PHYSMAS: PHYSICS-GROUNDED MULTI-AGENT SYNTHESIS OF COMPOSITIONAL 4D GAUSSIANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient, fully automatic, and physically plausible 4D Gaussian synthesis is an important goal for dynamic scene generation. Recent physics-based methods couple 3D Gaussians with the Material Point Method (MPM) to generate physically driven motion, but extending this paradigm to heterogeneous multi-part objects and interacting multi-object scenes remains challenging. Object-level physical assignment collapses distinct parts into a single material state, while one-shot predictions from large language models, vision-language models, or agents neither reliably bind different materials to identified parts nor verify that the resulting MPM configuration is executable. Score Distillation Sampling (SDS)-based parameter optimization, meanwhile, requires repeated per-scene score evaluations and gradient backpropagation, incurring lengthy optimization and potentially yielding suboptimal or unstable solutions. We therefore present PhysMAS, a physics-grounded multi-agent framework. From a motion prompt and four scene views, an Object-Part Scene Agent establishes persistent identities and calls a Material Reasoning Agent for part-wise profiles. It invokes solver-aware skills to bind these identities and profiles to per-particle MPM fields and execute all objects in a shared domain; the framework then screens candidate forward-simulation results. This supports heterogeneous multi-part and interacting multi-object scenes without per-scene diffusion-score backpropagation. Extensive experiments demonstrate that, compared with recent physics-based 4D Gaussian baselines that rely on SDS, PhysMAS achieves better semantic alignment and perceived physical plausibility while requiring less runtime.

1 INTRODUCTION

Efficient simulator-driven 4D Gaussian synthesis is an important goal in computer graphics and generative AI. 3D Gaussian Splatting (3DGS) provides an explicit representation for high-quality rendering (Kerbl et al., 2023), and physics-based methods couple its primitives with the Material Point Method (MPM) (Sulsky et al., 1994; Stomakhin et al., 2013) to generate physically driven motion (Xie et al., 2024). Yet many physics-based 4D Gaussian pipelines assign properties at the object or Gaussian-field level without an explicit hierarchy linking parts to their parent objects across perception and simulation. This is restrictive when objects interact or contain parts with different structural roles and material responses. In this work, *compositional* refers to scene state indexed jointly by object and part. Figure 1 illustrates the resulting dynamics.

This extension exposes two coupled problems. First, without persistent object–part identity, visual evidence cannot be bound consistently to the corresponding part and object. Object-wide assignment forces heterogeneous parts to share one material state, while co-occurring objects must remain distinct inside a shared simulation. Second, an LLM or VLM can propose labels and parameters, but a one-time prediction does not enforce the profiles and numerical constraints accepted by MPM. Alternatives based on Score Distillation Sampling (SDS) (Poole et al., 2023) optimize parameters with video priors, yet require repeated per-scene score evaluation and backpropagation (Huang et al., 2025; Lin et al., 2025). A practical system therefore requires compositional scene state and simulator-constrained execution without per-scene diffusion-score optimization.

We propose **PhysMAS**, a physics-grounded multi-agent framework that receives a motion prompt, four calibrated RGB views with their aligned static 3DGS, and prescribed output cameras. An **Object-**

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.

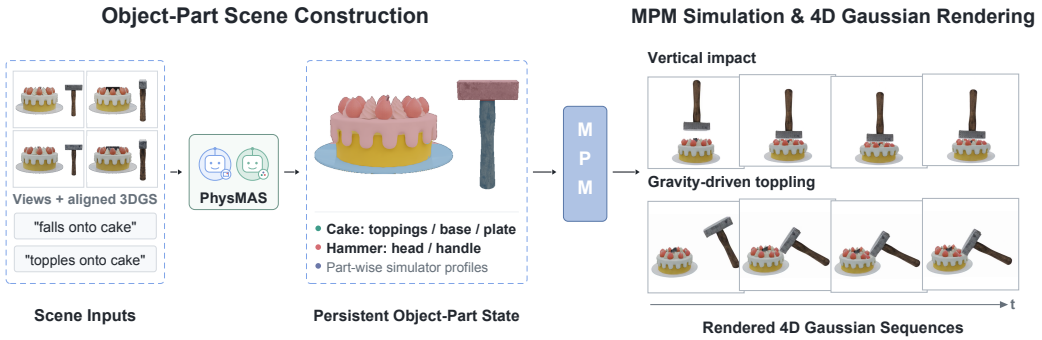

Figure 1: Overview of PhysMAS. Each run receives a motion prompt, four calibrated RGB views, their aligned static 3DGS, and prescribed output cameras. PhysMAS constructs persistent object–part state with part-wise simulator profiles; the compiled state is advanced in a shared MPM domain and rendered as a 4D Gaussian sequence. Shown are two runs of the same registered scene bundle: vertical impact and gravity-driven toppling.

Part Scene Agent grounds object instances and semantic parts, assigns compound object–part labels to the scene Gaussians, and calls a **Material Reasoning Agent** for part-wise simulator profiles and physical parameters. Deterministic, solver-aware **MPM skills** transfer these labels and profiles to particles, execute all objects in one domain, and return execution records for validity screening. Within a shared single-velocity grid, standard MPM updates coordinate momentum exchange among objects and support the multi-object interactions evaluated in our experiments. Thus, PhysMAS automates object–part grounding, material profiling, MPM compilation, simulation, and screening without per-scene SDS backpropagation.

In summary, our main contributions are as follows:

- We introduce persistent object–part state for representing heterogeneous parts and multiple objects in an aligned static 3DGS using four calibrated scene views.
- We formalize an auditable two-agent compiler contract around this typed state: the scene agent owns identity, planning, and acceptance; the material agent returns row-wise profile evidence; deterministic routines compile, execute, and screen part-wise candidates.
- Across the multi-part and multi-object evaluations, PhysMAS obtains the highest observed mean frame-wise CLIP$_{sim}$ and the lowest paired-reference image-plane trajectory, projected-area, and local-strain errors among the compared methods; it achieves a two-minute reported runtime and is substantially faster than the evaluated long per-scene SDS optimization baselines.

2 Related Work

2.1 Generative 4D Gaussian Splatting

Dynamic Gaussian representations include 4D-GS, Deformable 3D Gaussians, Spacetime Gaussians, and L4GM (Wu et al., 2024; Yang et al., 2024; Li et al., 2024; Ren et al., 2024). Diffusion-guided 4D generation builds on DreamFusion and includes Align Your Gaussians, Animate3D, 4Diffusion, and CAT4D (Poole et al., 2023; Ling et al., 2024; Jiang et al., 2024; Zhang et al., 2024a; Wu et al., 2025). They target reconstruction or plausible animation rather than material-conditioned object–part simulation.

2.2 Physics-Grounded 4D Gaussian Splatting

Physics-grounded methods connect Gaussian appearance to MPM (Sulsky et al., 1994; Stomakhin et al., 2013); PhysGaussian provides the bridge, while Gaussian Splashing studies rotation artifacts (Xie et al., 2024; Feng et al., 2025). PhysDreamer, DreamPhysics, MotionPhysics, OmniPhysGS,

PhysSplat, and PhysGM automate property inference (Zhang et al., 2024b; Huang et al., 2025; Wang et al., 2026; Lin et al., 2025; Zhao et al., 2025; Lv et al., 2026a). Feature Splatting, GaussianProperty, PhysGS, and GaussianFluent address semantic or physical attributes and mixed materials (Qiu et al., 2024; Xu et al., 2025; Chopra et al., 2026; Huang et al., 2026). PAC-NeRF, GIC, Vid2Sim, FreeGave, and PhysTwin infer dynamics, properties, or simulation-ready representations from observations (Li et al., 2023; Cai et al., 2024; Chen et al., 2025b; Li et al., 2025b; Jiang et al., 2025). PhysMAS introduces an object-scoped part table that remains addressable through profile proposal, particle compilation, and executable candidate screening.

2.3 Object-Part Scene Grounding and Agentic Physical Reasoning

PartSLIP and SAGA address open-vocabulary or promptable 3D segmentation (Liu et al., 2023; Cen et al., 2025); Gaussian Grouping and LangSplat attach identity or language semantics to Gaussians (Ye et al., 2024; Qin et al., 2024); Part^{2}GS recovers part structure (Yu et al., 2026). Physics-scene methods cover object worlds, contact, heterogeneous solvers, and multi-body evaluation (Liu et al., 2024a; Chen et al., 2025a; Wang et al., 2025; Jiang & Lee, 2026; Li et al., 2025a; Liu et al., 2026; Kim et al., 2026). Language agents support programmatic interaction with environments (Ma et al., 2024); PhysAgent refines simulator force fields in the loop (Lv et al., 2026b). PhysMAS maintains typed object-scoped part state from grounding through material candidates, particle compilation, and screening; its two-agent contract assigns identity, planning, and acceptance to the scene agent, while the material agent supplies row-wise evidence.

3 Method

PhysMAS maps a motion prompt, four calibrated views with an aligned 3DGS, and output cameras to a 4D Gaussian sequence (Figure 2).

3.1 Problem Setup

Let $\mathcal{I} = \{(I_v, \mathsf{C}_v)\}_{v=1}^{4}$ be the ordered RGB views and calibrated cameras, and $\mathcal{G}^0 = \{g_i^0\}_{i=1}^{N_G}$ the aligned scene-level 3DGS, with $g_i^0 = (\boldsymbol{\mu}_i^0, \boldsymbol{\Sigma}_i^0, \alpha_i, \boldsymbol{h}_i)$. A fixed run specification provides bounded gravity/colliders, supported velocity or position blends, and impulses. Given y and output cameras $\{\mathsf{C}_{\text{out}}^t\}_{t=1}^{T}$, PhysMAS advances constitutive MPM under compiled controls and returns $\mathcal{V} = \{\text{Render}(\mathcal{G}^t, \mathsf{C}_{\text{out}}^t)\}_{t=1}^{T}$. Each Gaussian carries an object identity and object-scoped part row $r = (o, p)$; semantic and residual rows form the active set $\mathcal{R}^{\text{act}}$ indexing material state. Identically named parts of different objects remain distinct, while a one-part object uses one semantic row.

3.2 Object-Part Scene Agent

The backbone serializes an object–part inventory with cross-view object IDs. GroundingDINO and SAM 2 (Liu et al., 2024b; Ravi et al., 2025) provide masks and confidences; Gaussian labels then persist through material reasoning and simulation. From the prompt and inventory, the agent proposes plan Π with roles, targets, directions, magnitudes, and intervals, which a deterministic compiler validates against the fixed run specification. Let $\mathcal{O}$ be the detected objects, $\mathcal{R}_o$ the rows of object o, and Ω_i the in-bounds positive-depth views of Gaussian i at $\boldsymbol{u}_{iv}$. We assign objects before comparing parts within each $\mathcal{R}_o$:

$$b_{io} = \sum_{v \in \Omega_i} e_{ov}^{\text{obj}}(\boldsymbol{u}_{iv}), \qquad \widetilde{o}_i = \arg\max_{o \in \mathcal{O}} b_{io}. \tag{1}$$

Conditioned on $\widetilde{o}_i$, the within-object part assignment is

$$a_{ir} = \sum_{v \in \Omega_i} e_{rv}^{\text{part}}(\boldsymbol{u}_{iv}), \qquad \widetilde{z}_i^G = \arg\max_{r \in \mathcal{R}_{\widetilde{o}_i}} a_{ir}. \tag{2}$$

Mask confidence and spatial support define the evidence; deterministic propagation fuses $(\widetilde{o}_i, \widetilde{z}_i^G)$ into fixed labels z_i^G and adds object-scoped residual rows for coverage (guards and ties in the supplement).

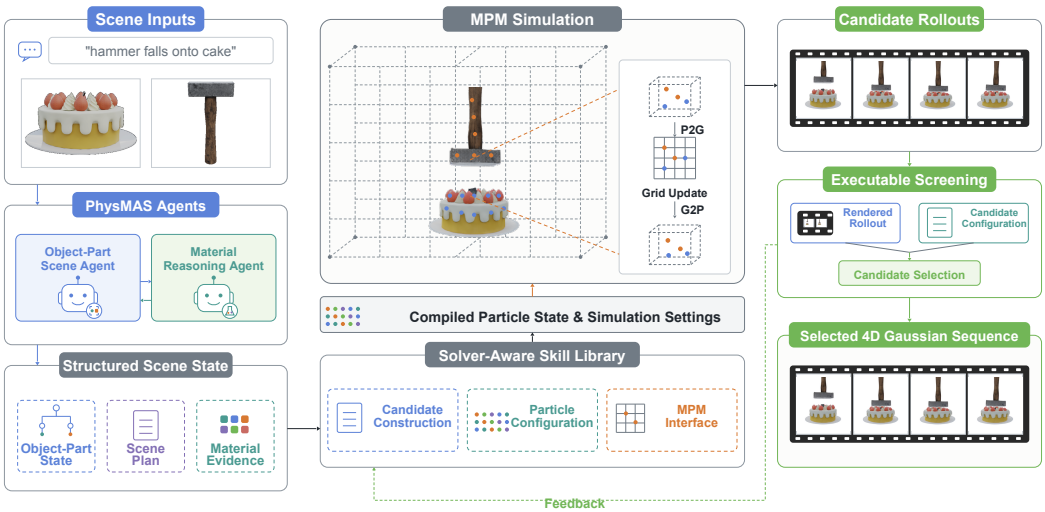

Figure 2: PhysMAS framework. The agents transform a motion prompt, four calibrated views, and an aligned static 3DGS into object–part state, a scene plan, and material evidence. Solver-aware skills compile and simulate candidates, screen their execution records, and return the accepted 4D Gaussian sequence.

