PHASEPLAY: CONTROLLABLE 3D SCENE SYNTHESIS WITH THERMAL-DRIVEN PHASE TRANSITIONS

Anonymous authors
Paper under double-blind review



Figure 1: **PhasePlay** generates dynamic 3D scenes on a single image and controllable user-defined actions, that reflect both force field-driven object interactions and material phase transitions induced by thermal simulation, firstly enabling user-interactive simulation of diverse material phase changes in a unified manner. See https://anonymous.4open.science/r/PhasePlay-388A for more video results.

ABSTRACT

Generating physically grounded videos of material transitions remains a significant challenge because both the geometry and appearance of objects evolve dynamically over time. Current approaches often face a trade-off: pre-trained diffusion models provide realistic appearance but frequently suffer from physical hallucinations and offer limited controllability for users, while traditional physics-based simulators excel at motion but struggle to capture complex visual state changes. To address this, we introduce PhasePlay, a framework for action-conditioned 3D scene dynamics that models realistic phase transitions by integrating specialized Material Point Method (MPM) and Position-Based Dynamics (PBD) solvers with an AI-assisted reasoning module. Our methodology combines the structural rigor of classical solvers with the adaptive capacity of AI to maintain precise control over both physical movement and evolving appearance. While we recognize the inherent complexity of full-scale multi-physics, evaluations against state-of-the-art baselines, such as Sora-2 and WonderPlay, demonstrate that PhasePlay achieves superior visual quality and physical fidelity across diverse interactive scenarios.

1 INTRODUCTION

The evolution of generative modeling has rapidly progressed from static imagery Baldridge et al. (2024) to the synthesis of high-fidelity, temporally coherent 3D and 4D scenes Liu et al. (2024b); Team et al. (2023); Agarwal et al. (2025). This advancement has catalyzed significant interest in generative world models—frameworks Singer et al. (2023) designed to predict future states based on specific action inputs—as foundational tools for simulating physical environments in AR/VR and embodied AI training Qureshi et al. (2025). In this context, a functional world model must satisfy three critical requirements: (a) providing consistent responses to physical interactions, (b)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ensuring these responses adhere to the laws of physics, and (c) maintaining photorealistic visual fidelity. Consequently, developing a model that simultaneously addresses these three pillars has emerged as a vital frontier in generative AI research Li et al. (2025b); Liu et al. (2024a); Chen et al. (2022); Zhang et al. (2024b).

While modern Video Diffusion Models (VDMs) can generate high-fidelity visual content Croitoru et al. (2023); Bar-Tal et al. (2024); Blattmann et al. (2023); Ho et al. (2020), they consistently struggle to replicate real-world physical dynamics, often producing motions that violate fundamental physical laws Motamed et al. (2026); Geng et al. (2025). Additionally, the lack of fine-grained user control impedes seamless editing via natural language instructions, as text prompts alone often fail to specify the precise physical forces or boundary conditions needed for meaningful content manipulation. Watson et al. (2024); Mou et al. (2024); Wang et al. (2024c). This deficiency, which persists even in state-of-the-art commercial models, stems from a significant disconnect between textual descriptions and complex motion, as well as a tendency for models to overfit training data rather than internalize generalizable physical principles Liu et al. (2024b). Although recent efforts have attempted to bridge this gap through detailed prompt engineering Xue et al. (2025) or by integrating graphics-based physical simulations, these solutions are often constrained by the specific capabilities of the underlying simulation engines and prohibitive computational costs. Consequently, developing a VDM that inherently understands and generalizes physical laws remains a critical challenge.

To improve physical plausibility, current approaches typically rely on physics simulations to compute dynamic 3D scenes based on user input Liu et al. (2024a); Zhang et al. (2024b); however, these methods are hindered by two fundamental limitations. First, they necessitate high-fidelity solvers for every dynamic interaction within a scene, yet complex behaviors—such as solid-fluid two-way coupling—remain unresolved challenges Chen et al. (2024). Second, these techniques require the full reconstruction of physical states from sparse observations Girdhar et al. (2024), a task that is often mathematically and practically infeasible when dealing with complex materials like snow, sand, cloth, or fluids from a single image Singer et al. (2022). Due to these constraints, existing methodologies are largely restricted to a narrow scope of dynamics, predominantly focusing on rigid body interactions and simplified elastic models.

Recent advancements, such as VLIPP Yang et al. (2025) and WonderPlay Li et al. (2025b), have made strides in enhancing realism, yet these systems still struggle to simulate complex physical transformations. VLIPP utilizes AI to plan object trajectories across a screen, while WonderPlay employs a 3D physics solver to calculate motion before generating the final video. Despite these improvements, both methods fail when an object's fundamental state or appearance changes. Because they are primarily designed to model objects moving or bending, they lack the necessary components to simulate phase changes, such as water freezing, butter melting, or wood igniting. This creates a major bottleneck: current models cannot accurately depict how an object's shape and visual appearance evolve during chemical or thermal transitions.

This motivates us to design a coupled framework that integrates AI reasoning with a 3D physics solver to generate both object trajectories and appearance priors for fine-tuned video synthesis. We propose PhasePlay, to the best of our knowledge, the first system that unifies force field-driven object dynamics and thermal simulation-induced material phase transitions within a fully controllable interactive environment. Ultimately, PhasePlay represents a new paradigm for bimodal video synthesis, where motion and appearance are controlled through a unified, physically consistent pipeline. Our core approach leverages AI reasoning to visualize phase changes, coupled with a novel 3D particle-based physics solver that supports phase-changing point cloud movement to generate optical flow and coarse video visualizations. In summary, our contributions are three-fold:

- We extend the range of action-based scenarios that can be realistically simulated from a static image to include complex material phase changes.
- We propose a novel 3D physics engine designed to mimic particle movement for general objects undergoing phase-transition simulations.
- We design a coupled schema that integrates physics engine outputs with AI reasoning to provide reliable, high-fidelity control over the final video generation.

2 Related Work

Physically Plausible Visual Content Generation. The generation of physically plausible videos provides immense value for real-world applications such as scientific simulation Sun et al. (2024a) and robotics Bar et al. (2025), yet conventional graphics pipelines remain limited by their dependence on complex, rule-based simulation systems Nealen et al. (2006). While recent research has explored dynamic simulations in image space or integrating physical priors into 3D representations Liu et al. (2024a); Xie et al. (2024), these solver-based approaches often lack generalizability. Similarly, while some Video Diffusion Models (VDMs) incorporate physical grounding—such as modeling oscillations in the frequency domain or using simulators like PhysDiff Yuan et al. (2023) — these methods typically focus on narrow, specific types of motion. In contrast, our work introduces a unified framework that strictly follows physical laws through a coupled AI and 3D physics solver, enabling the synthesis of complex scenarios.

