PhysOmni: Physics-Grounded Multi-Object Scene Generation from a Single Image with Real-Time Interaction

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 (TeleAI), China Telecom
[*]Equal contributions. [†]Work done during an internship at TeleAI. [✉]Corresponding authors.

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 PhysOmni, 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, PhysOmni resolves object penetration and alignment ambiguity. Unlike prior methods, this formulation enables accurate scene-level multi-object interactions and introduces richer, complex control types for advanced mechanics-based manipulation. By decoupling simulation from rendering, PhysOmni bypasses latency-heavy priors, achieving real-time physical interaction previews paired while preserving photorealistic visual fidelity. Experimental results demonstrate that PhysOmni substantially outperforms prior methods in physical fidelity, spatial coherence, and controllability.

Keywords: Video Generation; Physical Consistency; Scene-level 3D Reconstruction; Controllable Generation; Image-to-Video
Project Page: https://physomni.github.io/

1 Introduction

Advances in visual representation and adaptation Xiang *et al.* (2025b); Huang *et al.* (2025a), AI-flow systems An *et al.* (2026); Shao & Li (2025), and neural video compression Chen *et al.* (2026b); Yuan *et al.* (2026a) improve visual modeling and delivery, while generative video models Xiang *et al.* (2025c); Huang *et al.* (2025b); Wang *et al.* (2026); Xiang *et al.* (2026); Zhang *et al.* (2026) achieve strong quality and controllability; however, these developments remain appearance- or systems-driven and lack explicit mechanical reasoning. As summarized in Figure 2, generating physics-grounded scenes from a single image is hindered by five primary limitations. First, current video generators are physically uncontrollable (Figure 2a) 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.* (2025); Yang *et al.* (2025a); 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 fail to capture long-horizon, complex physical interactions.

To improve physical reasoning, recent work has explored 3D reconstruction, 4D world modeling Chen *et al.* (2025c, 2026a), 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.* (2025b); 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 often causes severe

Figure 1 PhysOmni reconstructs a holistic 3D scene from one image for controllable, physics-grounded multi-object video generation. Additional results are provided in Appendix J.

multi-object inter-penetration (Figure 2b), while directly combining 3D generation with physical control frequently yields cartoonish appearances (Figure 2d). Scene-level 3D generation methods jointly model object layouts and physical plausibility Paschalidou *et al.* (2021); Tang *et al.* (2024); Yang *et al.* (2024, 2025b); Zhou *et al.* (2024b); Chen *et al.* (2025b); Lin *et al.* (2025a), using scalable explicit representations like Gaussian splatting Zhou *et al.* (2024b); Yang *et al.* (2025b); Go *et al.* (2025); Zhou *et al.* (2024a). Nevertheless, aligning generated scenes with the input image remains challenging. Spatial misalignment (Figure 2c) in object pose, scale, and orientation is frequent, and enforcing physical constraints can conflict with image evidence under occlusion and limited viewpoints Yang *et al.* (2024); Yao *et al.* (2025); Wu *et al.* (2025a), introducing inconsistency with the input image (Figure 2e).

We present PhysOmni, a training-free unified framework for holistic 3D scene generation and physics simulation from a single image. Unlike fragmented object-centric pipelines, PhysOmni jointly models all elements within a shared 3D representation under consistent spatial constraints, resolving penetration and alignment ambiguities. This unified formulation enables accurate scene-level multi-object interactions and introduces richer control types for advanced mechanics-based manipulation (e.g., diverse force fields and 3D velocities). By decoupling simulation from rendering, PhysOmni bypasses latency-heavy priors, providing real-time interactive physics previews to transform static images into dynamic, physically coherent environments.

We evaluate PhysOmni 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, reducing artifacts like drifting and penetration over long-horizon simulations. In challenging multi-object scenes, PhysOmni 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, mitigating inter-penetration. Quantitative and qualitative re-

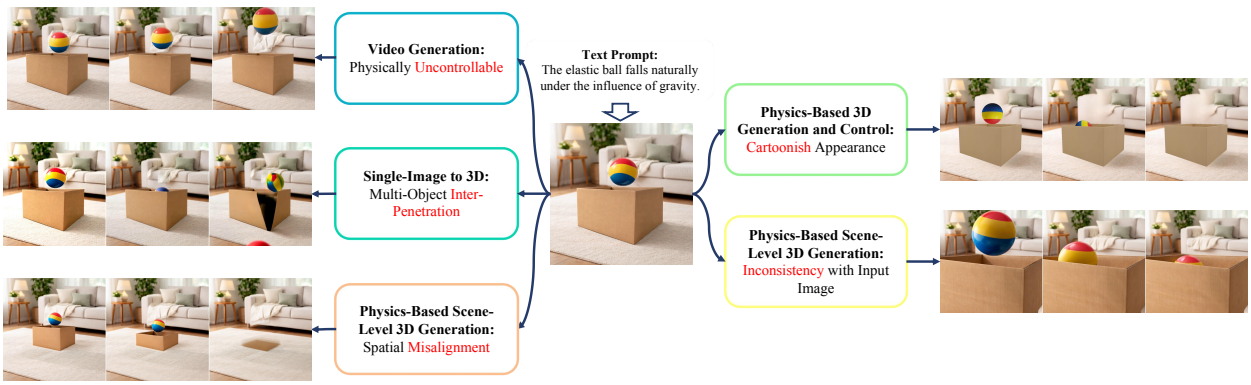

Figure 2 Five limitations of physics-grounded scene generation: (a) uncontrollable motion, (b) object interpenetration, (c) spatial misalignment, (d) cartoonish rendering, and (e) inconsistency with the input image.

sults show that PhysOmni outperforms prior methods in physical realism, temporal consistency, and visual fidelity.

Our contributions are summarized as follows:

- We introduce PhysOmni, a training-free framework enabling accurate scene-level multi-object interactions and complex control types. By decoupling simulation from rendering, it delivers real-time interactive physics previews while maintaining the visual fidelity of the input.
- We propose a Scene-Aware Pose Alignment mechanism that transforms representations of egocentric meshes into a unified world coordinate system and anchors them to a canonical ground plane, which enforces coherent and physically plausible configurations of the scene.
- We introduce a coarse-to-fine camera pose optimization strategy for perspective alignment, which ensures accurate photometric and geometric consistency between reconstructed objects and the input image.

2 Related Works

A brief overview of the related literature is provided below. A more comprehensive discussion and an extended review of related works can be found in Appendix A.

Controllable and Physics-Aware Video Generation. Recent video diffusion and autoregressive models, including reference-controllable long-form generation systems Huang *et al.* (2026), achieve remarkable visual realism but remain predominantly appearance-driven, lacking explicit 3D geometry and an understanding of mechanics. To bridge this gap, several methods incorporate physical principles through auxiliary priors or intermediate signals Liu *et al.* (2024b); Zhang *et al.* (2024b); Xie *et al.* (2025); Gillman *et al.* (2025); Wang *et al.* (2025); Li *et al.* (2025b); Yuan *et al.* (2026b); Zhang *et al.* (2025); Satish *et al.* (2026); Liu *et al.* (2026); Xiong *et al.* (2026). However, these approaches rely on soft constraints and manually specified parameters, which limits their applicability to short and simplistic motions. In contrast, PhysOmni enables real-time and explicit mechanical control over complex multi-object scenes using diverse force fields (*e.g.*, vortex, wind, and gravity) without relying on latency-heavy priors.

Single-Image to 3D Reconstruction. Significant advances have been made in lifting single images to 3D representations via SDS-based optimization Poole *et al.* (2022), feed-forward networks Xu *et al.* (2024); Li *et al.* (2025a); Zhang *et al.* (2024a); Lai *et al.* (2025); Xiang *et al.* (2025a), and video diffusion priors Chen *et al.* (2024c,b); Wu *et al.* (2025b). However, these pipelines reconstruct objects in isolation within egocentric coordinate systems. Even segmentation-aware methods Liu *et al.* (2024a); Chen *et al.* (2024a) lack global spatial awareness, which leads to severe interpenetration when assembling scenes. PhysOmni introduces a *Scene-Aware Pose Alignment* mechanism that anchors all elements to a canonical ground plane within a unified world coordinate system. This approach inherently resolves ambiguities related to penetration.

Because the contact-anchor cue is a generic geometric prior rather than a category-specific one, and material parameters are freely selected by the VLM, PhysOmni generalizes to new scenes and unseen object categories without any category-specific training.

Physics-Grounded Scene Generation. Recent works jointly model object layouts and physical plausibility by utilizing scalable representations such as Gaussian splatting Paschalidou *et al.* (2021); Tang *et al.* (2024); Yang *et al.* (2024, 2025b); Zhou *et al.* (2024b); Chen *et al.* (2025b); Lin *et al.* (2025b); Go *et al.* (2025); Zhou *et al.* (2024a); Yao *et al.* (2025); Wu *et al.* (2025a); Meng *et al.* (2025). However, the imposition of physical constraints often conflicts with visual evidence under occlusion Yang *et al.* (2024); Yao *et al.* (2025); Wu *et al.* (2025a), which causes spatial discrepancies or a loss of input fidelity. PhysOmni addresses this issue through coarse-to-fine camera pose optimization for precise photometric alignment. Crucially, it decouples the physical simulation from the rendering process to deliver real-time interactive physics previews while preserving the original appearance.

3 Methods

Existing pipelines for physics-grounded scene generation face several interconnected challenges (Fig. 2). Controlling generative video models to enforce physical constraints and target motions remains inherently difficult. Furthermore, current 3D lifting methods often yield physically implausible layouts with intersecting shapes, lack photorealistic details, and tend to drift from the input image, degrading both structural and visual fidelity.

We address these challenges with PhysOmni, 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 significantly improve rendering speed and 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. To prevent artifacts such as penetration between objects and fragmented geometry, we decompose the input image into instance-level foreground geometry and a globally consistent background(Fig. 3a).

Reconstruction of Instance-Level Meshes. Objects in the foreground are independently segmented using the Segment Anything Model 3 Carion *et al.* (2025) and reconstructed as explicit 3D meshes via SAM-3D-Objects Team *et al.* (2025). Unlike implicit representations, explicit meshes provide well-defined surface topologies, ensuring precise contact interfaces for downstream mechanical simulation and collision handling. These separate meshes are subsequently integrated through coordinate alignment (refer to Appendix B for detailed formulations and spatial reasoning).