3.3 Material Reasoning Agent

For each $r \in \mathcal{R}^{\mathrm{act}}$, evidence $\mathcal{H}_r$ contains highlighted observations, its name, and structural role. The Material Reasoning Agent combines vision–language material/branch scores with E/ν statistics from a frozen crop-conditioned predictor (Lv et al., 2026a) to return $\psi_r = \mathcal{A}_{\mathrm{mat}}(\mathcal{H}_r)$. After row-wise validation, candidate records $\Theta_r^{(c)} = (\beta_r^{(c)}, E_r^{(c)}, \nu_r^{(c)}, \rho_r^{(c)})$ are constrained to the accepted simulator-profile domain $\mathcal{K}_{\mathrm{sim}}$ (Jelly, Metal, Sand, Foam, Snow, and Plasticine). Candidate construction yields a part table and controls

$$\big(\{\Theta_r^{(c)}\}_r, \mathcal{U}_{\mathrm{aux}}^{(c)}\big) = \mathcal{T}_c\big(\{\psi_r\}_r; \mathcal{K}_{\mathrm{sim}}, \Pi\big). \tag{3}$$

The fixed scene state $\mathcal{Z}^0$ contains $\mathcal{G}^0$, Gaussian labels, initialized particles, numerical settings, and output cameras. The compiler then forms the schedule and simulator input

$$\begin{aligned} \Gamma^{(c)} &= \mathrm{Compile}(\Pi, \mathcal{U}_{\mathrm{aux}}^{(c)}), \\ \mathcal{C}^{(c)} &= \big(\mathcal{Z}^0, \{\Theta_r^{(c)}\}_r, \Gamma^{(c)}\big). \end{aligned} \tag{4}$$

Here β selects the constitutive branch; (E, ν, ρ) specify Young's modulus, Poisson's ratio, and density; and profile-specific constants remain fixed. Fixed transforms of posterior medians yield MAP, target-impact, support-bonded/stiff, solver-compatible, projection, soft/stiff, uncertainty-quantile, and second-profile variants in fixed order (supplement). All candidates preserve $\mathcal{Z}^0$, identities, and Π; $\mathcal{U}_{\mathrm{aux}}^{(c)}$ changes only structure/interface controls.

3.4 Solver-Aware Shared-Domain Simulation

Let $\mathcal{Q} = \{1, \ldots, N_Q\}$ index the particles in $\mathcal{Z}^0$. The map $\mathcal{N}(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{x} + \boldsymbol{d}$, with isotropic $\boldsymbol{L} = s\boldsymbol{I}_3$ and $s > 0$, converts scene coordinates to normalized MPM space. The first N_G particles are Gaussian means $\widehat{\boldsymbol{\mu}}_i^0 = \mathcal{N}(\boldsymbol{\mu}_i^0)$; support-filling particles retain source-object identity $o(q)$ and source Gaussian

$$\iota(q) = \begin{cases} q, & q \le N_G, \\ \arg\min_{i:\, o_i = o(q)} \|\boldsymbol{x}_q^0 - \widehat{\boldsymbol{\mu}}_i^0\|_2, & q > N_G, \end{cases} \tag{5}$$

The candidate is then bound to per-particle identity, parameters, and mass:

$$\begin{aligned} z_q^Q &= z_{\iota(q)}^G, \qquad \vartheta_q^{(c)} = \mathrm{Bind}(\Theta_{z_q^Q}^{(c)}), \\ M_q^{(c)} &= \rho_q^{(c)} V_q^0. \end{aligned} \tag{6}$$

Here V_q^0 is rest volume and Bind builds the branch record; residual rows ensure complete particle coverage. Affine Particle-In-Cell (APIC) transfers advance the MPM state (Sulsky et al., 1994; Stomakhin et al., 2013; Jiang et al., 2015; Xie et al., 2024), with parameters, gravity, and substeps defined in the coordinates induced by $\mathcal{N}$.

Omitting candidate superscript (c), APIC particle-to-grid (P2G) transfer at node a accumulates mass and momentum from every object:

$$\begin{aligned} m_a &= \sum_q w_{aq} M_q, \\ \boldsymbol{p}_a &= \sum_q w_{aq} M_q [\boldsymbol{v}_q + \boldsymbol{B}_q(\boldsymbol{x}_a - \boldsymbol{x}_q)] \\ &\quad + \Delta t\, \boldsymbol{f}_a^{\mathrm{int}}. \end{aligned} \tag{7}$$

The provisional grid velocity before gravity and grid-stage collider/action updates is

$$\boldsymbol{u}_a = \boldsymbol{p}_a / m_a, \qquad m_a > 0. \tag{8}$$

Here w_{aq} is the interpolation weight, $\boldsymbol{x}_a$ the node position, $\boldsymbol{B}_q$ the APIC affine velocity map, and $\boldsymbol{f}_a^{\mathrm{int}}$ the stress force. Gravity and collider/action updates precede G2P; all objects sample the common grid field, providing inter-object coupling. Configured collider friction handles collider interactions, and $\Gamma^{(c)}$ schedules substep actions and frame projections.

Kinematic structure-preserving projection for support-like parts. For support-like row r selected by a fixed role predicate, let $\mathcal{Q}_r = \{q \mid z_q^Q = r\}$ and $\boldsymbol{\delta}_{qr}^0 = \boldsymbol{x}_q^0 - \bar{\boldsymbol{x}}_r^0$. At solver step n_t for output frame t, tildes denote provisional states and bars means over $\mathcal{Q}_r$; $\boldsymbol{x}$, $\boldsymbol{v}$, $\boldsymbol{F}$, and $\boldsymbol{B}$ denote position, velocity, deformation gradient, and affine velocity matrix. The relaxed projection is

$$\begin{aligned} \boldsymbol{x}_q^{n_t} &= (1 - \eta_r^{(c)})\widetilde{\boldsymbol{x}}_q^{n_t} + \eta_r^{(c)}(\widetilde{\bar{\boldsymbol{x}}}_r^{n_t} + \boldsymbol{\delta}_{qr}^0), \\ \boldsymbol{v}_q^{n_t} &= (1 - \eta_r^{(c)})\widetilde{\boldsymbol{v}}_q^{n_t} + \eta_r^{(c)}\widetilde{\bar{\boldsymbol{v}}}_r^{n_t}. \end{aligned} \tag{9}$$

The same strength relaxes the deformation gradient and affine velocity state:

$$\begin{aligned} \boldsymbol{F}_q^{n_t} &= (1 - \eta_r^{(c)})\widetilde{\boldsymbol{F}}_q^{n_t} + \eta_r^{(c)}\boldsymbol{I}_3, \\ \boldsymbol{B}_q^{n_t} &= (1 - \eta_r^{(c)})\widetilde{\boldsymbol{B}}_q^{n_t}. \end{aligned} \tag{10}$$

With $\eta_r^{(c)} \in [0, 1]$ scheduled by $\Gamma^{(c)}$, this operator relaxes toward the rest shape, reducing stretch/shear and burr-like artifacts after Gaussian transport (Xie et al., 2024; Feng et al., 2025).

An optional frame-0-neighborhood projection retains attached-part interfaces (supplement). Both projections act after G2P at scheduled output frames, before Gaussian transport, and seed the next interval; multi-object cases use shared-grid interaction.

Transport to dynamic Gaussians. For Gaussian-associated particle i, let $\boldsymbol{F}_{i,\mathrm{sc}}^t = \boldsymbol{L}^{-1}\boldsymbol{F}_i^{n_t}\boldsymbol{L} = \boldsymbol{F}_i^{n_t}$ since $\boldsymbol{L} = s\boldsymbol{I}_3$. Rendering uses

$$\begin{aligned} \boldsymbol{\mu}_i^t &= \boldsymbol{L}^{-1}(\boldsymbol{x}_i^{n_t} - \boldsymbol{d}), \\ \boldsymbol{\Sigma}_i^t &= \boldsymbol{F}_{i,\mathrm{sc}}^t \boldsymbol{\Sigma}_i^0 (\boldsymbol{F}_{i,\mathrm{sc}}^t)^{\mathsf{T}}. \end{aligned} \tag{11}$$

Opacity and appearance stay fixed; the polar rotation of $\boldsymbol{F}_{i,\mathrm{sc}}^t$ updates spherical-harmonic viewing direction (Xie et al., 2024).

3.5 Executable Candidate Selection

For each candidate c, skills return process, numerical, artifact, and endpoint diagnostics $\mathcal{D}^{(c)}$, rollout $\mathcal{G}^{(c,t)}$, and video $\mathcal{V}^{(c)}$. Gate-admitted candidates above threshold τ are ranked by Q_{exec}. If none qualifies, diagnostics trigger fixed repairs when a round remains (Figure 2); after the final round, the run fails.

Only the scene agent may update $(\boldsymbol{z}^G, \Pi)$, invoke tools, or accept a rollout; the material agent returns $\boldsymbol{\psi}_r$. Fixed-budget skills handle compilation, simulation, repairs, gating, and ranking. The gate checks

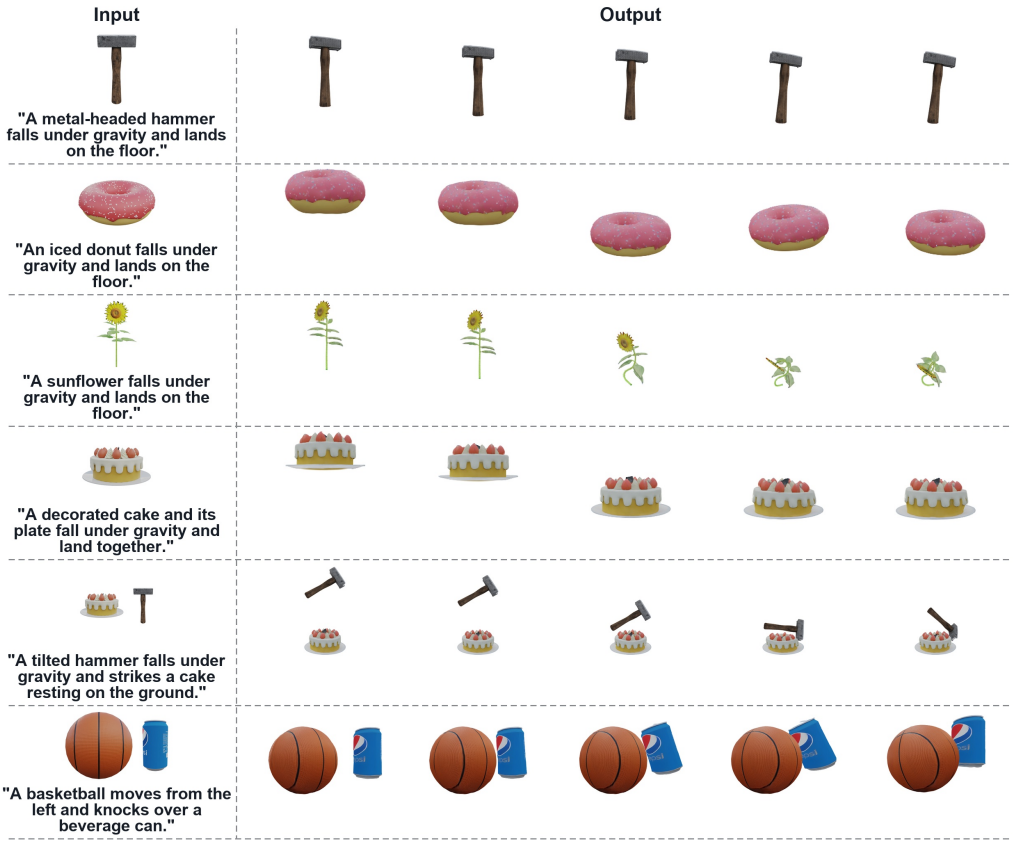

Figure 3: PhysMAS rollouts with an input, prompt, and five synchronized frames per row. Rows 1–4 highlight part-dependent deformation; Rows 5–6 show multi-object execution.

actions, termination, numerical/runtime stability, and support/spread/flattening criteria; Q_{exec} scores artifacts, confidence, material/role fit, numerical range, and projection use. Final audit requires a complete, decodable T-frame video and multi-object configurations with interface bonding disabled; success returns the selected rollout, while failure terminates the run. Budgets, weights, criteria, repair order, and ties appear in the supplement; representative rollouts appear in Figure 3.

4 Experiments

4.1 Implementation Details

We implement PhysMAS in PyTorch with Qwen3.7-Plus, Grounding DINO, and SAM 2. On one RTX 4090, MPM renders 50 frames at 800×800 on a 100^3 grid with at most 200,000 particles. The fixed run specification supplies gravity, colliders, and numerical settings; the agent supplies roles, targets, and supported actions. Candidates share this specification but are generated without evaluation reference videos. Further settings are in the supplement.

