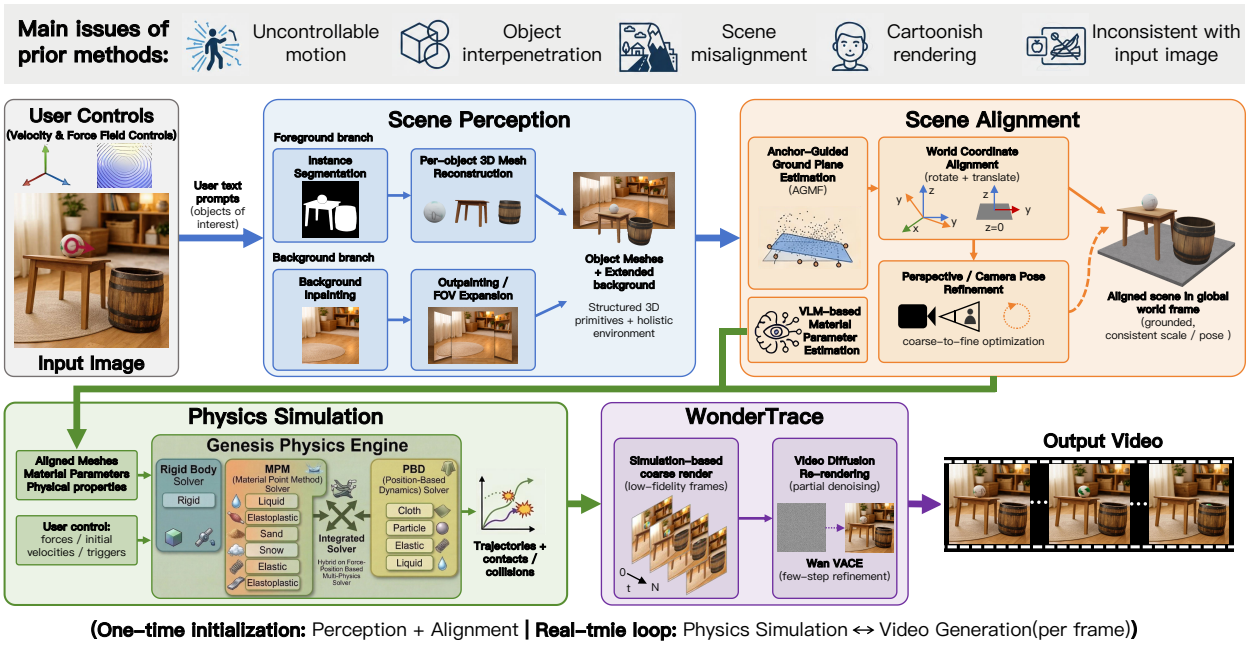

Figure 3 Overview of the TelePhysics framework. **(a)** Given a single input image and user controls, the pipeline applies Scene Perception to reconstruct 3D object meshes and synthesize a background environment. **(b)** These components are grounded in a unified global coordinate system through Scene Alignment to ensure geometric consistency, while a VLM-driven parameter estimation module concurrently deduces the physical properties of each entity. **(c)** Guided by these semantic priors, the Physics Simulation stage computes physically compliant trajectories and collision responses. **(d)** WonderTrace then bridges the visual domain gap by refining the coarse simulation renders into photorealistic video sequences.

instance masks $\{M_i\}_{i=1}^{N}$. These masks are utilized by SAM-3D-Objects Chen *et al.* (2025c) to lift 2D representations into 3D geometric primitives.

In contrast to implicit or point-based representations, which are prone to ghosting artifacts and numerical instability during contact, we adopt an explicit mesh-based representation with well-defined surface topology. This choice facilitates stable mechanical simulation by providing precise contact interfaces for the computation of collision manifolds and friction forces. By jointly processing masks and image features, our framework preserves the relative spatial configuration and scale of individual entities, ensuring that the recovered geometries $\mathcal{S}_i$ adhere to fundamental physical constraints.

Background Synthesis and Expansion. To ensure effective separation of the background and spatial consistency, we reconstruct the environment by eliminating foreground occlusions. We first synthesize the occluded regions using the LaMa inpainting model Suvorov *et al.* (2021) to derive a restored background:

$$B_{\text{rest}} = \mathcal{F}_{\text{inpaint}}(I, \bigcup_{i=1}^{N} M_i), \tag{1}$$

which provides an unobstructed view of the static environment.

To accommodate dynamic camera viewpoints and resolve scale ambiguity, we further expand the field of view via outpainting Filoni (2025):

$$B_{\text{ext}} = \mathcal{F}_{\text{outpaint}}(B_{\text{rest}}, \Phi), \tag{2}$$

where Φ denotes the target aspect ratio and expansion parameters. This extended panoramic context B_{ext} provides constraints on vanishing points and the global layout. Consequently, as the camera moves, the parallax between the foreground meshes $\mathcal{S}_i$ and the distant background B_{ext} remains physically consistent, effectively mitigating the "cardboard" effect often observed in localized reconstruction methods.

3.2 Scene Alignment

To ensure physically valid object interactions and maintain consistency with the input image, we introduce a Scene Alignment module(Fig. 3b) designed to resolve spatial inconsistencies stemming from egocentric reconstruction and inaccurate camera estimation (Fig. 2). This module comprises two primary components: *Pose Coordinate Alignment*, which maps reconstructed meshes into a unified world coordinate system and anchors them to the ground plane; and *Perspective Alignment*, which refines camera parameters to ensure geometric consistency between the reconstructed scene and the input image.

3.2.1 Pose Coordinate Alignment

Although the previous stage yields detailed mesh representations, these raw descriptors are ill-suited for downstream physical reasoning and interaction tasks. As illustrated in Fig. 2, a primary bottleneck arises from the inherent spatial misalignment introduced during the transformation of camera-centric observations into a global world coordinate system. As depicted in the figure, objects expected to adhere to basic physical constraints often exhibit implausible configurations. For instance, a ball positioned directly above a carton may drift laterally or penetrate the ground plane. To address these spatial misalignments and enforce physical plausibility, we introduce a Scene-Aware Pose Alignment mechanism. This module systematically transforms egocentric mesh representations into a unified world coordinate frame and anchors them to a canonical ground plane, thereby ensuring coherent spatial alignment and physically valid scene configurations.

To establish a consistent coordinate system for physical interactions, we implement a hierarchical alignment strategy that transforms meshes from the camera coordinate system $\mathcal{C}$ to the world coordinate system $\mathcal{W}$.

Single-Object Canonical Alignment. For isolated objects, we first perform Global Centroid Normalization. Given a mesh comprising N_v vertices $V = \{\mathbf{v}_i\}_{i=1}^{N_v}$, we translate the mesh to the world origin such that the centroid $\bar{\mathbf{v}}$ satisfies:

$$\bar{\mathbf{v}} = \frac{1}{N_v}\sum_{i=1}^{N_v}\mathbf{v}_i, \quad \mathbf{v}'_i = \mathbf{v}_i - \bar{\mathbf{v}}. \tag{3}$$

To determine the principal axes of the object, we apply Principal Component Analysis (PCA) to the zero-centered vertex distribution. The covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{3\times3}$ is computed as follows:

$$\mathbf{\Sigma} = \frac{1}{N_v}\sum_{i=1}^{N_v}\mathbf{v}'_i(\mathbf{v}'_i)^\top. \tag{4}$$

Through eigendecomposition $\mathbf{\Sigma}\mathbf{u}_k = \lambda_k\mathbf{u}_k$, we identify the primary eigenvector $\mathbf{u}_1$ corresponding to the largest eigenvalue λ_1. We then derive the rotation matrix $\mathbf{R}$ required to align $\mathbf{u}_1$ with the world vertical axis (Z-up) $[0,0,1]^\top$ to achieve a canonical pose. Finally, Ground-Contact Correction is applied:

$$\mathbf{v}_{\text{final}} = \mathbf{R}\mathbf{v}'_i - [0,0,Z_{\min}]^\top, \tag{5}$$

where $Z_{\min}$ denotes the minimum vertical extent of the rotated mesh, ensuring the object rests precisely on the $Z = 0$ plane.

Multi-Object Scene Alignment. Building upon the alignment of individual objects, we extend the approach to complex scenes involving multiple objects. To ensure physical consistency across such multi-object environments, we implement a multi-stage rectification pipeline, as illustrated in Fig. 4.

Starting from the constituent meshes of all objects in the scene, denoted as $\mathcal{V}_{\text{all}}$, we first aggregate their vertices to form a global unorganized point cloud $\mathcal{P}$ within the camera frustum. To identify the environmental layout, we employ the RANSAC algorithm to fit a dominant plane to $\mathcal{P}$, defined by the equation $ax + by + cz + d = 0$. The unit normal $\mathbf{n} = [a,b,c]^\top$ thus represents the estimated ground orientation (Fig. 4b-c). To standardize the reference frame for downstream physics-based tasks, we derive an orthogonal transformation $\mathbf{T}_{\text{ext}}$ that aligns the estimated normal $\mathbf{n}$ with the world vertical axis $\mathbf{z}_w = [0,0,1]^\top$. Specifically, the rotation $\mathbf{R} \in SO(3)$ is computed via Rodrigues' formula to map $\mathbf{n}$ to $\mathbf{z}_w$, while the translation $\mathbf{t}$ is determined by the minimum vertical coordinate of the rectified points. This effectively anchors the lowest point of the scene to the global ground level (Fig. 4e-f).

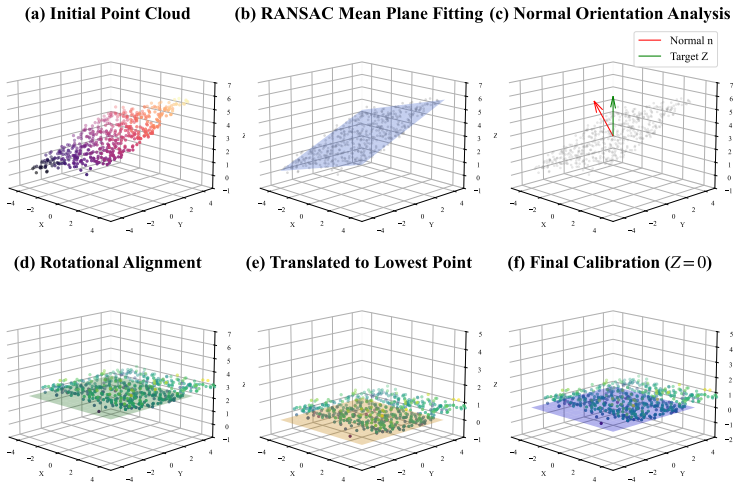

Figure 4 Ground Plane Alignment Pipeline. An overview of the proposed scene-level alignment procedure, which estimates the dominant ground plane from the aggregated point cloud and applies a rigid transformation to obtain a gravity-aligned world frame $\mathcal{W}$, where the ground plane coincides with $z = 0$.

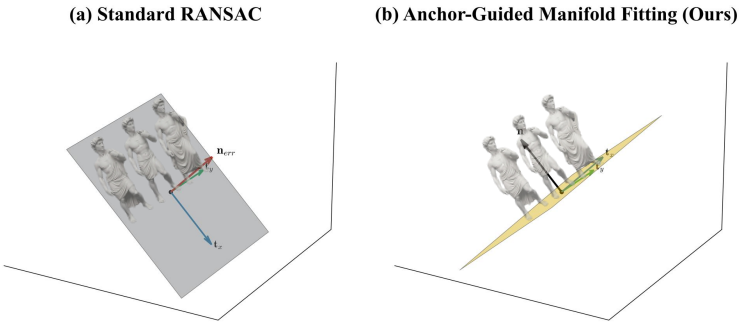

Figure 5 Comparison of plane estimation strategies. (a) Standard RANSAC: Global density-based fitting is susceptible to high-density vertical distractors (e.g., human torsos or walls), leading to erroneous orientation estimates. (b) Anchor-Guided Manifold Fitting (Ours): By selectively sampling local minima $\mathcal{P}_{\text{base}}$ from object extrema, our strategy effectively decouples the ground manifold from vertical structures, ensuring a physically plausible estimation of the surface normal $\mathbf{n}$ and the gravitational axis.