Synthesis and Expansion of Background. To eliminate foreground occlusions and ensure spatial consistency, we employ a two-stage background completion strategy. First, the LaMa inpainting model Suvorov *et al.* (2021) synthesizes occluded regions to yield an unobstructed static view. Second, we use a diffusion-based outpainting model Filoni (2025) to extend the scene beyond the original field of view. This panoramic context enables physically consistent parallax during camera movement. Detailed implementations of this background synthesis are provided in Appendix B.

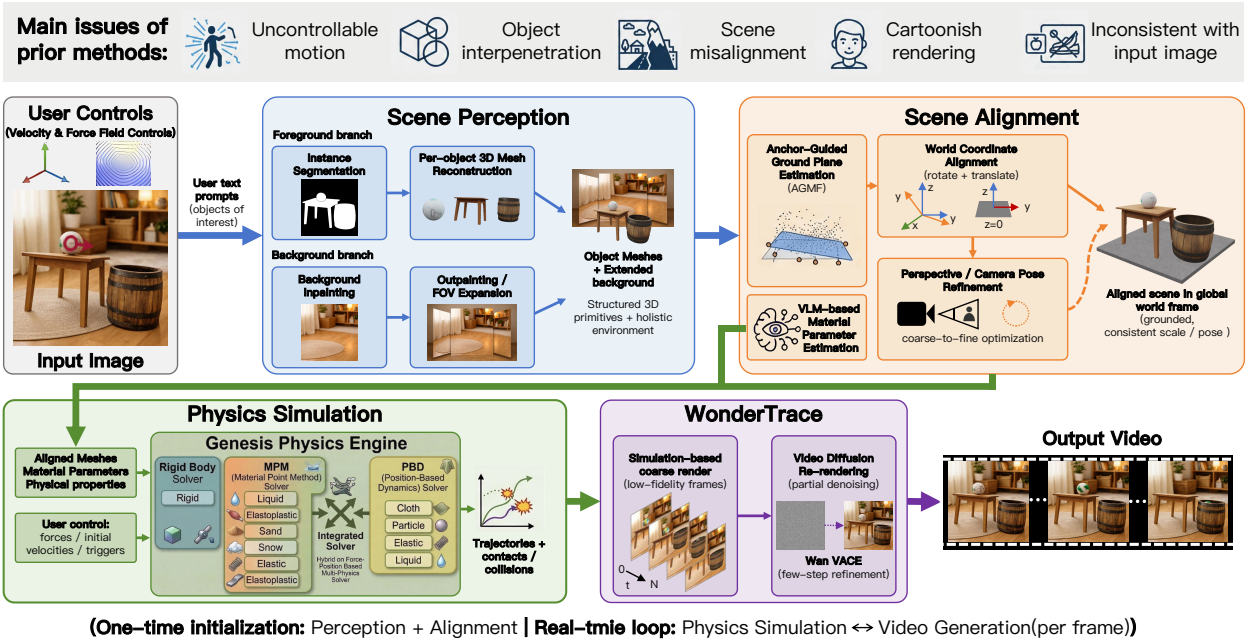

Figure 3 Overview of the PhysOmni framework. **(a)** Given a single input image and user controls, the pipeline applies Scene Perception to reconstruct 3D object meshes and synthesize a background environment. **(b)** These components are grounded in a unified global coordinate system through Scene Alignment 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.

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). This module is designed to resolve spatial inconsistencies that stem from egocentric reconstruction and inaccurate camera estimation (Fig. 2). It comprises two primary components: *Alignment of Pose Coordinates*, 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 observation.

3.2.1 Alignment of Pose Coordinates

Although the previous stage yields detailed mesh representations, these raw descriptors are ill-suited for downstream physical reasoning and interaction tasks. A primary bottleneck arises from the inherent spatial misalignment introduced during the transformation of camera-centric observations into a global world coordinate system (Fig. 2). 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 mechanism systematically transforms representations of egocentric meshes 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 $\mathcal{W}$.

Canonical Alignment for Single Objects. For isolated objects, we first perform Global Centroid Normalization. Given a mesh that comprises N_v vertices $V = \{\mathbf{v}_i\}_{i=1}^{N_v}$, the mesh is translated to the origin of

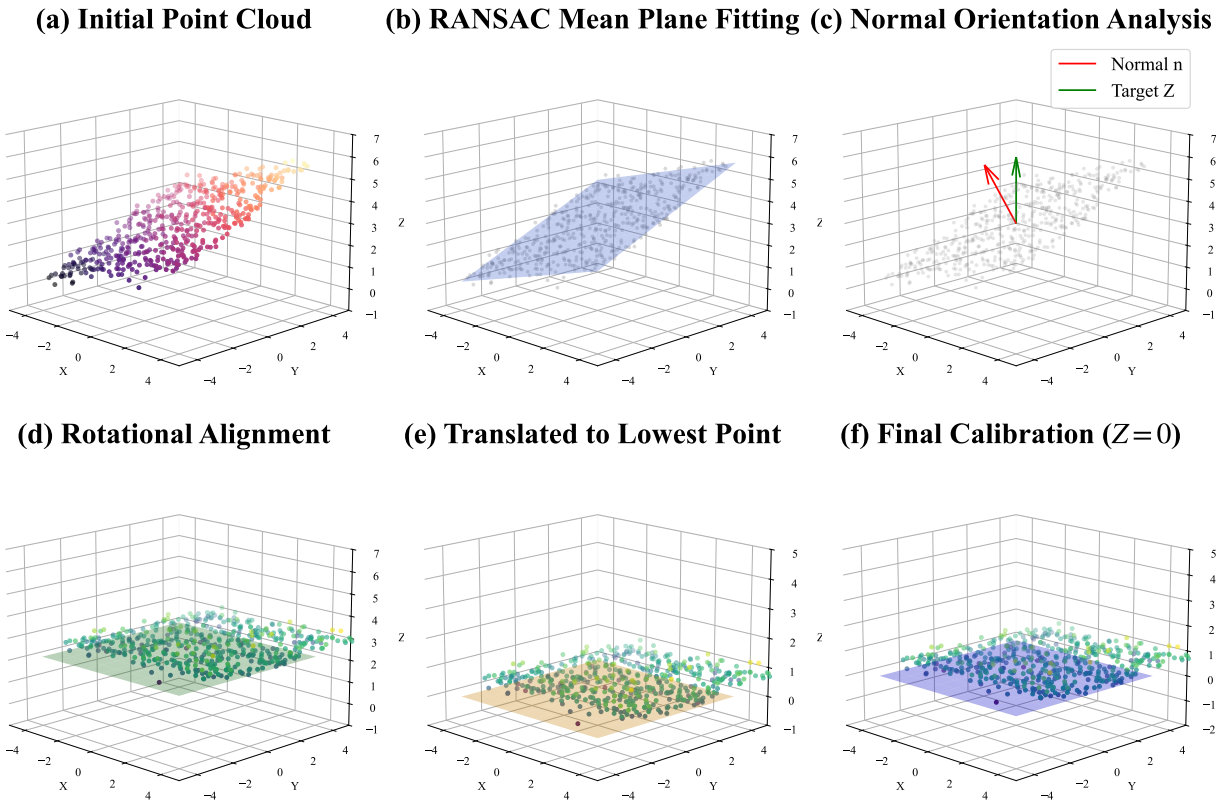

Figure 4 Ground-plane alignment. The dominant plane is estimated from the scene point cloud and rigidly transformed to $z = 0$ in the gravity-aligned frame $\mathcal{W}$.

the world such that the centroid $\bar{\mathbf{v}}$ satisfies:

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

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:

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

Through the eigendecomposition $\mathbf{\Sigma}\mathbf{u}_k = \lambda_k \mathbf{u}_k$, we identify the minor eigenvector $\mathbf{u}_3$ corresponding to the smallest eigenvalue λ_3. Assuming that most objects rest in a structurally stable configuration where the shortest physical dimension of the object aligns with the direction of gravity, we derive the rotation matrix $\mathbf{R}$ required to align $\mathbf{u}_3$ with the vertical axis of the world coordinate system (Z-up) $[0, 0, 1]^\top$. Finally, a Ground-Contact Correction step is applied:

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

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

Scene Alignment for Multiple Objects. 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

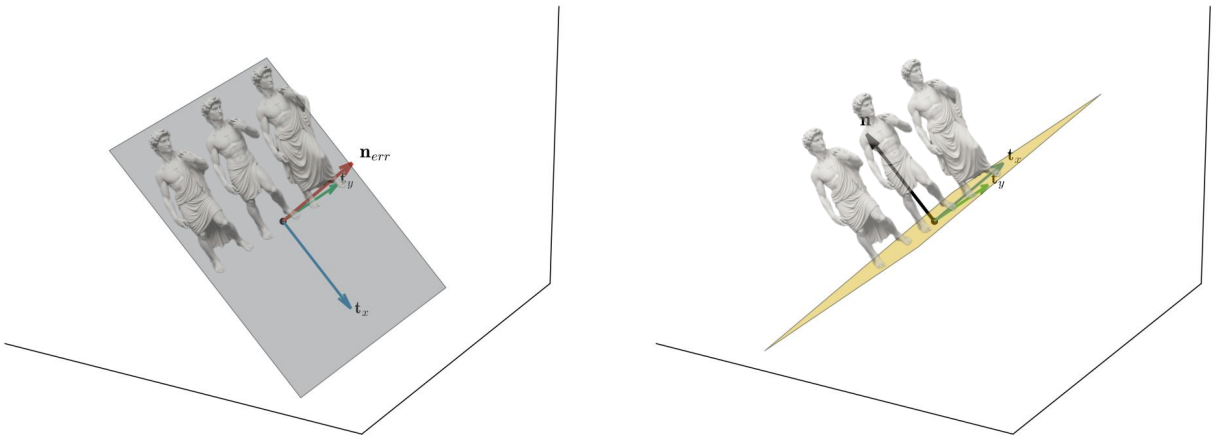

Figure 5 Ground-plane estimation. (a) RANSAC is biased by dense vertical structures. (b) Anchor-Guided Manifold Fitting samples object-base points $\mathcal{P}_{\text{base}}$ to recover the ground normal $\mathbf{n}$.

the constituent meshes of all objects in the scene, denoted as $\mathcal{V}_{\text{all}}$, we first aggregate the 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 the Rodrigues rotation 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 transformation effectively anchors the lowest point of the scene to the global ground level (Fig. 4e-f).