4.2 Datasets and Evaluation Protocol

We evaluate preconstructed static 3DGS scenes from **PhysGaussian** (Xie et al., 2024), **Objaverse** (Deitke et al., 2023), and **PhysAssets** (Lv et al., 2026a). **Multi-part** cases contain a single object with distinct semantic parts; the six simulator profiles are aggregation categories only (Table 1). A multi-profile case enters each applicable column, and "Avg." averages the six profiles.

Multi-object cases contain at least two identifiable objects in a visible impact or support interaction and are pooled for shared-grid evaluation.

4.3 Comparison with State-of-the-Art Methods

Baselines. We compare with OmniPhysGS (Lin et al., 2025), PhysSplat (Zhao et al., 2025), DreamPhysics (Huang et al., 2025), and PhysGM (Lv et al., 2026a). Each receives the same motion prompt, initial 3DGS, cameras, external action, ground, timestamps, duration, and resolution through its native interface. Every baseline retains its native input representation under this shared evaluation setup.

Perceptual metrics. Following prior evaluations (Lv et al., 2026a;b), we report CLIP similarity (CLIP_{sim}) and User Preference Rate (UPR). Frozen CLIP ViT-B/32 (Radford et al., 2021) scores each frame against the full prompt, then averages by case and evaluated set. UPR is a blinded, position-randomized forced choice by 20 CV/CG researchers for prompt-consistent, perceptually plausible motion. Main results use 5AFC; ablations use a separate 3AFC and are reported separately (full protocol in the supplement).

Motion and deformation metrics. Using paired references reserved for evaluation, we measure image-normalized CoTracker3 trajectory nRMSE (Karaev et al., 2025), mask-area log-RMSE, and frame-0-neighborhood Green–Lagrange local-strain RMSE (Kim et al., 2026). Table 3 multiplies raw RMSEs by 100 and macro-averages over profiles (multi-part) or cases (multi-object).

Table 1: Multi-part results across six material profiles. Entries are CLIP_{sim}/UPR; "Avg." is the six-profile macro-average. Multi-profile cases enter each applicable column. Higher is better.

Method	Jelly		Metal		Sand		Foam		Snow		Plasticine		Avg.	
	CLIP_{sim}	UPR	CLIP_{sim}	UPR	CLIP_{sim}	UPR	CLIP_{sim}	UPR	CLIP_{sim}	UPR	CLIP_{sim}	UPR	CLIP_{sim}	UPR
OmniPhysGS	0.2292	8%	0.2151	6%	0.2048	9%	0.2053	7%	0.1831	11%	0.2135	7%	0.2085	8.0%
PhysSplat	0.2528	17%	0.2426	15%	0.2419	13%	0.2390	16%	0.2218	17%	0.2516	19%	0.2416	16.2%
DreamPhysics	0.2459	12%	0.2272	10%	0.2216	8%	0.2326	10%	0.2071	13%	0.2437	15%	0.2297	11.3%
PhysGM	0.2774	28%	0.2732	29%	0.2997	34%	0.2609	23%	0.2548	27%	0.2691	25%	0.2725	27.7%
PhysMAS (Ours)	**0.2819**	**35%**	**0.2770**	**40%**	**0.3021**	**36%**	**0.2784**	**44%**	**0.2595**	**32%**	**0.2734**	**34%**	**0.2787**	**36.8%**

Multi-part results. Table 1 shows macro CLIP_{sim}/UPR increasing from OmniPhysGS (0.2085/8.0%) through DreamPhysics (0.2297/11.3%), PhysSplat (0.2416/16.2%), and PhysGM (0.2725/27.7%) to PhysMAS (0.2787/36.8%). PhysMAS leads both metrics in all six profiles; its UPR gains on Metal and Foam are notable despite modest CLIP_{sim} margins.

Table 2: Multi-object results; higher is better.

Method	$\text{CLIP}_{sim}\uparrow$	UPR (%) $\uparrow$
OmniPhysGS	0.2391	8.7%
PhysSplat	0.2434	13.3%
DreamPhysics	0.2437	15.0%
PhysGM	0.2631	25.7%
PhysMAS (Ours)	**0.2733**	**37.3%**

Multi-object results. On multi-object cases (Table 2), OmniPhysGS scores 0.2391/8.7%; PhysSplat and DreamPhysics score 0.2434/13.3% and 0.2437/15.0%; PhysGM reaches 0.2631/25.7%; and PhysMAS leads at 0.2733/37.3%, extending its advantage to interactions among distinct objects.

Motion, deformation, and efficiency. Table 3 reports multi-part/multi-object pairs. On multi-part cases, PhysMAS beats the strongest baseline, PhysGM, on trajectory, area, and local-strain errors (0.626, 2.641, 6.654). On multi-object cases, DreamPhysics leads baseline trajectory/area and OmniPhysGS baseline strain, but PhysMAS leads all three (1.085, 3.980, 15.361). Its rounded two-minute runtime matches PhysSplat, beats OmniPhysGS and DreamPhysics, and trails sub-minute PhysGM (Table 4).

Table 3: Image-plane motion and deformation errors for multi-part/multi-object results; lower is better.

Method	Trajectory nRMSE	Area Log RMSE	Local-strain RMSE
OmniPhysGS	1.594/3.395	8.318/9.704	12.871/43.060
PhysSplat	1.447/3.192	7.821/9.113	15.959/46.563
DreamPhysics	1.477/3.115	7.021/7.928	14.612/58.422
PhysGM	1.028/3.387	3.428/9.693	11.993/55.959
PhysMAS	**0.626/1.085**	**2.641/3.980**	**6.654/15.361**

Qualitative results. Figure 4 shows that OmniPhysGS flattens most, PhysSplat retains height but deforms uniformly, and DreamPhysics ends compressed. PhysGM better preserves silhouettes, while PhysMAS retains the cake's plate/layer separation and the sunflower's part-dependent bending. Figure 3 shows part-dependent responses in Rows 1–4 and cake–hammer and basketball–beverage interactions in Rows 5–6.

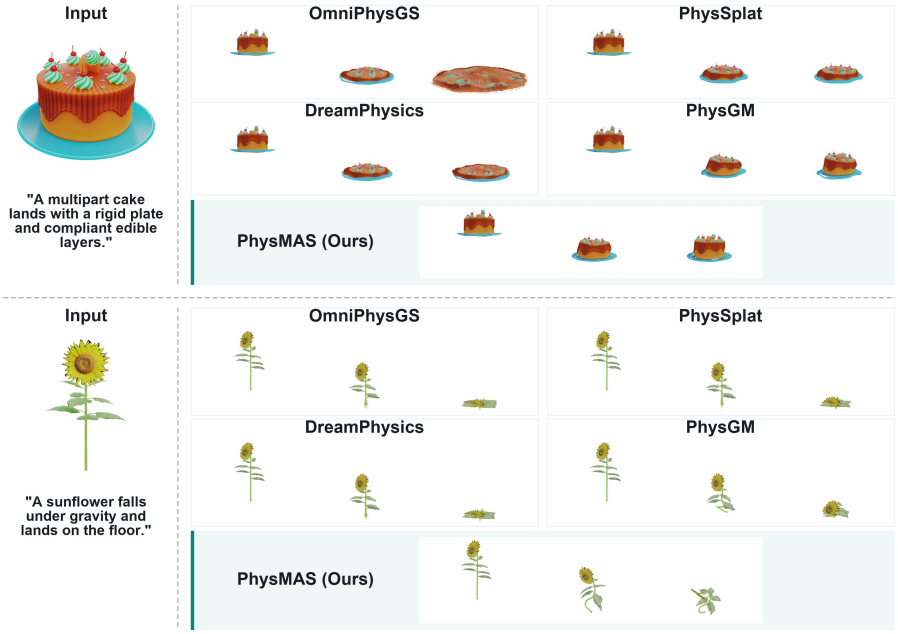

Figure 4: Multi-part comparisons on cake impact and sunflower falling. Synchronized frames show part preservation across methods.

Table 4: Post-reconstruction runtime per 50-frame sequence; lower is better.

Method	Runtime ↓
OmniPhysGS	> 12 h
PhysSplat	2 min
DreamPhysics	> 0.5 h
PhysGM	< 1 min
PhysMAS (Ours)	2 min

4.4 Ablation Study

Object-Part Scene Agent. This ablation replaces agent-owned object–part state and part-aware compilation with an object-level variant. Macro CLIP$_{sim}$/UPR falls from 0.2787/42.5% to 0.2616/25.5%, with declines in every profile. Foam changes most (0.2784/49% to 0.2184/14%) because object-wide assignment conflates its regions; Figure 5 shows the corresponding strawberry landing and contact. This supports the combined object–part state and compiler contribution.

Table 5: Six-profile ablation; UPR is in percent and "Avg." is the macro-average.

Profile	w/o O-P Agent		w/o S-P Projection		**Full**	
	CLIP$_{sim}$	UPR	CLIP$_{sim}$	UPR	CLIP$_{sim}$	UPR
Jelly	0.2727	31%	0.2696	26%	**0.2819**	**43%**
Metal	0.2693	27%	0.2703	31%	**0.2770**	**42%**
Sand	0.2921	27%	0.3015	35%	**0.3021**	**38%**
Foam	0.2184	14%	0.2684	37%	**0.2784**	**49%**
Snow	0.2484	26%	0.2487	30%	**0.2595**	**44%**
Plasticine	0.2688	28%	0.2719	33%	**0.2734**	**39%**
Avg.	0.2616	25.5%	0.2717	32.0%	**0.2787**	**42.5%**

Structure-Preserving Projection. Without structure-preserving projection, macro CLIP$_{sim}$/UPR falls to 0.2717/32.0%; the full model leads on all six profiles. UPR gains are largest on Jelly, Snow, and Foam (17, 14, and 12 points). Figure 5 shows the full model better preserving the pineapple's crown–body support and landing silhouette.

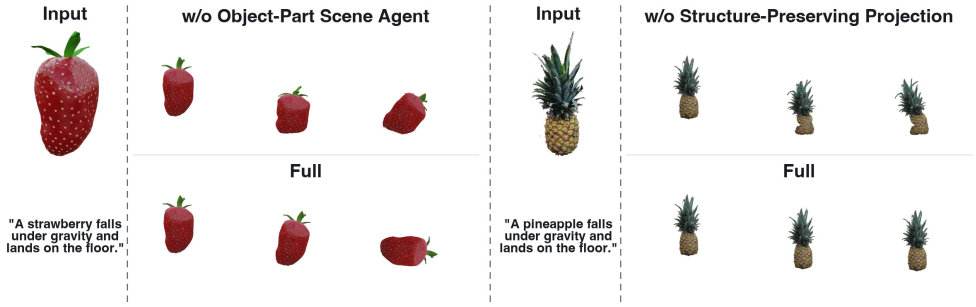

Figure 5: Qualitative ablation: each variant (top) versus Full PhysMAS (bottom).

5 Discussion

PhysMAS fills a key gap in physics-grounded 4D Gaussian synthesis by automating object–part scene organization and part-wise material assignment. Persistent object–part identities and material profiles provide a structured interface between multimodal reasoning and MPM simulation, while solver-aware skills compile and screen executable candidates. Our implementation uses a 100^3 grid to balance fidelity and efficiency. Future work will combine adaptive simulation with learned execution feedback to support real-time, large-scale compositional dynamics.

6 Conclusion

We introduce PhysMAS, a physics-grounded multi-agent framework for compositional 4D Gaussian synthesis. Its Object-Part Scene Agent, Material Reasoning Agent, and solver-aware skills carry persistent identities and part-wise profiles into shared-domain MPM simulation, supporting multi-part and multi-object dynamics without per-scene SDS backpropagation. Experiments demonstrate superior semantic alignment, perceived physical plausibility, and reference-video motion and deformation consistency among the compared methods while maintaining a short runtime.

AI USE STATEMENT

Generative AI tools were used to support literature discovery, manuscript drafting and editing, translation, figure preparation, software assistance, and feedback on research methodology and experiments. A large language model is also an explicit component of the proposed PhysMAS system and was used during its experimental execution. The authors reviewed the resulting manuscript, figures, citations, configurations, code changes, and reported results, and take responsibility for the final content, claims, and artifacts.

ETHICS STATEMENT

The human preference evaluation asks participants to compare rendered simulation videos and reports the resulting judgments only in aggregate. The manuscript and supplementary material contain no personally identifying participant information. The authors are responsible for ensuring that recruitment, consent, compensation, and data handling comply with applicable institutional policies.

REPRODUCIBILITY STATEMENT

The method definition, object–part and material-agent interfaces, deterministic candidate-selection procedure, simulator configuration, evaluation metrics, and experimental protocol are documented in Sections 3 and 4. The appendix provides the expanded agent contracts, material profiles, particle-compilation rules, execution-screening procedure, metric definitions, user-study protocol, and additional qualitative results needed to reproduce and audit the reported pipeline.

REFERENCES

Junhao Cai, Yuji Yang, Weihao Yuan, Yisheng He, Zilong Dong, Liefeng Bo, Hui Cheng, and Qifeng Chen. GIC: Gaussian-informed continuum for physical property identification and simulation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75035–75063, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/89379d5fc6eb34ff98488202fb52b9d0-Abstract-Conference.html.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3D gaussians. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2): 1971–1979, 2025. doi: 10.1609/aaai.v39i2.32193.

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, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_PhysGen3D_Crafting_a_Miniature_Interactive_World_from_a_Single_Image_CVPR_2025_paper.html.

