

PhyMAGIC: Physical Motion-Aware Generative Inference with Confidence-guided VLM

Siwei Meng¹, Yawei Luo², and Ping Liu¹

¹ Department of Computer Science & Engineering, University of Nevada, Reno, USA

² School of Software Technology, Zhejiang University, China

siweim@unr.edu, yaweiluo@zju.edu.cn, pino.pingliu@gmail.com

Abstract. Inferring physical properties from a single image is fundamentally under-constrained. Attributes such as density, elasticity, and yield stress govern how objects move, yet they are largely invisible in a static frame. Existing physics-aware methods attempt to resolve this ambiguity through task-specific fine-tuning or supervised property estimation, but both strategies struggle to generalize across diverse materials and scenes. We observe that different motions expose complementary physical cues. Building on this observation, we propose **PhyMAGIC**, a training-free framework that actively probes physical properties by synthesizing targeted motions from a single image. Specifically, PhyMAGIC uses a pretrained image-to-video model to construct *motion probes* that generate diverse dynamic sequences from the input image. A vision-language model then analyzes these sequences to estimate physical parameters, each accompanied by a confidence score. Parameters with low confidence trigger targeted prompt refinement, which generates additional probe motions to gather complementary evidence. Once all parameters reach sufficient confidence, PhyMAGIC compiles them into a complete physical specification and executes it in a differentiable Material Point Method simulator initialized from 3D Gaussian reconstructions. Experiments on diverse real-world scenes demonstrate that PhyMAGIC achieves stronger text-motion alignment and higher human-rated physical plausibility than state-of-the-art open-source video generators and physics-aware baselines. Code is available at: <https://mengsiwei.github.io/MAGIC/>.

Keywords: Dynamic 3D · Physical Simulation · Vision-Language Models · Material Point Method

1 Introduction

Recent advances in geometry and texture synthesis [5, 20, 27, 30, 32] have significantly improved the fidelity of static 3D assets from a single image. Extending these assets to physically consistent dynamics, however, remains an open

[✉] Corresponding authors.

challenge. Modern video diffusion models [22, 35] can produce visually realistic motion, but they lack explicit control over physical attributes such as mass, elasticity, and friction. As a result, their outputs frequently exhibit momentum violations, object interpenetration, and unrealistic material responses.

These limitations have motivated research on physics-aware generative methods [12, 19]. Existing approaches generally follow two paradigms. The first embeds physical priors directly into generative architectures [4, 34], which improves physical consistency but often reduces generalization to novel scenes or unseen materials. The second relies on vision-based refinement or learning-based property estimation [15, 16, 26], which offers more flexibility but depends on large annotated datasets and remains vulnerable to perceptual ambiguity. Both paradigms share a common bottleneck: they struggle to reliably infer intrinsic physical properties when visual evidence is scarce.

This bottleneck reflects a deeper issue: physical inference from a single static image is fundamentally under-constrained. Multiple motion trajectories can appear visually plausible for the same object while being consistent with very different underlying physical properties. Attributes such as density, elasticity, and yield stress are largely invisible in a static frame, yet they critically determine how an object moves, deforms, and interacts. Manually specified priors [33] lack adaptability across materials. Learning-based estimators [3, 13, 38] are susceptible to data scarcity and domain shift. Recent Vision-Language Models (VLMs)-based methods [15, 16] offer a promising direction by incorporating high-level physical reasoning, but they are less reliable when visual cues are sparse and are not coupled with generative models in a closed-loop fashion. What is missing is a mechanism that can actively gather physical evidence from a single image, rather than passively predicting properties from limited observations.

To address this challenge, we propose **PhyMAGIC**, a training-free framework that resolves the ambiguity of single-image physical reasoning through an active video evidence chain and generates physically controlled dynamics with a differentiable MPM simulator. Instead of directly estimating physical attributes from a single static observation, PhyMAGIC synthesizes motion probes that expose otherwise hidden physical cues, which provides temporal evidence about object behavior. Our key insight is that combining the reasoning ability of vision-language models (VLMs) with the motion diversity of pretrained image-to-video (I2V) diffusion models can actively reveal informative physical behaviors and progressively reduce uncertainty in physical inference. As illustrated in Figure 1, different motion patterns reveal complementary aspects of an object’s physical properties: free-fall and explosion motions expose material and mass cues that are difficult to observe from gentle perturbations alone. This formulation transforms physical inference from passive single-image prediction into an active probing process, where each generated motion serves as a targeted experiment to reduce uncertainty.

Concretely, PhyMAGIC operates in three tightly coupled stages. In the *Motion Probe Generation stage*, PhyMAGIC synthesizes motion probe videos using a pretrained I2V model [35], transforming a single frame into a rich source of

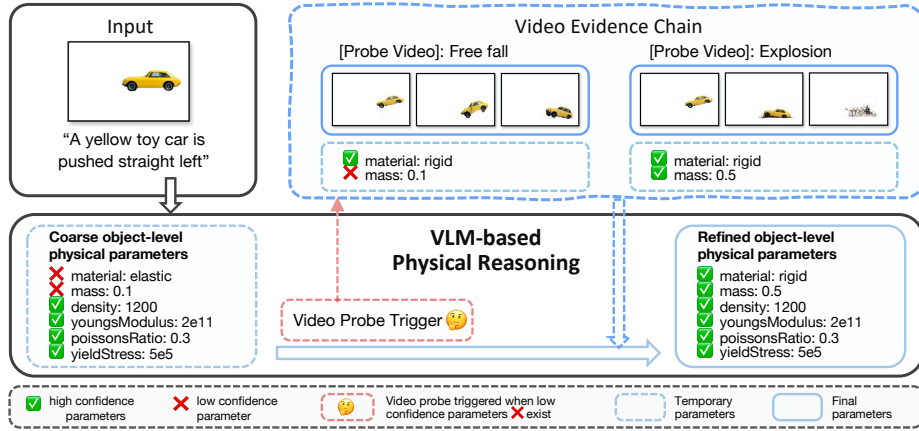


Fig. 1: PhyMAGIC: Active motion probing for physical identification. A single image and instruction are under-constrained for inferring motion-critical physical attributes. For example, coarse VLM reasoning on the static input correctly identifies the material as rigid but assigns low confidence to mass and yield stress. To resolve this uncertainty, PhyMAGIC synthesizes targeted motion probes that expose complementary physical cues. The VLM re-evaluates the probed videos across iterations, progressively refining low-confidence parameters until a complete set of physical attributes is obtained. The resulting parameters, together with simulator-native descriptors are then executed in an MPM simulator to produce physically consistent dynamics.