Estimation of Ground Plane Guided by Anchors. Standard techniques for plane estimation, 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, the RANSAC algorithm 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. Because the true world vertical is unknown at this stage, we initialize a heuristic gravity vector $\mathbf{g}_{\text{cam}}$ based on standard camera coordinate conventions (e.g., the positive Y-axis corresponding to the visual downward direction). 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 along this visual gravity prior. Specifically, we project the vertices onto $\mathbf{g}_{\text{cam}}$ and sample those within a proximity threshold δ of the maximal projection (i.e., visually lowest) for each object:

$$\mathcal{P}_{\text{base}}=\bigcup_{i=1}^{N}\left\{\mathbf{v}\in M_i \mid \mathbf{v}^\top\mathbf{g}_{\text{cam}}\geq\max_{\mathbf{u}\in M_i}(\mathbf{u}^\top\mathbf{g}_{\text{cam}})-\delta\right\}.$$

By restricting the input space to $\mathcal{P}_{\text{base}}$, we effectively eliminate the vertical distractors that typically cause the RANSAC method to yield erroneous estimates. We then determine the optimal parameters of the

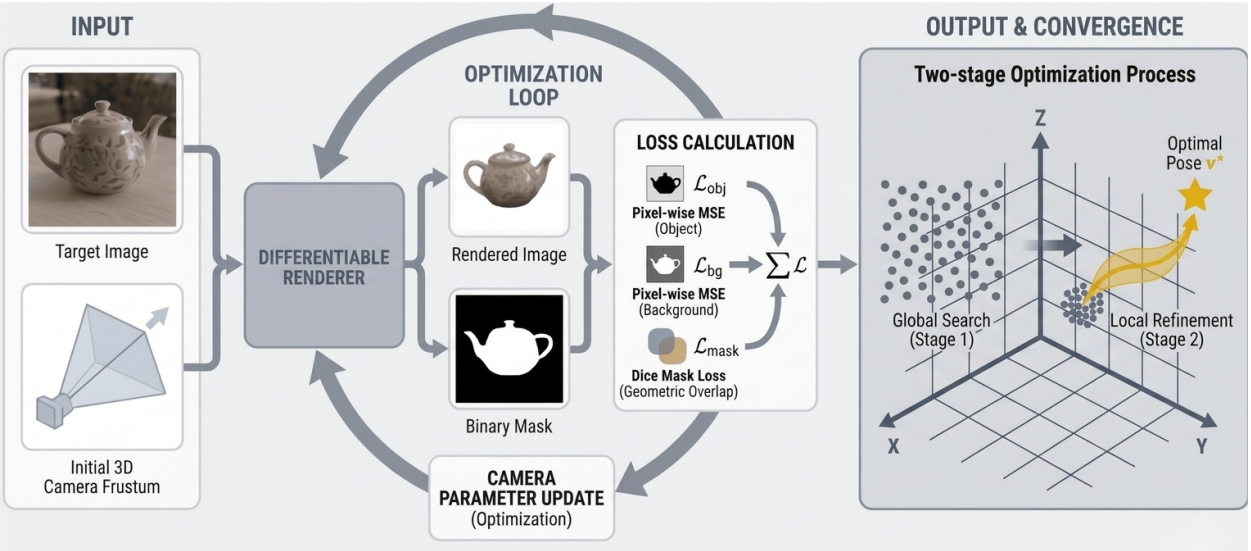

Figure 6 Coarse-to-fine camera pose optimization: global stochastic sampling initializes local refinement toward the converged pose v^*.

ground plane $\boldsymbol{\pi} = [a, b, c, d]^\top$ by minimizing a robust M-estimator over this refined subset:

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

where $\rho(\cdot)$ denotes a robust loss function (e.g., the Huber or Tukey loss) 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—such as inaccurate apparent scale and pose—between the reconstructed objects and the input image. To mitigate these issues, we formulate perspective alignment as a continuous optimization problem over camera parameters to ensure geometric and photometric consistency between the rendered synthetic outputs and the target image (Fig. 6). Given the non-convex nature of the rendering loss landscape, identifying the global optimum is non-trivial; therefore, we propose a robust coarse-to-fine optimization strategy. The three composite loss terms (an object-region loss, a background loss, and a Dice mask loss) jointly address the classic hard cases of symmetric layouts, textureless surfaces, and foreground–background overlap. This optimized approach effectively eliminates previously observed perspective distortions, yielding renderings that accurately align with the input in spatial configuration, silhouettes, and visual appearance. We refer the reader to Appendix C for the detailed algorithm, including the global stochastic sampling, local derivative-free refinement, and the composite objective functions guiding this process.

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. To resolve appearance–material ambiguities (*e.g.*,

solid wood versus hollow plastic), we additionally expose a lightweight manual interface that lets users override the per-object material prior before simulation. 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), which leverages video generation models as a rendering prior to translate 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 details while strictly adhering to the engine's underlying structural and temporal dynamics. Furthermore, because the background environment is expanded during the initial Scene Perception stage (Section 3.1), this rendering process seamlessly supports novel view synthesis and virtual camera movements without out-of-bounds artifacts, with comprehensive implementation details (including specific architectures like Wan2.1 and Wan2.2 VACE) provided in Appendix E.

4 Experiments

We evaluate PhysOmni in terms of controllability, physical consistency, visual quality, and runtime efficiency.

4.1 Experimental Setup

We evaluate all methods on 60 indoor and outdoor single-image-to-video scenes spanning single- and multi-object interactions, varied object counts, occlusions, and viewpoints. All experiments use one NVIDIA H100 GPU. PhysOmni uses SAM 3 Carion *et al.* (2025) for segmentation, SAM-3D-Objects Team *et al.* (2025) for 3D lifting, Genesis Authors (2024) for rigid and deformable simulation, Wan2.1 VACE Wan *et al.* (2025) by default, and Wan2.2 VACE 14B as an optional offline rerenderer. We compare against Sora2-pro OpenAI (2025), Veo3.1 Google DeepMind (2025), CogVideoX1.5 Yang *et al.* (2025c), Wan2.2-A14B Wan *et al.* (2025), WonderPlay Li *et al.* (2025b), and PhysCtrl Wang *et al.* (2025) using the same input image and event description. WonderPlay and PhysCtrl receive only the corresponding official control format with released default settings, while video-prior baselines additionally receive object-level motion text and 2D arrows to provide a comparable control budget. PhysOmni requires approximately 46 s of one-time initialization per scene and then provides an interactive physics preview at approximately 15 FPS (Table 7 in the Appendix).

4.2 Qualitative Comparison

Figure 7 shows qualitative comparisons with strong video generation baselines. While appearance-driven methods often produce visually appealing motion, they frequently lack reliable control over object trajectories and interactions. In scenes involving multiple objects, these methods tend to generate ambiguous collisions, inconsistent contacts, or motions that deviate from the specified force direction. Although PhysCtrl improves controllability, it still struggles with long-horizon stability and spatial consistency in multi-object settings.

In contrast, PhysOmni produces motions that remain temporally coherent and mechanically consistent with the intended events. The advantage is especially clear in scenarios involving support, collision, and

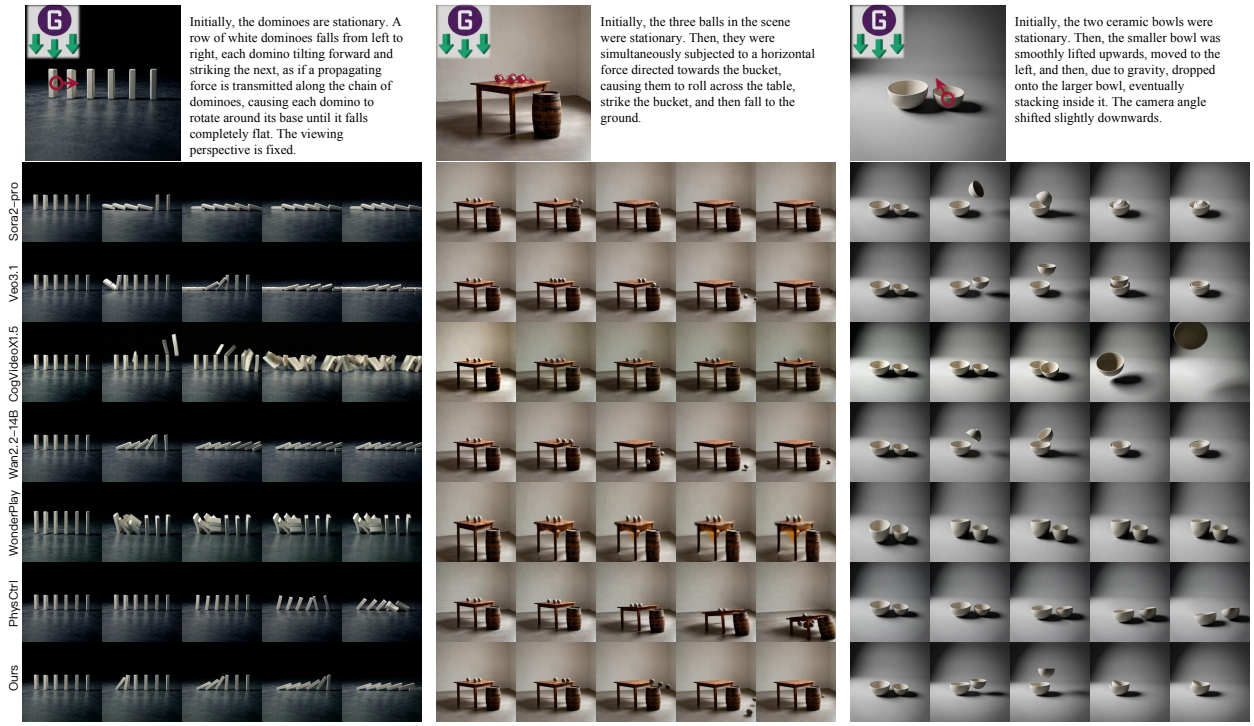

Figure 7 Qualitative comparisons between our approach and existing video generation methods. Additional results are provided in Appendix J.

sequential interactions among multiple objects. By reconstructing the scene in a shared world frame and explicitly simulating object dynamics, our method avoids the drift, interpenetration, and identity inconsistency commonly observed in video-prior baselines.