Chuhao Chen, Zhiyang Dou, Chen Wang, Yiming Huang, Anjun Chen, Qiao Feng, Jiatao Gu, and Lingjie Liu. Vid2Sim: Generalizable, video-based reconstruction of appearance, geometry and physics for mesh-free simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26545–26555, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Vid2Sim_Generalizable_Video-based_Reconstruction_of_Appearance_Geometry_and_Physics_for_CVPR_2025_paper.html.

Samarth Chopra, Jing Liang, Gershom Seneviratne, and Dinesh Manocha. PhysGS: Bayesian-inferred gaussian splatting for physical property estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18980–18990, 2026.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, and Yin Yang. Gaussian splashing: Unified particles for versatile motion synthesis and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 518–529, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Feng_Gaussian_Splashing_Unified_Particles_for_Versatile_Motion_Synthesis_and_Rendering_CVPR_2025_paper.html.

Bei Huang, Yixin Chen, Ruijie Lu, Gang Zeng, Hongbin Zha, Yuru Pei, and Siyuan Huang. GaussianFluent: Gaussian simulation for dynamic scenes with mixed materials. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21583–21593, 2026.

Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson W. H. Lau. DreamPhysics: Learning physics-based 3D dynamics with video diffusion priors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3733–3741, 2025. doi: 10.1609/aaai.v39i4.32389.

Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. The affine particle-in-cell method. *ACM Transactions on Graphics*, 34(4):1–10, 2015. doi: 10.1145/2766996.

Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. PhysTwin: Physics-informed reconstruction and simulation of deformable objects from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7219–7230, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jiang_PhysTwin_Physics-Informed_Reconstruction_and_Simulation_of_Deformable_Objects_from_Videos_ICCV_2025_paper.html.

Suyi Jiang and Gim Hee Lee. PhysHO: Physics-based dynamic 3D gaussian human and object from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32507–32517, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jiang_PhysHO_Physics-Based_Dynamic_3D_Gaussian_Human_and_Object_from_Monocular_CVPR_2026_paper.html.

Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. Animate3D: Animating any 3D model with multi-view video diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125879–125906, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e3b53f89136b1bc69a5714ea465f01b6-Abstract-Conference.html.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6013–6022, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Mijeong Kim, Gunhee Kim, Jungyoon Choi, Wonjae Roh, and Bohyung Han. PhysGaia: A physics-aware benchmark with multi-body interactions for dynamic novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22604–22614, 2026.

Jinxi Li, Ziyang Song, and Bo Yang. TRACE: Learning 3D gaussian physical dynamics from multi-view videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8820–8829, 2025a.

Jinxi Li, Ziyang Song, Siyuan Zhou, and Bo Yang. FreeGave: 3D physics learning from dynamic videos by gaussian velocity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12433–12443, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_FreeGave_3D_Physics_Learning_from_Dynamic_Videos_by_Gaussian_Velocity_CVPR_2025_paper.html.

Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming C. Lin, Chenfanfu Jiang, and Chuang Gan. PAC-NeRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tVkrbkz42vc.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8508–8520, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_Spacetime_Gaussian_Feature_Splatting_for_Real-Time_Dynamic_View_Synthesis_CVPR_2024_paper.html.

Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. OmniPhysGS: 3D constitutive gaussians for general physics-based dynamics generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9HZtP6I5lv.

Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4D with dynamic 3D gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8576–8588, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ling_Align_Your_Gaussians_Text-to-4D_with_Dynamic_3D_Gaussians_and_Composed_CVPR_2024_paper.html.

Chunjiang Liu, Xiaoyuan Wang, Qingran Lin, Albert Xiao, Haoyu Chen, Shizheng Wen, Hao Zhang, Lu Qi, Ming-Hsuan Yang, Laszlo A. Jeni, Min Xu, and Yizhou Zhao. MOSIV: Multi-object system identification from videos. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0ylAe3Orfy.

Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. PartSLIP: Low-shot part segmentation for 3D point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21736–21746, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Liu_PartSLIP_Low-Shot_Part_Segmentation_for_3D_Point_Clouds_via_Pretrained_CVPR_2023_paper.html.

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, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024b.

Chunji Lv, Zequn Chen, Donglin Di, Weinan Zhang, Hao Li, Chen Wei, Yinjie Lei, and Changsheng Li. PhysGM: Large physical gaussian model for feed-forward 4D synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29855–29865, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lv_PhysGM_Large_Physical_Gaussian_Model_for_Feed-Forward_4D_Synthesis_CVPR_2026_paper.html.

Chunji Lv, Jiaxi Ye, Yuchen Jiang, Rexar Lin, and Changsheng Li. PhysAgent: Automating physics-based 4D synthesis via trajectory-grounded multi-agent feedback, 2026b.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IEduRUO55F.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Language-driven physics-based scene synthesis and editing via feature splatting. In *European Conference on Computer Vision*, pp. 368–383, 2024. URL https://feature-splatting.github.io/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Raedle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, and Huan Ling. L4GM: Large 4D gaussian reconstruction model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56828–56858, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6808f2c57d9564a2639a4710e3bbd9b9-Abstract-Conference.html.

Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics*, 32(4):102:1–102:10, 2013. doi: 10.1145/2461912.2461948.

Deborah Sulsky, Zhen Chen, and Howard L. Schreyer. A particle method for history-dependent materials. *Computer Methods in Applied Mechanics and Engineering*, 118(1–2):179–196, 1994. doi: 10.1016/0045-7825(94)90112-0.

Miaowei Wang, Yibo Zhang, Weiwei Xu, Rui Ma, Changqing Zou, and Daniel Morris. DecoupledGaussian: Object-scene decoupling for physics-based interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11361–11372, 2025.

Miaowei Wang, Jakub Zadrozny, Oisin Mac Aodha, and Amir Vaxman. MotionPhysics: Learnable motion distillation for text-guided simulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):9993–10001, 2026. doi: 10.1609/aaai.v40i12.37965.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_4D_Gaussian_Splatting_for_Real-Time_Dynamic_Scene_Rendering_CVPR_2024_paper.html.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: Create anything in 4D with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26057–26068, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_CAT4D_Create_Anything_in_4D_with_Multi-View_Video_Diffusion_Models_CVPR_2025_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/

Xie_PhysGaussian_Physics-Integrated_3D_Gaussians_for_Generative_Dynamics_CVPR_2024_paper.html.

Xinli Xu, Wenhang Ge, Dicong Qiu, ZhiFei Chen, Dongyu Yan, Zhuoyun Liu, Haoyu Zhao, Hanfeng Zhao, Shunsi Zhang, Junwei Liang, and Ying-Cong Chen. GaussianProperty: Integrating physical properties to 3D gaussians with LMMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7231–7240, 2025.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20331–20341, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Deformable_3D_Gaussians_for_High-Fidelity_Monocular_Dynamic_Scene_Reconstruction_CVPR_2024_paper.html.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3D scenes. In *European Conference on Computer Vision*, pp. 162–179, 2024. doi: 10.1007/978-3-031-73397-0_10.

Tianjiao Yu, Vedant Shah, Muntasir Wahed, Ying Shen, Kiet A. Nguyen, and Ismini Lourentzou. Part^{2}GS: Part-aware modeling of articulated objects using 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18913–18923, 2026.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4Diffusion: Multi-view video diffusion model for 4D generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 15272–15295, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1bbfea488a8968e2d3c6565639b08e5e-Abstract-Conference.html.

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, 2024b. doi: 10.1007/978-3-031-72627-9_22.

Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. PhysSplat: Efficient physics simulation for 3D scenes via MLLM-guided gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5242–5252, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhao_PhysSplat_Efficient_Physics_Simulation_for_3D_Scenes_via_MLLM-Guided_Gaussian_ICCV_2025_paper.html.

A Agent Contracts and Object–Part State

This section details the two agent contracts and the deterministic object–part lifting stage. Only the Object-Part Scene Agent can update scene identity and plan Π, invoke tools, or accept a rollout. The Material Reasoning Agent returns row-wise evidence; candidate construction, compilation, simulation, repair generation, gating, ranking, and final audit are fixed routines.

A.1 Generalized Object-Part Scene Agent Prompt

The invariant instruction template below summarizes the prompt contract. Direction labels and calibrated cameras accompany the four-view sheet, and the aligned 3DGS is addressed through the persistent scene record.

> **Role:** scene-level object–part orchestrator.
> **Input:** motion prompt, four direction-labeled scene views, camera records, and the registered scene state.
> **Instructions:** identify objects before their parts; keep object IDs stable across views; create part IDs only within a parent object; preserve identically named parts from different objects as different rows; form a scene plan using only the supported action vocabulary;

Table A.1: Structured records and ownership in the PhysMAS inference protocol.

Record	Required fields	Owner / consumer
Object inventory	persistent object ID, name, scene role, view-linked queries	Scene Agent / perception and lifting
Object–part row	compound key $r = (o, p)$, part name, parent object, structural role, physics-group token, schema-expected visual labels, evidence references	Scene Agent / Material Agent and compiler
Scene plan Π	action type, source/target keys, direction, magnitude, active interval	Scene Agent / action compiler
Material evidence ψ_r	visual-label evidence, six-profile evidence, E/ν statistics, confidence	Material Agent / candidate builder
Whole-object evidence	full-object visual label, mapped profile, E/ν statistics, profile-derived density prior	Fixed evidence routine / candidate builder
Candidate row $\Theta_r^{(c)}$	profile branch $\beta_r^{(c)}$, $E_r^{(c)}$, $\nu_r^{(c)}$, $\rho_r^{(c)}$	Fixed builder / particle compiler
Candidate controls $\mathcal{U}_{\mathrm{aux}}^{(c)}$	support-projection weights, interface-retention flags	Fixed builder / schedule compiler
Candidate metadata	source-transform ID, parent candidate ID, fixed-order index	Fixed builder / repair and screen
Execution record $\mathcal{D}^{(c)}$	process, numerical, artifact, and end-point diagnostics	Simulator skills / deterministic screen and Scene Agent audit

request material evidence independently for every active row; and accept a rollout only after deterministic execution screening.
Output: an object inventory, object-scoped part rows, and a proposed structured scene plan. Do not emit constitutive parameters or modify fixed numerical settings.

The proposed scene plan records source and target identities, semantic roles, action type, direction, magnitude, and active interval. A deterministic compiler resolves the referenced rows, canonicalizes the action fields, and validates the plan before rollout. Gravity, colliders, cameras, grid resolution, and rendering settings remain in the fixed run specification rather than in the language-model output. A one-part object follows the same schema with one semantic row; a multi-object scene uses distinct object IDs while retaining one scene-level plan.

A.2 Generalized Material Reasoning Agent Prompt

The Material Reasoning Agent is called once per active object–part row. Its invariant instruction template is:

Role: row-wise material-evidence reasoner.
Input: one persistent object–part key, its stored name and structural role, highlighted observations from the calibrated views, and the six accepted simulator profiles.
Instructions: assess visual-material evidence for the specified row; distinguish the visible material label from the simulator profile; return confidence-bearing evidence for the profile and continuous E/ν statistics; preserve the supplied identity; and abstain when a required semantic row has no valid evidence.
Output: evidence over visual labels and simulator profiles, together with continuous E and ν statistics where available. Density is supplied by the selected visual-label prior during deterministic candidate construction. Do not change object/part identities, the scene plan, actions, cameras, or solver settings.

The row evidence is serialized as ψ_r and validated before candidate construction. Table A.1 lists the normalized records passed between the agents and fixed routines; it is the semantic schema used throughout the paper, independent of backend file names.

Table A.2: Frozen implementation settings for Gaussian object–part lifting.

Setting	Frozen value
Interior-distance support	range $[0.45, 1.00]$
Mask-confidence floor	0.05
Residual confidence cap	0.20
Area tie-break	$10^{-6}/\max(1, A_r)$
Ambiguity / reliable margin	0.18 / 0.30
Neighborhood sizes	12 and 16
Residual rescue target	$\min\big(n_r^{\mathrm{cand}}, \max(64, \mathrm{round}(0.003 N_G))\big)$

A.3 Perception Configuration

Perception allows at most two segmentation attempts. The initial schema query receives the front scene view; agent mode then uses one direction-labeled 2×2 grid of all four views for perception planning and repair. Material reasoning uses one highlighted multi-view sheet per row. Calls use deterministic inference, while pretrained Grounding DINO and SAM 2 provide object/part masks and confidence values under a fixed configuration. The front-view object inventory defines canonical query IDs. Projection-consistent Gaussian support and fixed query order provide deterministic matching for masks in the remaining views.

A.4 Object–Part Lifting, Guards, and Ties

The main paper writes object and part support as e_{ov}^{obj} and e_{rv}^{part}. In the implementation, support is zero outside the selected mask and otherwise combines the stored mask confidence with normalized distance to the mask boundary. An area-normalized positive tie-break prevents exact zero scores. Multi-view-consistent neighbors regularize ambiguous assignments, followed by a residual-label cleanup pass. Every Gaussian retains either its selected semantic row or an object-scoped residual row with a persistent part ID.

