PHYSFORGE: MATERIAL-CONTROLLABLE 3D RECONSTRUCTION WITH SEPARABLE COMPONENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in physical 3D generation have enabled visually faithful assets with simulatable mechanical behavior. However, existing methods provide limited control over material behavior beyond appearance-based inference and struggle to estimate physical properties from complex interactions between objects and semantic parts. Therefore, we present PHYSFORGE, a Video Diffusion Model (VDM)-based framework for addressing these challenges, comprising two complementary modules for material-controllable physical fitting and component-level interaction modeling. The first, PHYSFIT, leverages material-conditioned VDM dynamics as references for physical-property fitting, introducing windowed fitting to focus on material-revealing interactions and Hallucinated Forcing to disentangle nonphysical motion hallucinated by the VDM from material responses. The second, PHYSCOMP decomposes complex interactions into individual objects and semantic components, enabling their physical parameters to be optimized independently and flexibly composed for joint simulation. Experiments on PhysAssets and Google Scanned Objects demonstrate that, compared with existing methods, PHYSFORGE provides stronger control over material-dependent physical responses while supporting component-level physical modeling for complex interactions. It achieves the best CLIP and physical controllability scores on both common and uncommon materials, with qualitative comparisons, component-level ablations, and a user study further supporting these improvements.

1 INTRODUCTION

Recent advances in 3D generation and reconstruction have made it possible to create assets with increasingly realistic geometry and appearance (Xiang et al., 2026; Liang et al., 2026). For physical simulation, visual fidelity alone is insufficient: meaningful interaction also requires modeling how objects made of different materials respond to forces, contact, and deformation. This requires 3D assets to be endowed with physical properties that govern their mechanical behavior and enable physically meaningful responses to interaction (Zhong et al., 2024; Jiang et al., 2025).

Toward this goal, existing methods augment reconstructed or generated 3D assets with physical properties inferred using Vision–Language Models (VLMs) or discriminative models trained on physical-property data (Cao et al., 2026b;a). However, two key challenges remain. First, appearance-based inference provides limited control over material behavior. Such predictions are largely shaped by appearance–material correlations learned from training data, making it difficult to assign a user-specified material to the same object and obtain the corresponding physical response (Xu et al., 2025; Li et al., 2023). Second, existing methods have limited ability to infer physical properties from complex interactions between objects or semantic parts (Lin et al., 2025b; Xu et al., 2026). In most scenes, material behavior is often manifested through contact and deformation among these interacting entities, whereas existing methods mainly rely on object-level visual observations and struggle to capture such interaction-dependent behavior (Zhang et al., 2024).

To address these limitations, we propose PHYSFORGE, a video diffusion model (VDM)-based framework for constructing physically simulatable 3D assets with separable semantic components, as shown in Figure 1. Our framework consists of two modules, PHYSFIT and PHYSCOMP. PHYSFIT leverages the ability of VDMs to generate material-conditioned dynamics and introduces a robust generated-video fitting mechanism that mitigates the influence of nonphysical motion when estimat-

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.

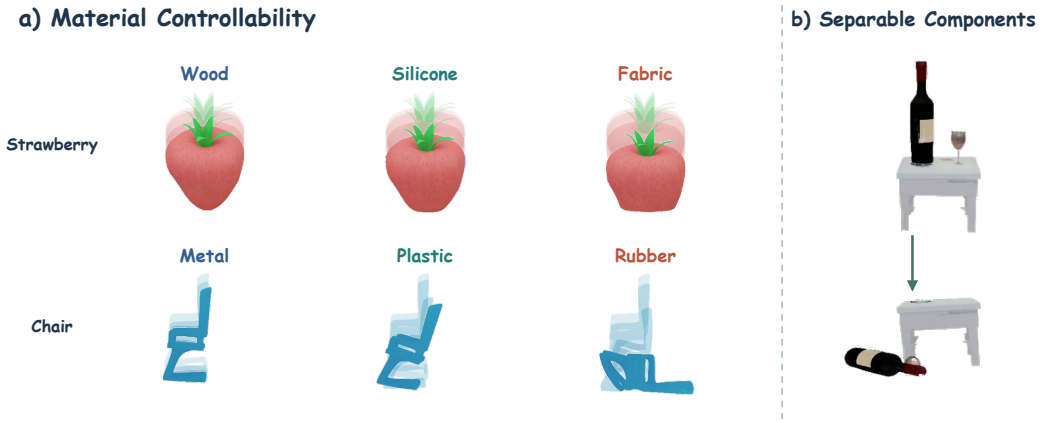


Figure 1: PHYSFORGE achieves material controllability and separable components.