Action-conditioned Dynamic Scene Generation. Early research in action-conditioned generation focused on modal bases for vibrating objects, which remains limited to vibration-like movements Yuan et al. (2023). Recent 3D physics-based methods have emerged, but they typically demand complete 3D geometric reconstructions from complex multi-view captures Chen et al. (2022; 2025); Ren et al. (2023); Wang et al. (2024a). While models like PhysMotion Tan et al. (2024) combine a single-image input with a physics solver, they are restricted to rigid and elastic dynamics because the video generator is only used for appearance refinement. WonderPlay Li et al. (2025b) improves upon this by using both a solver and a generator for dynamics. We further distinguish our work by introducing a novel 3D particle-based physics engine that specifically mimics material phase transitions, allowing for action-based simulation of phenomena like melting and freezing that remain unsupported by current solvers.

Controllable Video Generation. While models like Sora Liu et al. (2024b) generate diverse physical phenomena, they are primarily conditioned on text or images and lack direct controllability for general actions Singer et al. (2022). Substantial research has addressed camera and motion control—such as drag-based Li et al. (2025a) or trajectory-based approaches Mou et al. (2024); Niu et al. (2024); Wang et al. (2024c) — but these typically require the resulting dynamics of an action as input rather than the action itself Bahmani et al. (2024). Furthermore, many physical actions involving fluids or thermal effects are nearly impossible to represent through simple trajectory signals Sun et al. (2024b); Zhang et al. (2025). PhasePlay overcomes these limitations by taking physics-based 3D actions as direct input, utilizing a bimodal synthesis pipeline to control both the motion and the visual evolution of the material state.

World Models. Parallel to the advancement of video models, there has been increasing interest in interactive world models Brooks et al. (2024); Zhao et al. (2024) that recurrently generate states based on prior actions Bruce et al. (2024); Watson et al. (2024). Historically, this research has been concentrated on video game environments due to data availability Liu et al. (2025); Che et al. (2024), leaving real-world applications restricted to camera control or text-based interactions Zhao et al. (2023). Our work shifts this focus toward physics-based actions within realistic environments by providing a framework where force field-driven interactions and thermal-induced phase changes are handled in a unified, physically-consistent manner, offering a more robust simulator for real-world physics than gaming-centric models.

Phase Change Simulation. Simulating multi-phase transitions—such as melting, solidification, and fluid-solid dynamics—has evolved from traditional sharp-interface tracking Stomakhin et al. (2014) to robust Eulerian-Lagrangian Tu et al. (2024) and diffuse-interface frameworks Gao et al. (2017). While early Lagrangian meshes and Eulerian level sets struggled with complex topology and large deformations Wojtan et al. (2009); Terzopoulos et al. (1987), hybrid techniques like the Material Point Method (MPM) effectively handle large deformations and heat transport Stomakhin et al. (2014); Tu et al. (2024). Modern state-of-the-art approaches further unify these paradigms by embedding implicit phase-field evolution within background grids Yang et al. (2017); Gao et al. (2017); Wang et al. (2024b). Despite these advancements, existing methods still lack a fully unified framework for simultaneous gas-solid-liquid transformations, typically relying on separate, disjointed solvers and incurring prohibitive computational costs for high-precision physical accuracy. In contrast, tailored for visual-centric 3D scene generation, PhasePlay utilizes a unified triple-phase solver with lightweight physics approximations to efficiently model all six phase transitions within a single

framework, delivering the computational speed and sufficient accuracy required for realistic video generation.

3 PHASEPLAY

3.1 PIPELINE OVERVIEW

Objective. To synthesize action-conditioned 3D scene dynamics with a primary focus on material phase transitions, we define a framework that transforms a single input image I into a dynamic sequence of states $\{S_t\}_{t=0}^{T}$ over T simulation steps. Unlike traditional models that focus solely on spatial trajectories, our formulation treats each state S_t as a bimodal entity encompassing both geometric coordinates and thermal-dependent material properties. We model the input actions as a combination of mechanical forces—including gravity $\mathbf{f}_g$, spatial force fields $\mathbf{f}_w$, and localized point forces $\mathbf{f}_p(t)$—and thermal interactions $\Delta\mathcal{T}$ that trigger the phase-change mechanism within our novel 3D particle-based physics solver. The core objective is to derive a transition function that simultaneously computes the physical movement of the point cloud and the evolution of its visual appearance (RGB), ensuring that the final video synthesis is strictly constrained by both motion consistency and physically-grounded phase-change priors.

Overview. The illustration of the pipeline is shown in Figure 2. The generation pipeline begins with a Scene Generator that separates a single input image into object and background components to establish the initial 3D scene Ke et al. (2024); Wu et al. (2024). Simultaneously, an AI-based Physics Reasoning module uses a Vision-Language Model (VLM) to analyze text prompts Zhang et al. (2024a), such as "Wood pile on fire," to predict the final visual state and necessary phase changes between solid, liquid, and gas states. These predictions guide a Physics Simulator, which utilizes techniques like Material Point Method (MPM) and Position Based Dynamics (PBD) with specialized scheduling to compute physically accurate particle movement. A Gaussian Temporal Updater then maps these particle locations onto a Gaussian representation using optimal transport to track temporal trajectories. Finally, the Video Generator integrates the rendered RGB data and optical flow through Go-with-the-Flow Burgert et al. (2025) and SDEdit Meng et al. (2021) modules to synthesize the final physically consistent video sequence.

3.2 PIPELINE DETAILS

3D scene generation We represent the 3D scene at timestep t as $S_t = B_t \cup O_t$, where the initial state S_0 is formed by reconstructing the background B_0 and dynamic objects O_0 separately from the input image I. The background $B_t = \{p^B, q^B, s^B, o^B, c_t^B\}$ is modeled using N_B Gaussian surfels that act as static boundaries during simulation. Dynamic entities are represented as topological Gaussian surfels Dai et al. (2024) $O_t = \{E, v_t^O, p_t^O, q_t^O, s_t^O, o_t^O, c_t^O\}$, which extend standard surfels by incorporating a velocity vector v_t and a connectivity matrix $E \in \{0,1\}^{N_O \times N_O}$ derived from a mesh generated via InstantMesh Xu et al. (2024). Each object is further assigned material properties m according to the phase of the material: rigid, liquid, and smoke, enabling the physics simulator to apply material-specific constraints and handle the complex particle movements required for realistic scene dynamics.

AI-based physical reasoning We employ ChatGPT-image-1.5 Roumeliotis & Tselikas (2023) to perform high-level semantic analysis of the input image and text prompt. It predicts the final physical state visualization and identifies the specific material phase changes. This module provides the necessary guidance for the physics solver, and appearance of the phase change to create coarse rendered videos.