temporal evidence. In the *VLM-based Physical Reasoning stage*, a VLM then analyzes these videos to estimate physical parameters, each accompanied by a confidence score. Parameters with low confidence trigger targeted prompt refinement, which generates additional probe motions to gather complementary evidence. This loop repeats until all parameters reach sufficient confidence or an iteration budget is exhausted. This closed-loop design enables PhyMAGIC to iteratively reduce uncertainty in physical inference without manual intervention. In the *physics-grounded dynamic stage*, the VLM additionally infers simulator-native descriptors (e.g., boundary conditions, contact modes, and force directions) from the probe context and instruction. PhyMAGIC compiles these descriptors together with the inferred material parameters into a hybrid physical specification. This specification is executed in a differentiable Material Point Method (MPM) simulator [7], initialized from 3D Gaussian reconstructions [32], producing physically consistent dynamic 3D content without task-specific training or manual annotation.

Our main contributions are as follows:

- **Active motion probing for physical inference.** We formulate single-image physical reasoning as an active evidence acquisition problem. Pretrained I2V models serve as motion probes, and iterative VLM feedback progressively reveals complementary physical cues for motion-critical attributes.

- **Confidence-guided prompt refinement.** We introduce an iterative refinement strategy that uses per-attribute confidence scores to identify uncertain parameters and automatically generates targeted probe motions that reduce uncertainty.
- **Hybrid physical specification with differentiable simulation.** We introduce a unified representation that combines VLM-inferred material properties (density, elasticity, yield stress) with simulator-native descriptors (boundary conditions, contact modes, force configurations), and execute it in a differentiable MPM simulator for physics-consistent dynamic 3D generation.
- **Comprehensive evaluation.** Experiments on PhysGaussian, PhysGen, and Internet-collected scenes demonstrate that PhyMAGIC achieves stronger text-motion alignment and higher human-rated physical plausibility than state-of-the-art video generators and physics-aware baselines, while maintaining competitive visual quality.

2 Related Work

3D Dynamic Generation. Recent methods extend static 3D representations into temporal sequences via trajectory modeling or deformation fields [11, 25]. Neural implicit approaches [6, 23] achieve photorealistic dynamic novel-view synthesis but at high computational cost. 3D Gaussian Splatting (3DGS) [9] and its dynamic extensions [8, 37] enable real-time rendering with limited deformations, yet these approaches focus on visual fidelity rather than explicit physical modeling. As a result, the recovered dynamics may look plausible in pixel space but are not directly tied to material properties, forces, or boundary conditions. PhyMAGIC builds on 3DGS but differs in that it explicitly injects VLM-inferred physical attributes into a differentiable simulator. This turns 3DGS from a mainly appearance-driven representation into a simulation-ready physical carrier for downstream motion prediction and rendering.

Physics-Grounded Generative Models. Physically embedded networks [4, 34] integrate constraints such as elasticity and collision responses into architectures, but typically involve material-dependent designs that restrict generalization. PhysGaussian [33] simulates deformation via continuum mechanics with manually specified parameters. PhysDreamer [38] learns physical properties from video diffusion models, and Physics3D [14] combines viscoelastic MPM with Score Distillation Sampling for multi-material simulation. OmniPhysGS [13] learns constitutive relations to model diverse materials, but its generalization depends on training data coverage. Vision-based refinement frameworks [15, 16, 26] decouple generation from physical reasoning, offering flexibility but facing difficulties in precisely controlling physical attributes. In contrast, PhyMAGIC infers parameters at test time through active motion probing, requiring no task-specific training or manual tuning.

Vision-Language Model Reasoning for Physical Inference. VLMs have shown promise for physical reasoning by leveraging implicit knowledge of materials and motion [2, 31]. PhysGen [15] infers mass, elasticity, and force directions

from static images via a VLM, while PhysFlow [16] enriches motion context through auxiliary video generation [35] before property estimation. However, both perform inference in a single forward pass without mechanisms to refine uncertain predictions. Learning-based estimators such as GIC [3] train dedicated networks but are susceptible to data scarcity and domain shift. PhyMAGIC addresses these limitations through an iterative confidence-guided loop that actively generates targeted motion probes to gather complementary evidence for low-confidence attributes.

3 Preliminary

In this section, we introduce the mathematical foundations and notation underlying our framework, with a focus on Gaussian Splatting for 3D representation and the Material Point Method for differentiable physical simulation in 3D space.

3D Gaussian Splatting (3DGS). 3D Gaussian Splatting [9] represents a scene as a collection of anisotropic Gaussian kernels $\{x_k, \Sigma_k, \alpha_k, c_k\}, k \in \mathcal{K}$, encoding position, covariance matrix, opacity, and view-dependent color. Unlike implicit neural representations [21, 36], 3DGS features a fully differentiable rasterization rendering process, where the pixel color $C(u)$ is computed by alpha blending depth-sorted Gaussians:

$$C(u) = \sum_{k \in \mathcal{P}(u)} \alpha_k \mathcal{G}_k(u) \text{SH}(d_k; c_k) \prod_{j < k} (1 - \alpha_j \mathcal{G}_j(u)). \quad (1)$$

Here $\mathcal{G}_k(u)$ is the current Gaussian projection at pixel u , and SH denotes spherical harmonics for view-dependent colors c_k of viewing direction d_k . Crucially, 3DGS exposes per-Gaussian parameters that are directly optimizable, which we later exploit to reconstruct and simulate deformable objects from the single image input.