4.3 Quantitative Comparison

We quantitatively evaluate the 60-scene test set for controllability, physical plausibility, and overall video quality. Following the VideoPhy Bansal *et al.* (2024) protocol, we adopt a GPT-5-based 5-point Likert scoring framework. For each method, we evaluate 60 generated videos under three complementary criteria. Semantic adherence(SA) measures alignment between the generated video and text prompt, emphasizing whether specified initial forces or velocities and resulting motions match intended dynamics. Physical commonsense(PC) assesses whether generated motions and interactions are plausible under basic physical principles like gravity, inertia, and contact constraints. Video quality(VQ) evaluates visual fidelity, frame-level realism, and temporal smoothness.

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

4.4 Human Evaluation

Participants rank seven videos per scene by semantic adherence, physical commonsense, video quality, and overall preference, and we aggregate the rankings using the Borda score $s_i = 8 - \pi_i$.

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.* (2025c)	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.* (2025c)	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 $[1, 7]$.

Table 2 shows that PhysOmni ranks first on all four criteria, with particularly large margins in semantic adherence and physical commonsense, indicating improved controllability, physical plausibility, and overall preference.

4.5 Ablation Study

We perform ablation studies to isolate the contribution of key components in PhysOmni, including scene alignment, camera optimization, and simulation-aware initialization. Additional results and analyses are provided in Appendix H.

Table 3 Ablation on scene-aware pose alignment.

Variant	Mask IoU↑	PR↓	SVR↓	ISR↑
w/o alignment	0.11	**0.00**	1.00	0.00
+ single-obj. canonical alignment	0.35	0.50	0.15	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.

Scene-aware pose alignment. Table 3 shows that removing alignment yields poor grounding (Mask IoU 0.11), complete support failure (SVR 1.00), and no successful interactions (ISR 0.00); its zero penetration is therefore trivial. Canonical single-object alignment improves Mask IoU to 0.35 but raises PR to 0.50, while RANSAC reaches a Mask IoU of 0.54 and reduces PR to 0.01. AGMF retains this spatial accuracy, eliminates penetration, and achieves the highest ISR of 0.90, demonstrating robust physically valid initialization.

Perspective alignment. Table 4 shows that the initial camera has a 38.95-pixel reprojection error and 0.26 Mask IoU. Global search and local Powell optimization reduce the error to 20.97 and 14.77 pixels,

Table 4 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 the average reprojection error in pixels after optimization.

respectively, while our coarse-to-fine strategy achieves the lowest error of 14.33 pixels with comparable visual quality (SSIM 0.76; LPIPS 0.13) and Mask IoU (0.54). This confirms the benefit of combining global exploration with local refinement.

5 Conclusion

We present PhysOmni, a training-free framework that converts a single image into physically controllable video through holistic 3D reconstruction, scene-aware pose and perspective alignment, explicit simulation, and photorealistic rerendering. Experiments demonstrate improved multi-object interaction, spatial coherence, and long-horizon control, while ablations validate the alignment and camera optimization. Performance remains limited by monocular segmentation and reconstruction errors, particularly under severe occlusion and for fluids or thin shells.

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*.

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*.

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, Minglin, Wang, Longguang, Ao, Sheng, Zhang, Ye, Xu, Kai, & Guo, Yulan. 2025b. *Layout2Scene: 3D Semantic Layout Guided Scene Generation via Geometry and Appearance Diffusion Priors*.

Chen, Tingxi, Hao, Ke, Chen, Yabo, Cheng, Zhengxue, Xie, Rong, Song, Li, Huang, Haibin, Zhang, Chi, & Li, Xuelong. 2026a. *Full-4D: Generating Full-Scope 4D Scenes from a Single-View Video*.

Chen, Xiangyu, Luo, Jixiang, Xu, Jingyu, Yi, Fangqiu, Zhang, Chi, & Li, Xuelong. 2026b. *Generative Video Compression: Towards 0.01% Compression Rate for Video Transmission.*

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.*

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.*

Chen, Yabo, Liang, Yuanzhi, Wang, Jiepeng, Chen, Tingxi, Cheng, Junfei, Gu, Zixiao, Huang, Yuyang, Jiang, Zicheng, Li, Wei, Li, Tian, Li, Weichen, Li, Zuoxin, Liu, Guangce, Liu, Jialun, Liu, Junqi, Wang, Haoyuan, Weng, Qizhen, Wu, Xuan'er, Xiang, Xunzhi, Yang, Xiaoyan, Zhang, Xin, Zhang, Shiwen, Zhou, Junyu, Zhou, Chengcheng, Huang, Haibin, Zhang, Chi, & Li, Xuelong. 2025c. *TeleWorld: Towards Dynamic Multimodal Synthesis with a 4D World Model.*

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.

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.

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, Yuyang, Chen, Yabo, Zhou, Junyu, Dai, Wenrui, Zhang, Xiaopeng, Zou, Junni, Xiong, Hongkai, & Tian, Qi. 2025a. *Diffusion-Driven Progressive Target Manipulation for Source-Free Domain Adaptation.*

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.*

Huang, Yuyang, Chen, Yabo, Dai, Wenrui, Zheng, Ziyang, Huang, Haibin, Zhang, Chi, Zou, Junni, Xiong, Hongkai, & Li, Xuelong. 2026. *CineWeaver: Training-Free Reference-Controllable Multi-Shot Long Video Generation for Cinematic Storytelling.*

Jiang, Chenfanfu, Schroeder, Craig, & Teran, Joseph. 2016. The Material Point Method for simulating continuum materials. *ACM SIGGRAPH 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.

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*.

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

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*.

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*.

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, Shaowei, Ren, Zhongzheng, Gupta, Saurabh, & Wang, Shenlong. 2024b. *PhysGen: Rigid-Body Physics-Grounded Image-to-Video Generation*.

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

Meng, Yanxu, Wu, Haoning, Zhang, Ya, & Xie, Weidi. 2025. *SceneGen: Single-Image 3D Scene Generation in One Feedforward Pass*.

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

NVIDIA, :, Ali, Arslan, Bai, Junjie, Bala, Maciej, Balaji, Yogesh, Blakeman, Aaron, Cai, Tiffany, Cao, Jiaxin, Cao, Tianshi, Cha, Elizabeth, Chao, Yu-Wei, Chattopadhyay, Prithvijit, Chen, Mike, Chen, Yongxin, Chen, Yu, Cheng, Shuai, Cui, Yin, Diamond, Jenna, Ding, Yifan, Fan, Jiaojiao, Fan, Linxi, Feng, Liang, Ferroni, Francesco, Fidler, Sanja, Fu, Xiao, Gao, Ruiyuan, Ge, Yunhao, Gu, Jinwei, Gupta, Aryaman, Gururani, Siddharth, Hanafi, Imad El, Hassani, Ali, Hao, Zekun, Huffman, Jacob, Jang, Joel, Jannaty, Pooya, Kautz, Jan, Lam, Grace, Li, Xuan, Li, Zhaoshuo, Liao, Maosheng, Lin, Chen-Hsuan, Lin, Tsung-Yi, Lin, Yen-Chen, Ling, Huan, Liu, Ming-Yu, Liu, Xian, Lu, Yifan, Luo, Alice, Ma, Qianli, Mao, Hanzi, Mo, Kaichun, Nah, Seungjun, Narang, Yashraj, Panaskar, Abhijeet, Pavao, Lindsey, Pham, Trung, Ramezanali, Morteza, Reda, Fitsum, Reed, Scott, Ren, Xuanchi, Shao, Haonan, Shen, Yue, Shi, Stella, Song, Shuran, Stefaniak, Bartosz, Sun, Shangkun, Tang, Shitao, Tasmeen, Sameena, Tchapmi, Lyne, Tseng, Wei-Cheng, Varghese, Jibin, Wang, Andrew Z., Wang, Hao, Wang, Haoxiang, Wang, Heng, Wang, Ting-Chun, Wei, Fangyin, Xu, Jiashu, Yang, Dinghao, Yang, Xiaodong, Ye, Haotian, Ye, Seonghyeon, Zeng, Xiaohui, Zhang, Jing, Zhang, Qinsheng, Zheng, Kaiwen, Zhu, Andrew, & Zhu, Yuke. 2026. *World Simulation with Video Foundation Models for Physical AI*.

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

Paschalidou, Despoina, Kar, Amlan, Shugrina, Maria, Kreis, Karsten, Geiger, Andreas, & Fidler, Sanja. 2021. *ATISS: Autoregressive Transformers for Indoor Scene Synthesis*.

Poole, Ben, Jain, Ajay, Barron, Jonathan T., & Mildenhall, Ben. 2022. *DreamFusion: Text-to-3D using 2D Diffusion*.

Ram, Daniel, Gast, Theodore, Jiang, Chenfanfu, Schroeder, Craig, Stomakhin, Alexey, Teran, Joseph, & Kavehpour, Pirouz. 2015. A material point method for viscoelastic fluids, foams and sponges. *Pages 157–163 of: Proceedings of the 14th 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*.

Shao, Jiawei, & Li, Xuelong. 2025. *AI Flow at the Network Edge*.

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. 1994. 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*.

Tang, Jiapeng, Nie, Yinyu, Markhasin, Lev, Dai, Angela, Thies, Justus, & Nießner, Matthias. 2024. *DiffuScene: Denoising Diffusion Models for Generative Indoor Scene Synthesis*.

Team, SAM 3D, Chen, Xingyu, Chu, Fu-Jen, Gleize, Pierre, Liang, Kevin J, Sax, Alexander, Tang, Hao, Wang, Weiyao, Guo, Michelle, Hardin, Thibaut, Li, Xiang, Lin, Aohan, Liu, Jiawei, Ma, Ziqi, Sagar, Anushka, Song, Bowen, Wang, Xiaodong, Yang, Jianing, Zhang, Bowen, Dollár, Piotr, Gkioxari, Georgia, Feiszli, Matt, & Malik, Jitendra. 2025. *SAM 3D: 3Dfy Anything in Images*.