Approximated unified triple-phase physics simulator. The Physics Simulator advances the scene state by solving for dynamics attributes at each simulation step. Given the current scene representation $\tilde{S}_t$ and the set of user action, including force field $\mathbf{F}(t)$ and temperature field $\mathbf{T}(t)$, the solver computes the next state velocity $\mathbf{v}_{t+1}$, position $\mathbf{p}_{t+1}^O$, and orientation $\mathbf{q}_{t+1}^O$:

$$\mathbf{v}_{t+1}^O, \mathbf{p}_{t+1}^O, \mathbf{q}_{t+1}^O = \text{Solver}(\tilde{S}_t, \mathbf{F}(t), \mathbf{T}(t)). \tag{1}$$

The coarse scene $\tilde{S}_{t+1}$ is then constructed by updating these dynamic attributes while maintaining the initial non-dynamics properties such as scales $\mathbf{s}_0^O$, opacities $\mathbf{o}_0^O$, and colors $\mathbf{c}_0^O$ from S_0. To handle

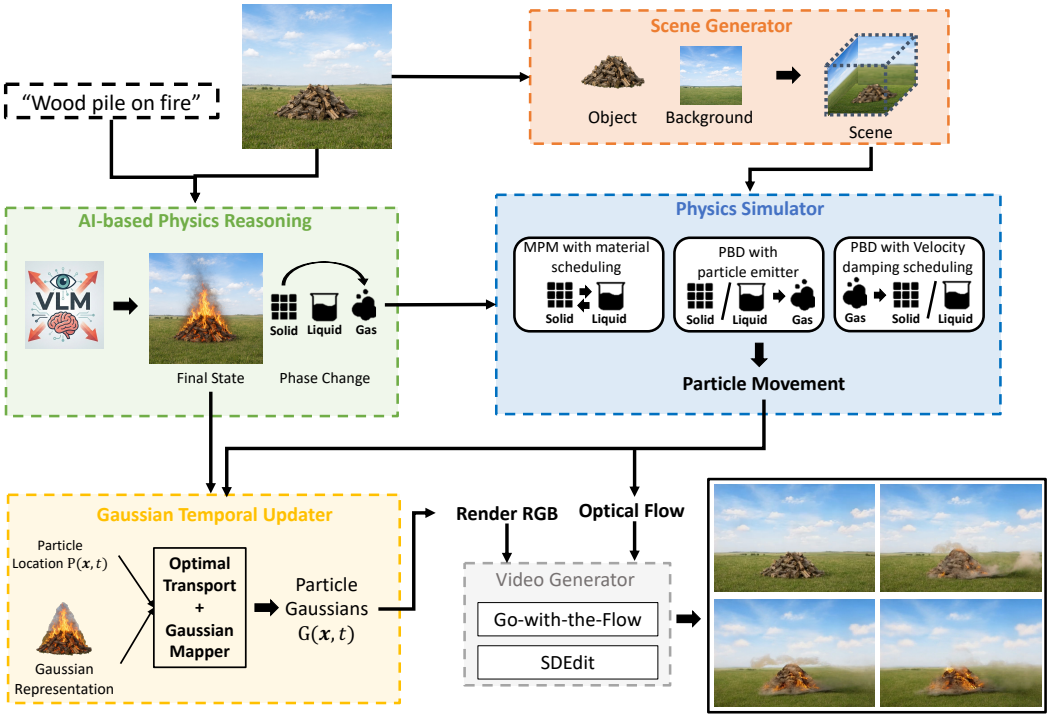


Figure 2: Overview of the PhasePlay framework. Given a single input image, PhasePlay reconstructs the 3D scene and extracts objects of interest. A VLM jointly reasons over the image and user prompts to generate an image of the desired material phase (e.g., melted, vaporized). To model temporal evolution, a novel particle-based physics simulator captures particle trajectories based on user actions and phase change types, and computes corresponding optical flow fields. A Gaussian temporal updater assigns and updates particle appearance based on the Gaussian representation of the desired state, enabling visually realistic and temporally coherent video synthesis. Finally, the dynamic 3D scene is refined using the synthesized video and optical flows to ensure consistency and realism.

complex, multi-physics environments involving rigid bodies, fluids, or gases, we employ a coupled solver framework utilizing both Material Point Method (MPM) and Position Based Dynamics (PBD) with specialized scheduling to ensure physical consistency across diverse material interactions.

Recognizing that standard physics engines often lack the native capability to explicitly resolve the thermodynamics of phase transitions, we adopt a phenomenological approach to approximate these transformations. Specifically, we implement three specialized mechanisms that govern the temporal evolution of particle positions $\mathbf{p}_t$ and velocities $\mathbf{v}_t$ by dynamically modulating physical constraints and material constitutive properties.

- **Fluid-Solid Transitions:** For transformations such as solidification and melting, the primary macroscopic characteristics are captured by modulating material stiffness and kinetic energy dissipation. Within our Material Point Method (MPM) framework, we define a time-varying Young's modulus $E(t)$ via a scheduling function $\mathcal{S}_E$, such that $E(t) = E_0 \cdot \mathcal{S}_E(t)$. This allows the material to transition across a spectrum of elasticities, providing a robust approximation of phase-induced structural changes. We employ an exponential scheduler $\mathcal{S}_E = e^{\lambda t}$, where the user-defined parameter λ dictates the transition rate: $\lambda > 0$ simulates rapid solidification, while $\lambda < 0$ models the liquefaction process.
- **Fluid/Solid-to-Air Transitions:** To represent the transition from condensed phases to a gaseous state, we treat the source geometry as a volumetric or surface-based emitter. We model the generation of gas particles $\mathcal{P}_{gas}$ from the boundary of a source object O. Newly instantiated particles are assigned an initial spawn velocity $\mathbf{v}_{spawn} = \alpha \cdot \mathbf{n}$, where $\mathbf{n}$

denotes the surface normal and α represents the emission scalar. The mass flux is governed by the particle injection rate per timestep, effectively transitioning the continuous mesh representation into a diffuse, stochastic gaseous state.

- **Air-to-Fluid/Solid Transitions:** A defining visual distinction between gaseous and condensed phases is the degree of internal turbulent fluctuation. To simulate transitions toward more ordered states, we implement a relative velocity damping mechanism. By activating particle fluctuation control, the system suppresses stochastic turbulent motion relative to the aggregate's barycentric velocity $\bar{v} = \frac{1}{N}\sum_{j=1}^{N} v_j$. This fosters coherent motion by updating the velocity of each individual surfel i according to:

$$v_i^{new} = \bar{v} + (1 - \eta\Delta t)(v_i - \bar{v}) \tag{2}$$

where η is the user-defined damping coefficient and Δt is the simulation time step. This formulation ensures that internal "wobble" decays as the relative velocity $(v_i - \bar{v}) \to 0$, forcing the aggregate mass to move as a unified, stabilized volume while strictly preserving global momentum.