Continuum Mechanics and Material Point Method. We model material behaviors using continuum mechanics [7, 24] that describes local deformation by a deformation map $\phi(\mathbf{X}, t)$, relating initial coordinates $\mathbf{X}$ to current positions $\mathbf{x} = \phi(\mathbf{X}, t)$. For numerical simulation, we employ the Material Point Method (MPM) [7, 40], a hybrid Lagrangian–Eulerian approach that discretizes materials into particles $\{\mathbf{x}_p, m_p, \mathbf{v}_p, \mathbf{F}_p\}_{p=1}^N$ carrying mass, velocity, and deformation state, while using a background grid to compute forces and spatial derivatives. In MPM, each simulation step alternates between: (1) particle-to-grid (P2G) transfer of mass and momentum, (2) grid update by solving discretized momentum equations under internal and external forces, and (3) grid-to-particle (G2P) transfer to update particle velocities, positions, and deformation gradients. This formulation naturally handles large deformations, collisions, and multi-material interactions with numerical stability, and its differentiable implementation allows seamless integration into our closed-loop generation framework.

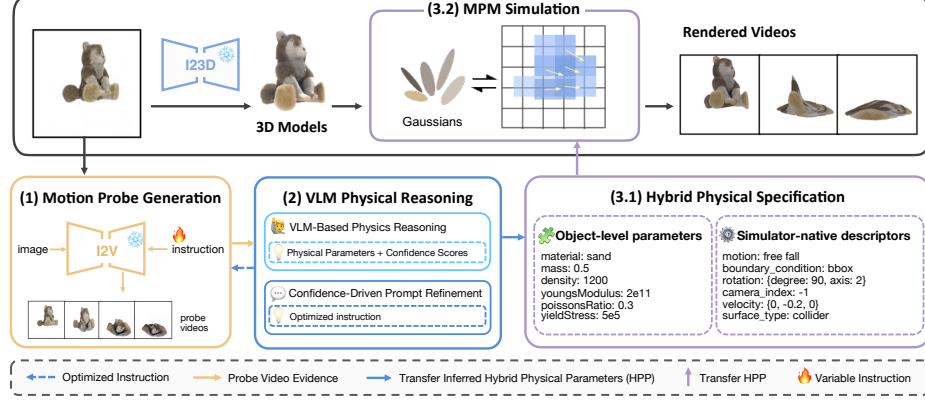


Fig. 2: Overview of PhyMAGIC framework. Given a single input image, PhyMAGIC first generates motion probe videos using a pretrained I2V diffusion model, which exposes diverse temporal evidence for physical reasoning. A VLM then estimates object-level physical parameters together with confidence scores. Low-confidence attributes trigger a prompt refinement loop that synthesizes additional motion probes to gather complementary evidence. Once parameters reach sufficient confidence, PhyMAGIC compiles a hybrid physical specification combining material properties with simulator-native descriptors. The resulting configuration is executed in a differentiable MPM simulator to produce physically consistent dynamic 3D behaviors.

4 Methodology

We propose **PhyMAGIC**, a training-free framework for inferring physical properties and simulating physically consistent 3D dynamics from a single image. As illustrated in Figure 2, PhyMAGIC consists of three components: (1) *Motion probe generation* (§4.1) synthesizes probe videos from the input image using a pretrained image-to-video diffusion model, providing temporal evidence that reveals otherwise unobservable physical cues. (2) *VLM-based physical reasoning* (§4.2) performs object-level physical inference within a confidence-driven refinement loop, where the VLM and I2V model alternate to progressively improve parameter estimates. (3) *Physics-grounded 3D dynamics* (§4.3) compiles the inferred attributes into a *Hybrid Physical Parameters (HPP)* specification that combines material properties with simulator-native descriptors, and executes it in a differentiable MPM simulator. The closed loop between motion probe generation and VLM inference iteratively reduces uncertainty in physical reasoning, while the simulation stage translates the converged estimates into physically consistent dynamic 3D content.

4.1 Motion Probe Generation

Physical attributes such as density, yield stress, and external forces are only revealed through temporal behavior [17]. To enrich the visual cues available

from a single static image, we use a pretrained image-to-video (I2V) diffusion model as a motion probe generator. Specifically, we employ CogVideoX-I2V³ [35] to synthesize a sequence of frames $V_0 = \{I_0, I_1, \dots, I_T\}$ from the input image I_0 . To reduce redundancy while preserving informative transitions, we apply motion-aware subsampling based on optical flow magnitude, retaining frames with significant motion changes. Rather than treating the generated video as a ground-truth simulation, we leverage it as an informative probe that transforms latent physical priors into explicit motion cues, such as temporal trajectories and deformation patterns, providing the subsequent VLM reasoning module with richer evidence for physical property estimation.

4.2 VLM-Based Physical Reasoning

Iterative Physical Inference. We design a structured, coarse-to-fine inference framework using GPT-4o [2], which combines synthesized probe videos with targeted descriptive prompts. As illustrated in Figure 3, the VLM first analyzes the input image and textual instruction to estimate a set of object-level physical parameters P_i along with confidence scores $c_i \in [0, 1]$. We estimate confidence via self-consistency: each attribute is inferred across up to three independent VLM samples, and c_i is computed as the agreement ratio among the sampled responses. Parameters with c_i below a confidence threshold γ are identified as uncertain and trigger the refinement mechanism described below.

Confidence-Driven Prompt Refinement. Direct VLM predictions can suffer from prompt-motion misalignment or insufficient motion cues, leading to inaccurate estimates for certain attributes. As illustrated in Figure 3, a toy horse is inferred with incorrect Young’s modulus and Poisson’s ratio due to vague descriptions and insufficient motion evidence. To address this, we introduce a confidence-guided refinement loop. We define a confidence threshold γ^4 and identify attributes P_i whose scores fall below it. These attributes are marked using a binary indicator:

$$m_i = \begin{cases} 1, & c_i < \gamma, \\ 0, & c_i \geq \gamma, \end{cases} \quad (2)$$

where all indices i with $m_i = 1$ form the set requiring refinement. Starting from the initial prompt $p^{(0)}$, we update it iteratively at step t as

$$p^{(t+1)} = \text{VLM}(p^{(t)}, \{P_i \mid m_i = 1\}), \quad (3)$$

where the VLM receives the current prompt and low-confidence attributes, then generates a refined prompt that explicitly requests more informative motion

³ We select CogVideoX-I2V for its open-source availability and support for reproducible inference with fixed seeds.