Anchor-Guided Ground Plane Estimation. Standard plane estimation techniques, such as vanilla RANSAC, often prove suboptimal in spatially complex scenes populated by multiple vertical structures (e.g., standing human subjects). As illustrated in Fig. 5a, RANSAC is prone to converging on regions of high point density—such as the torsos of human subjects or walls—rather than the true ground manifold (e.g., the soles of the feet). To address this density-driven bias, we introduce Anchor-Guided Manifold Fitting (AGMF). Unlike global fitting methods that treat all points with equal importance, AGMF shifts the optimization focus toward local geometric extrema that are likely to represent ground-contact points.

The core intuition behind AGMF is to isolate semantic "anchors"—points physically constrained to the ground—prior to performing the fit. For a scene containing N object meshes, we define an anchor set $\mathcal{P}_{\text{base}}$ by extracting vertices from the lower extrema of each individual mesh M_i. Specifically, we sample vertices within a proximity threshold δ of the local Z-minimum for each object:

$$\mathcal{P}_{\text{base}} = \bigcup_{i=1}^{N} \{\mathbf{v} \in M_i \mid v_z \leq \min_{\mathbf{u} \in M_i} u_z + \delta\}. \tag{6}$$

By restricting the input space to $\mathcal{P}_{\text{base}}$, we effectively eliminate the vertical distractors that typically

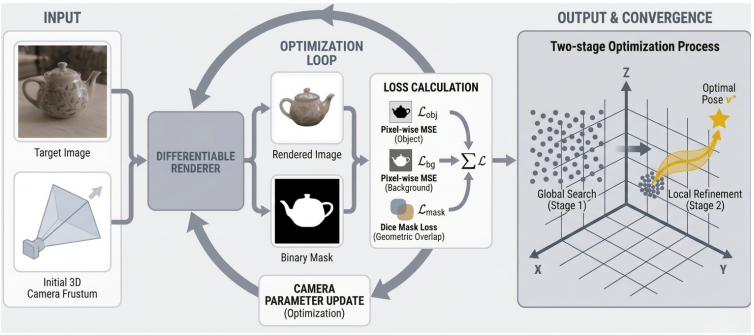

Figure 6 Differentiable rendering pipeline for explicit camera pose estimation. The framework optimizes camera spatial parameters by minimizing the discrepancy between the rendered output and the target image. The visualization (right) illustrates the transition from global stochastic sampling to local refinement, leading to the final converged pose v^*.

cause RANSAC to yield erroneous estimates. We then determine the optimal ground plane parameters $\boldsymbol{\pi} = [a, b, c, d]^\top$ by minimizing a robust M-estimator over this refined subset:

$$\min_{\boldsymbol{\pi}} \sum_{\mathbf{p} \in \mathcal{P}_{\text{base}}} \rho\left(\frac{|ax_p + by_p + cz_p + d|}{\sqrt{a^2 + b^2 + c^2}} \right), \tag{7}$$

where $\rho(\cdot)$ denotes a robust loss function (e.g., Huber or Tukey) employed to suppress outliers within the anchor set. This formulation ensures that the resulting surface normal $\mathbf{n} = [a, b, c]^\top$ aligns with the collective gravitational base of all objects in the scene. As demonstrated in Fig. 5b, AGMF yields a physically plausible world orientation, whereas standard RANSAC fails by fitting to dominant vertical geometries.

3.2.2 Perspective Alignment

As illustrated in Fig. 2, naive rendering often results in significant perspective discrepancies between the reconstructed objects and the input image, manifesting as inaccuracies in apparent scale and pose (see also Fig. 7).

To mitigate these issues, we formulate perspective alignment as a continuous optimization problem over the camera parameters. The objective is to recover a camera pose that ensures geometric consistency with the input observation. Fig. 6 depicts this process, wherein camera parameters are iteratively refined to align the rendered output with the input image. The pipeline accepts a target RGB image of the object and an initial estimate of the 3D camera pose (represented by a camera frustum) as input. Given the current camera parameters, a renderer generates a synthetic image of the 3D object and a corresponding binary silhouette mask. These outputs are compared to the target image to compute multiple loss terms, including the pixel-wise mean squared error for both the object and background regions, as well as a Dice-based mask loss that quantifies the geometric overlap between silhouettes. The aggregated loss serves as the objective function for updating the camera parameters. Given the non-convex nature of the rendering loss landscape, identifying the global optimum is non-trivial. Consequently, we propose a robust coarse-to-fine optimization strategy that integrates global stochastic sampling with local derivative-free refinement. Initially, we conduct a coarse global search by uniformly sampling candidate poses within a bounded space $\mathcal{B}$. This step identifies a reliable initialization, thereby mitigating the risk of convergence to poor local minima. Subsequently, starting from the optimal candidate identified in the coarse stage, we execute a local refinement step using the derivative-free Powell method under box constraints. This approach enables precise pose adjustment without reliance on potentially unstable gradients associated with the rendering process.

Problem Formulation. Let $\mathbf{P}_c = (X_c, Y_c, Z_c) \in \mathbb{R}^3$ denote the position of the camera (or object) within the world coordinate system. Given an initial estimate $\mathbf{P}_c^{\text{init}}$, we define a bounded search space $\mathcal{B}$ as follows:

$$\mathcal{B} = [x_0 - \Delta_x,\, x_0 + \Delta_x] \times [y_0 - \Delta_y,\, y_0 + \Delta_y] \times [z_0 - \Delta_z,\, z_0 + \Delta_z], \tag{8}$$

where (x_0, y_0, z_0) corresponds to $\mathbf{P}_c^{\text{init}}$. For any candidate position $\mathbf{P} \in \mathcal{B}$, we generate a rendered image $I(\mathbf{P})$ and its associated object mask $M(\mathbf{P})$. The goal is to minimize the discrepancy between these outputs and

the target image I^* through a set of region-aware constraints. To this end, we categorize the optimization objectives into photometric consistency and geometric alignment.

Region-Aware Appearance Loss. To capture local photometric details, we define an appearance loss for the object region using the target mask M^*:

$$\mathcal{L}_{\text{obj}}(\mathbf{P}) = \frac{\|(I(\mathbf{P}) - I^*) \odot M^*\|_1}{\|M^*\|_1 + \epsilon}, \tag{9}$$

where $\odot$ denotes element-wise multiplication and $\|\cdot\|_1$ represents the L_1 norm. The normalization term is critical to prevent the optimization from biasing toward larger foreground areas. Similarly, the background-region loss is formulated as:

$$\mathcal{L}_{\text{bg}}(\mathbf{P}) = \frac{\|(I(\mathbf{P}) - I^*) \odot (1 - M^*)\|_1}{\|1 - M^*\|_1 + \epsilon}. \tag{10}$$

Mask Alignment Loss. To explicitly enforce geometric consistency between the rendered silhouette and the target, we employ a Dice loss:

$$\mathcal{L}_{\text{mask}}(\mathbf{P}) = 1 - \frac{2\|M(\mathbf{P}) \odot M^*\|_1 + \epsilon}{\|M(\mathbf{P})\|_1 + \|M^*\|_1 + \epsilon}. \tag{11}$$

This formulation ensures robustness against scale variations and is particularly effective for small foreground objects.

Overall Objective. The final objective function is a weighted combination of the aforementioned losses, balanced by the hyperparameters λ:

$$\mathcal{L}(\mathbf{P}) = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}(\mathbf{P}) + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}}(\mathbf{P}) + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}(\mathbf{P}). \tag{12}$$

As demonstrated in Fig. 7, the proposed coarse-to-fine camera pose optimization substantially improves the geometric and perspective consistency between the rendered results and the input images. In contrast to the initial estimates, the optimized poses yield renderings that align well with the input in terms of both spatial configuration and viewpoint, effectively eliminating the perspective distortion and pose misalignment observed earlier. The close agreement regarding object silhouettes and visual appearance indicates that the proposed method successfully resolves inconsistencies with the input image and enables accurate camera pose estimation.

3.3 Physical Simulation

To ensure physically consistent 3D dynamics without relying on unscalable manual assignment, our framework utilizes a Vision-Language Model (VLM) to automatically estimate categorical and continuous material parameters—such as mass, friction, and Young's modulus (detailed in Appendix D.1)—for each reconstructed entity based on its visual appearance and scene context. Following this parameter evaluation, objects are routed to appropriate solvers within a multi-physics backend. While recent physics-based generation methods, such as PhysCtrl Wang *et al.* (2025) and PhysGen3D Chen *et al.* (2025a), primarily use the Material Point Method (MPM), this approach can be computationally expensive and may introduce artifacts when handling rigid bodies or stiff constraints. We instead adopt Genesis Authors (2024) as a unified backend(Fig. 3c) to address these limitations. Guided by the VLM classifications, the framework decouples material representations across three specialized solvers. Rigid Body Dynamics (RBD) handles non-deformable objects and robotic manipulators; MPM is reserved for fluids and hyperelastic materials undergoing large volumetric deformations; and Position-Based Dynamics (PBD) manages thin-shell structures like cloth and cables. These solvers operate concurrently with GPU-based parallelism and adaptive time-stepping, exchanging force and state information to enable efficient, high-fidelity multi-material interactions. The mathematical formulations and integration mechanisms for these solvers are detailed in Appendix D.2.

3.4 WonderTrace

To bridge the domain gap between visually coarse physical simulations and photorealistic real-world dynamics, we introduce the WonderTrace module (Fig. 3d). Videos rendered directly from physical simulators

typically exhibit a synthetic, "plastic" appearance resulting from simplified lighting models, unrealistic material responses, and the absence of environmental stochasticity, such as atmospheric scattering or organic texture variations (as highlighted in Fig. 2). WonderTrace addresses these limitations by leveraging the high-dimensional latent space of state-of-the-art video generation models as a rendering prior, translating simulated frames into high-fidelity temporal sequences.

By projecting simulated physical states and trajectories into conditional control signals—such as depth maps, optical flow, and object masks—the module synthesizes photorealistic, pixel-level semantic enhancements while strictly adhering to the engine's underlying structural and temporal dynamics. To balance computational efficiency with visual fidelity, we employ a partial denoising strategy. Rather than generating videos from pure Gaussian noise—a process that is computationally intensive and prone to deviating from the underlying simulation logic—we perturb the simulation frames with a moderate level of noise and apply only the final denoising stages of the diffusion process. This ensures that the generated video faithfully inherits the complex spatiotemporal structure of the simulation while enriching the scene with high-frequency visual details (e.g., sub-surface reflections and motion blur) that are challenging for real-time simulators to model accurately.

Specifically, our primary pipeline adopts the Wan2.1 VACE Wan *et al.* (2025) model, which incorporates a self-forcing DMD-based distillation strategy. This design enables the rerendering process to operate with a minimal number of refinement steps, significantly reducing inference latency and facilitating real-time performance without compromising visual quality. For offline rendering or scenarios demanding maximum perceptual realism, we provide an alternative configuration utilizing the larger Wan2.2 VACE 14B Wan *et al.* (2025) model, which delivers higher-resolution outputs and superior high-fidelity visualization.

Furthermore, a fundamental challenge in 3D scene modeling is the inherent sparsity of source data, which typically captures a limited field of view (FOV) and leads to voids or severe boundary artifacts when the virtual camera deviates from its original trajectory. To overcome this, WonderTrace seamlessly supports novel view synthesis and extensive virtual camera movements without out-of-bounds artifacts. Following the background handling approach discussed in Section 3.1, we employ a combined in-painting and out-painting strategy to decouple the foreground content from the environment. Because the background environment is expanded during the initial Scene Perception stage, we extend this design into the video rerendering stage to enable scene completion well beyond the original camera frustum. As illustrated in Fig. 8, the reconstructed background is spatially expanded to provide sufficient semantic context for severe viewpoint changes. Finally, the rerendered, expanded background is seamlessly composited with the dynamic foreground elements. This dual-stream processing ensures continuous visual fidelity and robust temporal consistency under entirely novel camera trajectories.

4 Experiments

We evaluate TelePhysics in terms of controllability, physical consistency, visual quality, and runtime efficiency. The qualitative results presented in Fig. 11 highlight the model's ability to maintain physical integrity even during complex multi-object interactions.