We retain the main-paper notation $r = (o, p)$ for a compound object–part row, where o and p denote its object and part components. Let $\mathcal{Q} = \{1, \ldots, N_Q\}$ be the particle index set and $\mathcal{Q}_r = \{q \in \mathcal{Q} \mid z_q^Q = r\}$ the particles assigned to row r. The method input and final rollout remain scene-level as defined in the main paper. In Table A.2, A_r is the accepted multi-view mask area for row r, and n_r^{cand} is the number of residual Gaussians eligible for its cleanup pass.

Table A.2 lists the frozen implementation settings used across all reported data sources.

B Material Profiles and Particle Compilation

B.1 Visual-to-Solver Mapping and Parameter Domain

Table B.1 gives the visual-material priors and initial map used during material reasoning and deterministic candidate construction. The serialized E and profile-derived ρ values carry the nominal Pa and $\mathrm{kg/m^3}$ labels expected by the backend. After uniform geometry normalization, these values serve as canonical solver coefficients. Ordinary part rows constrain E and ν to the selected visual material's supported domain and use the local solver ranges $\nu \in [0.05, 0.45]$ and $\rho \in [300, 3000]$. The explicitly listed composite transforms in Table B.3 instead use the whole-object record, allowing ν up to 0.485 and a profile-density prior in $[100, 7800]$. Thus $\mathcal{K}_{\mathrm{sim}}$ in the main paper is the union of the ordinary-row domain and these enumerated composite-transform domains, rather than the result of one contradictory global clamp. The adaptive wave-speed rule uses a minimum substep of 5×10^{-5} seconds and at most 800 substeps per output frame.

Table B.1: Exact initial mapping, visual-material E prior ranges, ν bounds, and density priors.

Visual label	Solver branch	E prior range (Pa)	ν bound	Raw ρ
Wood	Metal	10^8–2×10^{10}	0.25–0.45	700
Metal	Metal	10^9–3×10^{11}	0.20–0.35	7800
Plastic	Metal	10^6–5×10^9	0.30–0.45	1200
Glass	Metal	10^9–10^{11}	0.18–0.30	2500
Fabric	Foam	10^4–10^8	0.20–0.45	500
Leather	Foam	10^5–10^9	0.30–0.49	900
Ceramic	Metal	10^8–10^{11}	0.15–0.35	2500
Stone	Metal	10^8–10^{11}	0.10–0.35	2600
Rubber	Jelly	10^4–10^8	0.40–0.499	1100
Paper	Foam	10^5–10^9	0.20–0.45	800
Sand	Sand	10^3–10^7	0.20–0.45	1600
Snow	Snow	10^3–10^7	0.10–0.35	300
Plasticine	Plasticine	10^3–10^7	0.25–0.49	2000
Foam	Foam	10^3–10^7	0.10–0.45	100

Table B.2: Constitutive branches selected by the six simulator profiles. "None" means that the trial deformation gradient is retained before stress evaluation.

Profile	Elastic / trial stress	Inelastic deformation update
Jelly	Compressible Neo-Hookean	None
Metal	Fixed-corotational elasticity	None
Sand	Hencky trial stress with Drucker–Prager projection ($k = 0$, $\alpha = 0.16$)	Sand return mapping
Foam	Hencky elasticity	StVK-based viscoplastic return mapping
Snow	Hencky trial stress with Drucker–Prager projection ($k = 1000$, $\alpha = 0.16$)	Stress-space projection
Plasticine	Hencky trial stress with Drucker–Prager projection ($k = 5000$, $\alpha = 0.16$)	Von-Mises-with-damage return mapping

B.2 Six Solver Profiles

For every particle, the compiler converts (E, ν) to Lamé parameters

$$\mu = \frac{E}{2(1+\nu)}, \qquad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \tag{B.1}$$

while ρ determines particle mass through $M_q = \rho_q V_q^0$. Table B.2 lists the stress and inelastic-update branches used by the six profiles. These names identify the simulator-facing constitutive branches associated with the inferred visual-material evidence.

B.3 Evidence Aggregation and Continuous Parameters

For the evidence aggregation summarized by $\mathcal{A}_{\text{mat}}$ in the main paper, the frozen source coefficients for schema, crop predictor, VLM, role, and whole-object evidence are 0.25, 0.35, 0.30, 0.10, and 0.05, respectively. These are unnormalized evidence coefficients—their sum need not equal one—and the resulting label scores are normalized only after all sources are accumulated. Crop records are averaged within their source. Valid VLM records are confidence-weighted and divided by their count. Reported semantic rows use validated crop and predictor evidence before aggregation, while residual rows retain a unit Plastic fallback. The same fixed in-run evidence sources are used throughout the reported evaluation.

The role source scans the lowercase concatenation of the stored part name and physical role. The same vocabulary is used unchanged across all three reported data sources. Let $R_r(m)$ be the maximum confidence assigned to label m by all matching rules, or zero if none matches. The fixed lookup is: head/blade/tip/edge/impact/metal $\mapsto$ Metal at 0.75; handle/grip/shaft $\mapsto$ Wood at 0.55 and Plastic at 0.45; wheel/tire/tyre/sole $\mapsto$ Rubber at 0.80; cushion/pad/pillow/foam $\mapsto$ Foam at

0.75; cloth/fabric/lace/upper $\mapsto$ Fabric at 0.65; plate/bowl/ceramic $\mapsto$ Ceramic at 0.55 and Plastic at 0.35; and frosting/cream/cake/soft $\mapsto$ Foam at 0.55 and Plasticine at 0.35. Multiple matching rules contribute evidence, while R_r retains the maximum per label for the guard below.

Let $\mathcal{M}_{\text{vis}}$ be the 14 visual labels in Table B.1, let $\mathcal{M}_r^{\text{exp}}$ be the deduplicated schema-expected labels stored with row r, and let $s_r(m)$ be the resulting nonnegative weighted evidence. If its sum is zero, the fixed fallback replaces the score vector by a unit score at Plastic. Define $Z_r = \sum_{m \in \mathcal{M}_{\text{vis}}} s_r(m)$ after this fallback, so $Z_r > 0$. The normalized posterior and selected label are

$$\begin{aligned} P_r(m) &= \frac{s_r(m)}{Z_r}, \\ \widetilde{m}_r &= \operatorname*{arg\,max}_{m \in \mathcal{M}_{\text{vis}}} P_r(m). \end{aligned} \tag{B.2}$$

Thus $P_r \in \Delta^{13}$, and posterior ties follow the stored evidence-insertion order. The unconstrained mode is $m_r^{(0)} = \widetilde{m}_r$. When $\mathcal{M}_r^{\text{exp}}$ is nonempty, let m_r^{exp} be its highest-posterior label, with the deduplicated expected-label order breaking ties, and define

$$\begin{aligned} \text{Cap}_{\text{exp}}(r) = \; & \max\{1.75 P_r(m_r^{\text{exp}}), \\ & \qquad P_r(m_r^{\text{exp}}) + 0.20\}, \\ \text{ExpGuard}(r) = \; & [m_r^{(0)} \notin \mathcal{M}_r^{\text{exp}}] \wedge [P_r(m_r^{\text{exp}}) \geq 0.18] \\ & \wedge [P_r(m_r^{(0)}) \leq \text{Cap}_{\text{exp}}(r)], \\ m_r^{(1)} = \; & \begin{cases} m_r^{\text{exp}}, & \text{ExpGuard}(r), \\ m_r^{(0)}, & \text{otherwise.} \end{cases} \end{aligned} \tag{B.3}$$

For an empty expected set, $\text{ExpGuard}(r)$ is false.

Two guards then run in a fixed order. First, the multi-view VLM guard examines only records marked successful that do not abstain and provide normalized material probabilities, confidence at least 0.60, and at least three canonical visible views. Let v be a record's top label. The record is eligible only if its top-label probability is at least 0.75, $v \in \mathcal{M}_r^{\text{exp}}$, $v \neq m_r^{(1)}$, $R_r(v) \geq 0.50$, $R_r(v) \geq R_r(m_r^{(1)}) + 0.10$, $P_r(v) \geq 0.20$, $P_r(m_r^{(1)}) \leq 1.50 P_r(v)$, and $P_r(m_r^{(1)}) - P_r(v) \leq 0.18$. If eligible records exist, $m_r^{(2)}$ is the label maximizing the product of VLM confidence, within-record top-label probability, and $P_r(v)$; otherwise $m_r^{(2)} = m_r^{(1)}$. Record order breaks ties.

Second, let Map denote the visual-label-to-branch map in Table B.1, and let $\mathcal{M}_{\text{el}} = \{m \mid \text{Map}(m) = \text{Metal}\}$. A row is structural when its normalized physics-group and role tokens contain support, supporting, structural, rigid, or loadbearing, or contain both load and bearing. A strong elastic-solid VLM record has at least three canonical visible views, confidence at least 0.60, total probability on $\mathcal{M}_{\text{el}}$ at least 0.70, and maximum elastic-solid probability at least 0.50. Define the expected-label elastic-solid fraction by

$$f_r^{\text{el}} = \begin{cases} |\mathcal{M}_r^{\text{exp}} \cap \mathcal{M}_{\text{el}}| / |\mathcal{M}_r^{\text{exp}}|, & |\mathcal{M}_r^{\text{exp}}| > 0, \\ 1, & |\mathcal{M}_r^{\text{exp}}| = 0. \end{cases} \tag{B.4}$$

If $m_r^{(2)} \notin \mathcal{M}_{\text{el}}$, the row is structural, $f_r^{\text{el}} \geq 0.50$, and such a record exists, then $m_r = \operatorname{arg\,max}_{m \in \mathcal{M}_{\text{el}}} P_r(m)$; otherwise $m_r = m_r^{(2)}$. Stored posterior order breaks the final tie. The stored material confidence is $P_r(m_r)$.

Residual rows with assigned Gaussians remain active during candidate construction. In the reported configuration, deserialization maps every particle to its stored active row, while the solver-compatible candidate uses its stored global profile. Ties follow active-row order.

For crop proposal j of row r, the feed-forward model produces normalized means μ_{rj}^E, μ_{rj}^ν, uncertainty terms $\sigma_{rj}^{E,\text{model}}, \sigma_{rj}^\nu$, and visual-material evidence e_{rj}^{phys}. The implementation decodes

$$\begin{aligned} \widehat{E}_{rj} &= 0.1 \, 10^{2.456477 \mu_{rj}^E + 7.387210}, \\ \widehat{\nu}_{rj} &= 0.111 \mu_{rj}^\nu + 0.398. \end{aligned} \tag{B.5}$$

The Qwen query provides discrete visual-material evidence, while continuous E and ν statistics are pooled from validated crop predictions. Each crop contributes $x^E_{rj} = \log_{10} \widehat{E}_{rj}$, $x^\nu_{rj} = \widehat{\nu}_{rj}$, $w_{rj} = \max(0.05, e^{\text{phys}}_{rj}(m_r))$, using 0.1 when the selected-label entry is absent, and uncertainty $a_{rj} = \sigma^{E,\text{model}}_{rj}$. Let $\mathcal{L}^E_r$ and $\mathcal{L}^\nu_r$ contain the records finite for the E- and ν-statistics, respectively, and define $W^E_r = \sum_{j\in\mathcal{L}^E_r} w_{rj}$ and $W^\nu_r = \sum_{j\in\mathcal{L}^\nu_r} w_{rj}$. For positive corresponding normalizers,

$$\begin{aligned}
\bar{e}_r &= (W^E_r)^{-1} \sum_{j\in\mathcal{L}^E_r} w_{rj} x^E_{rj}, \\
\bar{\nu}_r &= (W^\nu_r)^{-1} \sum_{j\in\mathcal{L}^\nu_r} w_{rj} x^\nu_{rj}, \\
v^e_r &= (W^E_r)^{-1} \sum_{j\in\mathcal{L}^E_r} w_{rj} (x^E_{rj} - \bar{e}_r)^2, \\
s^e_r &= \max\bigl(0.25, \sqrt{v^e_r}\bigr)\,.
\end{aligned} \tag{B.6}$$

Let $E_{\text{def}}(m)$, $\nu_{\text{def}}(m)$, and $\rho_{\text{def}}(m)$ denote the geometric-midpoint E, midpoint ν, and raw density for visual label m in Table B.1. With fewer than two valid E records, $s^e_r = 0.25$; with no valid E record, the implementation sets $\bar{e}_r = \log_{10} E_{\text{def}}(m_r)$ without evaluating the E moments in Equation (B.6). With no valid ν record, it independently sets $\bar{\nu}_r = \nu_{\text{def}}(m_r)$. Let clip^E_m and clip^ν_m denote the E- and ν-domain clamps of visual label m. The posterior builder then stores

$$\begin{aligned}
e^{\text{post}}_r &= \log_{10}\Bigl[\text{clip}^E_{m_r}(10^{\bar{e}_r})\Bigr]\,, \\
\nu^{\text{post}}_r &= \text{clip}^\nu_{m_r}(\bar{\nu}_r).
\end{aligned} \tag{B.7}$$

Let $\bar{a}_r = |\mathcal{L}^E_r|^{-1} \sum_{j\in\mathcal{L}^E_r} a_{rj}$ when E records exist (default 0.35 otherwise), $H_r = -\sum_m P_r(m) \log P_r(m)$ with $0 \log 0 := 0$, and $h_r = \min(0.5, H_r / \max(1, \log 14))$. Let $\bar{\sigma}^\nu_r$ be the mean finite crop uncertainty over $\mathcal{L}^\nu_r$, with default 0.06 for an empty set. These defaults provide complete residual-row records. The final dispersions are

$$\begin{aligned}
\sigma^e_r &= \max(0.10, 0.50\bar{a}_r + 0.35 s^e_r + 0.15 h_r)\,, \\
\sigma^\nu_r &= \max(0.02, \bar{\sigma}^\nu_r)\,.
\end{aligned} \tag{B.8}$$