ing physical properties for user-specified materials. PHYSCOMP decomposes complex interactions into individual objects and semantic components, enabling their physical properties to be analyzed separately and flexibly composed for joint simulation.

Specifically, PHYSFIT exploits the ability of VDMs to generate different material-dependent dynamics for the same object and treats these material-conditioned outputs as reference dynamics for physical-property fitting. To further streamline the fitting process, we develop a windowed fitting paradigm that optimizes physical parameters only over informative intervals that reveal material properties, such as contact, collision, and deformation events. An unavoidable limitation is that VDM-generated dynamics does not fully obey physical laws, making direct fitting to the entire generated sequence unreliable. Therefore, we introduce **Hallucinated Forcing** to account for generated motion that cannot be explained solely by the underlying physical parameters while separating it from the material response used for fitting. For complex interactions, PHYSCOMP propagates object–part identities from video to reconstructed 3D geometry, allowing the physical parameters of individual objects and semantic parts to be optimized independently. It then determines simulation groups based on their relative motion, enabling independently optimized components to be composed for joint simulation while preserving their semantic identities.

We conduct extensive experiments on comparing PHYSFORGE to other baseline methods. Qualitative results show that the material controllability and component separability enabled by our design lead to more physically plausible responses and interactions; a human study further finds that participants prefer our results. In quantitative evaluations, PHYSFORGE achieves the best CLIP score and physical controllability score for both common and uncommon materials. Finally, ablation studies demonstrate the effectiveness of PHYSCOMP and PHYSFIT.

Our contributions are as follows:

- We introduce PHYSFORGE, a VDM-based framework for constructing physically simulatable 3D assets with controllable material behavior and separable semantic components, enabling material specification beyond appearance-based physical-property inference.
- We propose PHYSFIT, which leverages material-conditioned VDM dynamics for controllable physical-property estimation and introduces robust generated-video fitting to account for nonphysical motion and focus optimization on physically informative intervals.
- We develop PHYSCOMP to enable independent physical-parameter optimization for individual objects and semantic components, followed by flexible composition into simulation groups based on their relative motion.
- Extensive experiments show that PHYSFORGE generates visually faithful 3D assets with plausible material-dependent responses. Ablation studies further validate the effectiveness of PHYSFIT and PHYSCOMP.

2 RELATED WORK

2.1 PHYSICAL-PROPERTY ESTIMATION

Observation-Driven Inference. In common cases, physical properties can be inferred from an object's appearance or identified from its observed motion and deformation. GaussianProperty uses segmentation and multimodal reasoning to assign physical attributes, while PhysGM jointly predicts geometry and physical parameters (Xu et al., 2025; Lv et al., 2026). PhysX-3D and PhysX-Anything further incorporate physical attributes into asset generation and reconstruction (Cao et al., 2026a;b). Complementing these visual and semantic cues, PAC-NeRF couples neural reconstruction with continuum simulation for parameter identification, and Spring-Gaus recovers deformable dynamics with a spring-based Gaussian representation (Li et al., 2023; Zhong et al., 2024). Static images alone provide limited evidence of how an object responds to forces and contact. PHYSFORGE therefore uses material-conditioned motion, collisions, and deformation as dynamic evidence, providing additional constraints for more accurate physical-property estimation.