4.1 Experimental Setup

We consider the single-image-to-video setting and evaluate all methods on a test set comprising 59 scenes that span diverse indoor and outdoor environments. The benchmark includes single-object and multi-object interactions, such as dropping, pushing, rolling, collision, support, stacking, and long-horizon rigid-body motion. These scenes vary in object count, occlusion level, and camera viewpoint, thereby providing a challenging testbed for controllable physics-based video generation. Unless otherwise specified, all experiments are conducted on a single NVIDIA H100 GPU. For scene perception, we utilize SAM 3 Carion *et al.* (2025) for instance segmentation and SAM-3D-Objects Chen *et al.* (2025c) for instance-level 3D lifting. Genesis Authors (2024) performs the physics simulation, handling rigid objects via the rigid-body solver and deformable assets via MPM or PBD. For video rerendering, we employ Wan2.1 VACE Wan *et al.* (2025) as the default fast refinement model and Wan2.2 VACE 14B Wan *et al.* (2025) as an optional high-quality offline variant.

We compare TelePhysics against representative appearance-driven and physics-aware baselines, including

Sora2-pro OpenAI (2025), Veo3.1 Google DeepMind (2025), CogVideoX1.5 Yang *et al.* (2024b), Wan2.2-A14B Wan *et al.* (2025), WonderPlay Li *et al.* (2025b), and PhysCtrl Wang *et al.* (2025). All methods are given the same input image and event description for a fair comparison. For WonderPlay and PhysCtrl, we translate the shared event description into each method's native physics-control format while keeping all other settings at their official defaults. Specifically, for WonderPlay, we provide the input image together with a YAML configuration that maps the event description to per-object external forces (direction and magnitude), as well as a text prompt for its Stage 2 video refinement; all reconstruction, simulation, and rendering parameters are kept identical to those in the released codebase. For PhysCtrl, we provide the input image and a JSON configuration specifying the material type, force direction, and a text prompt derived from the event description, using the officially released base-model checkpoint and default inference hyperparameters. For video-prior baselines, we use the event description as the text prompt and additionally provide object-centric motion descriptions and 2D motion arrows overlaid on the input image to improve motion specification under the same supervision budget.

4.2 Qualitative Comparison

Figure 9 presents qualitative comparisons with strong video generation baselines. While appearance-driven methods often produce visually appealing motion, these approaches frequently lack reliable control over object trajectories and interactions. In scenes that involve multiple objects, these methods tend to generate ambiguous collisions, inconsistent contacts, or motions that deviate from the specified direction of force. Although PhysCtrl improves controllability, the model still struggles with long-horizon stability and spatial consistency in multi-object settings.

In contrast, TelePhysics produces motions that remain temporally coherent and mechanically consistent with the intended events. This advantage is especially evident in scenarios that involve support, collision, and sequential interactions among multiple objects. By reconstructing the scene in a shared world frame and explicitly simulating object dynamics, the proposed method avoids the drift, interpenetration, and identity inconsistency that are commonly observed in video-prior baselines.

4.3 Quantitative Comparison

We quantitatively evaluate the test set of 59 scenes for controllability, physical plausibility, and overall video quality. Following the VideoPhy Bansal *et al.* (2024) protocol, we adopt a 5-point Likert scoring framework based on GPT-5. For each method, we evaluate the 59 generated videos under three complementary criteria. Semantic adherence measures the alignment between the generated video and the text prompt, emphasizing whether the specified initial forces or velocities and the resulting motions match the intended dynamics. Physical commonsense assesses whether the generated motions and interactions are plausible under basic physical principles, such as gravity, inertia, and contact constraints. Video quality evaluates visual fidelity, frame-level realism, and temporal smoothness.

Table 1 reports the evaluation results from GPT-5. The proposed method achieves the best performance on semantic adherence and physical commonsense, which demonstrates that explicit scene reconstruction and simulator-based dynamics substantially improve controllability and physical plausibility. Notably, the lower scores of PhysCtrl and WonderPlay stem from the fact that the test set predominantly features multi-object scenes, in which earlier methods struggle to accurately model complex physical interactions. Although Veo3.1 attains the highest score for video quality, TelePhysics remains competitive while providing substantially stronger physical grounding and interaction correctness. These results support the core design of TelePhysics: the decoupling of physical reasoning from appearance synthesis leads to better control without sacrificing perceptual quality.

4.4 Human Evaluation

To complement the automatic evaluation, we conduct a ranking-based human study. In each trial, participants view seven videos generated from the same input image, with one video from each method, and rank the videos according to semantic adherence, physical commonsense, video quality, and overall preference.

Table 1 Quantitative comparison

Method	SA↑	PC↑	VQ↑
PhysCtrl Wang *et al.* (2025)	1.85	2.34	2.58
WonderPlay Li *et al.* (2025b)	2.24	2.49	2.97
CogVideoX1.5 Yang *et al.* (2024b)	2.36	2.02	2.42
Wan2.2-A14B Wan *et al.* (2025)	2.46	2.68	3.46
Sora2-pro OpenAI (2025)	2.82	2.35	3.58
Veo3.1 Google DeepMind (2025)	2.66	2.92	**3.71**
Ours	**3.29**	**3.15**	3.61

SA: Semantic Adherence. PC: Physical Commonsense. VQ: Video Quality.

Table 2 Human evaluation

Method	SA↑	PC↑	VQ↑	UP↑
PhysCtrl Wang *et al.* (2025)	3.45	3.12	3.29	3.31
WonderPlay Li *et al.* (2025b)	2.02	2.05	1.98	2.06
CogVideoX1.5 Yang *et al.* (2024b)	2.78	2.81	2.86	2.84
Wan2.2-A14B Wan *et al.* (2025)	3.72	3.85	3.73	3.76
Sora2-pro OpenAI (2025)	4.38	4.43	4.67	4.28
Veo3.1 Google DeepMind (2025)	4.72	4.84	4.83	4.84
Ours	**6.93**	**6.90**	**6.64**	**6.91**

Participants rank seven videos per scene, one per method. Scores are aggregated using Borda count ($8 - \text{rank}$), yielding a score range of $[1, 7]$.

We convert the rankings into Borda scores using $s_i = 8 - \pi_i$, where $\pi_i \in \{1, \ldots, 7\}$ denotes the rank of method i. Higher scores indicate a stronger human preference.

As shown in Table 2, TelePhysics achieves the highest scores across all four criteria. The large margins in semantic adherence and physical commonsense indicate that human annotators consistently perceive the results of the proposed method as more controllable and physically plausible. Moreover, TelePhysics obtains the best overall preference, which suggests that stronger physical grounding improves not only correctness but also the overall viewing experience.

4.5 Generation Speed

We further analyze the runtime of TelePhysics on an NVIDIA H100 GPU. As shown in Table 3, scene perception and scene alignment incur only a one-time initialization cost per scene. After initialization, the system supports real-time simulation and interactive rendering at ∼15 FPS. We report the runtime of the interactive pipeline here; optional high-quality offline rerendering is excluded from the FPS numbers.

4.6 Ablation Study

To isolate the contribution of key components in TelePhysics, we conduct extensive ablation studies. We evaluate the modular designs covering scene alignment, physical simulation constraints, and rerendering efficiency.

Scene-aware pose alignment. Accurately placing objects into the 3D scene without violating physical constraints is crucial for realistic simulation. Table 4 evaluates different pose alignment strategies. Removing alignment completely (*w/o alignment*) reliably avoids physical violations (Penetration Rate, PR = 0.00) but fails to place the object accurately, resulting in a drastically low Mask IoU of 0.11. Simply applying *single-object canonical alignment* improves the IoU (0.35) but introduces severe geometric conflicts, raising the

Table 3 Average runtime breakdown of TelePhysics (NVIDIA H100).

Component	Time / Throughput
One-time initialization (per scene)	
Segmentation (SAM3 + inpainting)	∼30 s
3D Mesh Reconstruction (SAM3D)	∼81 s
Physics Config Estimation (VLM)	∼227 s
Scene Alignment (pose + camera opt.)	∼15 s
Interactive simulation loop	
Physics Simulation	∼77 FPS
Interactive Rendering	∼25 FPS
End-to-end Interactive Loop	∼15 FPS

Initialization timings include model loading. Physics Config uses Qwen2.5-VL-72B for automatic material and force estimation. The interactive loop throughput corresponds to the real-time simulation with composited rendering. Optional diffusion-based offline rerendering can be applied afterward for higher perceptual quality.

Table 4 Ablation on scene-aware pose alignment.

Variant	Mask IoU↑	PR↓	SVR↓	ISR↑
w/o alignment	0.11	**0.00**	**0.00**	**1.00**
+ single-object canonical alignment	0.35	0.50	**0.00**	0.88
+ vanilla RANSAC plane fitting	**0.54**	0.01	0.15	0.88
+ AGMF (full pose alignment)	0.54	**0.00**	0.15	0.90

PR: Penetration Rate. SVR: Support Violation Rate. ISR: Interaction Success Rate.

PR to 0.50. Incorporating *vanilla RANSAC plane fitting* significantly improves spatial grounding (IoU reaches 0.54) and reduces PR, yet it still struggles with proper physical resting states, yielding a Support Violation Rate (SVR) of 0.15. In contrast, our full pose alignment using the AGMF strategy strictly resolves collisions (PR = 0.00) while maintaining the highest spatial alignment and achieving an Interaction Success Rate (ISR) of 0.90. This demonstrates that holistic scene-aware alignment is indispensable for valid initial physical states.

Perspective alignment. Table 5 ablates the camera parameter optimization module. Relying solely on the *initial camera* guess leads to poor geometric projection, as evidenced by a severe Reprojection Error of 38.95 pixels and low Mask IoU (0.26). Using either *global search* or *local Powell optimization* individually improves the alignment. However, our proposed *coarse-to-fine* approach consistently achieves the most robust geometric consistency, yielding the lowest Reprojection Error (14.33) and highly competitive visual metrics (SSIM 0.76, LPIPS 0.13). This confirms that combining global contextual search with local precision refinement is necessary to bridge the gap between 2D pixels and the 3D simulation space.

Force field configuration. Table 6 investigates the impact of force fields during the physical simulation stage. Interestingly, the *No forces* variant achieves the highest scores in automated 2D metrics such as Motion Amplitude (4.77) and Smoothness (0.45). However, we argue that these purely appearance-based metrics inherently favor unconstrained, linear, or "floating" motions. When our *Full forces* (including gravity and intended primary forces) are introduced, the objects are subjected to strict physical laws—such as sudden deceleration due to collisions, friction, and resting contacts. While these physical constraints naturally reduce the naive optical flow smoothness and overall motion amplitude, they are fundamentally required to prevent objects from drifting unnaturally. The parity between *Primary force only* and *Full forces* further suggests that explicit external driving forces dominate the valid dynamic interactions in our targeted events.

Inference efficiency in video synthesis. Finally, we study the trade-off between rendering quality and computational cost in the video synthesis stage (Table 7). We observe that executing the diffusion process

Table 5 Ablation on perspective alignment.

Variant	SSIM↑	LPIPS↓	Mask IoU↑	Reproj. Err.↓
Initial camera only	0.76	0.14	0.26	38.95
Global search only	0.76	0.13	0.44	20.97
Local Powell only	**0.77**	**0.13**	**0.55**	14.77
Coarse-to-fine (ours)	0.76	0.13	0.54	**14.33**

Reproj. Err. denotes average reprojection error in pixels after optimization.

Table 6 Ablation on force field configuration.

Variant	Motion Amp. ↑	Smoothness ↑	Aesthetic ↑
No forces	**4.77**	**0.45**	**0.22**
Gravity only	4.73	0.41	0.17
Primary force only	4.35	0.40	0.17
Full forces (ours)	4.35	0.40	0.17

Motion Amp.: mean optical flow magnitude. Smoothness: inverse flow variance.

for *10 steps* yields the optimal visual fidelity (SSIM 0.82, LPIPS 0.05, Aesthetic 0.20) while requiring the least wall-clock time (539.03 s). Increasing the inference to 30 or 50 steps exponentially increases the computational burden but slightly degrades the perceptual metrics. This indicates that our condition signals (rendered explicit dynamics) are highly informative, allowing the rerendering model to converge rapidly without requiring over-parameterized iterative refinement. Consequently, we adopt 10 steps as the default setting for optimal efficiency.

Geometric Fidelity of Re-rendering. We evaluate whether depth-conditioned video re-rendering (using 10 denoising steps) reliably preserves the underlying physical geometry of the raw simulation. As demonstrated in Tab. 8, the re-rendered outputs maintain robust scene layouts, achieving an overall SSIM of 0.72 and a background SSIM of 0.78. Fluid environments exhibit the highest structural fidelity (SSIM of 0.80, background SSIM of 0.84), indicating that depth maps provide strong geometric constraints for continuous surfaces. Additionally, deformable objects such as cloth demonstrate nearly zero silhouette drift (ΔIoU ≈ 0.00). This structural consistency is further corroborated by an average centroid displacement of merely 52.5 pixels (6.0% of the 880-pixel image width), confirming that precise spatial positioning is strictly maintained throughout the re-rendering process.