⁴ We select $\gamma=0.6$ using a held-out pilot set of 15 prompts (around 100 attributes queries). We plot the confidence-score histogram and choose $\gamma=0.6$ to separate the low-confidence sets that most often trigger video probe generation requests. See supplementary for the details.

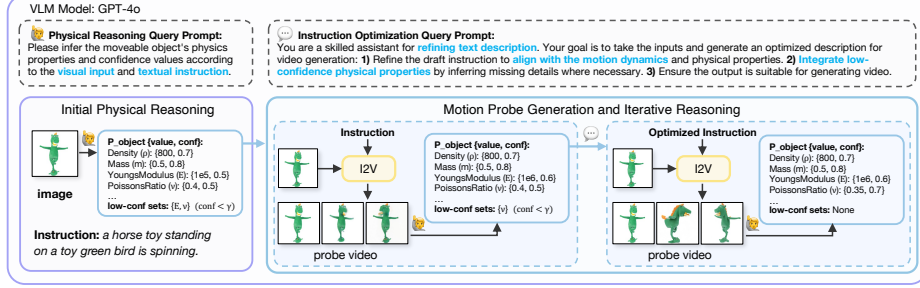


Fig. 3: Confidence-driven prompt refinement process. Given an image and an initial instruction, a VLM (GPT-4o [2]) first infers coarse object-level physical properties together with confidence scores. Parameters with confidence below a threshold γ are treated as uncertain and trigger motion probe generation. The probe videos provide complementary visual evidence, based on which the VLM refines the instruction to expose motion dynamics relevant to the uncertain properties. This process iteratively accumulates evidence until all parameters reach sufficient confidence or the maximum number of refinement iterations is reached, yielding a stable set of physical parameters for subsequent simulation.

patterns for those attributes. At each iteration, the refined prompt drives a new round of I2V probe generation, and the VLM re-estimates all parameters with updated confidence scores. This loop repeats until all attributes exceed γ or a maximum iteration budget is reached.

4.3 Physics-Grounded 3D Dynamics

While §4.2 focuses on identifying *what* physical properties an object has, this stage addresses *how* it moves and interacts over time. We combine the inferred attributes with simulator-native descriptors into a unified specification, reconstruct a 3D representation from the input image, and execute the resulting configuration in a differentiable MPM solver.

Hybrid Physical Parameters (HPP). A key challenge in connecting VLM reasoning to physical simulation is the semantic gap between their outputs. The VLM produces high-level material descriptions (e.g., “rigid body with density 1200 kg/m³”), while the MPM solver requires a complete, low-level executable configuration specifying not only material constants but also boundary conditions, contact modes, and applied forces. Neither component alone is sufficient for physically consistent simulation. Material properties such as density and elasticity define how an object responds to forces, but without specifying the force itself, the initial velocity, or the boundary geometry, the simulator has no well-posed initial-boundary value problem to solve. Conversely, execution context such as “apply gravity and drop” is meaningless without knowing whether the object is made of rubber or steel, as the same force profile produces qualitatively different trajectories depending on material stiffness and yield behavior.

To bridge this gap, we introduce a Hybrid Physical Parameters (HPP) specification that unifies both into a single simulator-ready configuration. Formally, we define HPP as:

$$\mathcal{H} = (\mathcal{P}_{\text{mat}}, \mathcal{D}_{\text{sim}}), \quad (4)$$

where $\mathcal{P}_{\text{mat}} = \{\rho, E, \nu, \sigma_y, \dots\}$ denotes the object-level material properties (density, Young’s modulus, Poisson’s ratio, yield stress, etc.) inferred during the perception stage, and $\mathcal{D}_{\text{sim}} = \{v_0, \mathcal{B}, f_{\text{ext}}, \dots\}$ denotes the simulator-native descriptors (initial velocity, boundary conditions, external forces, etc.). All material values in $\mathcal{P}_{\text{mat}}$ are expressed in SI units [29] and normalized into simulator ranges to maintain numerical stability. HPP is defined at the object level and propagated to all constituent particles during simulation initialization.

For the simulator-native descriptors $\mathcal{D}_{\text{sim}}$, the VLM analyzes the textual instruction and accumulated probe evidence to select a canonical motion template from a predefined set (e.g., free-fall, compression, spinning, rebound, side-push). Each template defines a structured combination of boundary type, initial velocity direction, force magnitude, and contact configuration. For example, a “side push” template configures a horizontal initial velocity, a bounding-box boundary condition, and a collider surface type, while a “free fall” template applies only gravitational force with no initial velocity. This template-based design ensures that the inferred physical state is translated into a complete, solver-compatible configuration while keeping the mapping interpretable and reproducible.

MPM-Based 3D Dynamic Simulation. We reconstruct the input image into a 3D Gaussian Splatting representation using Trellis [32]. Each Gaussian kernel serves as a material-carrying particle, initialized by augmenting the Gaussian state with the inferred HPP attributes:

$$\mathcal{G}_p^t = (x_p^t, \Sigma_p^t, \alpha_p, c_p, \theta_p^t), \quad (5)$$

where θ_p^t encodes physical properties such as density, elasticity, and yield stress. We execute the simulation using a differentiable MPM solver [7], which updates particle positions, velocities, and deformation gradients via standard particle-grid-particle transfers. Unlike existing differentiable simulators [1, 10, 39] that assume predefined material parameters, our approach initializes all physical attributes directly from VLM inference, enabling physically consistent dynamic 3D generation without manual parameter specification.