Video-Prior-Guided Fitting. Video generation provides dynamic priors for physical modeling. MotionCraft, PhysGen, and PSIVG incorporate simulation guidance (Montanaro et al., 2024; Liu et al., 2024b; Foo et al., 2026), while PhyT2V uses physics-oriented prompt refinement (Xue et al., 2025). For physical-property estimation, PhysDreamer fits parameters to generated motion, while DreamPhysics and Physics3D distill video diffusion priors through simulation (Zhang et al., 2024; Huang et al., 2025; Liu et al., 2024a). PhyMAGIC further incorporates motion-aware frame selection and confidence-guided reasoning (Meng et al., 2026). Fitting an entire generated sequence can mix material-dependent responses with nonphysical motion, since generated videos can violate physical constraints (Bansal et al., 2025). In particular, we observe Hallucinated Forcing that produces visually plausible but physically incorrect motion, which can bias the fitted material parameters. PHYSFORGE uses window selection to fit physical-property estimation which limits the influence of this nonphysical motion.

2.2 COMPONENT-AWARE 3D ASSET CONSTRUCTION

Semantic decomposition and component reconstruction. Component-aware reconstruction makes an object's internal structure available for modelling. PartField and P3-SAM recover semantic regions in 3D assets (Liu et al., 2025; Ma et al., 2025), while PartCrafter and PartGen generate decomposed geometry with coherent component structure (Lin et al., 2026; Chen et al., 2025). Geometric decomposition alone does not establish how the resulting components respond to physical interactions. PHYSFORGE therefore associates semantic components with their observed dynamic responses, enabling extracted components to retain both their geometry and corresponding physical properties from both semantic recognition and dynamic priors.

Physical modeling of object and component interactions. PhysGaussian couples Gaussian appearance with continuum mechanics (Xie et al., 2024), PhysTwin reconstructs interactive physical twins from observations (Jiang et al., 2025), and OmniPhysGS learns constitutive behavior under video guidance (Lin et al., 2025b). DecoupledGaussian separates objects from contacting scene surfaces to support physical interactions (Wang et al., 2025). When multiple objects or components interact, their coupled motion and deformation can obscure which material properties produce the observed response. PHYSFORGE therefore combines separate material estimation with joint physical fitting to account for both individual component behavior and interactions.

3 METHOD

Our goal is to reconstruct physically simulatable 3D assets with controllable material behavior and separable semantic components. To achieve this goal, PHYSFORGE couples two complementary modules: PHYSFIT uses material-conditioned VDM dynamics to fit physical properties while accounting for nonphysical motion, and PHYSCOMP decomposes the scene into persistent objects and semantic components for separate estimation and joint simulation. Their outputs are combined to produce components that can be composed in a shared physical simulation. We first introduce 3DGS and the particle–grid mechanics underlying the component-level simulator, then