5 Conclusion and Limitations

We presented TelePhysics, a training-free framework that converts a single image into a physically grounded and controllable video by reconstructing a holistic 3D scene, aligning all entities in a shared world coordinate system, simulating object dynamics with an explicit physics engine, and rerendering the result into a realistic video. Compared with appearance-driven video generation, TelePhysics offers substantially stronger control over object interactions and long-horizon dynamics. Compared with object-centric 3D lifting pipelines, it reduces inter-object penetration and spatial ambiguity through scene-aware pose alignment and perspective optimization. Our experiments and ablations show that each component—scene perception, anchor-guided alignment, coarse-to-fine camera refinement, multi-solver physics, and diffusion-based rerendering—contributes meaningfully to the final quality.

Limitations. Despite these advantages, several limitations remain. First, the overall quality still depends on the accuracy of monocular segmentation and 3D reconstruction; failures in instance decomposition or heavily occluded geometry can propagate to later stages. Second, the current system handles rigid and moderately deformable objects well, but highly articulated objects, transparent surfaces, and complex material

Table 7 Ablation on video synthesis inference steps.

Variant	SSIM↑	LPIPS↓	Aesthetic↑	Time (s)↓
10 steps	**0.82**	**0.05**	**0.20**	**539.03**
30 steps	0.79	0.06	0.16	1247.30
50 steps	0.79	0.06	0.15	1922.37

Time: wall-clock inference time per video in seconds.

Table 8 Geometric fidelity of video re-rendering. We measure structural similarity between raw physics simulation and depth-conditioned re-rendered output (10 denoising steps) to verify that the re-rendering preserves underlying physical geometry.

Material	SSIM↑	BG-SSIM↑	Obj-SSIM↑	Centroid↓	ΔIoU
Fluid (7)	0.80	0.84	0.56	42.3	-0.34
Rigid/Elastic (5)	0.65	0.71	0.40	72.7	-0.27
Cloth (3)	0.67	0.74	0.44	42.8	-0.00
Overall (15)	**0.72**	**0.78**	**0.48**	**52.5**	**-0.25**

SSIM: per-frame structural similarity (raw sim. ↔ re-rendered). BG/Obj-SSIM: SSIM within background/object regions. Centroid: mean object centroid displacement (px, image size 880×880). ΔIoU: mask IoU change vs. ground truth after re-rendering.

behavior remain challenging. Third, although the interactive simulation loop is real-time after initialization, high-quality offline rerendering still incurs additional latency. Finally, material parameters and contact properties are not always directly observable from a single image and may require heuristic initialization.

Future work. A promising direction is to jointly infer geometry, material, and contact attributes from image and language supervision, enabling richer physical interaction and better generalization to unseen scenes. Another important avenue is uncertainty-aware scene reconstruction, which would allow the simulator and renderer to reason explicitly about ambiguous geometry under severe occlusion. We believe TelePhysics provides a practical foundation for controllable, image-faithful, and physically consistent video generation from minimal visual input.

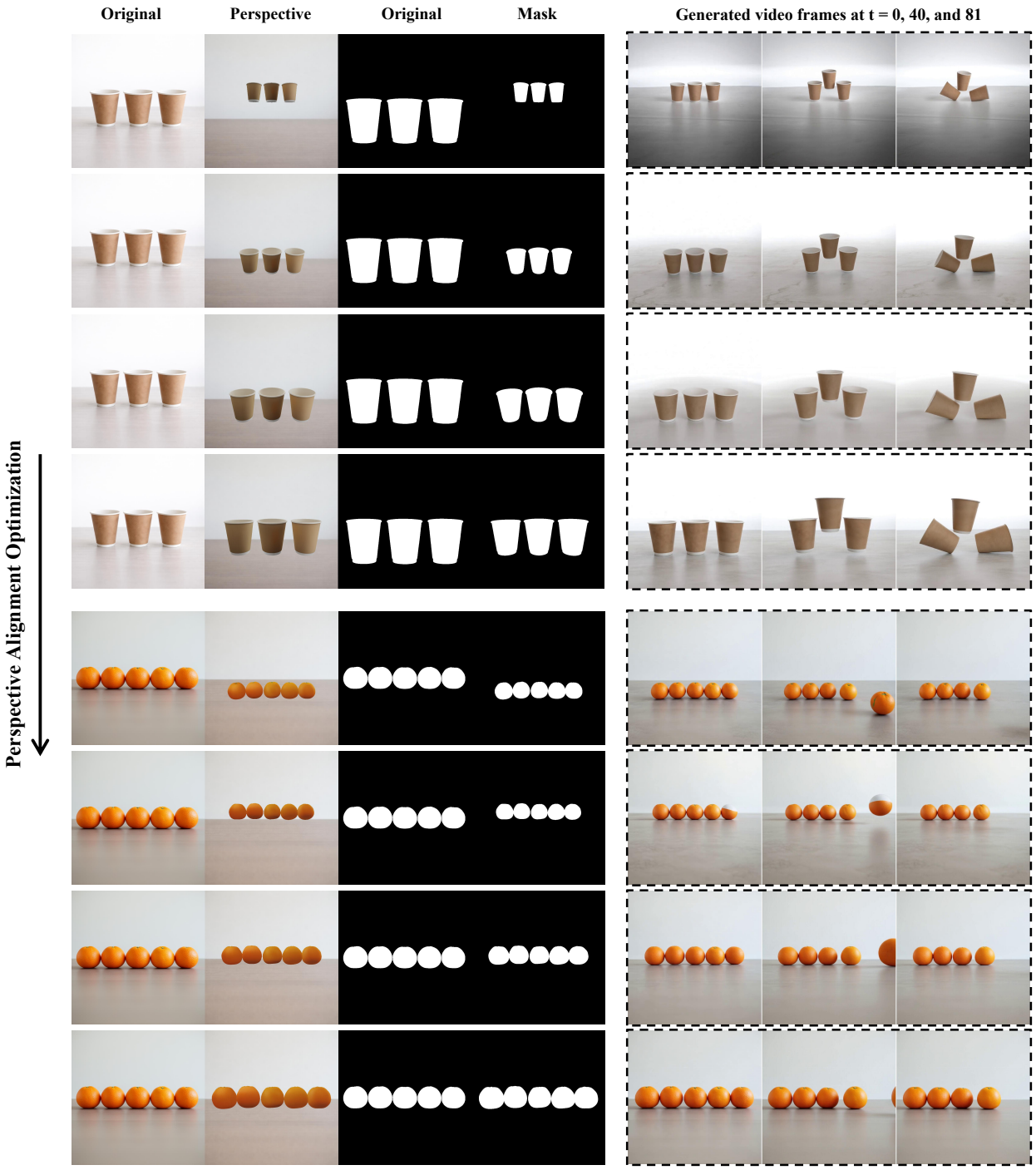

Figure 7 Qualitative visualization of the perspective alignment optimization process. From top to bottom, the rendering loss gradually decreases as the camera parameters are refined. Each row displays the original image, the rendered perspective under the current camera parameters, and the corresponding binary mask. Due to the non-convex nature of the rendering loss, directly identifying a global optimum is challenging. Our proposed coarse-to-fine optimization strategy combines global random sampling with local derivative-free optimization, enabling robust convergence and progressively improved alignment. Suboptimal alignment at this stage would otherwise lead to significant geometric distortion and rendering artifacts in downstream synthesis.