5 Experiments

We evaluate PhyMAGIC in a training-free setting by integrating pretrained Trellis [32] for 3D Gaussian reconstruction, CogVideoX-5b [35] for video generation (720p, 50 frames), GPT-4o [2] for physical reasoning, and a differentiable MPM solver [33] implemented in Warp [18]. Simulations utilize 128–200 steps, depending on scenario complexity, and run on a single NVIDIA RTX 4090 GPU. We benchmark on PhysGaussian [33], PhysGen [15], and Internet-collected single-image scenes. Evaluation metrics include CLIP similarity for text–motion alignment, aesthetic score for visual quality, Image-Motion-FID for evaluating the

Table 1: Quantitative evaluation on CLIP similarity and Aesthetic score with video generation methods across eight scenarios. Best results in each scenario are in **bold**, and the second-best results are in underline.

CLIP similarity $\uparrow$									
Method	Swing figus	Sand wolf	Driving car	Rolling basketball	Tear toast	Sway tree	Lifting hat	Swing carnation	Average
OpenSora2.0	0.270	0.264	0.195	0.230	0.274	0.174	0.239	0.217	0.233
CogVideoX	0.279	0.225	0.241	0.226	0.250	0.186	0.252	0.254	0.239
CogVideoX*	0.257	0.263	0.234	0.233	0.279	0.195	0.241	0.254	<u>0.240</u>
Ours	0.294	0.200	0.252	0.266	0.231	0.243	0.270	0.256	0.251
Aesthetic score $\uparrow$									
Method	Swing figus	Sand wolf	Driving car	Rolling basketball	Tear toast	Sway tree	Lifting hat	Swing carnation	Average
OpenSora2.0	27.26	14.58	5.51	17.74	23.85	17.44	2.53	26.97	16.98
CogVideoX	33.19	22.55	38.55	11.91	47.18	21.51	16.37	54.02	30.66
CogVideoX*	34.89	23.94	41.20	17.95	46.49	22.53	16.41	51.47	31.86
Ours	31.00	18.38	29.40	26.82	49.85	39.67	26.74	23.70	<u>30.69</u>

Table 2: Quantitative results on CLIP similarity with physics-aware generation methods across six representative scenes. Best results are highlighted in **bold**.

Method	Swing figus	Sand wolf	Lifting hat	Sway tree	Swing carnation	Rolling basketball	Average
OMNIPHYSGS	0.227	0.167	-	-	-	-	0.197
PhysDreamer	0.223	0.145	0.239	0.186	0.272	-	0.213
Physics3D	0.225	0.147	0.229	0.147	0.269	0.283	0.217
Ours	0.294	0.200	0.270	0.226	0.256	0.266	0.252

distribution of ground-truth images with generated videos, and a 61-participant user study for physical plausibility and text consistency. We compare with state-of-the-art video generation models (CogVideoX, Open-Sora 2.0) and physics-aware approaches (PhysDreamer, Physics3D, OMNIPHYSGS). CogVideoX* denotes CogVideoX augmented with our confidence-driven prompt refinement for ablation analysis. Additional implementation settings, video generator choice, and user study details are provided in the supplementary material.

5.1 Quantitative Results

Comparison with image-to-video generation models. By leveraging iterative VLM-guided reasoning and a physically grounded MPM simulator, PhyMAGIC generates physically plausible 3D dynamics from only a single input image and text prompt. Table 1 reports quantitative comparisons with state-of-the-art image-to-video generation models. Our method achieves the highest CLIP similarity scores in most scenarios, indicating stronger semantic alignment between generated motions and textual descriptions. On average, PhyMAGIC reaches 0.251, outperforming OpenSora2.0 (0.233), CogVideoX (0.239), and CogVideoX* (0.240). Despite operating with only a single-image input, Phy-

Table 3: Quantitative results on Image and Motion-FID of generated videos with the GT input images.

Method	Image-Motion-FID↓
PhysDremer	107.47
OMNIPHYSGS	106.60
Physics3D	98.91
Ours	94.69

Table 4: Human evaluation on physical plausibility and text consistency across eight real-world scenes compared to three I2V generation models. Ours achieves the highest human scores in both terms.

Method	Physical Plausibility↑	Text Consistency↑
OpenSora2.0	2.02	2.18
CogVideoX	2.58	2.49
CogVideoX*	2.97	3.00
Ours	3.00	3.07

Table 5: Representative motion scenarios including ficus swing (elastic), car driving (rigid-body), and basketball rolling (deformable). We report key governing parameters (density, modulus, and friction angle) and their accuracy across iterations, showing progressive improvement and convergence toward ground truth.

Parameter	Swing ficus				Rolling basketball				Driving car			
	GT	Iter1	Iter2	Iter3 (final)	GT	Iter1	Iter2	Iter3 (final)	GT	Iter1	Iter2	Iter3 (final)
Material	Elastic	Elastic	Elastic	Elastic	Elastic	Elastic	Elastic	Elastic	Rigid	Elastic	Rigid	Rigid
Density	400	600	250	250	1000	85.71	200	600	1200	1000	1200	1200
Young's Modulus	3e6	5e4	3e6	3e6	1e5	1e4	1e5	1e5	2e9	2.5e4	2e9	2e9
Poisson's Ratio	0.30	0.30	0.40	0.30	0.40	0.47	0.50	0.40	0.40	0.25	0.75	0.50
Yield Stress	—	—	—	—	—	—	—	—	6e7	—	3e6	3e7
Accuracy (%) ↑	—	62.92	82.29	90.63	—	50.27	73.75	90.00	—	29.17	44.50	93.75

MAGIC also achieves competitive performance on the Aesthetic score, surpassing other models in the *rolling basketball*, *tear toast*, *sway tree*, and *lifting hat* scenarios. These improvements stem from PhyMAGIC’s explicit integration of physical priors, which guides motion synthesis toward trajectories that are not only visually appealing but also physically consistent.