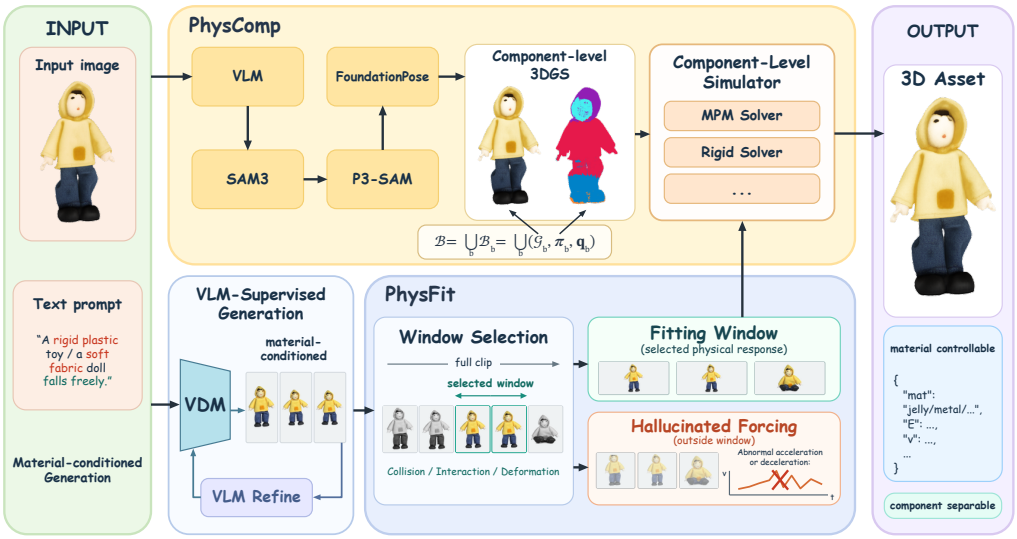


Figure 2: **Overview of PHYSFORGE.** The framework comprises PHYSFIT and PHYSCOMP. PHYSFIT receives a material-conditioned video generated by a VLM-supervised VDM. The generated video is analyzed to select windows for physical-property fitting. PHYSCOMP constructs component-level 3DGS through semantic recognition, 3D reconstruction, component segmentation, and scene placement. Each component is associated with material properties and an initial state, and is then fitted within the windows selected by PHYSFIT.

present video-guided physical-property fitting and component-aware physical modeling; Figure 2 summarizes the framework.

3.1 PRELIMINARY

3.1.1 3D GAUSSIAN SPLATTING

We use 3D Gaussian Splatting (3DGS) as the appearance representation for every reconstructed object (Kerbl et al., 2023). A 3D Gaussian primitive $g_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \boldsymbol{c}_i)$ has center $\boldsymbol{\mu}_i \in \mathbb{R}^3$, positive-definite covariance $\boldsymbol{\Sigma}_i$, opacity α_i, and view-dependent color coefficients $\boldsymbol{c}_i$. With rotation $\mathbf{R}_i$ and diagonal scale $\mathbf{S}_i$,

$$\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^\top \mathbf{R}_i^\top. \tag{1}$$

For a camera projection π, the projected covariance is $\boldsymbol{\Sigma}_i^{2\mathrm{D}} = \mathbf{J}_i \boldsymbol{\Sigma}_i \mathbf{J}_i^\top$, where $\mathbf{J}_i$ is the projection Jacobian. With $G_i(\mathbf{p})$ denoting the projected Gaussian kernel, front-to-back compositing gives

$$I(\mathbf{p}) = \sum_i T_i(\mathbf{p}) \alpha_i G_i(\mathbf{p}) \mathbf{c}_i, \qquad T_i(\mathbf{p}) = \prod_{j<i} \big(1 - \alpha_j G_j(\mathbf{p})\big). \tag{2}$$

Each Gaussian stores both an object identifier and a semantic component identifier. A separate mapping assigns these labelled components to simulated bodies, allowing their semantic identities to persist under different simulation groupings.

3.1.2 MATERIAL POINT METHOD

The material point method (MPM) represents a deformable body by particles that carry position $\mathbf{x}_p$, velocity $\mathbf{v}_p$, mass m_p, and deformation gradient $\mathbf{F}_p$ (Sulsky et al., 1994; Hu et al., 2018). At each time step, particle-to-grid (P2G) transfer maps particle mass and momentum to grid nodes, where material forces, external forces, and contact update the nodal velocities. Grid-to-particle (G2P) transfer then interpolates the updated grid velocities back to the particles as $\mathbf{v}_p^{t+1} = \sum_i w_{ip}^t \mathbf{v}_i^{t+1}$

and advances their positions as $\mathbf{x}_p^{t+1} = \mathbf{x}_p^t + \Delta t\, \mathbf{v}_p^{t+1}$, where w_{ip}^t is the interpolation weight between particle p and grid node i, and Δt is the time-step duration. The deformation gradient is updated by

$$\mathbf{F}_p^{t+1} = \left(\mathbf{I} + \Delta t \sum_i \mathbf{v}_i^{t+1} (\nabla w_{ip}^t)^\top \right) \mathbf{F}_p^t. \tag{3}$$

The resulting particle positions, velocities, and deformation gradients become the physical state for the next time step.

3.2 PHYSFIT: VIDEO-GUIDED PHYSICAL-PROPERTY FITTING

For a requested material and interaction, a pretrained video diffusion model (Wan et al., 2025) generates a material-conditioned sequence $\mathcal{V} = \{I_t\}_{t=0}^{T-1}$. A VLM assesses the generated video and provides feedback to refine the generation (Long et al., 2026; Feng et al., 2026). Existing methods fit entire video sequences, including intervals that provide little evidence of the object's physical properties, which leads to suboptimal results. Therefore, PHYSFIT focuses on selecting physically informative windows and assigning an initial state to each window.