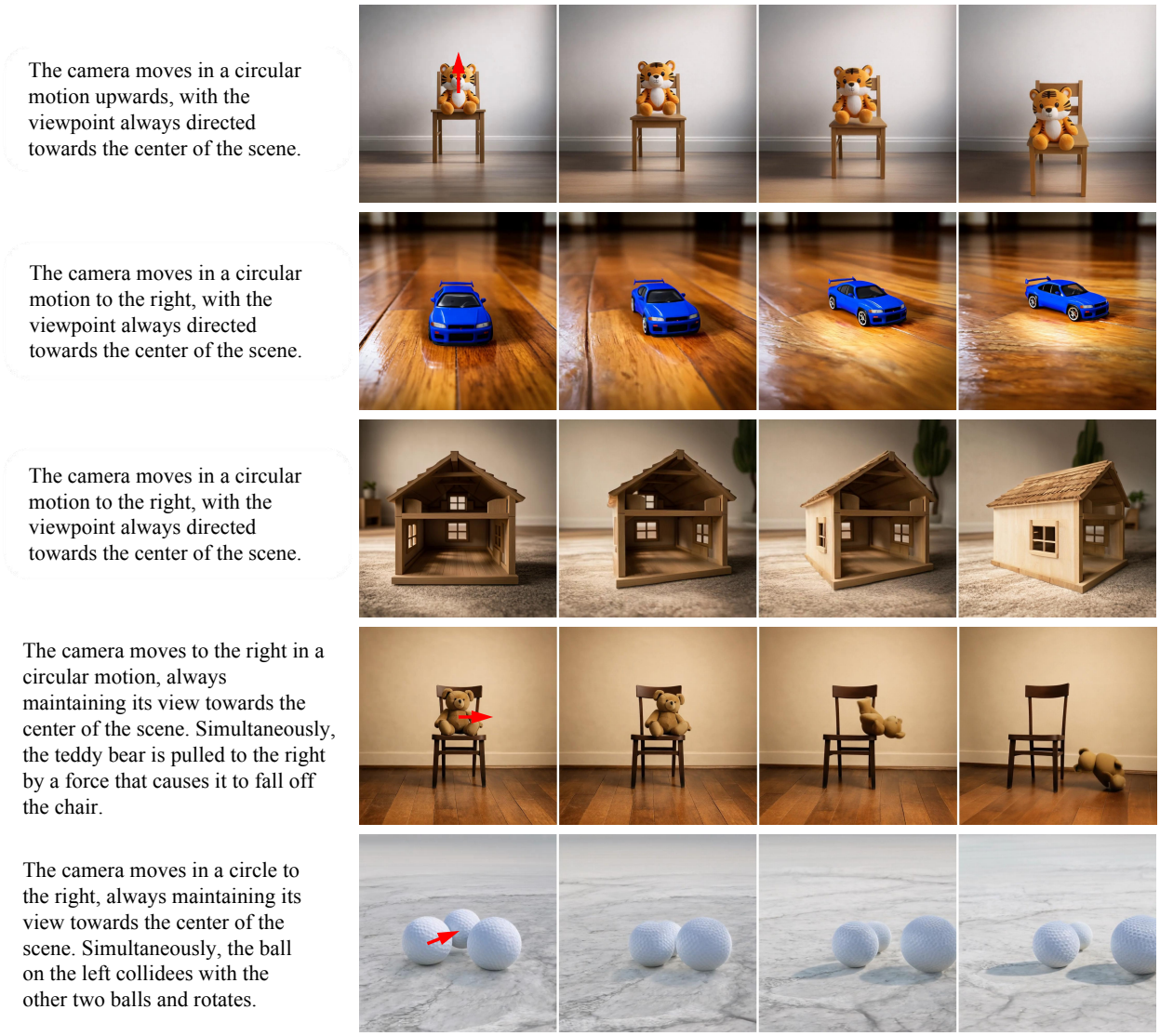

Figure 8 Illustration of camera motion and viewpoint variation

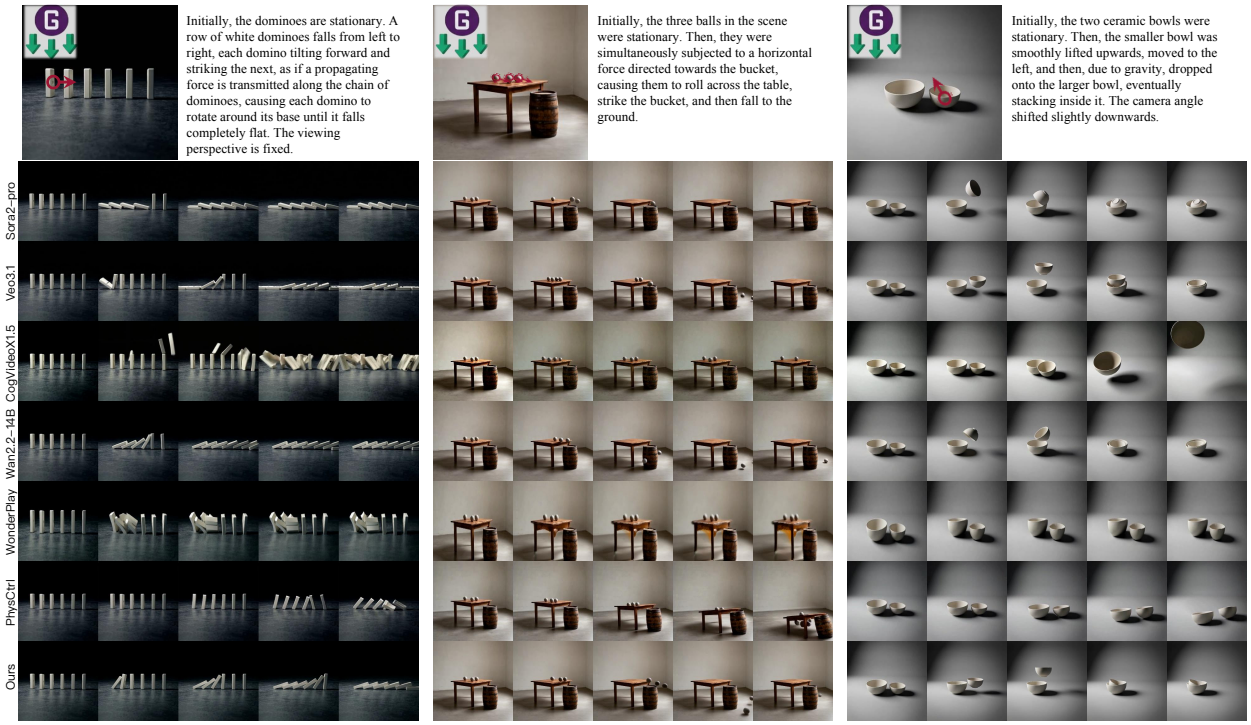

Figure 9 Qualitative comparisons between the proposed approach and existing video generation methods.

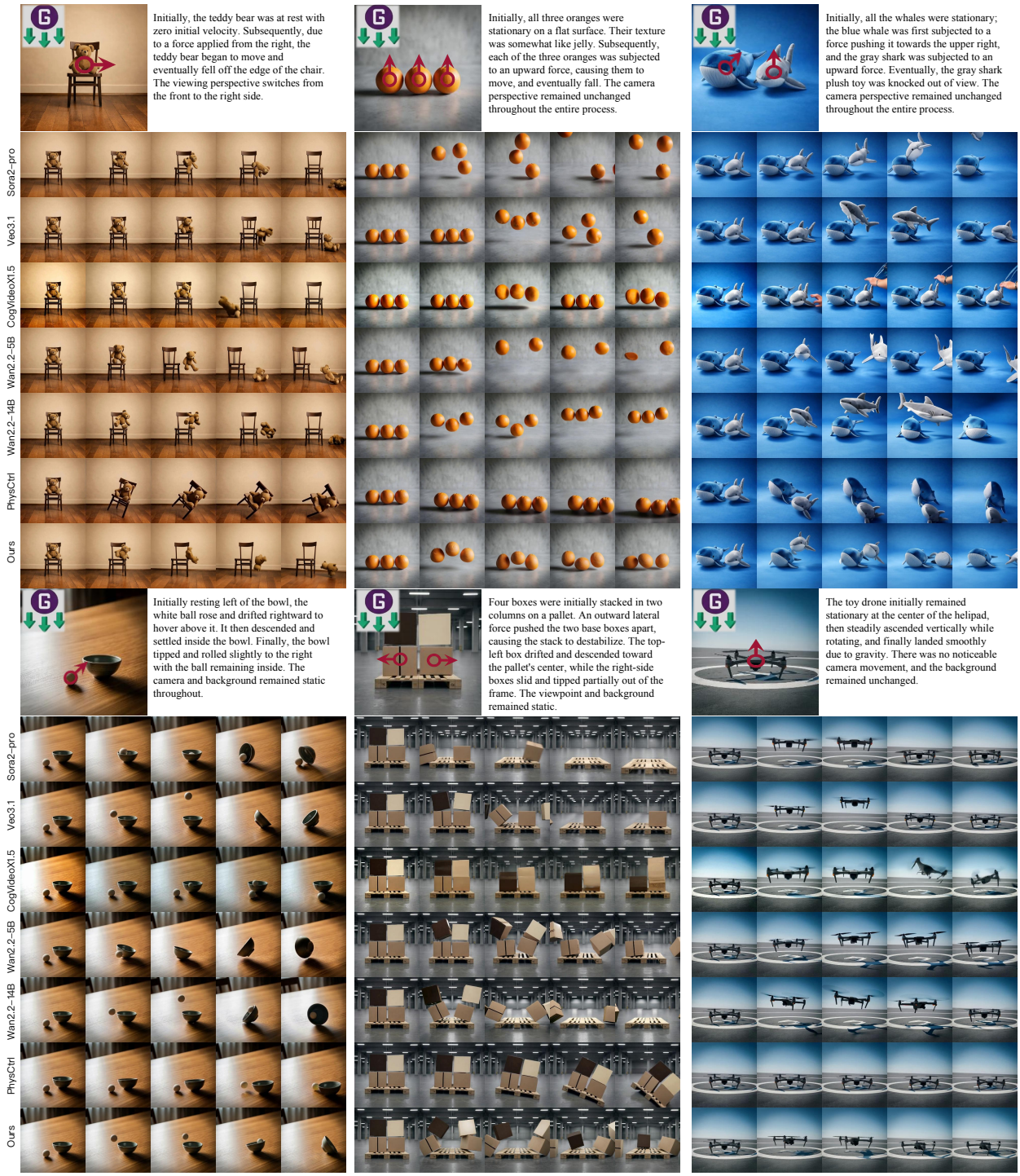

Figure 10 Qualitative comparison between the proposed method and existing video generation methods.

Figure 11 Qualitative results of the proposed method. The approach robustly handles complex interactions among multiple objects, such as simultaneous collisions, thereby demonstrating its effectiveness in challenging dynamic scenarios.

References

An, Hongjun, Hu, Wenhan, Huang, Sida, Huang, Siqi, Li, Ruanjun, Liang, Yuanzhi, Shao, Jiawei, Song, Yiliang, Wang, Zihan, Yuan, Cheng, *et al.* 2026. Ai flow: Perspectives, scenarios, and approaches. *Vicinagearth*, **3**(1), 1.

Authors, Genesis. 2024 (December). *Genesis: A Generative and Universal Physics Engine for Robotics and Beyond*.

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

Baraff, David. 1997. An introduction to physically based modeling: rigid body simulation I—unconstrained rigid body dynamics. *SIGGRAPH course notes*, **82**.

Blattmann, Andreas, Rombach, Robin, Oktay, Deniz, & Esser, Patrick. 2023. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2304.08818*.

Carion, Nicolas, Gustafson, Laura, Hu, Yuan-Ting, Debnath, Shoubhik, Hu, Ronghang, Suris, Didac, Ryali, Chaitanya, Alwala, Kalyan Vasudev, Khedr, Haitham, Huang, Andrew, Lei, Jie, Ma, Tengyu, Guo, Baishan, Kalla, Arpit, Marks, Markus, Greer, Joseph, Wang, Meng, Sun, Peize, Rädle, Roman, Afouras, Triantafyllos, Mavroudi, Effrosyni, Xu, Katherine, Wu, Tsung-Han, Zhou, Yu, Momeni, Liliane, Hazra, Rishi, Ding, Shuangrui, Vaze, Sagar, Porcher, Francois, Li, Feng, Li, Siyuan, Kamath, Aishwarya, Cheng, Ho Kei, Dollár, Piotr, Ravi, Nikhila, Saenko, Kate, Zhang, Pengchuan, & Feichtenhofer, Christoph. 2025. *SAM 3: Segment Anything with Concepts*.

Chen, Boyuan, Jiang, Hanxiao, Liu, Shaowei, Gupta, Saurabh, Li, Yunzhu, Zhao, Hao, & Wang, Shenlong. 2025a. Physgen3d: Crafting a miniature interactive world from a single image. *Pages 6178–6189 of: Proceedings of the Computer Vision and Pattern Recognition Conference*.

Chen, Haonan, *et al.* 2025b. Layout2Scene: Controllable Scene Generation from Structured Layouts. *In: International Conference on Learning Representations (ICLR)*.

Chen, Minghao, Shapovalov, Roman, Laina, Iro, Monnier, Tom, Wang, Jianyuan, Novotny, David, & Vedaldi, Andrea. 2024a. *PartGen: Part-level 3D Generation and Reconstruction with Multi-View Diffusion Models*.

Chen, Xingyu, Chu, Fu-Jen, Gleize, Pierre, Liang, Kevin J, Sax, Alexander, Tang, Hao, Wang, Weiyao, Guo, Michelle, Hardin, Thibaut, Li, Xiang, *et al.* 2025c. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*.

Chen, Yabo, Fang, Jiemin, Huang, Yuyang, Yi, Taoran, Zhang, Xiaopeng, Xie, Lingxi, Wang, Xinggang, Dai, Wenrui, Xiong, Hongkai, & Tian, Qi. 2024b. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. *Pages 311–330 of: European Conference on Computer Vision*. Springer.

Chen, Yabo, Yang, Chen, Fang, Jiemin, Zhang, Xiaopeng, Xie, Lingxi, Shen, Wei, Dai, Wenrui, Xiong, Hongkai, & Tian, Qi. 2024c. LiftImage3D: Lifting any single image to 3D Gaussians with video generation priors. *arXiv preprint arXiv:2412.09597*.

Featherstone, Roy. 2014. *Rigid body dynamics algorithms*. Springer.

Filoni, Sylvain (fffiloni). 2025. *Diffusers Image Outpaint*. Hugging Face Space. https://huggingface.co/spaces/fffiloni/diffusers-image-outpaint.

Fischler, Martin A, & Bolles, Robert C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, **24**(6), 381–395.

Gillman, Nate, Herrmann, Charles, Freeman, Michael, Aggarwal, Daksh, Luo, Evan, Sun, Deqing, & Sun, Chen. 2025. *Force Prompting: Video Generation Models Can Learn and Generalize Physics-based Control Signals*.

Go, Hyojun, Park, Byeongjun, Nam, Hyelin, Kim, Byung-Hoon, Chung, Hyungjin, & Kim, Changick. 2025. *VideoRFSplat: Direct Scene-Level Text-to-3D Gaussian Splatting Generation with Flexible Pose and Multi-View Joint Modeling*.

Google DeepMind. 2025. *Veo 3 Technical Report*. Tech. rept. Google DeepMind. Technical report.

Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., & Fleet, D.J. 2022. *Video diffusion models*. ICLR Workshop on Deep Generative Models for Highly Structured Data.

Hou, Yuenan, Huang, Xiaoshui, Tang, Shixiang, He, Tong, & Ouyang, Wanli. 2024. Advances in 3d pre-training and downstream tasks: a survey. *Vicinagearth*, **1**(1), 6.

Hu, Yuanming, Fang, Yuanhao, Ge, Ziheng, Qu, Zheng, Zhu, Yixin, Pradhana, Arman, & Jiang, Chenfanfu. 2018. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Transactions on Graphics (TOG)*, **37**(4), 1–14.

Huang, Tianyu, Zhang, Haoze, Zeng, Yihan, Zhang, Zhilu, Li, Hui, Zuo, Wangmeng, & Lau, Rynson WH. 2025a. DreamPhysics: Learning Physics-Based 3D Dynamics with Video Diffusion Priors. *Pages 3733–3741 of: Proceedings of the AAAI Conference on Artificial Intelligence*.

Huang, Yuyang, Chen, Yabo, Ding, Li, Zhang, Xiaopeng, Dai, Wenrui, Zou, Junni, Xiong, Hongkai, & Tian, Qi. 2025b. IM-Zero: Instance-level Motion Controllable Video Generation in a Zero-shot Manner. *Pages 7265–7275 of: Proceedings of the Computer Vision and Pattern Recognition Conference*.