Gaussian Temporal Updater: To maintain high-fidelity appearance during dynamic transitions, we implement a dual-anchor assignment strategy that maps simulation particles to two boundary Gaussian representations: the initial state $\mathcal{G}_{init}$ and the predicted final state $\mathcal{G}_{final}$. At specific synchronization intervals $t = K\Delta t$, we establish a one-to-one correspondence by solving a discrete Optimal Transport (OT) problem. Let $\mathcal{P}_t = \{\mathbf{x}_i\}_{i=1}^{N}$ be the simulated point cloud and $\mathcal{P}_G = \{\mathbf{y}_j\}_{j=1}^{N}$ be the target Gaussian centers. We seek a permutation matrix $\pi \in \Pi(N)$ that minimizes the total quadratic transport cost:

$$\min_{\pi \in \Pi(N)} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{y}_{\pi(i)}\|^2 \tag{3}$$

This assignment ensures that each simulation particle i is uniquely associated with a specific Gaussian identity from both boundary states. To mitigate the computational overhead of the OT solver, the permutation π is updated periodically every K steps, while attributes for intermediate timesteps $m \in [0, M]$ are determined via linear interpolation. Specifically, the appearance attributes $\mathcal{A}_t$ (comprising orientation $\mathbf{q}$, scale $\mathbf{s}$, opacity o, and color $\mathbf{c}$) are computed as a weighted sum:

$$\mathcal{A}_t(i) = \left(1 - \frac{m}{M}\right)\mathcal{A}_{init}(\pi(i)) + \left(\frac{m}{M}\right)\mathcal{A}_{final}(\pi(i)) \tag{4}$$

Furthermore, to handle particle volatility—such as the recycling, expiration, or emission of particles where the underlying topology is non-consistent—we utilize a distance-weighted spatial interpolation. For any newly spawned particle p^* that lacks a direct OT mapping from the previous synchronization step, its attributes are estimated from its k-nearest neighbors $\mathcal{N}(p^*)$ that possess valid mappings:

$$\mathcal{A}_t(p^*) = \frac{\sum_{j \in \mathcal{N}(p^*)} w_j \mathcal{A}_t(j)}{\sum w_j}, \quad w_j = \exp\left(-\frac{\|\mathbf{x}_{p^*} - \mathbf{x}_j\|^2}{\sigma^2}\right) \tag{5}$$

This formulation ensures a spatially continuous and temporally smooth transition of the 3D scene representation, effectively bridging the discrete physical simulation with the continuous Gaussian field even during complex phase changes. By applying this mapping logic to both initial state and final state, each particle i at timestep m effectively references attributes from both $\mathcal{G}_{init}$ and $\mathcal{G}_{final}$. The final rendered Gaussian attributes $\mathcal{A}_t$ are then computed as a temporal weighted sum:

$$\mathcal{A}_t(i) = \left(1 - \frac{m}{M}\right)\mathcal{A}_{init}(\pi_{init}(i)) + \left(\frac{m}{M}\right)\mathcal{A}_{final}(\pi_{final}(i)) \tag{6}$$

where M is the total simulation span. This formulation ensures a spatially continuous and temporally smooth transition of the 3D scene representation, effectively bridging the discrete physical simulation with a continuous Gaussian field that evolves in both geometry and appearance.

Video Generator. The video generation module g synthesizes the final high-fidelity video V by integrating two control modalities: motion, represented by the pixel-space flow F rendered from particle velocities, and appearance, represented by the coarse video $\tilde{V}$. The process utilizes a pre-trained diffusion model where motion is enforced through structured noise $N(F)$ via noise warping, and appearance is refined using SDEdit by initiating denoising at a specific timestep s_1. To maintain background stability and prevent hallucinations, a spatially varying bimodal control is applied using a binary mask M, following the implementation of WonderPlay Li et al. (2025b). For a comprehensive breakdown of the diffusion coefficients and the photometric optimization process, please refer to WonderPlay for more details.

4 EXPERIMENT

4.1 IMPLEMENTATION

We implement our framework using the Genesis Authors (2024) framework to unify various physics solvers. We compare PhasePlay against four baselines: a) Sora-2 Liu et al. (2024b): the second-generation video foundation model developed by OpenAI; b) Google Flow Masalkhi et al. (2024): the video generation model developed by Google with flexible control, combined with the final state prediction as additional inputs; c) CogVideoX-I2V Yang et al. (2024): an open-source text-conditioned video generation model; d) VLIPP Yang et al. (2025): a two-stage image-to-video generation framework that explicitly incorporates physics with vision and language informed physical prior; e) WonderPlay Li et al. (2025b): A framework integrating physics simulation with video generation for generating action conditioned dynamic 3D scenes from a single image. All baseline methods are evaluated using only an image and a text prompt as input, ensuring a fair and consistent comparison across the different generative and simulation-based approaches. All the computation is carried out in single node of H100 GPU. The details of the simulation setup for each example is documented in Appendix A and Appendix B.

Quantitative evaluation is conducted across 10 diverse scenarios encompassing various phase transitions, including melting, solidification, evaporation, deposition, sublimation, and combustion. To assess these outputs, we utilize the VBench metric suite Huang et al. (2024) to evaluate imaging quality, aesthetic appeal, motion fidelity, and temporal consistency. To complement these automated scores, we implement a GPT-4o-based physical realism metric to assess the conceptual accuracy of the simulated phenomena. Finally, we conduct a human preference study involving 25 participants to evaluate perceptual quality and overall satisfaction. In a multi-model evaluation, we report the percentage of evaluators who selected our model as the best among all competing baselines.