Wan, Team, Wang, Ang, Ai, Baole, Wen, Bin, Mao, Chaojie, Xie, Chen-Wei, Chen, Di, Yu, Feiwu, Zhao, Haiming, Yang, Jianxiao, Zeng, Jianyuan, Wang, Jiayu, Zhang, Jingfeng, Zhou, Jingren, Wang, Jinkai, Chen, Jixuan, Zhu, Kai, Zhao, Kang, Yan, Keyu, Huang, Lianghua, Feng, Mengyang, Zhang, Ningyi, Li, Pandeng, Wu, Pingyu, Chu, Ruihang, Feng, Ruili, Zhang, Shiwei, Sun, Siyang, Fang, Tao, Wang, Tianxing, Gui, Tianyi, Weng, Tingyu, Shen, Tong, Lin, Wei, Wang, Wei, Wang, Wei, Zhou, Wenmeng, Wang, Wente, Shen, Wenting, Yu, Wenyuan, Shi, Xianzhong, Huang, Xiaoming, Xu, Xin, Kou, Yan, Lv, Yangyu, Li, Yifei, Liu, Yijing, Wang, Yiming, Zhang, Yingya, Huang, Yitong, Li, Yong, Wu, You, Liu, Yu, Pan, Yulin, Zheng, Yun, Hong, Yuntao, Shi, Yupeng, Feng, Yutong, Jiang, Zeyinzi, Han, Zhen, Wu, Zhi-Fan, & Liu, Ziyu. 2025. *Wan: Open and Advanced Large-Scale Video Generative Models*.

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*.

Wang, Haoyuan, Chen, Yabo, Huang, Haibin, Zhang, Chi, & Li, Xuelong. 2026. *Directing the World: Fast Autoregressive Video Generation with Compositional Human-Camera Control*.

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*.

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, Qirui, Iliash, Denys, Ritchie, Daniel, Savva, Manolis, & Chang, Angel X. 2025a. *Diorama: Unleashing Zero-shot Single-view 3D Indoor Scene Modeling*.

Wu, Tianhao, Zheng, Chuanxia, Guan, Frank, Vedaldi, Andrea, & Cham, Tat-Jen. 2025b. *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, Tian, Xingye, Zhang, Guiyu, Chen, Yabo, Zhang, Shaofeng, Wang, Xuebo, Tao, Xin, & Fan, Qi. 2025b. *Denoising Vision Transformer Autoencoder with Spectral Self-Regularization.*

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. 2025c. *Macro-from-Micro Planning for High-Quality and Parallelized Autoregressive Long Video Generation.*

Xiang, Xunzhi, Duan, Zixuan, Chen, Yabo, Wei, Zhengxuan, Zhang, Guiyu, Gu, Zixiao, Gao, Zhe, Huang, Haibin, Zhang, Chi, Fan, Qi, & Li, Xuelong. 2026. *VideoWeave: Unlocking Geometric Consistency in Video Generation via Joint Geometry-Video Modeling.*

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

Xiong, Zhexiao, Song, Yizhi, He, Liu, Xiong, Wei, Yuan, Yu, Qiao, Feng, & Jacobs, Nathan. 2026. *PhysAlign: Physics-Coherent Image-to-Video Generation through Feature and 3D Representation Alignment.*

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.*

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

Yang, Xiuyu, Man, Yunze, Chen, Jun-Kun, & Wang, Yu-Xiong. 2025b. *SceneCraft: Layout-Guided 3D Scene Generation.*

Yang, Yandan, Jia, Baoxiong, Zhi, Peiyuan, & Huang, Siyuan. 2024. *PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI.*

Yang, Zhuoyi, Teng, Jiayan, Zheng, Wendi, Ding, Ming, Huang, Shiyu, Xu, Jiazheng, Yang, Yuanming, Hong, Wenyi, Zhang, Xiaohan, Feng, Guanyu, Yin, Da, Zhang, Yuxuan, Wang, Weihan, Cheng, Yean, Xu, Bin, Gu, Xiaotao, Dong, Yuxiao, & Tang, Jie. 2025c. *CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer.*

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.*

Yuan, Cheng, Jia, Zhenyu, Shao, Jiawei, & Li, Xuelong. 2026a. *Enhancing Neural Video Compression of Static Scenes with Positive-Incentive Noise.*

Yuan, Jianhao, Zhang, Xiaofeng, Friedrich, Felix, Beltran-Velez, Nicolas, Hall, Melissa, Askari-Hemmat, Reyhane, Han, Xiaochuang, Ballas, Nicolas, Drozdzal, Michal, & Romero-Soriano, Adriana. 2026b. *Inference-time Physics Alignment of Video Generative Models with Latent World Models.*

Zhang, Guiyu, Chen, Yabo, Xiang, Xunzhi, Huang, Junchao, Wang, Zhongyu, & Jiang, Li. 2026. *Sympho-Motion: Joint Control of Camera Motion and Object Dynamics for Coherent Video Generation.*

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.*

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.*

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

Zhou, Shijie, Fan, Zhiwen, Xu, Dejia, Chang, Haoran, Chari, Pradyumna, Bharadwaj, Tejas, You, Suya, Wang, Zhangyang, & Kadambi, Achuta. 2024a. *DreamScene360: Unconstrained Text-to-3D Scene Generation with Panoramic Gaussian Splatting*.

Zhou, Xiaoyu, Ran, Xingjian, Xiong, Yajiao, He, Jinlin, Lin, Zhiwei, Wang, Yongtao, Sun, Deqing, & Yang, Ming-Hsuan. 2024b. *GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting*.

PhysOmni: Physics-Grounded Multi-Object Scene Generation from a Single Image with Real-Time Interaction

Supplementary Material

A Extended Related Work

A.1 Controllable and Physics-Aware Video Generation

Recent generative video models based on diffusion Ho *et al.* (2022) and autoregressive architectures achieve unprecedented visual realism in the synthesis of dynamic content. However, these models are predominantly appearance-driven and lack an explicit understanding of three-dimensional geometry and mechanical laws. To address this limitation, several methods attempt to incorporate physics through auxiliary priors, drag-based controls, or intermediate physical signals. PhysGen Liu *et al.* (2024b) integrates rigid-body simulation with video diffusion to generate physically plausible image-to-video content conditioned on applied forces. PhysDreamer Zhang *et al.* (2024b) endows static 3D Gaussian objects with interactive dynamics by leveraging video generation priors for the estimation of material properties. PhysAnimator Xie *et al.* (2025) combines deformable-body simulation with sketch-guided video diffusion for physically grounded cartoon animation. Force Prompting Gillman *et al.* (2025) finetunes video models on simulator-synthesized data to follow localized point forces and global wind fields. PhysCtrl Wang *et al.* (2025) learns a generative physics network that produces 3D point trajectories conditioned on physical parameters across multiple material types. WonderPlay Li *et al.* (2025b) introduces a hybrid generative simulator that couples a physics solver with a video generator for action-conditioned dynamic 3D scene synthesis. VLIPP Yuan *et al.* (2026b) proposes a two-stage framework that first predicts coarse physical trajectories via vision-language reasoning, and subsequently conditions a controllable diffusion model on the resulting optical flow. VideoREPA Zhang *et al.* (2025) distills physics understanding from self-supervised video encoders into text-to-video diffusion models through token-level relational alignment. PhysVideoGenerator Satish *et al.* (2026) embeds a learnable physics prior by aligning diffusion latents with V-JEPA 2 representations via cross-attention. RealWonder Liu *et al.* (2026) achieves real-time action-conditioned generation at 13.2 frames per second by bridging single-image 3D reconstruction, physics simulation, and a distilled four-step video generator. PhysAlign Xiong *et al.* (2026) constructs a fully controllable synthetic data pipeline based on rigid-body simulation and proposes a unified physical latent space that couples 3D geometry constraints with Gram-based spatio-temporal relational alignment. More recently, large-scale video foundation models for Physical AI, such as Cosmos NVIDIA *et al.* (2026), learn world simulation directly from massive video corpora and demonstrate strong general-purpose dynamics priors. These world models, however, primarily target learned, implicit dynamics for data generation and policy learning, and do not expose explicit, interactive control over per-object forces or provide the deterministic, real-time physical previews that PhysOmni offers.

Despite these advances, existing approaches rely heavily on soft constraints and manually specified parameters. This reliance limits the respective methods to short, simplistic motions that fail to support fine-grained, long-horizon physical interactions. In contrast, PhysOmni enables real-time, explicit mechanical control over complex multi-object scenes using diverse force fields (*e.g.*, vortex, wind, gravity) without relying on latency-heavy priors.

A.2 Single-Image to 3D Reconstruction

Significant progress has been made in the conversion of single images to 3D models. Score Distillation Sampling (SDS) Poole *et al.* (2022) pioneered the lifting of 2D diffusion priors into 3D optimization, inspiring a family of methods that employ triplanes, large-scale feed-forward networks, and rectified flow transformers. CLAY Zhang *et al.* (2024a) trains a 1.5B-parameter latent diffusion transformer on diverse 3D geometries with multi-resolution VAE encoding. Hunyuan3D 2.5 Lai *et al.* (2025) scales a shape foundation model (LATTICE) to 10B parameters for the generation of sharp, detailed meshes with support for physically based rendering textures. TRELLIS Xiang *et al.* (2025a) introduces Structured 3D Latents (SLAT) decoded into radiance fields, Gaussians, or meshes via 2B-parameter rectified flow transformers trained on 500K objects. InstantMesh Xu *et al.* (2024) combines multi-view diffusion with a transformer-based sparse-view reconstruction model for the generation of meshes within 10 seconds. Amodal3R Wu *et al.* (2025b) addresses occluded objects through mask-weighted cross-attention and occlusion-aware conditioning for amodal 3D reconstruction. Cascade-Zero123 Chen *et al.* (2024b) proposes a self-prompted cascade framework that progressively generates nearby views to improve multi-view consistency. LiftImage3D Chen *et al.* (2024c) leverages latent video diffusion models with articulated camera trajectories and distortion-aware 3D Gaussian splatting. TripoSG Li *et al.* (2025a) employs large-scale rectified flow models with an advanced SDF-based VAE to achieve state-of-the-art shape fidelity.