Comparison with physics-aware 3D dynamic generation models. Table 2 compares PhyMAGIC with three representative physics-aware baselines across multiple challenging scenarios. PhyMAGIC achieves substantial gains, attaining scores of 0.294 and 0.200 on the elastic (*swing ficus*) and sand (*sand wolf*) cases, respectively. We present the distribution of the collected GT images with the generated videos in Table 3. Our method achieves the lowest Image-Motion-FID score, surpassing existing physics-aware baselines. These results highlight PhyMAGIC’s ability to preserve visual fidelity while maintaining accurate text–motion coherence. Notably, PhyMAGIC improves semantic similarity by 16.1% over Physics3D, underscoring the advantage of our iterative VLM-guided refinement in producing dynamics that are both physically plausible and semantically aligned.

Human Evaluation. We conducted a user study to assess physical plausibility and text–motion alignment. Participants rated 32 videos on a 4-point Likert scale (1 = least, 4 = most consistent). As shown in Table 4, PhyMAGIC achieved the highest mean ratings (3.00 for plausibility, 3.07 for alignment), followed by CogVideoX* (2.97/3.01), which still outperformed CogVideoX and OpenSora2.0. These results confirm that confidence-guided refinement improves

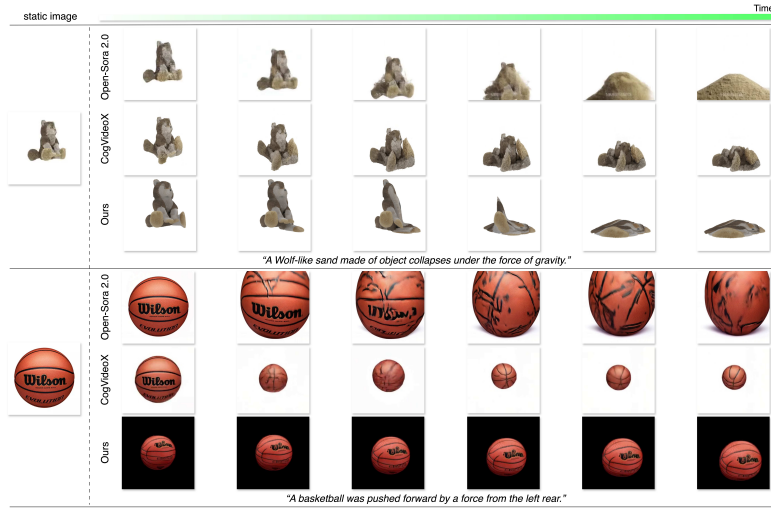


Fig. 4: Qualitative comparisons of dynamic scene generation between our method and state-of-the-art video generation models [22,35]. Given only a single static image, PhyMAGIC effectively infers intrinsic physical properties. It generates highly realistic dynamics over time, demonstrating superior physical realism and temporal consistency.

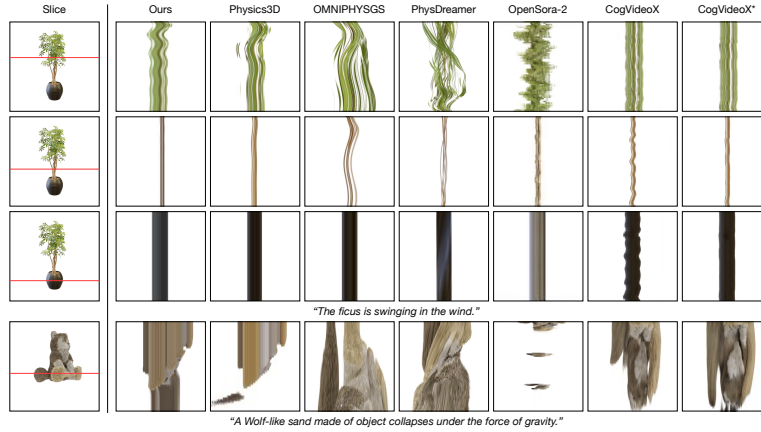


Fig. 5: Space-time slices comparison with physical-aware and I2V generation methods.

dynamic realism, which enables PhyMAGIC to deliver motions preferred by human evaluators.

5.2 Qualitative Results

Visual Comparison with Video Generation Models. As illustrated in Figure 4, our method exhibits clear advantages in modeling physically realistic dynamics given only a static image, whereas OpenSora 2.0 [22] and CogVideoX [35]

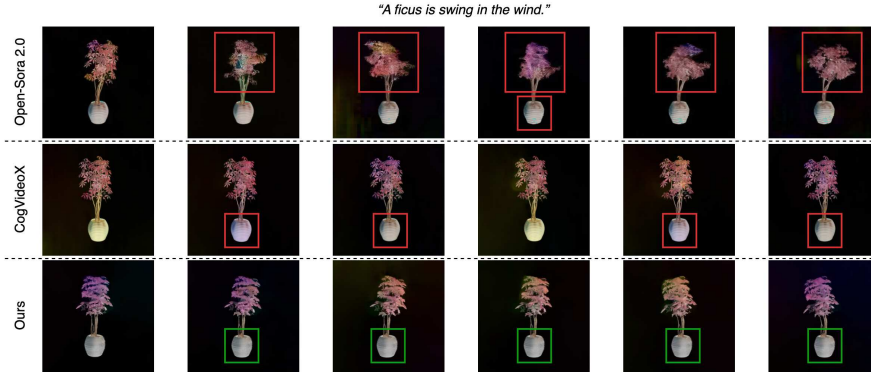


Fig. 6: Qualitative comparison of optical flow visualizations on the *swing ficus* scene. PhyMAGIC maintains coherent motion and structural stability over time (green boxes), while competing methods [22, 35] show structural inconsistencies and unnatural distortions (red boxes).

often produce visual distortions or physically implausible motions. For the wolf-like sand collapse, Open-Sora 2.0 and CogVideoX exhibit unnatural particle scattering behaviors, whereas our results accurately simulate the behavior of sand material under gravity. In the *rolling basketball* scene, baseline models produce implausible deformations under external forces or distorted appearance, whereas our method maintains consistent appearance and motion.