4.2 MAIN RESULTS

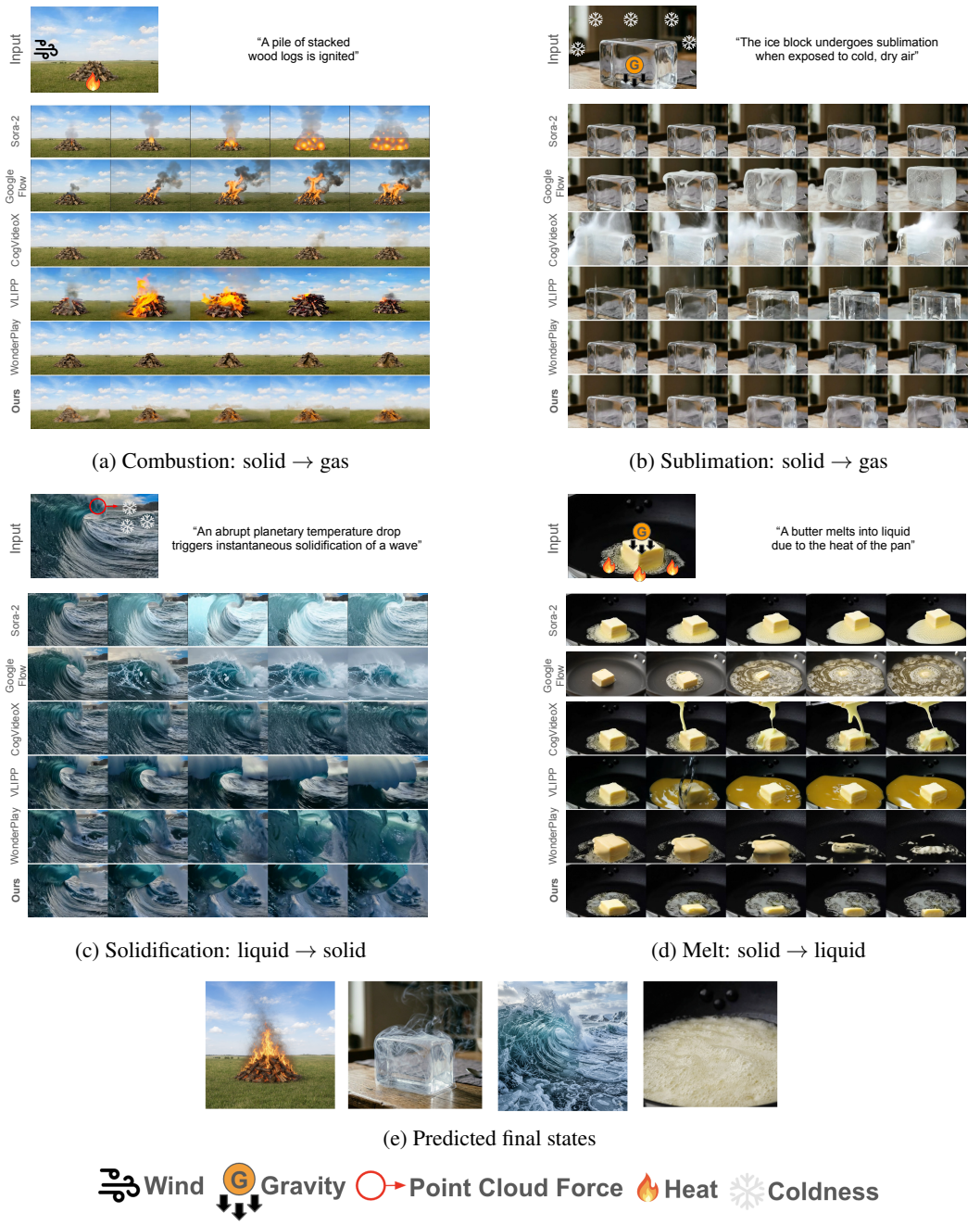
The quantitative results presented in Table 1. highlight the significant advantages of the PhasePlay framework. While generative baselines achieve competitive scores in imaging and motion, they lack the underlying particle-level solvers required for high physical fidelity. Specifically, our model achieves a significant improvement in physical realism (PhysReal), outperforming the strongest baseline by over 40% in the score. Furthermore, our generated results achieved the highest user preference rate across all evaluated baselines.

Table 1: Quantitative comparison to baselines.

Methods	Imaging (↑)	Motion (↑)	Consistency (↑)	PhysReal (↑)	UserPrefer (↑)
Sora-2	0.692	0.987	0.201	0.347	0.12
Google-Flow	0.694	0.995	**0.218**	0.617	0.24
CogVideoX	0.639	0.990	0.206	0.475	0.00
VLIPP	0.638	0.992	0.211	0.498	0.08
WonderPlay	0.647	0.994	0.214	0.542	0.08
Ours	**0.695**	**0.996**	0.217	**0.768**	**0.48**

As illustrated in Figure 3, PhasePlay demonstrates a significant advancement over existing models by successfully synthesizing the complex morphological transformations inherent in phase changes while maintaining rigorous physical grounding.

In the case of combustion, our method generates a fluid, expanding fire front where the increasing volume of fire is coupled with the stochastic emission of new particles. In contrast, Sora-2 generates physically inconsistent fire frames, CogVideoX-I2V is limited to producing sparse smoke effects, and WonderPlay only captures subtle deformations in the source geometry (e.g., the wood pipe) without rendering the combustion process itself. This highlights a fundamental limitation of existing methods that lack an AI-reasoning module dedicated to synthesizing the visual appearance of newly emerged phases. While Sora-2 and VLIPP leverage large-scale pre-trained priors for visualization, their outputs often lack physical grounding. This is further evidenced in the sublimation case, where our method correctly renders newly emitted gas particles around the ice, providing a far more convincing visual transition.



(a) Combustion: solid → gas

(b) Sublimation: solid → gas

(c) Solidification: liquid → solid

(d) Melt: solid → liquid

(e) Predicted final states

Figure 3: **Qualitative results** of the proposed PhasePlay. Different scenes of coupled user interaction and phase change are presented: combustion (top left), sublimation (top right), solidification (bottom left), and melt (bottom right).

Regarding liquid-to-solid transformations, our framework is the only method capable of rendering a realistic ice texture while ensuring that the aggregate motion strictly adheres to physical laws. While most baselines—with the exception of Sora-2—can produce plausible wave motion, they universally fail to capture the realistic transition of a liquid wave freezing into a solid structure. Conversely, in the melting scenario, PhasePlay successfully synthesizes the material softening and appearance change of melting butter. Among the baselines, while WonderPlay maintains physical plausibility, the appearance of the butter remains statically "solid," resulting in a visual disconnect.

Collectively, these comparisons demonstrate that our approach—anchored by AI-driven appearance reasoning and Gaussian Splatting—is essential for realistic phase-change visualization. By guiding the 3D scene representation with explicit physical constraints (e.g., gravity, thermal gradients), we ensure the resulting sequences adhere to scientific logic, enabling the high-fidelity replication of complex phenomena like instant wave solidification or ice sublimation that purely data-driven models cannot reliably reconstruct. Furthermore, comparisons with Google Flow indicate that providing the final state as an additional input is insufficient on its own; because our framework extracts appearance references strictly for visual guidance rather than physics prediction, the inclusion of a rigorous physics solver remains critical to accurately "interpolate" the transitional dynamics between the initial and final states. To further demonstrate our model's capabilities, we visualize predictions of unusual physical phenomena in Appendix C, along with failure cases under extreme scenarios in Appendix D.