3.2.1 WINDOWED PHYSICAL-PROPERTY FITTING

To select physically informative windows, we identify intervals in the generated video where interactions between semantic components produce motion or deformation that reveals their material properties. We treat the scene boundary as a component, so every candidate interaction of $\mathcal{C}$ is described by the mask intersection between $M_{\mathcal{C}}$ and the mask $M_{\mathcal{C}'}$ of another component $\mathcal{C}'$.

For each candidate pair $(\mathcal{C}, \mathcal{C}')$ at frame t, we define the contact indicator $\chi_t^{\mathcal{C}\mathcal{C}'} = 1$ when $M_{t,\mathcal{C}} \cap M_{t,\mathcal{C}'} \neq \emptyset$ and zero otherwise. The first transition to $\chi_t^{\mathcal{C}\mathcal{C}'} = 1$ identifies contact; the window starts at the last frame in the preceding approaching run for which the masks are disjoint and the relative center motion indicates approach, thereby preserving the incoming state. The window ends when the interacting objects come to rest or the contact indicator returns to $\chi_t^{\mathcal{C}\mathcal{C}'} = 0$.

For such an interaction window w, we define its temporal frames as $\mathcal{W}_w = [t_w^{\text{in}}, t_w^{\text{out}}]$. The component set involved in the interaction is $\{\mathcal{C}\} \cup \bigcup_{t=t_w^{\text{in}}}^{t_w^{\text{out}}} \{\mathcal{C}' \,|\, \mathcal{C}' \neq \mathcal{C},\ M_{t,\mathcal{C}} \cap M_{t,\mathcal{C}'} \neq \emptyset\}$. These components are organized into simulation bodies by PHYSCOMP, and the resulting body set is denoted by $\mathcal{B}_w$. A complete fitting event can thus be defined as $\big(\mathcal{W}_w, \mathcal{B}_w\big)$, which is used to fit the physical parameters through joint material refinement in Section 3.3.2.

3.2.2 REFERENCE MOTION AND HALLUCINATED FORCING

Window selection determines where material parameters are fitted, but does not by itself provide a reference-consistent incoming state. We therefore recover the reference motion explicitly: for each semantic object, SAM3 (Carion et al., 2026) provides the temporally aligned mask and Depth Anything (Lin et al., 2025a) provides the corresponding depth, from which we track the center trajectory $\widehat{\mathbf{c}}_t$. Given the video frame rate fps, the frame interval is $\Delta t = 1/\text{fps}$, the tracked velocity is $\widehat{\mathbf{v}}_t = (\widehat{\mathbf{c}}_{t+1} - \widehat{\mathbf{c}}_t)/\Delta t$, and the apparent acceleration used to reproduce the next center is $\mathbf{a}_t^{\text{ref}} = (\widehat{\mathbf{c}}_{t+1} - \mathbf{c}_t - \Delta t \mathbf{v}_t)/(\Delta t)^2$.

Based on the above analysis, we use Hallucinated Forcing ($\mathbf{a}_t^{\text{ref}}$) to denote abnormal velocity changes introduced by a nonphysical generation process, such as unexplained acceleration or deceleration. Applying the corresponding velocity update produces a reference trajectory whose tracked positions and velocities follow the generated video, while the nonphysical motion remains distinguishable from the material response used for fitting.

3.3 PHYSCOMP: COMPONENT-AWARE PHYSICAL MODELING

Existing methods often simulate each segmented object as a single body, suppressing relative motion and interactions among its semantic components. Therefore, PHYSCOMP focuses on component-level identification and segmentation, organizing the resulting components into simulation bodies to preserve their distinct motion and physical interactions.