Spatiotemporal Slice Analysis. To evaluate temporal coherence, we compare space-time slices in Figure 5. Our method produces precise and continuous motion trajectories that match expected physical laws. By contrast, Physics3D, OMNIPHYSGS, PhysDreamer, and OpenSora-2 often show blurred edges and fragmented strips, indicating weak temporal consistency and inaccurate physical dynamics. CogVideoX and CogVideoX* exhibit regular motion patterns but fail to control static regions, resulting in unintended motion in non-moving components such as the base of the ficus.

Optical Flow Visualization. Figure 6 visualizes optical flow [28] on a sample *swing ficus* scene. OpenSora2.0 exhibits spurious edges and temporal jitter around movable regions, indicating weak motion coherence, and the bottom vase also shows slight deformation. Both PhyMAGIC and CogVideoX illustrate stable optical flow; however, CogVideoX lacks precise spatial control and generates unintended motion in static regions due to its global motion modeling strategy. Our method achieves accurate motion localization through physics-aware constraints, ensuring that only the target regions exhibit dynamic deformation while static objects remain unaffected.

5.3 Ablation Study

To assess the contribution of VLM-based physics reasoning and determine an effective number of refinement iterations, we report results in Table 5. With lim-

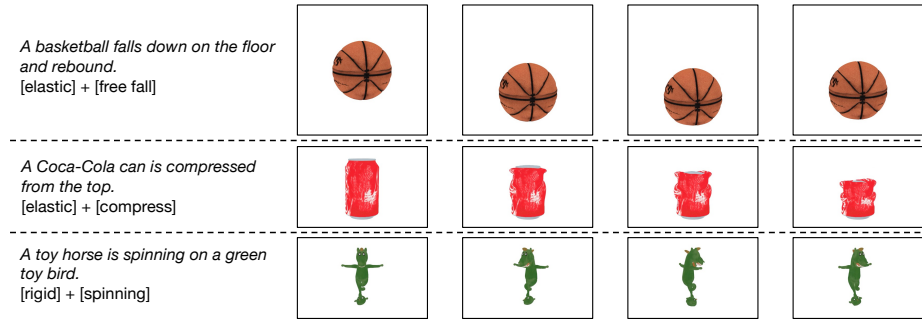


Fig. 7: The generalization ability of our PhyMAGIC across different materials and motion types. Given a single image (the first column frame), PhyMAGIC produces plausible and controlled dynamics under diverse motion conditions.

Table 6: Iterative refinement ablation of material accuracy, boundary-condition accuracy, and the average number of probes per scene on the five-scene subset.

Method	Mat. acc. (%) $\uparrow$	BC acc. (%) $\uparrow$	Probes/scene $\downarrow$
One-shot VLM + fixed boundary	60.0	60.0	0.00
One-shot VLM + best template	76.0	80.0	0.00
Iterative probing, no confidence gating	83.0	80.0	1.00
PhyMAGIC, full	90.0	90.0	0.80

ited initial motion context, the first iteration often misclassifies materials (e.g., predicting a driving car as elastic) or yields inaccurate physical parameters. Subsequent iterations progressively improve parameter estimation by refining text prompts and regenerating motions. Performance stabilizes at three iterations, where the overall accuracy exceeds 90%, indicating that iterative reasoning effectively resolves early ambiguities and converges to a reliable physical understanding. More results are provided in the supplementary material.

5.4 Analysis

Generalization. Due to the training-free and modular nature of PhyMAGIC, it exhibits robust generalization across distinct material-behavior combinations, ranging from elastic and rigid to granular material. We showcase this diversity through a variety of motion patterns in Figure 7, including free-fall, collision, spinning, compression, and side-push.

Necessity of Iterative Probing. A one-shot VLM observes only a static image, while density, modulus, and yield stress are ambiguous. As shown in Table 6, one-shot reaches 60.0%/60.0% material/BC accuracy, while PhyMAGIC-full reaches 90.0%/90.0% with 20% less probe budget than ungated probing. This gain shows that iterative video probing provides missing dynamic evidence for physical inference, rather than simply increasing computation. The confidence

Table 7: Comparison of OpenSora-2 and CogVideoX-5B as I2V models in our pipeline on accuracy gain over the single-image baseline, inference time, and GPU need.

Video Generators	Inference Accuracy Gain ($\pm\%$)	Inference Time (mins)	Max GPU Memory Required (GB)
OpenSora-2	+10.9	53	~ 83
CogVideoX-5B	+12.2	21	~ 9

gate further avoids unnecessary probes when the initial prediction is already reliable, leading to better accuracy with a lower average probe cost.

Video Generator Choice. Considering the limitation of computing resources, we experimented with both OpenSora-2 [22] and CogVideoX [35] as candidate open-source image-to-video generators for our framework. We found that both models yielded comparable improvements in physical inference accuracy than single image, as shown in Table 7, with a difference of less than 1.3%. However, CogVideoX was substantially more resource-efficient, requiring only a single 9 GB GPU and 21 minutes for inference, whereas OpenSora-2 approximately needed 86 GB of GPU memory and 53 minutes of inference time. Based on this comparison, we selected CogVideoX as our primary video generation model due to its efficiency and open-source availability, which are key requirements for our iterative physical reasoning pipeline.

Limitation. Despite its generalization, the simulation fidelity is inherently coupled with the quality of the initial 3D reconstruction. For instance, thin-shell structures or complex topologies generated by TRELLIS [32] may result in suboptimal particle sampling for the MPM solver. Second, while our iterative motion-probing loop significantly reduces ambiguity, the current VLM-based reasoning is primarily qualitative. This may lead to discrepancies in high-precision scenarios, such as determining exact friction coefficients or modeling high-frequency elastic vibrations. Finally, the current MPM simulation focuses on foreground object dynamics and does not yet support complex articulated objects or dense multi-object interactions. Future iterations could integrate differentiable rendering gradients to fine-tune these physical parameters directly from the visual residuals between the probe videos and the simulated results.