4.3 Ablation studies

Our analysis of the framework's hyperparameters reveals two critical factors influencing generation quality:

- **Synchronization Interval (K):** This parameter controls the trade-off between visual fidelity and speed. Smaller values of K (e.g., $K = 10$) ensure smooth transitions in motion and appearance but require more computation (up to 10 minutes). Larger intervals (e.g., $K = 250$) are significantly faster ($\sim$10 seconds) but result in sparse control signals, causing the model to lose track of evolving phases and bias toward the initial state.
- **Simulation Time Step (dt):** The precision of the video guidance depends heavily on the time step. Finer steps ($dt = 10^{-5}$) are essential for maintaining structural consistency. Coarser steps ($dt \geq 10^{-4}$) lead to rapid particle displacement, which degrades the Optimal Transport (OT) mapping accuracy and introduces visual artifacts due to topological divergence.

For a comprehensive breakdown of these trade-offs, qualitative comparisons, and additional performance metrics, please refer to the Appendix E.

5 Conclusion

We presented PhasePlay, a unified framework for action-conditioned 3D scene dynamics that bridges the gap between physics-based simulation and generative visual synthesis. By integrating our novel physics solvers with AI-driven physical reasoning, PhasePlay enables realistic simulation of complex material phase transitions—such as melting, freezing, and evaporation—from a single image. Our Gaussian Temporal Updater ensures that visual appearances evolve in synchrony with physical trajectories. Experimental results demonstrate that PhasePlay significantly outperforms current state-of-the-art models in physical plausibility while maintaining high aesthetic quality.

Limitations. While PhasePlay successfully models diverse and complex phase transitions, several avenues remain open for future improvement. First, computational efficiency can be further optimized by developing a more advanced Gaussian temporal updater to generate the reference video. Second, our current "phenomenological" approach to thermodynamics relies on approximating temperature effects rather than solving full, rigorous heat transfer equations; this approximation may introduce inaccuracies in scenarios where precise, localized thermal gradients are critical, suggesting that a more sophisticated approximated physics simulator with similar computation cost will be useful in this framework. Third, expanding the scope of supported physical phenomena represents a major objective, with the primary current limitation being the exclusion of fracture physics within our simulation framework. Finally, maintaining robust performance and physical plausibility when handling unusual physical phenomena or out-of-distribution, unreal scene predictions—such as those required for cinematic content generation—remains an important challenge for future exploration.

Broader Impact. PhasePlay has the potential to significantly advance training environments for embodied AI and robotics by providing physically grounded synthetic data for rare material interactions. In the creative industries, it lowers the barrier for high-fidelity VFX and AR/VR content creation. However, as with all high-fidelity generative models, there is a risk of misuse in creating deceptive "deepfake" physical evidence or highly realistic misinformation.

AI USE STATEMENT

In this work, we have not used generative AI tools for research ideation, literature review, data collection, or core scientific content generation. We used generative AI tools solely to polish and refine the English phrasing and wording of the text. We have reviewed all AI-assisted work, ensuring that every sentence accurately reflects our original ideas and scientific claims. We checked the polished text to preserve technical precision and correctness. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

In this work, we introduce PhasePlay, a framework designed to synthesize action-conditioned 3D scene dynamics and material phase transitions from a single image. While our primary motivation is to advance physically grounded simulation for applications such as robotic training environments, virtual reality, and scientific visualization, we recognize that high-fidelity video generation models can raise ethical concerns regarding potential misuse—specifically the creation of deceptive visual media or misinformation.

To mitigate these risks, we have evaluated the broader societal impact of our framework. PhasePlay is explicitly structured around physics-driven reasoning and solver-backed constraints, which inherently tie outputs to scientific laws rather than unconstrained synthetic generation. Furthermore, we strongly advocate for responsible deployment, robust provenance tracking, and watermarking standards within the generative video community to prevent misuse. We have complied with all applicable research integrity guidelines, and we remain committed to ensuring that our methodologies are transparently documented to support safe, reliable applications in creative and technical industries.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we have made all source codes and sample data publicly accessible via our anonymous repository at https://anonymous.4open.science/r/PhasePlay-388A. Detailed specifications of the physics simulation settings, parameters, and hyperparameters for each evaluated scenario are comprehensively documented in Appendix A. Furthermore, video refinement and implementation configurations are provided in Appendix B, and a thorough analysis of hyperparameter trade-offs—including the synchronization interval K and simulation time step dt-is detailed in Appendix E.

REFERENCES

NVIDIA: Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/Genesis.

Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pp. 53–72. Springer, 2024.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13–23, 2025.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024.

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

Duowen Chen, Zhiqi Li, Junwei Zhou, Fan Feng, Tao Du, and Bo Zhu. Solid-fluid interaction on particle flow maps. *ACM Transactions on Graphics (TOG)*, 43(6):1–20, 2024.

Hsiao-yu Chen, Edith Tretschk, Tuur Stuyck, Petr Kadlecek, Ladislav Kavan, Etienne Vouga, and Christoph Lassner. Virtual elastic objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15827–15837, 2022.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10850–10869, 2023.

Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

Yang Gao, Shuai Li, Lipeng Yang, Hong Qin, and Aimin Hao. An efficient heat-based model for solid-liquid-gas phase transition and dynamic interaction. *Graphical Models*, 94:14–24, 2017.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1–12, 2025.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *European Conference on Computer Vision*, pp. 205–224. Springer, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5031–5038, 2025a.

Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9080–9090, 2025b.

Jinxiu Liu, Shaoheng Lin, Yinxiao Li, and Ming-Hsuan Yang. Dynamicscaler: Seamless and scalable video generation for panoramic scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6144–6153, 2025.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024a.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024b.

Mouayad Masalkhi, Joshua Ong, Ethan Waisberg, and Andrew G Lee. Google deepmind's gemini ai versus chatgpt: a comparative analysis in ophthalmology. *Eye*, 38(8):1412–1417, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026.

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505, 2024.

Andrew Nealen, Matthias Müller, Richard Keiser, Eddy Boxerman, and Mark Carlson. Physically based deformable models in computer graphics. In *Computer graphics forum*, volume 25, pp. 809–836. Wiley Online Library, 2006.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European conference on computer vision*, pp. 111–128. Springer, 2024.

M Nomaan Qureshi, Sparsh Garg, Francisco Yandun, David Held, George Kantor, and Abhisesh Silwal. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6502–6509. IEEE, 2025.

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023.

Alexey Stomakhin, Craig Schroeder, Chenfanfu Jiang, Lawrence Chai, Joseph Teran, and Andrew Selle. Augmented mpm for phase-change and varied materials. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014.

Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: biomedical generalist video generation model. *arXiv preprint arXiv:2407.08944*, 2024a.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024b.

Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. Physmotion: Physics-grounded dynamics from a single image. *arXiv preprint arXiv:2411.17189*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Demetri Terzopoulos, John Platt, Alan Barr, and Kurt Fleischer. Elastically deformable models. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pp. 205–214, 1987.

Zaili Tu, Chen Li, Zipeng Zhao, Long Liu, Chenhui Wang, Changbo Wang, and Hong Qin. A unified mpm framework supporting phase-field models and elastic-viscoplastic phase transition. *ACM Transactions on Graphics*, 43(2):1–19, 2024.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024a.

Xiaokun Wang, Yanrui Xu, Sinuo Liu, Bo Ren, Jirí Kosinka, Alexandru C Telea, Jiamin Wang, Chongming Song, Jian Chang, Chenfeng Li, et al. Physics-based fluid simulation in computer graphics: Survey, research trends, and challenges. *Computational Visual Media*, 10(5):803–858, 2024b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024c.

Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024.

Chris Wojtan, Nils Thürey, Markus Gross, and Greg Turk. Deforming meshes that split and merge. In *ACM SIGGRAPH 2009 papers*, pp. 1–10. 2009.

Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18826–18836, 2025.

Tao Yang, Jian Chang, Ming C Lin, Ralph R Martin, Jian J Zhang, and Shi-Min Hu. A unified particle system framework for multi-phase, multi-material visual simulations. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12360–12370, 2025.

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

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16010–16021, 2023.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024a.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pp. 388–406. Springer, 2024b.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2063–2073, 2025.

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *arXiv preprint arXiv:2411.02319*, 2024.

A PHYSICS SIMUALTION SETTING

desublimation example uses a particle gas that is initially very light and moderately compressible (mass-density setting 0.03, particle radius 0.018, density correction strength 0.2, viscosity damping 0.01, upward lift +0.16), then transitions to a falling solid-like cloud through a staged schedule: lift is canceled halfway (0.5), reversed after halfway (0.5) with immediate gravity, and density increasing from 0.03 to 10 over time. The solver runs 4 constraint iterations per step for 1000 steps at timestep 0.001 with 16 internal substeps; collision responses are disabled for force terms but enabled as hard constraints against both environment and objects, and motion is clipped by a floor at y = -0.4.

Fire example uses a rigid source that continuously emits fluid-like particles: 8000 particles are seeded near the surface with tiny offset 3e-7, each particle has low density scale 0.01, radius 0.022, zero buoyant lift, and a small initial upward velocity of 0.01 (no outward radial kick). Emission continues during simulation by respawning 400 particles per step starting at step 1. Visual intensity is ramped with a cap of 0.92, time exponent 1.0, particle-density gain 1.2, and EMA smoothing 0.88; edge blending band is dilate 15, erode 5, blend strength 1.0. Dynamics run for 1000 steps, 4 solver iterations each, timestep 1e-5; all collision force and constraint toggles are off, with a floor at y = -0.4.

Butter example is simulated as elastoplastic material (configured as two elastic bodies) with stiffness 1.0e6, Poisson ratio 0.25, material density 1100.0, lower/upper yield thresholds 0.005 and 0.001, von Mises yield stress 200.0, grid resolution 96, and particle radius 0.008. To mimic melting/softening, stiffness is linearly reduced from 1.0e6 to 1.0e2 between steps 0 and 60 for object 0. The run uses 250 steps, 4 iterations per step, timestep 0.0009; collision forces are off but hard collision constraints are on for environment and objects, and the same floor bound y = -0.4 is enforced.

Wave example starts as flowing liquid and then freezes progressively: density scale 1.0, particle radius 0.014, softer incompressibility correction 0.18, low viscosity 0.002, and explicit downward gravity -4.0 on vertical axis. Transition sharpness is sigmoid with steepness 7.0, max velocity damping 0.90, and max lock strength 0.82. Numerical integration uses 80 steps, 4 iterations per step, very small timestep 1e-7, plus 16 substeps for stability. Collision forces stay off while geometric constraints stay on; debug thresholds flag near-static behavior at vertical-vertex displacement below 1e-4 and particle displacement below 1e-6 after step 120. Floor bound is y = -0.4.

Ice sublimation example also uses rigid-surface emission of fluid-like particles, but with heavier setting than fire: 8000 emitted particles, surface offset 3e-7, density scale 1.0, particle radius 0.022, zero buoyancy, upward initial speed 0.01, no outward burst, and continuous respawn of 400 particles per step starting from step 1. Simulation runs 80 steps with 4 solver iterations per step and timestep 0.0001; all collision forces and hard collision constraints are disabled, and objects are still clamped above floor y = -0.4.

B VIDEO REFINEMENT DETAILS

In the video-refinement stage, we take the coarse physics trajectory and turn it into motion-guided latent noise, then run a diffusion video generator to sharpen realism while staying consistent with simulated flow: we crop each frame starting 120 pixels from the top, standardize the clip to 49 frames at 720x480, and generate latent motion noise with 16 channels; when optical-flow arrays are available, we keep frame-rate scaling at 1x, use flow scaling at 4x, and use an effective latent scale of 32, while the fallback video-input path uses 0.5x frame scaling and 8x flow scaling. The generative model is executed in memory-saving mode and denoises for 25 steps with corruption strength 0.4, and an edit-strength sweep over 0.5 / 0.6 / 0.7 / 0.8 (or a user-selected subset), with mask-specific editing disabled in this setup. Initial-frame anchoring is enabled, and the anchor frame is taken from the first frame of the simulated input video.

C OUT-OF-DISTRIBUTION SCENARIOS

Desublimation—the direct transition of water vapor into solid ice—represents a significant "out-of-distribution" (OOD) challenge, as this phenomenon is rarely recorded in natural datasets or synthesized in standard computer graphics. We omit pre-trained AI baseline comparisons using final-

frame conditioning because we found that the extreme nature of these out-of-distribution physical phenomena makes predicting an accurate final state infeasible without prior domain knowledge, thereby degrading the realism of the model's predictions. Consequently, pre-trained generative models struggle to generalize to this rare event. As shown in Figure 4, while our model does not yet produce a perfectly photorealistic simulation of this complex process, our results clearly demonstrate a physically grounded transition from a fluctuating gaseous state to a unified, solid structure, complete with a distinct change in material appearance. In contrast, baseline methods typically suffer from physical hallucinations, often incorrectly rendering the vapor as smoke that turns into liquid water. These baselines fail to capture any meaningful change in the material's underlying behavior or visual properties, highlighting the necessity of our physics-reasoning approach for handling rare physical transitions.