However, these pipelines typically reconstruct individual objects in isolation within egocentric coordinate systems. Even with segmentation-aware formulations such as Part123 Liu *et al.* (2024a) and PartGen Chen *et al.* (2024a), the methods lack global spatial awareness. When independently reconstructed objects are composed into a single scene, severe spatial misalignment and interpenetration inevitably arise. Unlike these fragmented object-centric methods, PhysOmni introduces a *Scene-Aware Pose Alignment* mechanism that anchors all elements to a canonical ground plane and transforms the elements into a unified world coordinate system. This approach inherently resolves penetration ambiguities and guarantees geometrically coherent configurations.

A.3 Physics-Grounded Scene Generation and Simulation

Recent literature explores physics-based scene-level generation, jointly modeling object layouts and physical plausibility. ATISS Paschalidou *et al.* (2021) formulates indoor scene synthesis as the autoregressive generation of object attributes via transformers. DiffuScene Tang *et al.* (2024) extends this approach with denoising diffusion models operating on unordered object sets, synthesizing physically plausible indoor scenes. PhyScene Yang *et al.* (2024) incorporates physical interaction constraints into 3D scene synthesis for embodied AI. SceneCraft Yang *et al.* (2025b) leverages layout-guided diffusion with 3D bounding boxes for the generation of complex indoor scenes. GALA3D Zhou *et al.* (2024b) proposes layout-guided generative Gaussian splatting with priors generated by LLMs and compositional diffusion optimization. Layout2Scene Chen *et al.* (2025b) employs semantic-guided geometry and appearance diffusion priors conditioned on 3D semantic layouts. PAT3D Lin *et al.* (2025b) integrates differentiable rigid-body contact simulation into the text-to-3D pipeline, introducing simulation-in-the-loop optimization for intersection-free, physically stable scenes. VideoRFSplat Go *et al.* (2025) jointly models multi-view images and camera poses via a dual-stream architecture for direct text-to-3D Gaussian generation. DreamScene360 Zhou *et al.* (2024a) generates 360-degree scenes by lifting panoramic images into 3D Gaussians with a globally optimized point cloud initialization. CAST Yao *et al.* (2025) reconstructs component-aligned 3D scenes from single RGB images using physics-aware correction with fine-grained relation graphs, and was nominated for the Best Paper award at SIGGRAPH 2025. Diorama Wu *et al.* (2025a) proposes a zero-shot pipeline that decomposes single-view scene modeling into perception and CAD-based assembly with semantic-aware layout optimization. SceneGen Meng *et al.* (2025) generates multiple 3D assets from a scene image in a single feedforward pass with local-global feature aggregation.

However, accurately aligning the generated scenes with the original input image remains a critical challenge. Imposing physical constraints often conflicts with visual evidence under occlusion or limited viewpoints Yang *et al.* (2024); Yao *et al.* (2025); Wu *et al.* (2025a), which results in spatial discrepancies, stylized artifacts, or a loss of input fidelity. To overcome this challenge, PhysOmni employs a coarse-to-fine camera pose optimization strategy to ensure precise photometric and geometric alignment. Crucially, by decoupling the underlying physical simulation from the rendering pipeline, the proposed framework uniquely bypasses

the latency bottlenecks of previous methods. This decoupling delivers real-time interactive physics previews while perfectly preserving the high-fidelity appearance of the original input.

B Details of Scene Perception

This section provides the implementation details, mathematical formulations, and detailed justifications for the Scene Perception module introduced in the main paper.

Instance-level Mesh Reconstruction. To prevent geometric interference, objects are processed within local coordinate frames. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a set of semantic prompts P, we employ the Segment Anything Model 3 (SAM 3) Carion *et al.* (2025) to generate high-fidelity instance masks $\{M_i\}_{i=1}^{N}$. These masks are then utilized by SAM-3D-Objects Team *et al.* (2025) to lift the 2D representations into 3D geometric primitives, outputting explicit meshes $\mathcal{S}_i$. The joint processing of instance masks and image features at this stage is crucial, as it preserves the relative spatial configuration and scale of individual entities.

We explicitly avoid implicit or point-based representations because they are prone to ghosting artifacts and numerical instability during contact resolution. Explicit meshes allow for the accurate computation of collision manifolds and friction forces, thereby satisfying fundamental physical constraints during mechanical simulation.

Background Synthesis and Expansion. To accommodate dynamic camera viewpoints, resolve scale ambiguity, and mitigate the "cardboard effect" common in localized reconstruction methods, we reconstruct the background sequentially.

For the first stage, we synthesize the occluded regions using the LaMa inpainting model Suvorov *et al.* (2021) to derive a restored background B_{rest}:

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

This operation effectively removes the foreground masks $\{M_i\}_{i=1}^{N}$ from the input image I, yielding a clean canvas.

In the second stage, we expand the field of view using a diffusion-based outpainting model Filoni (2025):

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

where Φ denotes the target aspect ratio and structural expansion parameters. The resulting extended background B_{ext} provides necessary constraints on vanishing points and global layout, enabling consistent foreground-background parallax.

C Details of Perspective Alignment

This section provides the detailed mathematical formulations and procedural steps for the perspective alignment optimization introduced in the main paper. The success of this coarse-to-fine approach in eliminating perspective distortions and aligning spatial configurations is comprehensively demonstrated in Fig. 8.

Coarse-to-Fine Optimization Strategy. Initially, we conduct a coarse global search by uniformly sampling candidate poses within a bounded space to identify a reliable initialization, thereby mitigating the risk of convergence to poor local minima. Subsequently, we execute a local refinement step using the derivative-free Powell optimization method under box constraints. This approach enables precise pose adjustment without reliance on potentially unstable gradients associated with the rendering process. The optimization is guided by a composite objective function that includes region-aware appearance losses for both the object and background, as well as a Dice-based mask loss for explicit geometric alignment.

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 $\mathcal{B}$ for the coarse global sampling step as follows:

$$\mathcal{B} = [x_0 - \Delta_x, x_0 + \Delta_x] \times [y_0 - \Delta_y, y_0 + \Delta_y] \times [z_0 - \Delta_z, z_0 + \Delta_z], \tag{1}$$

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.

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

$$\mathcal{L}_{\text{obj}}(\mathbf{P}) = \frac{\|(I(\mathbf{P}) - I^*) \odot M^*\|_1}{\|M^*\|_1 + \epsilon}, \tag{2}$$

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

$$\mathcal{L}_{\text{bg}}(\mathbf{P}) = \frac{\|(I(\mathbf{P}) - I^*) \odot (1 - M^*)\|_1}{\|1 - M^*\|_1 + \epsilon}. \tag{3}$$

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

$$\mathcal{L}_{\text{mask}}(\mathbf{P}) = 1 - \frac{2\|M(\mathbf{P}) \odot M^*\|_1 + \epsilon}{\|M(\mathbf{P})\|_1 + \|M^*\|_1 + \epsilon}. \tag{4}$$

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

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

$$\mathcal{L}(\mathbf{P}) = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}(\mathbf{P}) + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}}(\mathbf{P}) + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}(\mathbf{P}). \tag{5}$$

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 10.

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. The parser accepts both the object form `{"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 5).

Table 5 lists all supported physics materials and their default parameters. Table 6 lists the available force field types.

D.2 Multi-Physics Solvers Formulation

As discussed in the Physical Simulation section of the main paper, our system relies on a unified multi-physics backend, Genesis Authors (2024), which integrates three specialized solvers. The detailed formulations and integration strategies are described below.

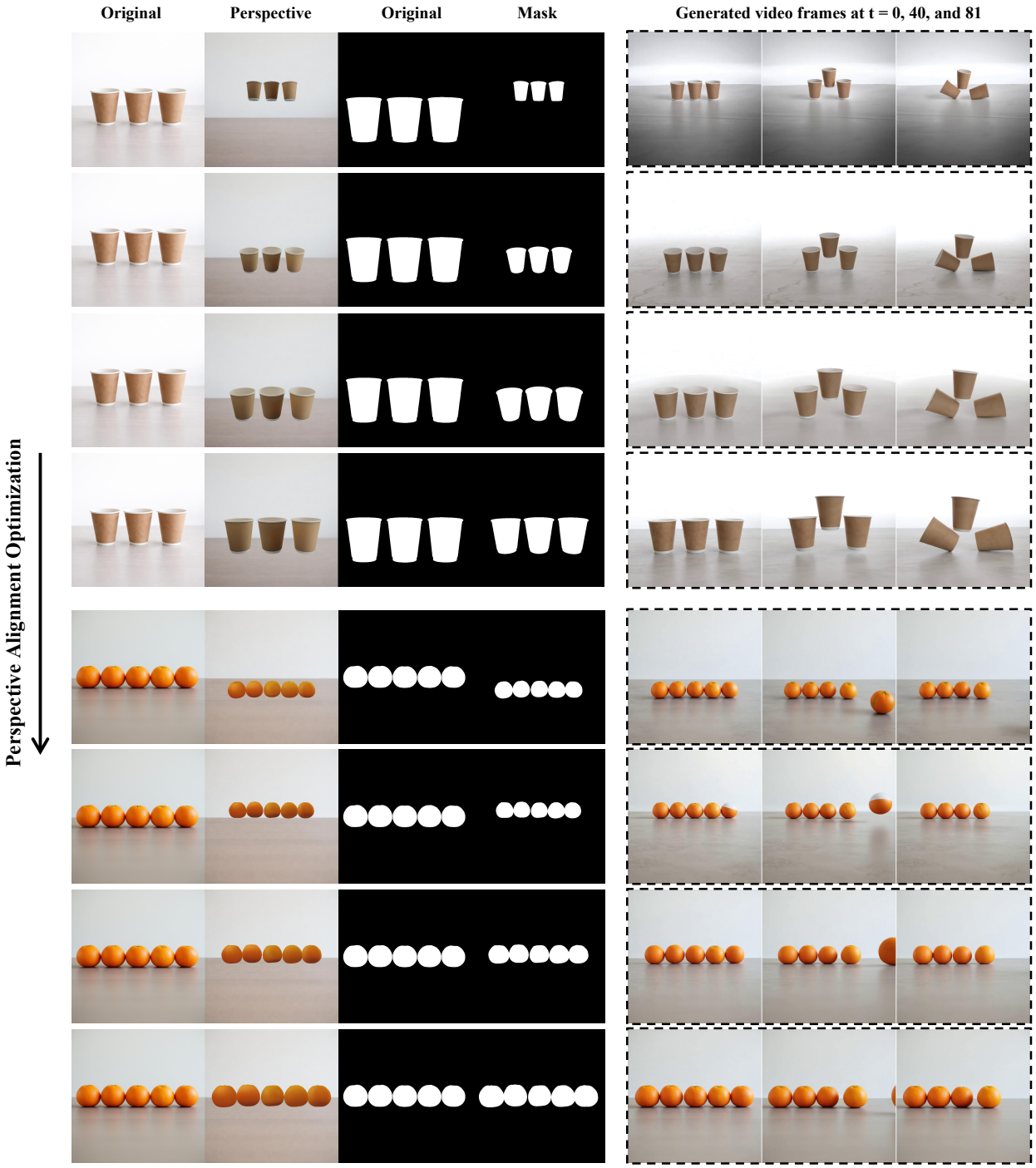

Figure 8 Qualitative visualization of the perspective alignment optimization process. From top to bottom, the rendering loss decreases gradually as the camera parameters undergo refinement. Each row displays the original input image, the rendered perspective under the current camera parameters, and the corresponding binary mask.