3.3.1 OBJECT–COMPONENT TRACKING, RECONSTRUCTION, AND PLACEMENT

To obtain component-level semantic information, a vision–language agent converts generated frames into object-level and component inventories. SAM3 (Carion et al., 2026) initializes and propagates the corresponding masks across frames and views, preserving identities for repeated or attached components. The masked views are converted into a 3DGS representation $\mathcal{G}$ using SAM3D (Chen et al., 2026), TRELLIS (Xiang et al., 2025), or other 3D generation or reconstruction backends. Depth Anything (Lin et al., 2025a) provides depth and camera-consistent scale; P3-SAM (Ma et al., 2025) decomposes the reconstructed geometry into semantic component regions and transfers their labels to the corresponding Gaussians. FoundationPose (Wen et al., 2024) restores the object pose and scale and replaces the labelled representation in the shared scene coordinate system, yielding a 3DGS asset from which each semantic component can be extracted.

3.3.2 SIMULATION BODY ASSIGNMENT AND JOINT MATERIAL REFINEMENT

To prepare the reconstructed semantic components for physical simulation, we organize them into simulation bodies according to their attachment relationships and relative motion. Each simulation body is represented as $\mathcal{B}_b = (\mathcal{G}_b, \boldsymbol{\pi}_b, \mathbf{q}_b)$, where $\mathcal{G}_b$ is its labelled 3DGS subset, $\boldsymbol{\pi}_b$ is its material–physical-parameter pair, and $\mathbf{q}_b$ is its dynamic state. The labelled Gaussians are converted to surface and interior particles, all bodies are assembled in the shared component-level simulator, and each window initializes $\mathbf{q}_b$ before the state is propagated through the response.

To fit the observed physical event within each window and discourage interpenetration, we use a physical regulation loss. Let $c_{w,b}$ and $\widehat{c}_{w,b}$ denote the observed and simulated deformation responses, and let T_w, $\widehat{T}_w$, and p_w denote the observed contact duration, simulated contact duration, and simulated penetration. The physical regulation loss is

$$\mathcal{L}_{\text{phys}} = \sum_{w,b} (\widehat{c}_{w,b} - c_{w,b})^2 + \lambda_T \sum_{w} (\widehat{T}_w - T_w)^2 + \lambda_P \sum_{w} p_w^2. \tag{4}$$

We further use the initial material estimate predicted by VLM as a prior while refining physical parameters from collision responses in the video. Let $\boldsymbol{\theta}_b$ denote the physical-parameter vector in $\boldsymbol{\pi}_b$ and $\boldsymbol{\theta}_b^0$ its initial estimate. The parameter regularizer softly penalizes deviations from this prior:

$$\mathcal{L}_{\text{reg}} = \sum_{b} \left\| \boldsymbol{\theta}_b - \boldsymbol{\theta}_b^0 \right\|_2^2. \tag{5}$$

A low weight on this term discourages unreasonable parameter changes caused by noise or insufficient observations while allowing deviations from the initial estimate when supported by the observed response. We optimize the weighted total loss

$$\mathcal{L}_{\text{total}} = \lambda_{\text{phys}} \mathcal{L}_{\text{phys}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \tag{6}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Implementation. We evaluate PHYSFORGE on PhysAssets (Lv et al., 2026) and Google Scanned Objects (GSO) (Downs et al., 2022). PhysAssets contains over 24,000 3D assets annotated with physical properties and accompanied by guiding videos, providing diverse objects for evaluating material-dependent dynamics. GSO provides over 1,000 high-quality 3D scans of household objects, capturing realistic geometry and appearance across diverse object categories. Our framework is implemented on two NVIDIA RTX 4090 GPUs.

Baselines. Our qualitative and quantitative comparisons include 6 baselines: DreamPhysics (Huang et al., 2025), PSIVG (Foo et al., 2026), PhysRVG (Zhang et al., 2026), WMReward (Yuan et al., 2026), DeforM (Li et al., 2026), and PhysGM (Lv et al., 2026). DreamPhysics learns material properties from video diffusion priors through Score Distillation Sampling (SDS) and MPM simulation. PhysGM jointly predicts 3D Gaussians and physical properties from multiview images, whereas PSIVG guides video generation with simulated trajectories reconstructed from an initial generated video. DeforM uses VLM reasoning and spatial–temporal masks to focus generation on regions

Table 1: Comparison between PHYSFORGE and baseline methods.