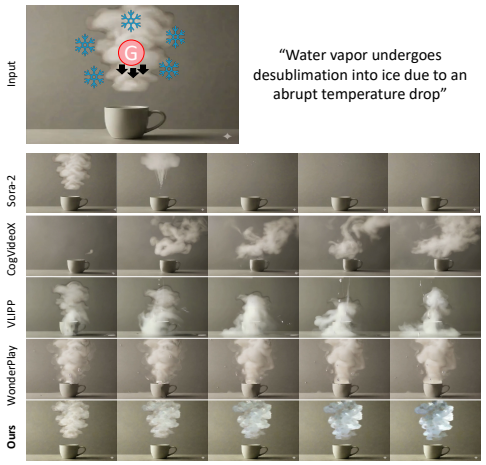


Figure 4: **Qualitative results** of the proposed PhasePlay for out-of-distribution scenarios.

D Failure Scenarios

In this section, we examine failure examples of our proposed framework to better understand its performance limitations in Figure 5.

A primary failure mode involves scenarios requiring material fracture, as demonstrated in the mug cracking experiment. This limitation stems from the current constraints of our physics simulation module, which relies on continuous particle and field solvers and does not explicitly support fracture mechanics, topological splitting, or fragmentation dynamics. When stresses exceed standard deformation thresholds and result in structural rupture, the solver fails to capture the abrupt physical discontinuity, producing visual artifacts and implausible deformations in the synthesized output. Integrating advanced fracture mechanics models represents an important direction for future work.

Another failure category arises in complex thermal transitions, such as the ship-to-ash sequence. Although the physics-guided reference video establishes a plausible trajectory, the pretrained video diffusion model used for appearance refinement can struggle to maintain this physical fidelity when the initial visual boundaries or scene cues are ambiguous. Under these conditions, the generator falls back on unconstrained generative priors, leading to structural hallucinations, flickering, or temporal inconsistencies that deviate from rigorous thermodynamic principles.

E Ablation Study Details

Effect of Synchronization Interval K. The primary hyperparameter in our framework is the synchronization interval K, which dictates the balance between visual fidelity and computational efficiency. We evaluate our video generation pipeline across varying values of K. As illustrated in Figure 6, smaller values of K yield superior appearance transitions from the initial to the final

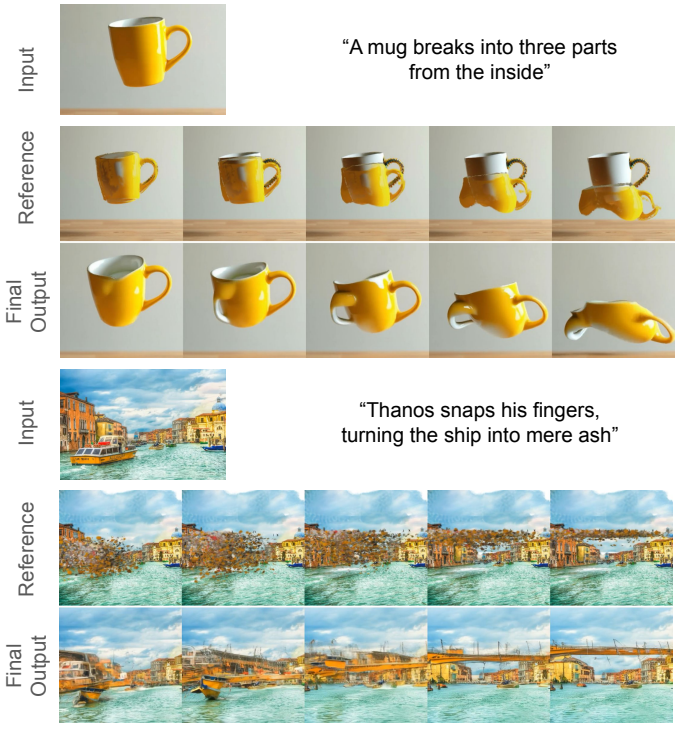


Figure 5: Failure cases of the proposed framework. First, material fracture scenarios (e.g., mug cracking) fail because continuous particle and field solvers lack explicit fracture mechanics for topological splitting. Second, complex thermal transitions (e.g., ship-to-ash) suffer from temporal inconsistencies and hallucinations when ambiguous visual boundaries cause the video diffusion model to rely on unconstrained priors.

state. Conversely, as K increases, the pipeline struggles to maintain smooth phase transitions in both appearance and motion. This occurs because the video refinement stage—leveraging the pre-trained priors from WonderPlay Li et al. (2025b)—relies heavily on the first frame of each segment for temporal anchoring. Consequently, when the RGB control signals are too sparse (high K), the pipeline tends to bias the generation toward the material's initial state, failing to capture the evolving phase. In this study, the total computational time for synchronization ranges from approximately 10 seconds for $K = 250$ up to 10 minutes for $K = 20$, representing the trade-off between speed and accuracy.

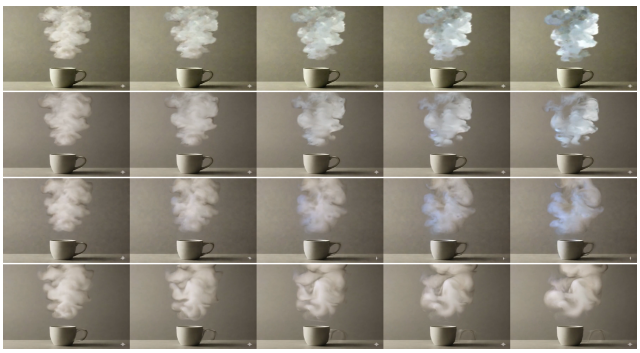


Figure 6: **Qualitative comparison** of various synchronization interval sizes K for Gaussian assignments. Smaller K values facilitate smoother transitions between the anchor states. For a simulation of 1,000 steps, we evaluate $K \in \{20, 50, 100, 250\}$, shown from the top to bottom rows respectively.

Sensitivity to Simulation Time Step dt. The quality of the final video is fundamentally governed by the consistency of the rendered RGB guidance. We analyze the framework's performance across different simulation time steps: the baseline $dt = 10^{-5}$, and coarser steps of 10^{-4} and 10^{-3}. As shown in Figure 7, our observations indicate that smaller time steps are significantly preferred; at larger dt, the generated frames exhibit reduced clarity and visual artifacts. We attribute this to the rapid displacement of particles between subsequent guidance frames, which results in significant structural divergence between point clouds. This divergence degrades the accuracy of the Optimal Transport (OT) mapping, as the solver cannot reliably establish correspondences when the underlying topology changes too drastically between intervals.



Figure 7: **Qualitative results** demonstrating the impact of different simulation time steps dt on video consistency and mapping accuracy. We evaluate the same scene with $dt = 10^{-5}$ (top), 10^{-4} (middle), and 10^{-3} (bottom).