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:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) = \boldsymbol{\tau} + \mathbf{J}(\mathbf{q})^T \mathbf{f}_{\text{ext}}, \tag{6}$$

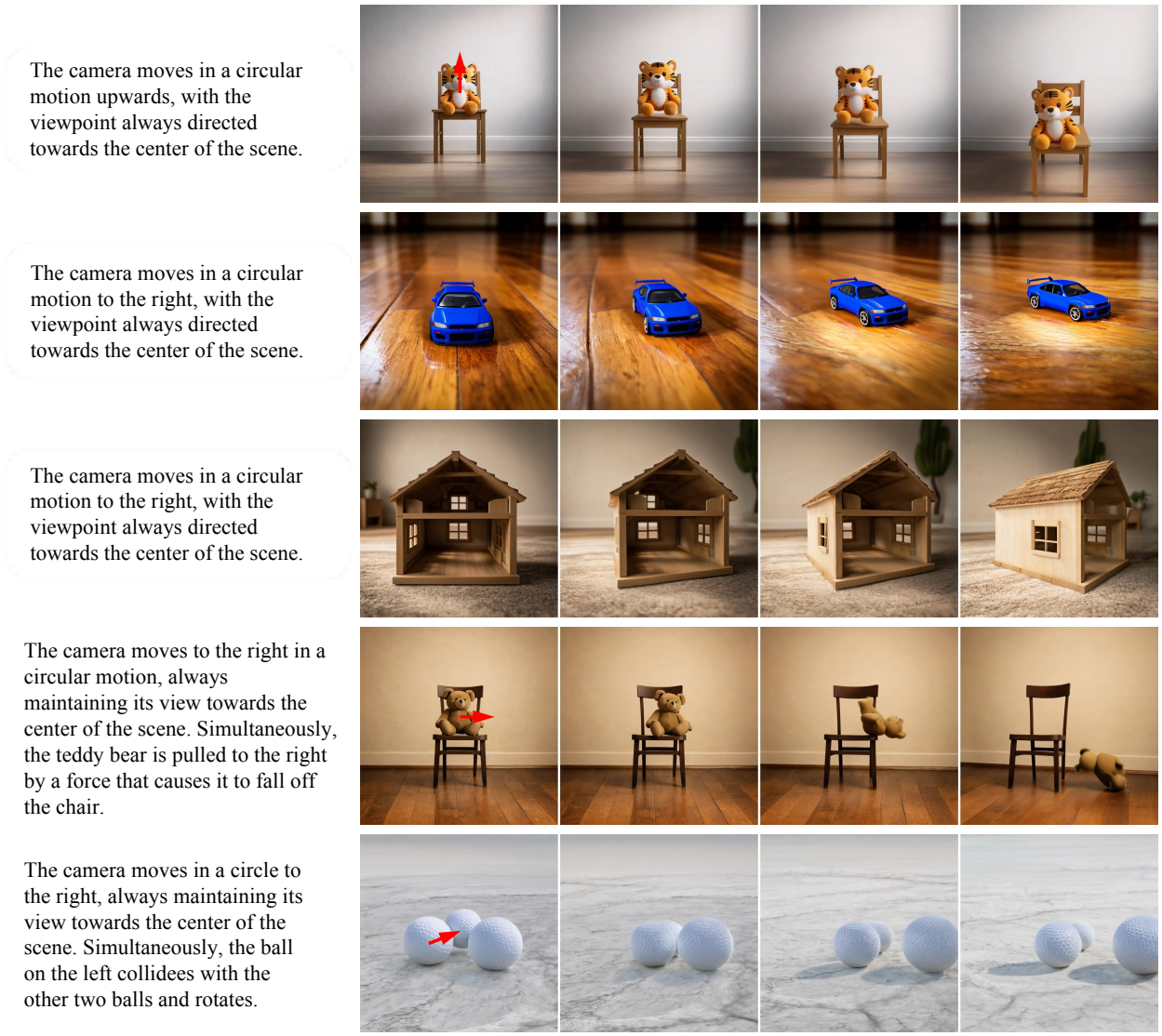

Figure 9 Illustration of camera motion and viewpoint variation

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.

Material Point Method (MPM). We employ MPM Sulsky *et al.* (1994); 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 $\boldsymbol{\sigma}_p$:

$$\mathbf{f}_i = \sum_p V_p \boldsymbol{\sigma}_p \nabla w_{ip} + \mathbf{f}_{\text{ext}}, \tag{7}$$

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

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 10 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.

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.* (2007) 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 $C(\mathbf{x}) = 0$:

$$\Delta \mathbf{x}_i = -w_i \frac{C(\mathbf{x})}{\sum_j w_j |\nabla_{\mathbf{x}_j} C(\mathbf{x})|^2} \nabla_{\mathbf{x}_i} C(\mathbf{x}), \tag{8}$$

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 WonderTrace

This section provides comprehensive details regarding the implementation of our video generation module, which consists of the WonderTrace rerendering pipeline and the scene construction strategy introduced in the main paper.

Table 5 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 6 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

WonderTrace Rerendering Details. As highlighted in the main paper, videos rendered directly from physical simulations often exhibit a synthetic appearance. This visual deficiency stems from simplified lighting models, unrealistic material responses, and the absence of environmental noise. Consequently, such videos display a "plastic" texture that fails to capture the stochastic richness of real-world scenes, such as subtle atmospheric scattering and organic texture variations. To address this limitation, we employ video generation models to rerender simulation outputs, utilizing the physical simulation as a strict structural prior. By leveraging the high-dimensional latent space of state-of-the-art video generative models, we perform pixel-level semantic enhancement to synthesize photorealistic details onto the simulated structure.

To balance computational efficiency with visual fidelity, our implementation utilizes a partial denoising strategy. Rather than generating videos from pure Gaussian noise—a process that is computationally intensive and prone to deviation 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.

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, facilitating real-time performance without compromising visual quality. This approach significantly reduces inference latency while ensuring that the generated video faithfully inherits the complex motion trajectories and spatiotemporal structure of the simulation. During these final refinement stages, the model enriches the scene with high-frequency visual details—such as atmospheric scattering, sub-surface reflections, and motion blur—that are typically challenging for real-time physical simulators to model accurately.

Furthermore, we provide an alternative rendering configuration based on the larger Wan2.2 VACE 14B Wan *et al.* (2025) model. This variant is designed to maximize perceptual realism, delivering higher-resolution outputs with improved visual quality. It is particularly suitable for offline rendering or high-fidelity visualization scenarios where real-time constraints are relaxed.

Scene Construction and Expansion. A fundamental challenge in 3D scene modeling is the inherent sparsity of source data, which typically captures a limited field of view (FOV). Conventional reconstruction pipelines often fail to maintain consistent visual fidelity across unconstrained viewing angles; the absence of wide-field contextual imagery results in voids or severe boundary artifacts when the virtual camera trajectory deviates from the original path. This restricted FOV constrains the synthesis of immersive environments and limits the degrees of freedom for virtual camera movement.

Following the background handling approach described in the main paper, we employ a combined inpainting and out-painting strategy to effectively decouple the foreground content from the environment. We extend this design to the video rerendering stage to enable scene completion well beyond the original camera frustum. As illustrated in Fig. 9, the reconstructed background is spatially expanded to provide sufficient semantic context for extensive camera motions and 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.

F Details of Experiments

F.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.

Runtime. Table 7 reports the runtime breakdown on a single NVIDIA H100 GPU.

F.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.

Table 7 Average runtime breakdown of PhysOmni.

Component	Time / Throughput
One-time processing (per scene, pure inference)	
Segmentation (SAM3 + inpainting)	∼5 s
3D Mesh Reconstruction (SAM3D)	∼15 s
Physics Config Estimation (VLM)	∼10 s
Scene Alignment (pose + camera opt.)	∼16 s
Interactive simulation loop	
End-to-end Interactive Loop	∼15 FPS

Processing timings represent highly optimized pure model inference, excluding model loading overhead. Physics configuration estimation uses Qwen2.5-VL-72B for automatic material and force prediction. The interactive loop throughput corresponds to a real-time interactive physics preview. Optional high-fidelity offline diffusion rerendering can be applied afterward for enhanced photorealism.

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:

$$\mathbf{x}_{\text{world}} = s \cdot \mathbf{R}_q \, \mathbf{x}_{\text{local}} + \mathbf{t}, \tag{9}$$

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.

F.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: $A = 2/(1 + e^{-\bar{\sigma}_L^2/500}) - 1$.

G Quantitative Evaluation via GPT-5

To quantitatively evaluate the 60-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.

G.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.

G.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*

G.3 GPT-Free Evaluation on VBench-2.0

The GPT-5-based protocol above follows the widely used VideoPhy Bansal *et al.* (2024) VLM-as-judge paradigm and is not a metric introduced by this work. To further remove any dependence on a proprietary

Table 8 GPT-free evaluation on VBench-2.0. We report four physics-related dimensions and their average. Rigid: rigidity; NoPen: non-penetration; Perm: object permanence; MotCoh: motion coherence.