Method	Common Materials		Uncommon Materials	
	CLIP ↑	Physical Controllability ↑	CLIP ↑	Physical Controllability ↑
DreamPhysics	0.250	3.2	0.248	2.6
PSIVG	0.233	2.0	0.241	1.8
PhysRVG	0.243	3.8	0.232	3.0
WMReward	0.246	4.4	0.244	2.6
DeforM	0.242	4.4	0.235	3.2
PhysGM	0.246	3.8	0.247	1.8
PHYSFORGE	**0.258**	**4.6**	**0.264**	**4.0**

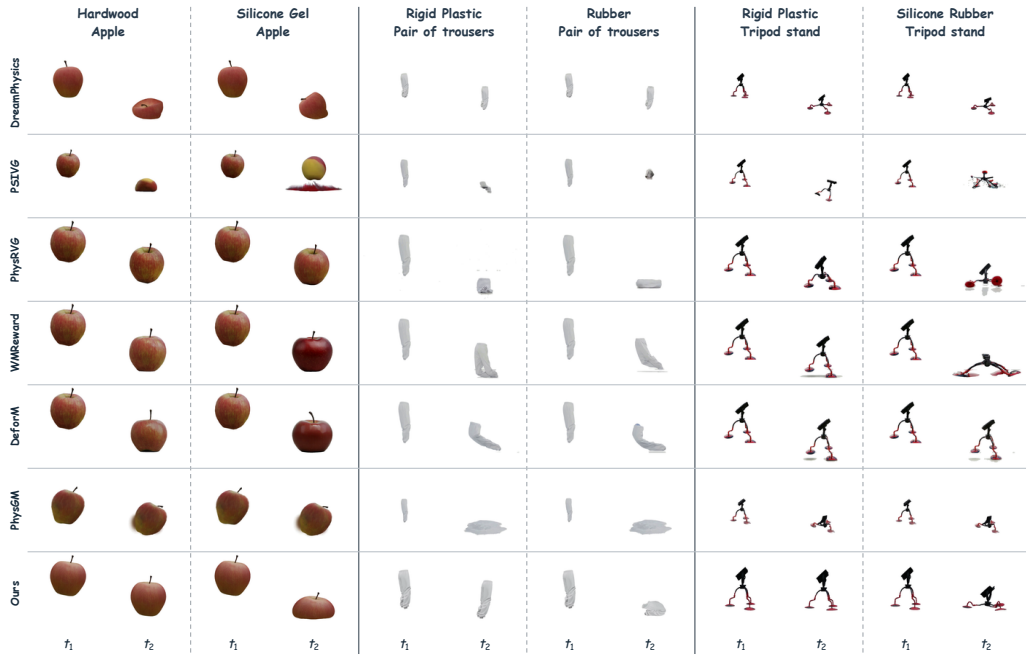


Figure 3: Comparison between PHYSFORGE and baseline methods on material conditions.

undergoing physical deformation. PhysRVG incorporates physics-aware reinforcement learning to improve collision dynamics, and WMReward uses a latent world model to score and steer candidate denoising trajectories at inference time.

Simulation. Our component-level simulator uses Taichi-based MPM and rigid-body solvers (Hu et al., 2018; 2019). For rigid bodies, contact friction controls sliding and restitution controls rebound while the shape remains fixed. For MPM, we generate 150-frame videos at 30 FPS with a maximum of 8 substeps per frame. Elastic bodies use Young's modulus to control stiffness and Poisson's ratio to control lateral deformation under compression or stretching. Sand uses Young's modulus and Poisson's ratio for its elastic response and an internal friction angle for Drucker–Prager yielding, which determines its resistance to shearing and spreading. Liquids use a weakly compressible fluid model in which the bulk modulus controls the pressure response to compression.