Jiang, Chenfanfu, Schroeder, Craig, Teran, Joseph, Stomakhin, Alexey, & Selle, Andrew. 2016. The material point method for simulating continuum materials. *Pages 1–52 of: ACM SIGGRAPH 2016 Courses*.

Klár, Gergely, Gast, Theodore, Pradhana, Arman, Fu, Chuyuan, Schroeder, Craig, Jiang, Chenfanfu, & Teran, Joseph. 2016. Drucker–Prager elastoplasticity for sand animation. *ACM Transactions on Graphics (TOG)*, **35**(4), 1–12.

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

Lai, Zeqiang, Zhao, Yunfei, Liu, Haolin, Zhao, Zibo, Lin, Qingxiang, Shi, Huiwen, Yang, Xianghui, Yang, Mingxin, Yang, Shuhui, Feng, Yifei, Zhang, Sheng, Huang, Xin, Luo, Di, Yang, Fan, Yang, Fang, Wang, Lifu, Liu, Sicong, Tang, Yixuan, Cai, Yulin, He, Zebin, Liu, Tian, Liu, Yuhong, Jiang, Jie, Linus, Huang, Jingwei, & Guo, Chunchao. 2025. *Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details*.

Lee, Jaeho, *et al.* 2024. DreamScene360: Consistent 360-Degree Scene Generation. *In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Xuelong. 2022. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, **35**(6), 8708–8714.

Li, Yangguang, Zou, Zi-Xin, Liu, Zexiang, Wang, Dehu, Liang, Yuan, Yu, Zhipeng, Liu, Xingchao, Guo, Yuan-Chen, Liang, Ding, Ouyang, Wanli, *et al.* 2025a. TripoSG: High-Fidelity 3D Shape Synthesis using Large-Scale Rectified Flow Models. *arXiv preprint arXiv:2502.06608*.

Li, Yunzhu, Wu, Jiajun, & Zhu, Yixin. 2024. PhysDiff: Physics-Guided Video Generation with Diffusion Models. *arXiv preprint arXiv:2401.05363*.

Li, Zizhang, Yu, Hong-Xing, Liu, Wei, Yang, Yin, Herrmann, Charles, Wetzstein, Gordon, & Wu, Jiajun. 2025b. WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions. *In: Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liang, Yuanzhi, Fang, Yijie, Li, Rui, Ni, Ziqi, Su, Ruijie, & Zhang, Chi. 2025. Integrating Reinforcement Learning with Visual Generative Models: Foundations and Advances. *arXiv preprint arXiv:2508.10316*.

Lin, Guying, Huang, Kemeng, Liu, Michael, Gao, Ruihan, Chen, Hanke, Chen, Lyuhao, Lu, Beijia, Komura, Taku, Liu, Yuan, Zhu, Jun-Yan, & Li, Minchen. 2025a. *PAT3D: Physics-Augmented Text-to-3D Scene Generation*.

Lin, Yuchen, Lin, Chenguo, Xu, Jianjin, & Mu, Yadong. 2025b. OmniPhysGS: 3D Constitutive Gaussians for General Physics-Based Dynamics Generation. *In: International Conference on Learning Representations*.

Liu, Anran, Lin, Cheng, Liu, Yuan, Long, Xiaoxiao, Dou, Zhiyang, Guo, Hao-Xiang, Luo, Ping, & Wang, Wenping. 2024a. *Part123: Part-aware 3D Reconstruction from a Single-view Image*.

Liu, Jingren, Wang, Yun, Zhang, Long, Wang, Yiheng, Xu, Shuning, Wang, Ling, Yan, Jiaqi, Zhang, Dell, & Chen, Xiangyu. 2025a. Towards training-free long video understanding: methods, benchmarks, and open challenges. *Vicinagearth*, **2**(1), 6.

Liu, Shaowei, Ren, Zhongzheng, Gupta, Saurabh, & Wang, Shenlong. 2024b. PhysGen: Rigid-Body Physics-Grounded Image-to-Video Generation. *In: European Conference on Computer Vision ECCV*.

Liu, Wei, Chen, Ziyu, Li, Zizhang, Wang, Yue, Yu, Hong-Xing, & Wu, Jiajun. 2026. *RealWonder: Real-Time Physical Action-Conditioned Video Generation*.

Liu, Zhuoman, Ye, Weicai, Luximon, Yan, Wan, Pengfei, & Zhang, Di. 2025b. *PhysFlow: Unleashing the Potential of Multi-modal Foundation Models and Video Diffusion for 4D Dynamic Physical Scene Simulation*.

Mittal, Himangi, Zhuang, Peiye, Lee, Hsin-Ying, & Tulsiani, Shubham. 2025. *UniPhy: Learning a Unified Constitutive Model for Inverse Physics Simulation*.

Müller, Matthias, Heidelberger, Bruno, Hennix, Marcus, & Ratcliff, John. 2007a. Position Based Dynamics. *Journal of Visual Communication and Image Representation*, **18**(2), 109–118.

Müller, Matthias, Heidelberger, Bruno, Hennix, Marcus, & Ratcliff, John. 2007b. Position based dynamics. *Pages 109–118 of: Journal of Visual Communication and Image Representation*, vol. 18. Elsevier.

OpenAI. 2025. *Sora 2 is here*. https://openai.com/index/sora-2/. OpenAI Research Blog.

Paschalidou, Despoina, Ulbrich, Stefan, Schult, Jonas, & Geiger, Andreas. 2021. ATISS: Autoregressive Transformers for Indoor Scene Synthesis. *In: Advances in Neural Information Processing Systems (NeurIPS)*.

Powell, Michael JD. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, **7**(2), 155–162.

Ram, Dinesh, Gast, Theodore, Jiang, Chenfanfu, Schroeder, Craig, Stomakhin, Alexey, Teran, Joseph, & Kavehpour, Hooman. 2015. A material point method for viscoelastic fluids, foams and sponges. *Pages 157–163 of: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*.

Satish, Siddarth Nilol Kundur, Jaiswal, Devesh, Chen, Hongyu, & Bakshi, Abhishek. 2026. *PhysVideoGenerator: Towards Physically Aware Video Generation via Latent Physics Guidance*.

Stomakhin, Alexey, Schroeder, Craig, Chai, Lawrence, Teran, Joseph, & Selle, Andrew. 2013. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, **32**(4), 1–10.

Sulsky, Deborah, Chen, Zhen, & Schreyer, Howard L. 1994a. A particle method for history-dependent materials. *Computer Methods in Applied Mechanics and Engineering*, **118**(1–2), 179–196.

Sulsky, Deborah, Chen, Zhen, & Schreyer, Howard L. 1994b. A particle method for history-dependent materials. *Computer methods in applied mechanics and engineering*, **118**(1-2), 179–196.

Suvorov, Roman, Logacheva, Elizaveta, Mashikhin, Anton, Remizova, Anastasia, Ashukha, Arsenii, Silvestrov, Aleksei, Kong, Naejin, Goka, Harshith, Park, Kiwoong, & Lempitsky, Victor. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. *arXiv preprint arXiv:2109.07161*.

Tan, Xiyang, Jiang, Ying, Li, Xuan, Zong, Zeshun, Xie, Tianyi, Yang, Yin, & Jiang, Chenfanfu. 2024. PhysMotion: Physics-Grounded Dynamics from a Single Image. *arXiv preprint arXiv:2411.17189*.

Tang, Jiaxiang, *et al.* 2024. DiffuScene: Scene Synthesis via Diffusion Models. *In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Wang, Chen, Chen, Chuhao, Huang, Yiming, Dou, Zhiyang, Liu, Yuan, Gu, Jiatao, & Liu, Lingjie. 2025. Physctrl: Generative physics for controllable and physics-grounded video generation. *arXiv preprint arXiv:2509.20358*.

Wang, Peng, Bai, Shuai, Tan, Sinan, Wang, Shijie, Fan, Zhihao, Bai, Jinze, Chen, Keqin, Liu, Xuejing, Wang, Jialin, Ge, Wenbin, Fan, Yang, Dang, Kai, Du, Mengfei, Ren, Xuancheng, Men, Rui, Liu, Dayiheng, Zhou, Chang, Zhou, Jingren, & Lin, Junyang. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Zhou, Bovik, Alan C, Sheikh, Hamid R, & Simoncelli, Eero P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, **13**(4), 600–612.

Wu, Jiajun, Sun, Jiaming, & Li, Yunzhu. 2023. Generative Neural Rendering with Physics Priors. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Tianhao, Zheng, Chuanxia, Guan, Frank, Vedaldi, Andrea, & Cham, Tat-Jen. 2025. *Amodal3R: Amodal 3D Reconstruction from Occluded 2D Images*.

Xiang, Jianfeng, Lv, Zelong, Xu, Sicheng, Deng, Yu, Wang, Ruicheng, Zhang, Bowen, Chen, Dong, Tong, Xin, & Yang, Jiaolong. 2025a. *Structured 3D Latents for Scalable and Versatile 3D Generation*.

Xiang, Xunzhi, Chen, Yabo, Zhang, Guiyu, Wang, Zhongyu, Gao, Zhe, Xiang, Quanming, Shang, Gonghu, Liu, Junqi, Huang, Haibin, Gao, Yang, Zhang, Chi, Fan, Qi, & Li, Xuelong. 2025b. *Macro-from-Micro Planning for High-Quality and Parallelized Autoregressive Long Video Generation*.

Xie, Tianyi, Zong, Zeshun, Qiu, Yuxing, Li, Xuan, Feng, Yutao, Yang, Yin, & Jiang, Chenfanfu. 2024. PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics. *Pages 4389–4398 of: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xie, Tianyi, Zhao, Yiwei, Jiang, Ying, & Jiang, Chenfanfu. 2025. PhysAnimator: Physics-Guided Generative Cartoon Animation. *arXiv preprint arXiv:2501.16550*.

Xu, Jiale, Cheng, Weihao, Gao, Yiming, Wang, Xintao, Gao, Shenghua, & Shan, Ying. 2024. InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models. *arXiv preprint arXiv:2404.07191*.

Yang, Runyu, *et al.* 2024a. SceneCraft: Layout-to-3D Scene Generation with Compositional Priors. *In: European Conference on Computer Vision (ECCV)*.

Yang, Xindi, Li, Baolu, Zhang, Yiming, Yin, Zhenfei, Bai, Lei, Ma, Liqian, Wang, Zhiyong, Cai, Jianfei, Wong, Tien-Tsin, Lu, Huchuan, & Jia, Xu. 2025. *VLIPP: Towards Physically Plausible Video Generation with Vision and Language Informed Physical Prior*.

Yang, Zhuoyi, Teng, Jiayan, Zheng, Wendi, Ding, Ming, Huang, Shiyu, Xu, Jiazheng, Yang, Yuanming, Hong, Wenyi, Zhang, Xiaohan, Feng, Guanyu, *et al.* 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Yang, Ziyi, *et al.* 2024c. PhyScene: Physically Plausible Scene Generation. *In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yao, Kaixin, Zhang, Longwen, Yan, Xinhao, Zeng, Yan, Zhang, Qixuan, Yang, Wei, Xu, Lan, Gu, Jiayuan, & Yu, Jingyi. 2025. *CAST: Component-Aligned 3D Scene Reconstruction from an RGB Image*.

Zhang, Haoze, Huang, Tianyu, Wan, Zichen, Jin, Xiaowei, Zhang, Hongzhi, Li, Hui, & Zuo, Wangmeng. 2025a. PhysChoreo: Physics-Controllable Video Generation with Part-Aware Semantic Grounding. *arXiv preprint arXiv:2511.20562*.

Zhang, Hongyuan, Huang, Sida, Guo, Yubin, & Li, Xuelong. 2025b. Variational positive-incentive noise: How noise benefits models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Longwen, Wang, Ziyu, Zhang, Qixuan, Qiu, Qiwei, Pang, Anqi, Jiang, Haoran, Yang, Wei, Xu, Lan, & Yu, Jingyi. 2024a. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, **43**(4), 1–20.

Zhang, Richard, Isola, Phillip, Efros, Alexei A, Shechtman, Eli, & Wang, Oliver. 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Pages 586–595 of: Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhang, Tianyuan, Yu, Hong-Xing, Wu, Rundi, Feng, Brandon Y., Zheng, Changxi, Snavely, Noah, Wu, Jiajun, & Freeman, William T. 2024b. PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation. *In: European Conference on Computer Vision*. Springer.