Method	Rigid	NoPen	Perm	MotCoh	Avg
Ours	**0.98**	0.84	**1.00**	0.98	**0.95**
Veo3.1	**0.98**	0.83	0.98	0.98	0.94
Wan2.2-A14B	0.95	0.82	**1.00**	0.98	0.94
Sora2-pro	**0.98**	**0.86**	0.98	0.93	0.94
CogVideoX1.5	0.77	0.82	**1.00**	**1.00**	0.90
Wan2.2-5B	0.93	0.81	0.95	0.91	0.90
PhysCtrl	0.80	0.78	**1.00**	0.97	0.89

Table 9 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

judge, we additionally report an open-source, GPT-free evaluation using VBench-2.0. Table 8 reports four physics-related dimensions—rigidity (Rigid), non-penetration (NoPen), object permanence (Perm), and motion coherence (MotCoh)—together with their average. PhysOmni attains the highest average physics score and leads on rigidity and permanence, corroborating the GPT-5-based results with a fully reproducible, open-source metric.

G.4 Hyperparameter Settings

Table 9 lists the key hyperparameters and their values used throughout the experiments.

H Additional Ablation Study

Force field configuration. Table 10 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

Table 10 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.

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.

Table 11 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.

Inference efficiency in video synthesis. Finally, we study the trade-off between rendering quality and computational cost in the video synthesis stage (Table 11). We observe that executing the diffusion process 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. 12, 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.

I Robustness and Failure Analysis

To quantify the robustness of PhysOmni and characterize its failure modes, we execute the full pipeline on 100 diverse scenes and manually inspect every intermediate and final result. Table 13 reports the per-stage failure counts and success rates. The dominant source of error is VLM-based material estimation (9 failures), followed by segmentation/decomposition and pose/scene alignment (4 each) and 3D lifting (3). These failures correspond to the challenges raised by reviewers—heavy occlusion, mis-segmentation, VLM material errors, and depth ambiguity—and remain relatively rare, yielding an overall usable rate of 0.89. We further stress-test the pipeline on highly cluttered kitchen-counter scenes (Fig. 13), where it reliably segments, reconstructs, and simulates the scene while remaining interactive.

Table 12 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.

Table 13 Per-stage failure analysis on 100 scenes. We manually inspect each result and attribute failures to the earliest stage at which they occur.

Stage (100 scenes)	# Failures	Success
Segmentation / decomposition (SAM3)	4	0.96
3D lifting (SAM-3D-Objects)	3	0.97
Pose / scene alignment	4	0.96
VLM material estimation	9	0.91
Overall usable results	**11**	**0.89**

J Additional Qualitative Results

In this section, we provide further qualitative results to complement the main text.

We further provide a direct comparison against PhysGen Liu *et al.* (2024b) in Fig. 14. PhysGen targets 2D image-space manipulation and therefore cannot express the 3D force-field scene manipulation that PhysOmni performs; consequently, the stronger physics-aware baselines PhysCtrl and WonderPlay are used for the main comparisons. On a shared domino-toppling event, PhysOmni produces physically coherent sequential collapse consistent with the applied force, whereas PhysGen exhibits less faithful 3D dynamics.

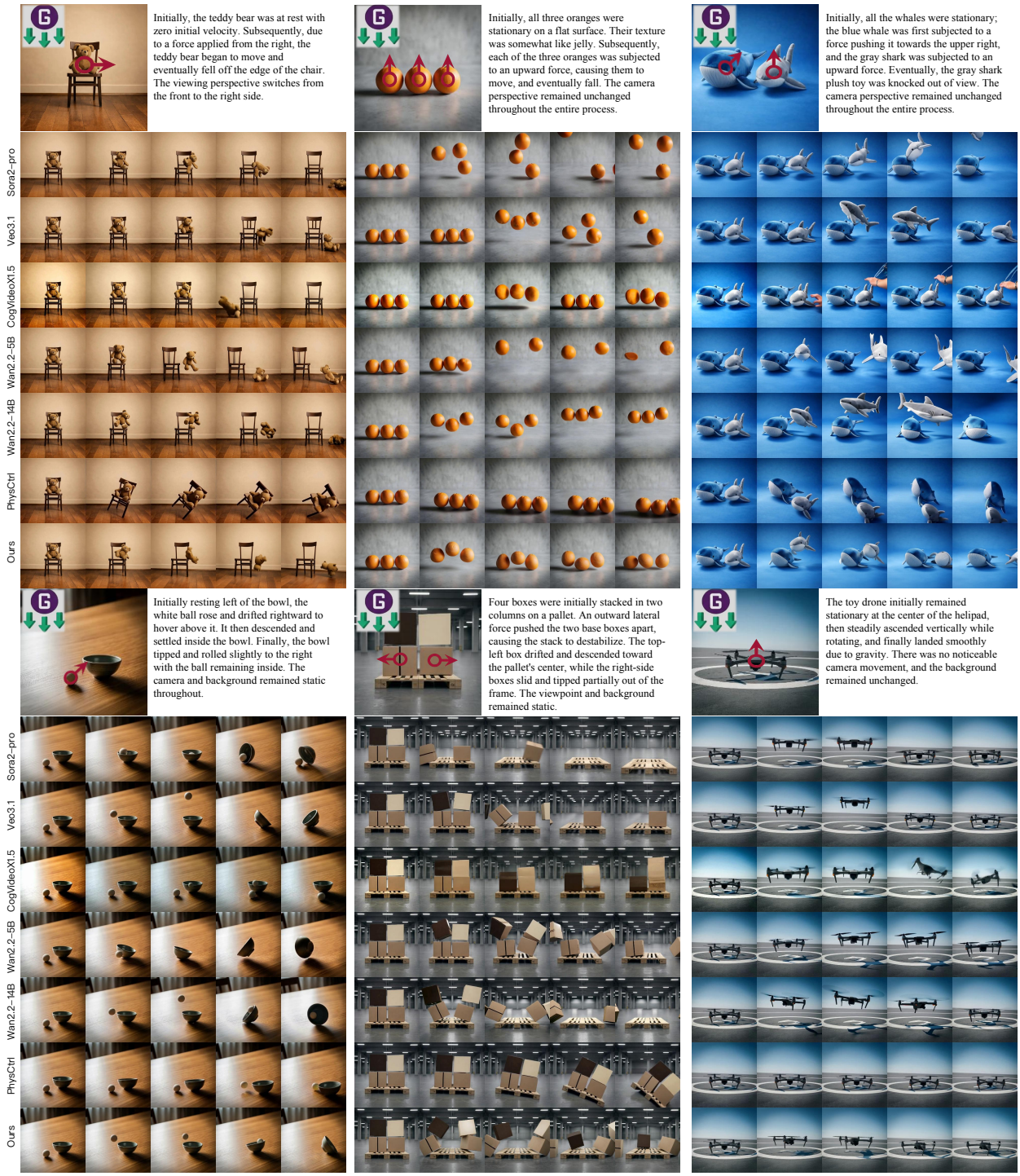

Figure 11 Qualitative comparison between the proposed method and existing video generation methods.

Figure 12 Qualitative results of the proposed method.

Figure 13 Robustness on a highly cluttered scene (a kitchen counter). Top-to-bottom, left-to-right shows the temporal evolution of the interactive physics preview. Despite dense clutter and heavy occlusion, PhysOmni reliably segments, reconstructs, aligns, and simulates the scene while remaining interactive.

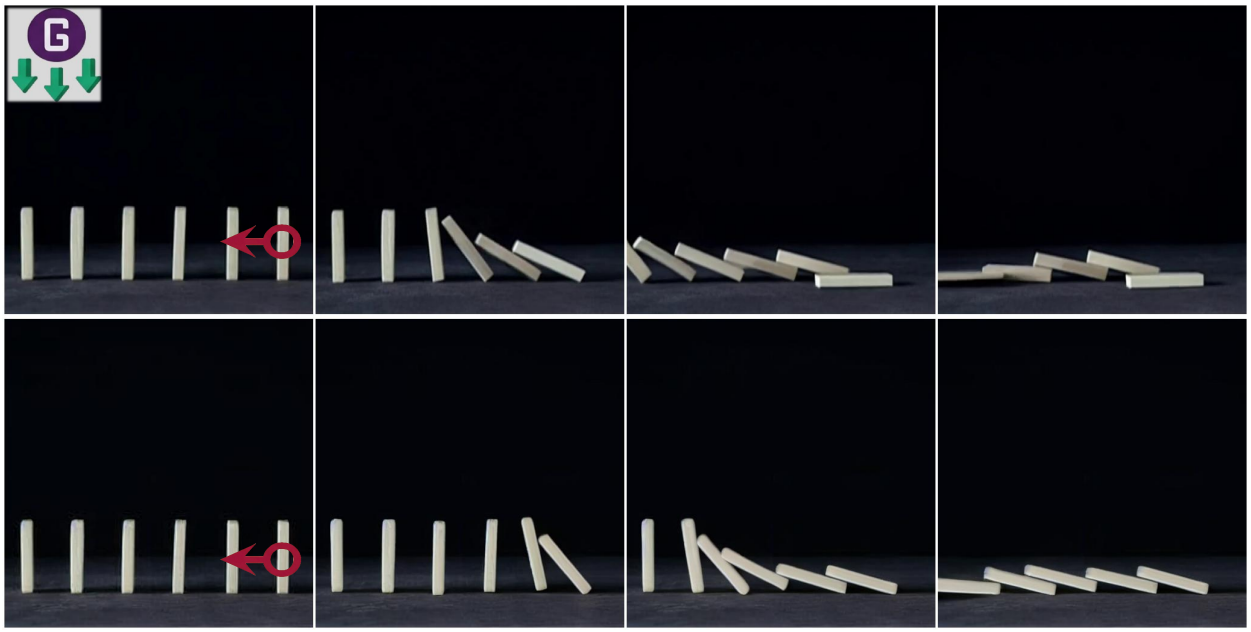

Figure 14 Direct qualitative comparison with PhysGen Liu *et al.* (2024b) on a domino-toppling event under the same applied force. Top: PhysOmni (Ours); bottom: PhysGen. PhysOmni yields a physically coherent sequential collapse, while PhysGen, operating in 2D image space, is less faithful to the underlying 3D dynamics.