Let $\gamma^{\text{mask}}_r \in [0,1]$ be the stored part-mask confidence. The lifting stage applies the positive floor in Table A.2; consequently, a nonpositive serialized value is the sentinel for an unavailable confidence rather than an observed zero-confidence mask. Define

$$\widetilde{\gamma}^{\text{mask}}_r = \begin{cases} \gamma^{\text{mask}}_r, & \gamma^{\text{mask}}_r > 0, \\ 0.7, & \text{otherwise.} \end{cases} \tag{B.9}$$

The stored posterior confidence used by candidate construction is

$$\gamma^{\text{part}}_r = \text{clip}_{[0,1]}\left(\frac{P_r(m_r)}{1+\sigma^e_r} \max(0.2, \widetilde{\gamma}^{\text{mask}}_r)\right). \tag{B.10}$$

Here $\text{clip}_{[0,1]}(x) = \min(1, \max(0, x))$. For independent $\xi_E, \xi_\nu \in \{0.15, 0.25, 0.50, 0.75, 0.85\}$, the implementation uses the corresponding $\kappa_\xi \in \{-1.04, -0.67, 0, 0.67, 1.04\}$ and computes

$$\begin{aligned}
E_r(\xi_E) &= \text{clip}^E_{m_r}\Bigl(10^{e^{\text{post}}_r + \kappa_{\xi_E}\sigma^e_r}\Bigr)\,, \\
\nu_r(\xi_\nu) &= \text{clip}^\nu_{m_r}\bigl(\nu^{\text{post}}_r + \kappa_{\xi_\nu}\sigma^\nu_r\bigr)\,.
\end{aligned} \tag{B.11}$$

A changed visual label uses its corresponding defaults before the same constraints are applied.

B.4 Candidate Construction and Repair Transforms

The initial row is $(\beta^0_r, E^0_r, \nu^0_r, \rho^0_r) = (\text{Map}(m_r), E_r(0.50), \nu_r(0.50), \text{clip}(\rho_{\text{def}}(m_r); 300, 3000))$. Thus ρ^0_r comes from the selected visual label's density prior rather than from a language-model prediction. The auxiliary whole-object record $(\beta_g, E_g, \nu_g, \rho_g)$ is produced from the full-object crop

by the same visual mapping and crop-conditioned E/ν predictor, with density again supplied by its selected visual-label prior and all fields passed through the backend guard. When that record is absent, the fixed fallback is $(\mathrm{Plasticine}, 5\times10^5, 0.476, 5000)$. All predicate tests below are case-insensitive substring matches. Let S_r denote the candidate-construction support predicate. A support token in the concatenated name, physics-group token, and role makes $S_r = 1$. Only when none matches, a soft token makes $S_r = 0$; if neither set matches, Wood, Metal, Glass, Ceramic, or Stone visual evidence makes $S_r = 1$; it is zero otherwise. The support tokens are plate, dish, tray, stand, support, holder, structural, rigid, load bearing, load-bearing, frame, shell, blade, head, tip, and impact part. The soft tokens are soft, deformable, compressible, cushion, foam, sponge, flesh, fruit, food, filling, cream, frosting, and icing.

The candidate-construction cohesion predicate C_r is evaluated only when $S_r = 0$. It is one when the concatenated name, physics-group token, role, and visual label contains body, base, core, main, bulk, filling, coating, layer, cream, frosting, icing, fruit, organic, soft, deformable, foam, sponge, rubber, gel, cloth, fabric, leather, or pad. Otherwise it is one when $\beta_r^0 \in \{\mathrm{Foam}, \mathrm{Plasticine}, \mathrm{Jelly}\}$ and the visual label is not Wood, Metal, Glass, Ceramic, or Stone; it is zero in every remaining case, including $S_r = 1$. No other Sand or Snow exclusion is applied during candidate construction. The first matching rule defines ω_r: it is 1.6 when $S_r = 1$, 1.45 for candle/stem/stick/handle, 1.25 for fruit/strawberry/berry/decoration, 1.05 for cream/frosting/icing, and 1 otherwise.

Let m_r^{sec} be the second-ranked label in the stored posterior order and $E_r^{\mathrm{post}} = 10^{e_r^{\mathrm{post}}}$. A row is alternative-feasible when $P_r(m_r^{\mathrm{sec}}) \geq 0.12$ and $E_{\mathrm{def}}(m_r^{\mathrm{sec}}) \leq \max(10E_r^{\mathrm{post}}, 10^7)$. Among feasible rows, $r^\dagger$ is the row with the largest $P_r(m_r^{\mathrm{sec}})$; ties use the fixed active-row order, and no alternative candidate is emitted when this set is empty. All rows begin from their posterior medians. Row $r^\dagger$ receives m_r^{sec}; if this differs from the stored label m_r, its mapped branch and default E, ν, and density are used, whereas equality retains the posterior median. The ordinary physical and solver clamps follow in either case. The uncertainty sweep independently chooses the minimum-γ_r^{part} row with the fixed order as a tie-break.

In Table B.3, Map and the three def functions use the branch and defaults just defined. Define $\beta_g^{\mathrm{coh}} = \beta_g$ when $\beta_g \in \{\mathrm{Plasticine}, \mathrm{Jelly}, \mathrm{Foam}\}$, and $\beta_g^{\mathrm{coh}} = \mathrm{Plasticine}$ otherwise. Initial-candidate rows use S_r and C_r. A repair clones one rejected base candidate; let $(\beta_r^{(b)}, E_r^{(b)}, \nu_r^{(b)}, \rho_r^{(b)})$ denote that clone's current row before the repair. Repair rows use the evaluator predicates $\widehat{S}_r(k)$ and $\widehat{C}_r(k)$ defined below. Table B.3 records the transforms; fields not named by a transform retain their current initial- or base-candidate values. Ordinary rows retain the local ranges above, while hard-support-bonded and solver-compatible rows use the explicitly shown whole-object ν and ρ fields.

The target-impact candidate is emitted only when Π marks the interaction role as target, the supported action type specifies an impact, and at least one non-residual target row satisfies $S_r = 0$ and $C_r = 1$. Residual rows start from the same candidate quantile as other active rows but are skipped by the target-impact softening rule. The solver-compatible and hard-support-bonded candidates require at least one row with $S_r = 1$ and another with $S_r = 0, C_r = 1$. Solver-compatible assigns the uniquely defined β_g^{coh} to every row and disables both projection and interface bonding. Active rows comprise all non-residual rows plus residual rows with assigned Gaussians. The initial order is posterior MAP, applicable target-impact, applicable hard-support bonded, support-stiff, solver-compatible, rigid-support, soft, stiff, uncertainty sweep, and alternative material; the first six applicable candidates are retained. Every retained candidate contains every active row. "Rigid-support" and "Repair: rigid support" are serialized candidate names retained to match the released artifacts; both instantiate the support-shape projection. Deterministic screening diagnostics map support failure to the support repairs, spread or flattening to the cohesion repairs, numerical failure to solver stability, and an otherwise low score to role balancing. At most three unique repairs are emitted in a repair round.

B.5 Gaussian and Filled-Particle Initialization

The first N_G particles are initialized at the normalized Gaussian means and retain the corresponding Gaussian indices. Additional particles provide interior support for MPM. Filling is performed within an object-scoped region, so each added particle first receives a source-object identity and only then searches for its nearest Gaussian among that object's Gaussians, as formalized by Equations (5)–(6)

Table B.3: Implemented material-candidate transforms. $\mathrm{clip}(x; l, u) = \min(\max(x, l), u)$.

Candidate	Eligible row	Final solver row/action
Posterior	all	$\beta = \beta^0$; $(\xi_E, \xi_\nu) = (0.50, 0.50)$
Soft / stiff	all	$\beta = \beta^0$; $(\xi_E, \xi_\nu) = (0.25, 0.75)$ / $(0.75, 0.50)$
Uncertainty sweep	all	selected weakest-confidence row uses $(0.15, 0.75)$; others use $(0.50, 0.75)$
Alternative material	$r = r^\dagger$	use m_r^{sec}; use its defaults only when $m_r^{\mathrm{sec}} \neq m_r$; apply the ordinary physical and solver clamps
Target-impact	target row with $S_r = 1$	start from $(\xi_E, \xi_\nu) = (0.50, 0.50)$; $\beta = \mathrm{Metal}$, $E = 10^7$; $\nu = \mathrm{clip}(\nu^0; 0.05, 0.35)$, $\rho = \max(\rho^0, 1000)$; disable projection and bonding
Target-impact	target row with $S_r = 0, C_r = 1$	start from $(0.50, 0.50)$; $E = \mathrm{clip}(E^0; 2\times10^4, 7.5\times10^4)$; $\beta = \beta^0$ if $\beta^0 \in \{\mathrm{Foam}, \mathrm{Plasticine}, \mathrm{Jelly}\}$, else Foam; $\nu = \mathrm{clip}(\nu^0; 0.2, 0.4)$; $\rho = \min(450, \max(300, \rho^0))$
Target-impact	target row with $S_r = C_r = 0$	retain the posterior-median row and disable projection and bonding
Support-stiff	$S_r = 1$	$\beta = \mathrm{Metal}$, $E = 10^7$; $\nu = \min(\nu^0, 0.35)$, $\rho = \max(\rho^0, 1000)$; no projection
Hard-support bonded	$S_r = 1$	$\beta = \mathrm{Metal}$, $(E, \nu, \rho) = (10^7, 0.35, 2500)$; $\eta_r = 0.85$ and enable interface retention
Hard-support bonded	$S_r = 0$	$\beta = \mathrm{Plasticine}$, $E = \max(E^0, \omega_r E_g)$; $\nu = \mathrm{clip}(\nu_g; 0.2, 0.485)$, $\rho = \rho_g$
Solver-compatible	all	$\beta = \beta_g^{\mathrm{coh}}$; $E = \mathrm{clip}(\omega_r E_g; 0.85E_g, 1.7E_g)$; $\nu = \mathrm{clip}(\nu_g; 0.2, 0.485)$, $\rho = \rho_g$
Rigid-support	all	start every row from $(\xi_E, \xi_\nu) = (0.25, 0.75)$; for $S_r = 1$, retain β and current ρ, set $E = 10^7$, $\eta_r = 1$
Repair: support stiffness	$\widehat{S}_r = 1$	$\beta = \mathrm{Metal}$, $E = 10^7$, $\nu = \min(\nu^{(b)}, 0.35)$, $\rho = \max(\rho^{(b)}, 1000)$
Repair: rigid support	$\widehat{S}_r = 1$	support-stiffness repair plus $\eta_r = 1$
Repair: elastic cohesion	$\widehat{S}_r = 1$	apply the support-stiffness repair
Repair: elastic cohesion	$\widehat{S}_r = 0, \widehat{C}_r = 1$	$\beta = \mathrm{Jelly}$, $E = \max(E^{(b)}, 5\times10^5)$; $\nu = \mathrm{clip}(\nu^{(b)}; 0.2, 0.42)$, $\rho = \max(\rho^{(b)}, 500)$
Repair: higher cohesion	$\widehat{S}_r = 1$	apply the support-stiffness repair
Repair: higher cohesion	$\widehat{S}_r = 0, \widehat{C}_r = 1$	retain $\beta^{(b)}, \nu^{(b)}$; $E = \max(E^{(b)}, 5\times10^5)$, $\rho = \max(\rho^{(b)}, 500)$
Repair: solver stability	all	retain $\beta^{(b)}, E^{(b)}$; $\nu = \min(\nu^{(b)}, 0.40)$, $\rho = \max(\rho^{(b)}, 300)$
Repair: role balanced	role-dependent	$\widehat{S}_r = 1$ uses the support repair; $\widehat{S}_r = 0, \widehat{C}_r = 1$ with $\beta^{(b)} \in \{\mathrm{Foam}, \mathrm{Plasticine}\}$ uses elastic-cohesion repair; others unchanged

in the main paper. This ordering prevents a filled particle near an object boundary from inheriting the row of a different object. The source Gaussian transfers the persistent object–part row, and the selected candidate row supplies the constitutive branch and continuous coefficients.

Each particle stores its initial position and rest volume V_q^0. Candidate-specific mass is formed as $M_q^{(c)} = \rho_q^{(c)} V_q^0$; therefore candidate changes affect material state without changing the shared particle geometry. Gaussian-associated particles are transported back to dynamic 3DGS primitives for rendering, whereas filled particles contribute mass, stress, and momentum during simulation but are not rendered directly. Residual rows ensure that every associated or filled particle receives a valid candidate record.