Zhang, Xiangdong, Liao, Jiaqi, Zhang, Shaofeng, Meng, Fanqing, Wan, Xiangpeng, Yan, Junchi, & Cheng, Yu. 2025c. *VideoREPA: Learning Physics for Video Generation through Relational Alignment with Foundation Models*.

Zheng, Kaizhi, Fan, Yue, Gu, Jing, Xu, Zishuo, He, Xuehai, & Wang, Xin Eric. 2025. *Self-Evolving 3D Scene Generation from a Single Image*.

Zhou, Xiaoyu, Ran, Xingjian, Xiong, Yajiao, He, Jinlin, Lin, Zhiwei, Wang, Yongtao, Sun, Deqing, & Yang, Ming-Hsuan. 2024. Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. *arXiv preprint arXiv:2402.07207*.

Supplementary Material

A Full Pipeline Algorithm

Algorithm 1 TelePhysics Full Pipeline

Require: RGB image $I \in \mathbb{R}^{H \times W \times 3}$
Ensure: Interactive physics simulation video $\mathcal{V}$

Stage 1: Scene Perception
$\{M_k\}_{k=0}^{K-1} \leftarrow \text{SAM3}(I, \textit{text\_prompts})$ ▷ Instance segmentation
$I_{\text{bg}} \leftarrow \textsc{LaMa}(I, \bigcup_k M_k)$ ▷ Background inpainting
for $k = 0, \ldots, K-1$ **do**
  $\mathcal{M}_k \leftarrow \text{SAM3D}(I, M_k)$ ▷ 3D mesh reconstruction
end for
$\mathcal{C} \leftarrow \text{VLM}(I, \{M_k\})$ ▷ Physics config estimation

Stage 2: Scene Alignment
$\hat{\mathbf{n}}, d, \mathbf{c} \leftarrow \text{RANSAC\_}\textsc{PlaneFit}(\{\mathcal{M}_k\})$ ▷ Ground plane estimation
$\mathbf{R} \leftarrow \textsc{Rodrigues}(\hat{\mathbf{n}}, [0,0,1]^\top)$ ▷ Rotation to z-up frame
for $k = 0, \ldots, K-1$ **do**
  $\mathcal{M}_k^{\text{norm}} \leftarrow \mathbf{R}(\mathcal{M}_k - \mathbf{c}); \quad z \leftarrow z - z_{\min}$ ▷ Ground normalization
end for
$\textsc{ResolvePenetration}(\{\mathcal{M}_k^{\text{norm}}\})$ ▷ AABB de-penetration
$\mathbf{p}_0, \mathbf{l}_0, \phi \leftarrow \textsc{CameraInit}(\{\mathcal{M}_k^{\text{norm}}\})$ ▷ Initial camera pose
$\mathbf{p}^* \leftarrow \textsc{CoarseToFine}(\mathbf{p}_0, \mathcal{L}_{\text{cam}})$ ▷ Camera optimization

Stage 3: Interactive Simulation
Build Genesis scene with materials from $\mathcal{C}$
for $t = 0, \ldots, T-1$ **do**
  $\textsc{Genesis.Step}()$ ▷ Physics step ($\Delta t = 4\,\text{ms}$)
  Activate scheduled forces at frame t
  $F_t \leftarrow \textsc{Composite}(\textsc{Render}(\mathbf{p}^*, \mathbf{l}_0), I_{\text{bg}})$ ▷ Shadow-aware rendering
end for

Stage 4: WonderTrace (Optional)
$\mathcal{D} \leftarrow \textsc{VideoDepthAnything}(\{F_t\})$ ▷ Depth estimation
$\mathcal{V} \leftarrow \textsc{Wan2.2-VACE}(\{F_t\}, \mathcal{D}, I)$ ▷ Diffusion rerendering

B Scene-Aware Pose Alignment

RANSAC Ground Plane Fitting. Given K reconstructed meshes $\{\mathcal{M}_k\}$, we uniformly sample 5,000 surface points per mesh and concatenate them into a point cloud $\mathcal{P} \in \mathbb{R}^{N \times 3}$. We select the lowest $p\%$ of points along the gravity direction ($p = 5$ by default, with fallback to $p = 10$) to form a ground-candidate set $\mathcal{P}_{\text{low}}$. We then fit a plane $\hat{\mathbf{n}}^\top \mathbf{x} + d = 0$ using RANSAC (Fischler & Bolles, 1981) with:

- Distance threshold: $\tau_d = 0.01$
- Minimum sample size: $n_{\text{ransac}} = 3$
- Iterations: $N_{\text{iter}} = 2{,}000$

The fitted normal $\hat{\mathbf{n}}$ is validated against the gravity hint $\mathbf{u}_g = [0,0,1]^\top$ by checking $|\hat{\mathbf{n}} \cdot \mathbf{u}_g| \geq \tau_{\cos}$ ($\tau_{\cos} = 0.8$ default). If validation fails, we apply an adaptive retry strategy (up to 8 attempts), progressively relaxing

p, τ_d, and $\tau_{\cos}$.

Scene Normalization. We compute the rotation matrix $\mathbf{R}$ that aligns $\hat{\mathbf{n}}$ with $[0,0,1]^\top$ via the Rodrigues formula:

$$\mathbf{R} = \mathbf{I} + [\mathbf{v}]_\times + [\mathbf{v}]_\times^2 \cdot \frac{1-c}{\|\mathbf{v}\|^2}, \quad \mathbf{v} = \hat{\mathbf{n}} \times [0,0,1]^\top, \quad c = \hat{\mathbf{n}} \cdot [0,0,1]^\top. \tag{13}$$

where $[\mathbf{v}]_\times$ is the skew-symmetric matrix of $\mathbf{v}$. All meshes are transformed as $\mathbf{x}' = \mathbf{R}(\mathbf{x} - \bar{\mathbf{x}})$ and then shifted so $z_{\min} = 0$.

Penetration Resolution. For multi-object scenes, we detect pairwise AABB overlaps and iteratively push objects apart along the axis of minimum overlap:

$$\Delta_i = \frac{1}{2} \min_{\text{axis}} \left(\frac{e_i^a + e_j^a}{2} - |c_i^a - c_j^a| \right), \quad \|\Delta_i\| \le \delta_{\max} = 0.05. \tag{14}$$

where e^a and c^a denote the extent and center along axis a. Two iterations of pairwise repulsion suffice to resolve most interpenetrations.

C Camera Pose Optimization

Given an initial camera position $\mathbf{p}_0$ (from heuristic ground-plane fitting or DA3 estimation), we optimize the camera pose to minimize a composite loss:

$$\mathcal{L}_{\text{cam}}(\mathbf{p}) = w_{\text{obj}} \cdot \mathcal{L}_{\text{obj}}(\mathbf{p}) + w_{\text{bg}} \cdot \mathcal{L}_{\text{bg}}(\mathbf{p}) + w_{\text{mask}} \cdot \mathcal{L}_{\text{mask}}(\mathbf{p}), \tag{15}$$

with default weights $w_{\text{obj}} = 1.0$, $w_{\text{bg}} = 0.2$, $w_{\text{mask}} = 1.0$.

Object Region Loss. Measures pixel-wise MSE within the ground-truth object mask Ω_{obj}:

$$\mathcal{L}_{\text{obj}} = \frac{\sum_{(u,v)\in\Omega_{\text{obj}}} \|\hat{I}(u,v) - I^*(u,v)\|^2}{\sum_{(u,v)} \Omega_{\text{obj}}(u,v) \cdot C + \epsilon}, \tag{16}$$

where $\hat{I}$ is the rendered image, I^* is the target, and C is the number of channels.

Background Region Loss. Same formulation as $\mathcal{L}_{\text{obj}}$ but evaluated over $1 - \Omega_{\text{obj}}$.

Mask Alignment Loss. Dice loss between the rendered silhouette $\hat{\Omega}$ and Ω_{obj}:

$$\mathcal{L}_{\text{mask}} = 1 - \frac{2\sum \hat{\Omega} \cdot \Omega_{\text{obj}} + \epsilon}{\sum \hat{\Omega} + \sum \Omega_{\text{obj}} + \epsilon}. \tag{17}$$

Coarse-to-Fine Optimization. We apply a two-stage strategy:

1. **Global search**: 60 uniform random samples within the search bounds $[\mathbf{p}_0 \pm \boldsymbol{\delta}]$, where $\boldsymbol{\delta} = (\delta_x, \delta_y, \delta_z)$ are per-axis search radii (default $\delta_y = \delta_z = 0.5$).
2. **Local refinement**: Powell's conjugate direction method (Powell, 1964) initialized at the best random sample, with a maximum of 80 iterations.

D Details of Physical Simulation

D.1 VLM-Based Physics Configuration

We use Qwen2.5-VL-72B-Instruct (Wang *et al.*, 2024) to automatically estimate per-object physics materials and scene force fields from a single image. The VLM receives: (1) the full scene image, and (2) per-object crops extracted via segmentation masks. The complete system prompt is shown in Figure 12.

VLM System Prompt for Physics Configuration

```
You are an expert physics material analyst and simulation designer. I will show you a scene image followed by cropped
images of individual objects.
Task A: Object Materials
For each object (numbered starting from 0), identify:
1. What the object is (e.g., sand castle, rubber duck)
2. Best-matching physics material type for INTERESTING DEFORMABLE simulation.
IMPORTANT: Prefer non-rigid materials --- choose the MOST DEFORMABLE plausible interpretation.
Available material types:
  MPM materials (particle-based, fluids/deformation):
- "mpm_liquid": liquids, viscous fluids. Params: E, nu, rho, viscous
- "mpm_elastoplastic": permanent deformation (clay, cream). Params: E, nu, rho, use_von_mises, yield_stress
- "mpm_elastic": elastic deformable (rubber, jelly). Params: E, nu, rho, model
- "mpm_sand": granular (sand, powder). Params: E, nu, rho, friction_angle
- "mpm_snow": snow/ice. Params: E, nu, rho, yield_lower, yield_higher
  PBD materials (position-based dynamics):
- "pbd_cloth": thin sheet (fabric, paper). Params: rho, friction, compliance, air_resistance
- "pbd_elastic": 3D soft body (sponge). Params: rho, friction, compliance
- "pbd_liquid": position-based fluid. Params: rho, density/viscosity relaxation
  Rigid (use sparingly, ONLY for immovable structures):
- "rigid": ground, wall, table. Do NOT use for small or deformable objects.
3. material_params: E (Young modulus), rho (density), nu (Poisson ratio)
4. fixed: true ONLY if truly static
5. surface_color: RGB float [0--1]
Task B: Force Fields
Suggest 1--3 force fields for interesting dynamics. Types:
constant, wind, point, drag, turbulence, vortex
Each with: direction, strength, start_frame (-1 = immediate).
Respond with ONLY a JSON object: {"objects": [...], "forces": [...]}
```

Figure 12 Complete VLM prompt for automatic physics configuration. The prompt instructs Qwen2.5-VL-72B to identify per-object materials and suggest force fields from the scene image and per-object crops.

Parameter Guidance. We provide the VLM with parameter ranges to ensure physically plausible outputs:
- Young's modulus E: foam $\sim 10^4$, jelly $\sim 10^3$, rubber $\sim 10^6$, glass $\sim 10^5$
- Density ρ (kg/m^3): foam ~ 50, cream ~ 500, rubber ~ 1100, clay ~ 1800, glass ~ 2500
- Poisson's ratio ν: typical 0.2–0.45; nearly incompressible ≈ 0.45

Postprocessing. The VLM output JSON is parsed with regex-based extraction (handling both `{"objects":...}` and legacy array formats). Unknown material types fall back to `mpm_elastic`. Invalid force types are discarded. All parameters are validated against type-specific defaults (Table 9).

Table 9 lists all supported physics materials and their default parameters. Table 10 lists the available force field types.

D.2 Multi-Physics Solvers Formulation

As discussed in Sec. 3.3, our physical simulation relies on a unified multi-physics backend, Genesis Authors (2024), which integrates three specialized solvers. The detailed formulations and integration strategies are described below.