6 Conclusions

We present PhyMAGIC, a training-free framework that achieves motion-aware generation through iterative VLM-guided video generation and differentiable MPM simulation. Our innovations lie in integrating a pretrained video generation model and a confidence-driven VLM feedback mechanism for physics reasoning optimization in the iteration. We also leverage a real-time MPM engine that simulates multi-material behaviors, guided by parameters inferred from robust physics reasoning. Extensive experimental results demonstrate that PhyMAGIC significantly improves physical plausibility in dynamic 3D assets, while maintaining high-quality text-video consistency.

References

1. Abou-Chakra, J., Dayoub, F., Sünderhauf, N.: Particlenerf: Particle based encoding for online neural radiance fields. In: WACV (2024)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Cai, J., Yang, Y., Yuan, W., He, Y., Dong, Z., Bo, L., Cheng, H., Chen, Q.: Gic: Gaussian-informed continuum for physical property identification and simulation. In: NeurIPS (2024)
4. Cao, J., Guan, S., Ge, Y., Li, W., Yang, X., Ma, C.: Neuma: Neural material adaptor for visual grounding of intrinsic dynamics. In: NeurIPS (2024)
5. Chen, Z., Wang, F., Wang, Y., Liu, H.: Text-to-3d using gaussian splatting. In: CVPR (2024)
6. Du, Y., Zhang, Y., Yu, H.X., Tenenbaum, J.B., Wu, J.: Neural radiance flow for 4d view synthesis and video processing. In: ICCV (2021)
7. Jiang, C., Schroeder, C., Selle, A., Teran, J., Stomakhin, A.: The affine particle-in-cell method. ACM Transactions on Graphics (TOG) (2015)
8. Katsumata, K., Vo, D.M., Nakayama, H.: A compact dynamic 3d gaussian representation for real-time dynamic view synthesis. In: ECCV (2024)
9. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (2023)
10. Le Cleac’h, S., Yu, H.X., Guo, M., Howell, T., Gao, R., Wu, J., Manchester, Z., Schwager, M.: Differentiable physics simulation of dynamics-augmented neural objects. IEEE Robotics and Automation Letters (2023)
11. Li, Z., Chen, Y., Liu, P.: Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. In: NeurIPS (2024)
12. Lin, M., Wang, X., Wang, Y., Wang, S., Dai, F., Ding, P., Wang, C., Zuo, Z., Sang, N., Huang, S., et al.: Exploring the evolution of physics cognition in video generation: A survey. arXiv preprint arXiv:2503.21765 (2025)
13. Lin, Y., Lin, C., Xu, J., MU, Y.: OmniphysGS: 3d constitutive gaussians for general physics-based dynamics generation. In: ICLR (2025)
14. Liu, F., Wang, H., Yao, S., Zhang, S., Zhou, J., Duan, Y.: Physics3d: Learning physical properties of 3d gaussians via video diffusion. arXiv preprint arXiv:2406.04338 (2024)
15. Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: ECCV (2024)
16. Liu, Z., Ye, W., Luximon, Y., Wan, P., Zhang, D.: Unleashing the potential of multi-modal foundation models and video diffusion for 4d dynamic physical scene simulation. CVPR (2025)
17. Louis, N., Corso, J.J., Templin, T.N., Eliason, T.D., Nicolella, D.P.: Learning to estimate external forces of human motion in video. In: ACMMM (2022)
18. Macklin, M.: Warp: A high-performance python framework for gpu simulation and graphics. In: NVIDIA GPU Technology Conference (GTC) (2022)
19. Meng, S., Luo, Y., Liu, P.: Grounding creativity in physics: A brief survey of physical priors in aigc. In: IJCAI (2025)
20. Miao, Q., Li, K., Quan, J., Min, Z., Ma, S., Xu, Y., Yang, Y., Liu, P., Luo, Y.: Advances in 4d generation: A survey. arXiv preprint arXiv:2503.14501 (2025)
21. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

22. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., et al.: Open-sora 2.0: Training a commercial-level video generation model in \$200 k. arXiv preprint arXiv:2503.09642 (2025)
23. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR (2021)
24. Reddy, J.N.: An introduction to continuum mechanics. Cambridge university press (2013)
25. Ren, J., Xie, C., Mirzaei, A., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H., et al.: L4gm: Large 4d gaussian reconstruction model. In: NeurIPS (2024)
26. Tan, X., Jiang, Y., Li, X., Zong, Z., Xie, T., Yang, Y., Jiang, C.: Physmotion: Physics-grounded dynamics from a single image. arXiv preprint arXiv:2411.17189 (2024)
27. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: ECCV (2024)
28. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020)
29. Thompson, A., Taylor, B.N.: Use of the international system of units (si). NIST Special Publication, Gaithersburg (2008)
30. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: ECCV (2024)
31. Wang, Q., He, J., Pan, Y., Yeo, S.Y., Yang, X., Li, S.: Monosr: Open-vocabulary spatial reasoning from monocular images. arXiv preprint arXiv:2511.19119 (2025)
32. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: CVPR (2025)
33. Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In: CVPR (2024)
34. Xu, Q., Liu, J., Wong, M., Chen, C., Ong, Y.S.: Precise-physics driven text-to-3d generation. arXiv preprint arXiv:2403.12438 (2024)
35. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
36. Zhang, J., Yao, Y., Quan, L.: Learning signed distance field for multi-view surface reconstruction. In: ICCV (2021)
37. Zhang, M., Zhang, K., Li, Y.: Dynamic 3d gaussian tracking for graph-based neural dynamics modeling. In: CoRL (2024)
38. Zhang, T., Yu, H.X., Wu, R., Feng, B.Y., Zheng, C., Snavely, N., Wu, J., Freeman, W.T.: Physdreamer: Physics-based interaction with 3d objects via video generation. In: ECCV (2024)
39. Zhong, L., Yu, H.X., Wu, J., Li, Y.: Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In: ECCV (2024)
40. Zong, Z., Li, X., Li, M., Chiaramonte, M.M., Matusik, W., Grinspun, E., Carlberg, K., Jiang, C., Chen, P.Y.: Neural stress fields for reduced-order elastoplasticity and fracture. In: SIGGRAPH Asia 2023 Conference Papers (2023)