B.6 Heterogeneous Step-Size Rule

Let $\mathfrak{R}_\Theta$ contain the selected part rows together with the auxiliary whole-object row $(\beta_g, E_g, \nu_g, \rho_g)$ defined above (or its stated fixed fallback), and let Δt_0 denote the base integration step. For each $a \in \mathfrak{R}_\Theta$, the compiler forms the elastic wave-speed proxy

$$\begin{aligned} G_a &= \frac{E_a}{2(1+\nu_a)}, \\ \lambda_a &= \frac{E_a \nu_a}{(1+\nu_a)(1-2\nu_a)}, \\ c_a &= \sqrt{\frac{\lambda_a + 2G_a}{\rho_a}}. \end{aligned} \tag{B.12}$$

Let $c_{\max} = \max_{a \in \mathfrak{R}_\Theta} c_a$, and let c_{ref} be computed from the nominal tuple $(E, \nu, \rho) = (5{\times}10^6, 0.35, 2500)$. With the fixed wave-speed scaling coefficient $\kappa_{\mathrm{CFL}} = 0.7$, the target integration step is

$$\Delta t_{\mathrm{target}} = \begin{cases} \Delta t_0, & c_{\max} \le \kappa_{\mathrm{CFL}} c_{\mathrm{ref}}, \\ \kappa_{\mathrm{CFL}} \Delta t_0 \, c_{\mathrm{ref}} / c_{\max}, & c_{\max} > \kappa_{\mathrm{CFL}} c_{\mathrm{ref}}. \end{cases} \tag{B.13}$$

The implementation converts this target into a ceiling-rounded number of substeps and then applies the configured minimum-step and maximum-substep caps. This CFL-inspired wave-speed scaling is the implemented substep-selection heuristic used consistently across candidates.

B.7 Structure-Preserving Projection

The main paper gives the complete implemented post-step update. Applied after each output-frame MPM update and before Gaussian transport, the operator acts on support-like rows to reduce local deformation and suppress burr-like rendering artifacts. The optional interface-retention update defined below follows it.

C Solver-Aware Skills and Executable Screening

C.1 Supported Scene Actions

The scene plan represents object roles, action-source/target identities, and the operations in Table C.1. The deterministic compiler canonicalizes directions, bounds magnitudes, validates targets, instantiates the actions over explicit spatial regions and intervals, and combines them with the gravity and floor/collider entries supplied by the fixed run specification.

C.2 Candidate Budget and Execution Records

Under the complete Problem Setup contract—motion prompt, four calibrated RGB views and cameras, their aligned static $\mathcal{G}^0$, and prescribed output cameras—the reported evaluation configuration combines scene-agent-proposed object–part rows and a deterministically validated plan, Qwen3.7-Plus material evidence, and a crop-conditioned feed-forward physical-property distribution for every semantic part. After the Material Reasoning Agent returns row-wise evidence posteriors, the Object-Part Scene Agent invokes candidate construction, retains at most six initial candidates, and controls at most three candidate–execute–observe rounds. Each repair round produces at most three new candidate IDs, and fixed gate/rank rules use acceptance threshold 0.72. The artifact audit requires a complete, decodable T-frame sequence; multi-object rollouts use configurations with interface bonding disabled. Hard-part structure preservation remains candidate-controlled. Each retained artifact contains the object–part records, masks, Gaussian identities, evidence records, complete scene-candidate configurations, execution logs, scores, selected table, runtime, and final flags.

Round 0 evaluates the retained initial candidates; rounds 1 and 2, when reached, evaluate repairs of the current base. Candidates accumulate across rounds, and each candidate ID is evaluated once. At each round, gate-admitted candidates are ranked by Q_{exec} over the accumulated set, with first

Table C.1: Supported action vocabulary and its implemented solver operation. "Enforced velocity" denotes a bounded particle boundary condition over the action interval; position fields blend the current state toward a locally constructed deformation target. The serialized `torque` token denotes a rotation request implemented by a tangential velocity field, rather than a directly applied mechanical moment.

Action	Implemented operation and spatial scope
`translation`	Enforced constant velocity on the selected object or particle region.
`top_press`	Enforced downward velocity on the selected target's top slab.
`radial_spread`	Outward horizontal velocity blend plus a bounded position increment on selected deformable particles.
`vertical_compress`	Position blend toward a reduced vertical extent, with the induced velocity capped, on the selected deformable region.
`side_compress`	Inward position and velocity blend on the contact-facing slab of the selected target.
`drag_pair`	Enforced puller velocity; optional serialized flags add interface or whole-target follow velocities.
`scale`	Opposing enforced velocities on two slabs of one object for compression or stretching.
`impulse`	A short particle-force impulse, capped at four base substeps.
`torque`	Bounded tangential particle-velocity field about the selected axis.
`support_hold`	Zero-velocity boundary condition on an eligible support region.

appearance in the fixed candidate order breaking ties. If the admitted set is nonempty, the highest-ranked candidate is submitted to the Object-Part Scene Agent for final artifact audit: a successful audit returns that rollout, whereas a failed audit terminates the run with failure. Only when the admitted set is empty and another round remains does the current highest-ranked executed candidate supply the repair base. Repairs are emitted in this fixed order: support stiffness, rigid support, elastic cohesion, higher cohesion, solver stability, and role balancing; the first three remaining unique IDs are retained. An empty admitted set after the final round returns failure.

C.3 Initial-Neighborhood Interface Retention

For a support row r, the hard-support-bonded candidate uses support-shape weight $\eta_r = 0.85$ and considers particles assigned to a valid compound row other than r. It computes each candidate's initial distance to the nearest particle assigned to r, keeps distances no greater than $d_b = 0.045$ in normalized solver coordinates, orders them by distance, and takes the first 26,000. For each retained particle, it stores $\boldsymbol{\delta}_{qr}^{b,0} = \boldsymbol{x}_q^0 - \bar{\boldsymbol{x}}_r^0$. Each frame applies

$$\boldsymbol{x}_q \leftarrow (1-\eta_b)\boldsymbol{x}_q + \eta_b(\bar{\boldsymbol{x}}_r + \boldsymbol{\delta}_{qr}^{b,0}),$$
$$\boldsymbol{v}_q \leftarrow (1-\zeta_b)\boldsymbol{v}_q + \zeta_b\bar{\boldsymbol{v}}_r.$$

with $\eta_b = \zeta_b = 0.95$. Here η_r controls the support-shape projection in the main paper, whereas the distinct weights η_b and ζ_b control position and velocity retention at the selected support–part interface. This initial-neighborhood projection preserves selected support–part interfaces during single-object rollouts. Multi-object evaluation uses configurations with interface bonding disabled, retaining object-scoped interaction through the MPM grid.

C.4 Validity Gate, Ranking, and Tie Rules

Let $n_{\mathrm{proj}}(k)$ and $n_{\mathrm{bond}}(k)$ be the numbers of rows in candidate k using support-shape projection and interface bonding, respectively. The implementation score Q_{exec} used to rank rollout candidate k is

$$\begin{aligned}\boldsymbol{\varphi}(k) &= (s_{\mathrm{vid}}(k), \bar{\gamma}_k, s_{\mathrm{div}}(k), s_{\mathrm{role}}(k), s_{\mathrm{base}}(k),\\ &\qquad s_\rho(k), s_{\mathrm{solver}}(k))^{\mathsf{T}},\\ \boldsymbol{w} &= (1, 0.27, 0.10, 0.25, 0.15, 0.08, 0.15)^{\mathsf{T}},\\ \ell(k) &= \min(0.12, 0.025 n_{\mathrm{proj}}(k) + 0.035 n_{\mathrm{bond}}(k))\,,\\ Q_{\mathrm{exec}}(k) &= \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\varphi}(k) - \ell(k). \end{aligned} \tag{C.1}$$

The video feature $s_{\mathrm{vid}}(k)$ is initialized to zero. A missing video subtracts 0.15; an existing video smaller than 4096 bytes subtracts 0.20; and an existing video of at least 4096 bytes adds 0.20. These terms order executed candidates and repair bases. Independently of this score, the final artifact audit requires the selected artifact to exist, decode successfully, and contain the complete T-frame sequence, so no missing or incomplete video can be returned. When valid first/last foreground measurements exist, the score adds 0.12, then subtracts 0.25 if foreground-box area grows above $2.0\times$ (or 0.10 above $1.45\times$), 0.22 if height falls below $0.45\times$ while width exceeds $1.25\times$, and 0.08 if the dark-foreground ratio exceeds 0.18. A nonzero process exit or a numerical/runtime-error token fails the validity gate. Let $\mathcal{R}_k^{\mathrm{act}}$ be candidate k's active-row set and $N_k^{\mathrm{part}} = |\mathcal{R}_k^{\mathrm{act}}| > 0$. Then $\bar{\gamma}_k = (N_k^{\mathrm{part}})^{-1} \sum_{r \in \mathcal{R}_k^{\mathrm{act}}} \gamma_r^{\mathrm{part}}$ averages Equation (B.10), and $s_{\mathrm{div}}(k)$ is the number of distinct visual labels divided by N_k^{part}. Candidate construction rejects an empty part table before execution. For every initial candidate, including target-impact, $s_{\mathrm{base}}(k) = \bar{\gamma}_k$; a repair inherits this value from its base candidate. The serialized base-confidence term is inherited by repairs, while the part-confidence term summarizes aggregated part-level evidence.

The role term uses a separate name-only lookup. A name containing head, blade, tip, plate, support, stand, tray, dish, shell, or frame receives 1 only when raw $E \geq 10^6$ Pa and $\beta = \mathrm{Metal}$, and 0.35 otherwise. A name containing rubber, sole, tire, wheel, foam, or cushion receives 1 when raw $E \leq 10^8$ Pa and 0.5 otherwise. Every other row receives 0.7. The term $s_{\mathrm{role}}(k)$ averages these values. The density term $s_\rho(k)$ averages 0.35 below 250, 0.55 above 5000, and 1 otherwise.

The evaluator uses separate frozen predicates $\widehat{S}_r(k)$ and $\widehat{C}_r(k)$. A row satisfies $\widehat{S}_r(k) = 1$ when its name contains plate, dish, tray, bowl, base support, support, stand, holder, case, shell, or frame, or when its visual label is Wood, Metal, Glass, Ceramic, or Stone; it is zero otherwise. It satisfies $\widehat{C}_r(k) = 1$ only when $\widehat{S}_r(k) = 0$, its branch is non-granular, and either the concatenation of its stored name, physics-group token, role, and visual label contains body, base, core, filling, cream, frosting, icing, soft, cushion, pad, foam, sponge, rubber, gel, fruit, or food, or its branch is Foam, Plasticine, or Jelly and its visual label is not one of the five rigid labels above. It is zero otherwise.

For rows with $\widehat{C}_r(k) = 1$, the first matching rule gives the solver score. A Plasticine visual label whose name contains clay, putty, plasticine, or dough receives 1 for Plasticine and 0.7 otherwise. The remaining rows receive 1 for Jelly, 0.55 for Foam or Plasticine, and 0.7 otherwise. The term $s_{\mathrm{solver}}(k)$ averages this lookup over $\widehat{C}_r(k) = 1$ rows and equals 0.7 when that set is empty.

Candidate admission requires a resolved scene plan; successful process termination; no numerical/runtime-error token; foreground-box area growth no greater than $2.0\times$; and no simultaneous height ratio below $0.45\times$ and width ratio above $1.25\times$. In addition, when the set $\{r : \widehat{S}_r(k) = 1\}$ is nonempty, every such support row must have $E_r^{(k)} \geq 10^7$ Pa or the candidate must use the solver-compatible regularization transform. This support condition is vacuously satisfied when no support row exists. Candidates satisfying these gates and $Q_{\mathrm{exec}}(k) \geq \tau$, with $\tau = 0.72$, are ordered by the same score and then undergo the mandatory artifact audit above. Action references are validated by the scene-plan compiler before rollout. The fixed implementation score provides a consistent criterion for candidate admission and ordering.

D Evaluation Protocol

D.1 CLIP Similarity and Aggregation

Let ϕ_I and ϕ_T be the frozen CLIP image and text encoders, $I_{c,t}$ frame t of case c, and y_c its complete motion prompt. Frame similarity and the case score are

$$s_{c,t} = \frac{\phi_I(I_{c,t})^\top \phi_T(y_c)}{\|\phi_I(I_{c,t})\|_2 \, \|\phi_T(y_c)\|_2},$$
$$\mathrm{CLIP}_{\mathrm{sim}}(c) = \frac{1}{T}\sum_{t=1}^{T} s_{c,t}. \tag{D.1}$$

For any reported set, scores are first computed per case and then averaged over its cases. A multi-part case contributes to every material-profile set represented by one of its active semantic rows; the six-profile average is the unweighted mean of those six set scores. Multi-object results are averaged once over the multi-object set without a material subdivision.

D.2 Reference-Video Motion and Deformation Metrics

The evaluation manifest fixes one synthetic MPM reference video for every case. Each reference is generated once from a preregistered, hand-authored MPM configuration using the same static 3DGS, initial pose, camera, frame count, frame time, renderer, and interaction as the compared outputs; its parameters are not changed after method videos have been scored. References are accessed only by the evaluation code—not by material reasoning, candidate construction, screening, or repair. The reported units for all three diagnostics equal 100 times the corresponding raw RMSE. Let $(x_{t,i}, y_{t,i})$ and $(\widehat{x}_{t,i}, \widehat{y}_{t,i})$ denote reference and predicted CoTracker3 coordinates for point i at frame t. With frame width W, height H, and $\mathcal{J}$ the point–frame pairs visible in both videos, reported trajectory nRMSE is

$$E_{\mathrm{traj}} = 100\left[\frac{1}{|\mathcal{J}|}\sum_{(t,i)\in\mathcal{J}}\left(\frac{(\widehat{x}_{t,i}-x_{t,i})^2}{W^2} + \frac{(\widehat{y}_{t,i}-y_{t,i})^2}{H^2}\right)\right]^{1/2}. \tag{D.2}$$

The reference is tracked once from a fixed 16×16 query grid restricted to its eroded frame-0 target mask, and every method reuses those same frame-0 queries. A pair belongs to $\mathcal{J}$ only when the same initialized point is visible, remains inside the registered target mask, and has a finite track in both videos at that frame. Frames with fewer than 16 such points are excluded. Restricting the sum to jointly visible pairs prevents a missing measurement from being treated as zero displacement. The metric record is valid only when at least one frame remains; otherwise the output is recorded as invalid rather than assigned zero error.