Rigid Body Dynamics (RBD). We utilize an RBD solver based on articulated body algorithms Featherstone (2014) for non-deformable objects and robotic manipulators. This method ensures stable, penetration-free interactions and accurate friction handling. The dynamics are governed by the generalized equation of motion:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) = \boldsymbol{\tau} + \mathbf{J}(\mathbf{q})^T \mathbf{f}_{\text{ext}}, \tag{18}$$

where $\mathbf{q}$ and $\dot{\mathbf{q}}$ denote the generalized coordinates and velocities, respectively. $\mathbf{M}$ represents the inertia matrix, $\mathbf{C}$ accounts for Coriolis and centrifugal forces, $\boldsymbol{\tau}$ denotes actuation torques, and $\mathbf{f}_{\text{ext}}$ represents external forces (e.g., contact) mapped into the joint space via the Jacobian $\mathbf{J}$. By solving for the acceleration $\ddot{\mathbf{q}}$, the solver updates the kinematic state of rigid entities without the numerical dissipation typical of particle-based methods.

Table 9 Supported physics material types and default parameters.

Material	Solver	Default Parameters
`rigid`	Rigid Body	$\rho = 200$, $\mu = 0.7$
`mpm_elastic`	MPM	$E = 3 \times 10^5$, $\nu = 0.2$, $\rho = 1000$, model = corotation
`mpm_elastoplastic`	MPM	$E = 3 \times 10^4$, $\nu = 0.4$, $\rho = 100$, von Mises yield = 10^4
`mpm_sand`	MPM	$E = 5 \times 10^5$, $\nu = 0.2$, $\rho = 1800$, friction angle = 45°
`mpm_liquid`	MPM	$E = 10^6$, $\nu = 0.2$, $\rho = 1000$, viscous = false
`mpm_snow`	MPM	$E = 10^6$, $\nu = 0.2$, $\rho = 1000$, yield $\in [0.025, 0.0045]$
`mpm_muscle`	MPM	$E = 10^6$, $\nu = 0.2$, $\rho = 1000$, model = Neo-Hookean
`pbd_elastic`	PBD	$\rho = 1000$, stretch/bending/volume compliance = 0, relaxation = 0.1
`pbd_cloth`	PBD	$\rho = 4\,\mathrm{kg/m^2}$, stretch compliance = 10^{-7}, bending = 10^{-5}, air resistance = 10^{-3}
`pbd_liquid`	PBD	$\rho = 1000$, density relaxation = 0.2, viscosity relaxation = 0.01
`pbd_particle`	PBD	$\rho = 1000$

Table 10 Supported force field types and default parameters.

Type	Default Parameters
`constant`	direction = $[0, 0, -1]$, strength = 9.8
`wind`	direction = $[1, 0, 0]$, strength = 1.0, radius = 1.0
`point`	strength = 1.0, position = $[0, 0, 0]$, falloff power = 0
`drag`	linear = 0, quadratic = 0
`vortex`	direction = $[0, 0, 1]$, perpendicular strength = 20.0
`turbulence`	strength = 1.0, frequency = 3
`noise`	strength = 1.0

Material Point Method (MPM). We employ MPM Sulsky *et al.* (1994b); Hu *et al.* (2018); Jiang *et al.* (2016); Klár *et al.* (2016); Ram *et al.* (2015); Stomakhin *et al.* (2013) to simulate fluids and hyperelastic materials. As a hybrid Lagrangian-Eulerian method, MPM tracks mass and momentum on Lagrangian particles (p) while computing forces on a background Eulerian grid (i). The governing equation for the grid node force $\mathbf{f}_i$ is derived from the divergence of particle stress $\boldsymbol{\sigma}_p$:

$$\mathbf{f}_i = \sum_p V_p \boldsymbol{\sigma}_p \nabla w_{ip} + \mathbf{f}_{\text{ext}}, \tag{19}$$

where V_p is the particle volume and ∇w_{ip} is the gradient of the interpolation kernel function connecting particle p to grid node i. In each time step, particle states are transferred to the grid to solve the momentum equation, and the updated grid velocities are interpolated back to deform the particles, naturally facilitating the simulation of fracture, splashing, and plastic flow.

Position-Based Dynamics (PBD). We apply PBD Müller *et al.* (2007b) for thin-shell structures such as cloth and cables. Unlike force-based solvers that require small time steps for stability in stiff systems, PBD directly modifies object positions to satisfy geometric constraints. The position correction $\Delta \mathbf{x}_i$ for a particle

i is computed to satisfy a constraint function $C(\mathbf{x}) = 0$:

$$\Delta \mathbf{x}_i = -w_i \frac{C(\mathbf{x})}{\sum_j w_j |\nabla_{\mathbf{x}_j} C(\mathbf{x})|^2} \nabla_{\mathbf{x}_i} C(\mathbf{x}), \tag{20}$$

where w_i is the inverse mass of particle i. By iteratively projecting positions onto the constraint manifold, PBD ensures that cloth remains inextensible and drapes naturally.

Solver Integration and Optimization. The integration of these three solvers is critical for high-fidelity multi-material simulations. The solvers operate concurrently, with interactions managed by transferring forces and state information at each simulation step. For instance, contact forces from the RBD solver are propagated to the MPM solver to induce material deformation, while the PBD solver enforces constraints during interactions with rigid bodies. To address computational challenges, we leverage GPU-based parallelism. Each solver operates independently on its respective domain subset, utilizing spatial partitioning to optimize interaction computations. Furthermore, adaptive time-stepping dynamically adjusts the simulation frequency based on local material properties, ensuring a balance between accuracy and efficiency.

E Details of Experiments

E.1 Simulation Details

Physics Engine. We build on Genesis (Authors, 2024), a GPU-accelerated differentiable physics engine. The simulation operates at $\Delta t = 4\,\text{ms}$ with 10 substeps per step, using the MPM (Material Point Method) solver for continuum materials and PBD (Position-Based Dynamics) for cloth and particles. Rigid–MPM and Rigid–PBD coupling is enabled via the legacy coupler. The MPM domain spans $[-2, 2]^3$ with particle size 0.01.

Rendering. Each simulation frame is composited via a shadow-aware pipeline:
1. Render with segmentation: obtain RGB, segmentation IDs per pixel.
2. Extract object mask (seg_id ≥ 2) and plane shadow mask (seg_id $= 1$ and brightness < 0.3).
3. Composite: $F = I_{\text{bg}} \cdot (1 - \alpha_{\text{obj}}) + I_{\text{render}} \cdot \alpha_{\text{obj}}$, then apply shadow darkening with strength 0.3.
Resolution is fixed at 880×880.

PBD Cloth Fixation. For cloth-like objects (e.g., dresses), we support a `fix_top_ratio` parameter that pins the topmost $r\%$ of particles by z-coordinate after scene building, simulating hanging or attachment points.

Camera Motion. Six camera motion modes are supported: four orbital (around XY or YZ axes, clockwise/counterclockwise), one lateral, and one descent. The angular velocity is $v = 0.001$ rad/frame at 60 FPS.

E.2 Mesh Reconstruction Details

Segmentation. We use SAM3 for text-prompted instance segmentation. Given text prompts (e.g., "glass ball", "sand castle"), SAM3 produces per-object binary masks $\{M_k\}$. A combined binary mask $\bigcup_k M_k$ is computed for background inpainting.

Background Inpainting. LaMa inpaints the masked regions to produce a clean background I_{bg}. We apply cumulative inpainting: objects are removed sequentially to handle overlapping masks. Dilation kernel size is 100 pixels.

3D Reconstruction. SAM3D reconstructs per-object 3D meshes from single-view images with per-object masks. The model predicts rotation (quaternion), scale, and translation for each object, which are composed into an affine transform:

$$\mathbf{x}_{\text{world}} = s \cdot \mathbf{R}_q \, \mathbf{x}_{\text{local}} + \mathbf{t}, \tag{21}$$

where $\mathbf{R}_q$ is the rotation matrix from the predicted quaternion, s is the scale factor, and $\mathbf{t}$ is the translation. A coordinate frame conversion (Y-up to Z-up) is applied before exporting as OBJ files.

E.3 Evaluation Metrics

Image Quality.

- **SSIM**: Structural Similarity (Wang *et al.*, 2004) between rendered first frame and input image (grayscale, data range 255).
- **LPIPS**: Learned Perceptual Image Patch Similarity (Zhang *et al.*, 2018) using AlexNet backbone. Falls back to normalized MSE ($\min(4 \cdot \text{MSE}, 1)$) when model weights are unavailable.

Alignment Metrics.

- **Mask IoU**: Binary intersection-over-union between the rendered object silhouette and the GT segmentation mask.
- **Reproj. Error**: L_2 distance (pixels) between the centroids of the predicted and GT masks.

Physics Metrics.

- **Penetration Rate (PR)**: Fraction of object pairs with AABB overlap. Lower is better.
- **Support Violation Rate (SVR)**: Fraction of objects with $z_{\min} > 0.02$ (floating above ground). Lower is better.
- **Interaction Success Rate (ISR)**: Fraction of frames where the object mask covers $> 0.1\%$ of the image area (880×880). Higher is better.

Video Quality.

- **Motion Amplitude**: Mean optical flow magnitude (Farneback) across consecutive frames.
- **Motion Smoothness**: Inverse of frame-to-frame flow variance: $S = 1/(1 + \text{Var}(\{\bar{f}_t\}))$.
- **Aesthetic**: CLIP-based (ViT-B/32) cosine similarity with "a beautiful high quality scene". Falls back to sigmoid-normalized Laplacian variance: $A = 2/(1 + e^{-\bar{\sigma}_L^2/500}) - 1$.

F Quantitative Evaluation via GPT-5

To quantitatively evaluate the 59-scene test set for controllability, physical plausibility, and overall video quality, we adopt an automated Vision-Language Model (VLM) scoring framework inspired by the VideoPhy Bansal *et al.* (2024) protocol. Specifically, we utilize a GPT-5-based evaluator to assign 5-point Likert scores to the generated videos.

F.1 Data Preparation and Frame Sampling

To construct the evaluation queries while adhering to the context limits of the VLM, we preprocess the input images and generated videos as follows:

- **Frame Extraction:** For each generated video, we uniformly sample 10 evenly spaced frames across the temporal axis to represent the full sequence.
- **Resolution Scaling:** Both the initial input image and the sampled video frames are downscaled such that their longest edge is 880 pixels, preserving the aspect ratio.
- **Visual Prompts:** The input image explicitly visualizes the initial motion position and direction using a red arrow, providing the model with a clear spatial reference for the applied physical constraints. All images and frames are converted into base64 JPEG format before being parsed to the VLM.

F.2 Evaluation Prompt and Criteria

The VLM is provided with the text prompt, the initial condition image, and the sequences of 10 sampled frames from the compared models (presented in chronological order). The model is instructed to assess the videos on a scale of 1 (poor) to 5 (excellent) based on three complementary criteria: Semantic Adherence, Physical Commonsense, and Video Quality.

The exact prompt template used for the evaluation is provided below:

> *You are tasked with evaluating the quality of image-to-video generation produced by a model. For each test case, you will be given:*
> *1. A text prompt describes one or more objects along with the initial motion direction. The motion's position and direction are visualized as a red arrow in the input image.*
> *2. An input image of the object.*
> *3. Eight sets of 10 evenly spaced frames—each set corresponds to a video generated by a different model from the same input.*
>
> *Please evaluate this video based on the following three criteria using a 5-point Likert scale (1 = poor, 5 = excellent):*
> - **Semantic Adherence:** *How well the content and motion in the video match the description in the text prompt, especially the alignment with the motion direction and position. Note that the video should starts with the input image.*
> - **Physical Commonsense:** *Whether the object's motion follows intuitive, physically plausible dynamics given the applied force direction and position.*
> - **Video Quality:** *The overall visual and temporal quality of the video (note that static or nearly-static sequences are less preferred).*
>
> *Provide your evaluation for each video strictly in the following one-line format:*
> *Model i, Semantic Adherence score, Physical Commonsense score, Video Quality score*

F.3 Hyperparameter Settings

Table 11 Key hyperparameters.

Component	Parameter	Value
Physics	Time step Δt	4 ms
	Substeps	10
	Default simulation steps T	300
	Render FPS	60
RANSAC	Distance threshold τ_d	0.01
	Min. samples n_{ransac}	3
	Iterations	2,000
	Horizontality threshold $\tau_{\cos}$	0.8
Camera Opt.	Random search samples N_{rand}	60
	Powell max iterations	80
	Object loss weight w_{obj}	1.0
	Background loss weight w_{bg}	0.2
	Mask loss weight w_{mask}	1.0
Video Synthesis	Denoising steps	10
	Number of frames	81
	Output FPS	15
Penetration	AABB padding	0.01
	Max displacement $\delta_{\max}$	0.05