The target masks are evaluation-only visible-target silhouettes rendered with all scene Gaussians retained as occluders. For target-mask areas A_t and $\widehat{A}_t$, we compute projected-area log-RMSE as

$$E_{\mathrm{area}} = 100\left[\frac{1}{T}\sum_{t=1}^{T}\big(\log(\widehat{A}_t+1)-\log(A_t+1)\big)^2\right]^{1/2}. \tag{D.3}$$

The logarithm measures relative area change; the $+1$ term keeps an associated empty mask finite, while failed target association yields an invalid record.

For local deformation, each frame-0 tracked point uses its eight nearest tracked neighbors. A least-squares local 2D deformation gradient $\boldsymbol{F}_{t,i}$ maps the initial neighborhood offsets to their offsets at frame t, giving Green–Lagrange strain

$$\boldsymbol{E}_{t,i} = \tfrac{1}{2}\big(\boldsymbol{F}_{t,i}^\top \boldsymbol{F}_{t,i} - \boldsymbol{I}_2\big). \tag{D.4}$$

The reference-defined eight-neighbor graph is reused for every method. The set $\mathcal{K}$ contains only point–frame pairs for which the center and at least four of its eight stored neighbors are jointly visible

with finite tracks and the regularized local least-squares system is nondegenerate. For this jointly valid set, local-strain RMSE is

$$E_{\text{strain}} = 100 \left[\frac{1}{|\mathcal{K}|} \sum_{(t,i)\in\mathcal{K}} \|\widehat{\boldsymbol{E}}_{t,i} - \boldsymbol{E}_{t,i}\|_F^2 \right]^{1/2}. \tag{D.5}$$

This removes rigid translation while retaining local stretch, compression, and shear discrepancies. The local-strain record is valid only when $|\mathcal{K}| > 0$; an empty valid-neighborhood set marks the output invalid rather than assigning it zero. Invalid runs remain in the evaluation registry as failures rather than receiving an artificial zero error. The finalizer aborts rather than dropping or imputing any required invalid record; the reported reference-metric tables were finalized only after their registries contained zero invalid records. Multi-part results are averaged within each applicable profile and then macro-averaged over the six profiles. For multi-object motion, registered objects are evaluated separately, averaged with equal weight within a scene, and then averaged across scenes. Together, the three diagnostics quantify image-plane trajectory, projected deformation scale, and local strain relative to synchronized reference videos.

D.3 User Preference Evaluation

Study design. The reported UPR aggregates responses from 20 researchers with Computer Vision or Computer Graphics expertise. The main comparison uses a Five-Alternative Forced-Choice (5AFC) study: each trial presents synchronized outputs from OmniPhysGS, PhysSplat, DreamPhysics, PhysGM, and PhysMAS. The ablation uses a separate Three-Alternative Forced-Choice (3AFC) study containing the two ablated variants and the full model. Method names are hidden, and the videos are shown in randomized positions. For every case, a respondent selects exactly one video that best matches the motion prompt while exhibiting physically plausible and visually natural dynamics. Every case remains in the study through an explicit output-status card where applicable. The 5AFC and 3AFC protocols are reported separately according to their respective choice sets and chance levels.

Aggregation. Let $\mathcal{S}$ be a reported set of evaluation cases, R_c the number of valid responses for case c, and J_{ch} the method selected by respondent h for that case. The per-case preference share and the reported User Preference Rate are

$$u_c(m) = \frac{100}{R_c} \sum_{h=1}^{R_c} \mathbf{1}[J_{ch} = m],$$
$$\text{UPR}_{\mathcal{S}}(m) = \frac{1}{|\mathcal{S}|} \sum_{c\in\mathcal{S}} u_c(m). \tag{D.6}$$

Thus every case receives equal weight even if valid-response counts differ. For the multi-object evaluation, $\mathcal{S}$ contains all evaluated multi-object cases. For a material-profile column, $\mathcal{S}$ contains every evaluated multi-part case carrying that profile; a case with several profiles contributes independently to each applicable set. The six-profile average is the unweighted mean of the six profile-level UPR values. UPR reports human preference for prompt consistency, physically plausible motion, and visual naturalness under the stated choice protocol.

E Limitations and Future Work

Cross-View Grounding and Scene Inputs. Future work can strengthen persistent object–part assignment through occlusion-aware cross-view association, confidence-calibrated residual handling, and additional observations.

Constitutive and Interaction Coverage. New material branches, joints, and topology-aware updates can extend the six-profile interface to a broader range of part behaviors and interactions.

Spatial and Temporal Resolution. Adaptive particle/grid resolution and temporal planning can support finer interactions and longer action schedules while preserving the current agent–compiler interface.

F Additional Results

Figures 6–11 present 42 additional PhysMAS rollouts. Each row shows five temporally ordered frames, covering heterogeneous multi-part objects and multi-object scenes under varied motion configurations.



Figure 6: Additional results (I).

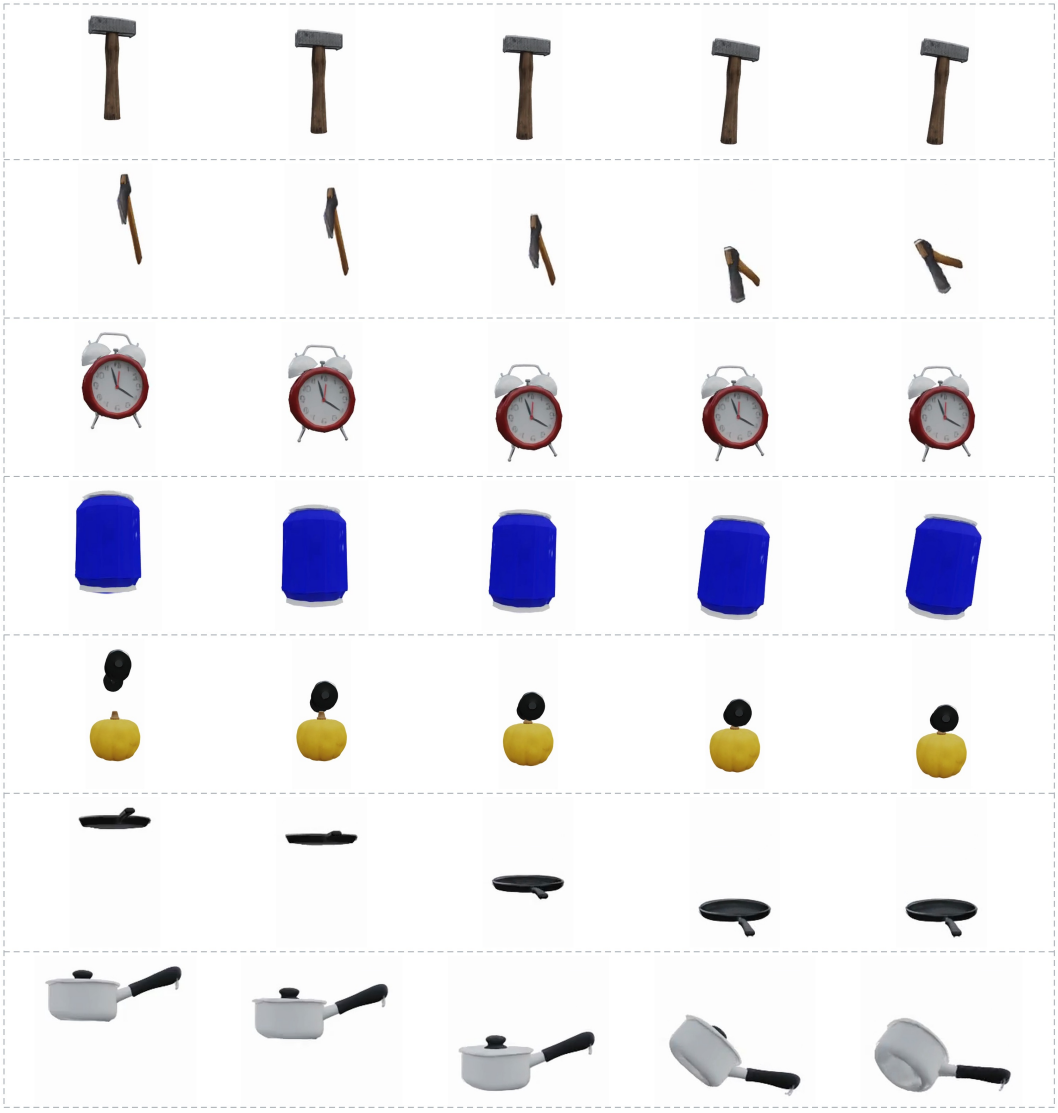

Figure 7: Additional results (II).

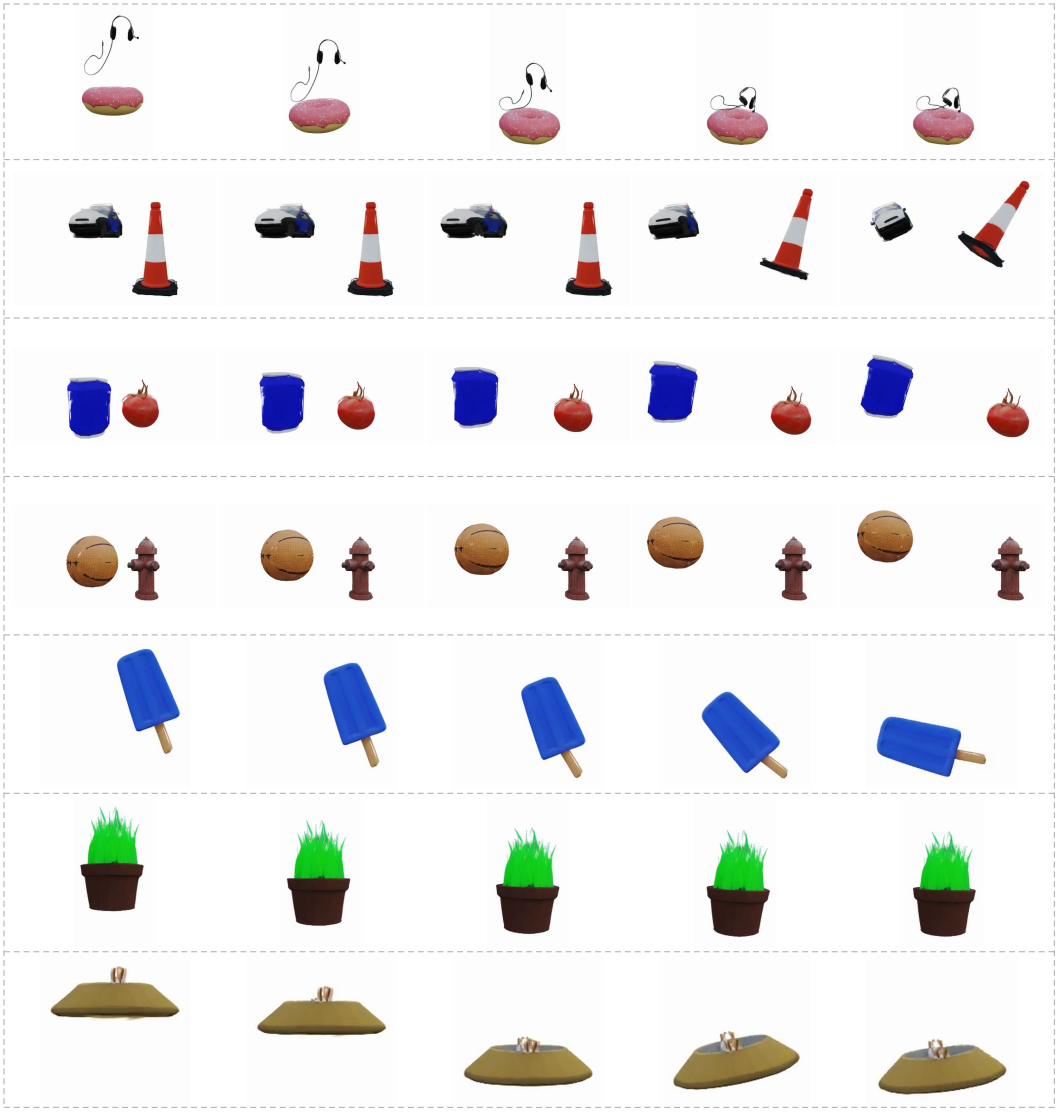

Figure 8: Additional results (III).



Figure 9: Additional results (IV).



Figure 10: Additional results (V).



Figure 11: Additional results (VI).