Evaluation Metrics. We evaluate text alignment, physical material controllability, and human preference. The CLIP score uses ViT-L/14 to encode the text prompt and video frames, and measures their cosine similarity averaged over sampled frames to assess semantic agreement with the requested content. Physical controllability (PC), rated on a five-point scale by VLM, assesses whether the simulated motion, contact, and deformation express the requested material behavior. For the user study, each trial presents seven anonymized videos from PHYSFORGE, DreamPhysics (Huang et al., 2025), PSIVG (Foo et al., 2026), PhysRVG (Zhang et al., 2026), WMReward (Yuan et al., 2026), DeforM (Li et al., 2026), and PhysGM (Lv et al., 2026) in randomized order, together with the text



Figure 4: Comparison between PHYSFORGE and baseline methods on component separation.

prompts and input images. We also conduct user study on qualitive results. Participants rank the videos by how well they match the requested material while maintaining plausible dynamics.

4.2 RESULTS AND ANALYSIS

Qualitative Comparison with Baselines. We compare PHYSFORGE with the baselines on material conditions in Figure 3 and component separation in Figure 4. Figure 3 shows that the baselines often exhibit similar physical responses under different material conditions. In the tripod case, although PhysRVG exhibits different material behaviors, it fails to maintain appearance consistency. In the silicone-gel apple case, WMReward and DeforM show some deformation, but the apple's appearance also changes. By fitting physical parameters within selected windows, PHYSFIT extracts material-conditioned physical priors from VDM-generated videos while excluding nonphysical motion, yielding 3D assets whose simulated behavior better matches the requested material.

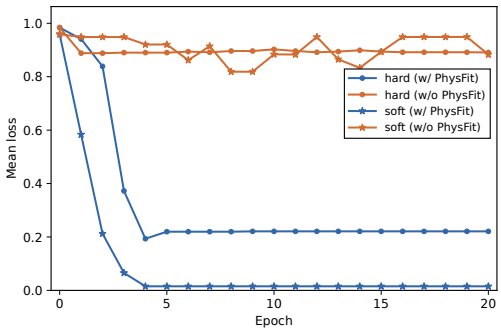
Figure 4 shows that the baselines struggle to preserve the separability and relative motion of components, often treating them as a single body during motion and losing their relative movement. PhysGM, which infers material from the input image alone, incorrectly assigns sand-like behavior to the pillow. PhysRVG shows some relative motion in the same case, but introduces an extra component. PSIVG generates multiple apparent components but fails to preserve their original appearance, making its output inconsistent with the input image and prompt. PHYSCOMP identifies objects and segments their semantic components, allowing each component to retain its own physical properties and motion state during simulation which produces more realistic interactions.

Quantitative Comparison with Baselines. Table 1 reports CLIP scores and PC scores for common and uncommon materials. PHYSFORGE leads all four reported metrics, with an average PC of 4.3 versus 3.8 for the strongest baseline and an average CLIP score of 0.261 versus 0.249 for the strongest baseline. Baseline methods such as WMReward and DeforM show high physical controllability for common materials, however, their performance drops sharply for uncommon user-specified materials. The user study in Table 2 also reflects the advantages of PHYSFORGE in material controllability and component separability. PHYSFIT selects physically informative fitting windows, while PHYSCOMP segments component-level 3D assets; together, they enable PHYSFORGE to produce physically controllable 3D assets with separable components.

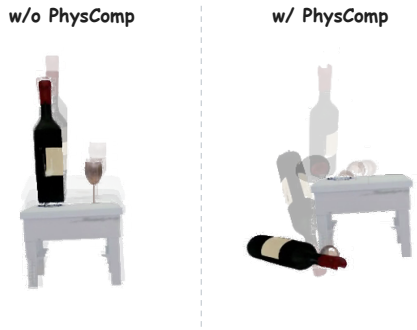
Ablation Studies. We conduct two ablations on our key mechanisms: PHYSFIT and PHYSCOMP.

Table 2: User-study rankings across PHYSFORGE and baseline methods.

	DreamPhysics	PSIVG	PhysRVG	WMReward	DeforM	PhysGM	PHYSFORGE
Win Rate (%) ↑	5.56	0.56	8.33	12.22	2.78	10.00	**60.56**
Average Rank ↓	4.2778	6.2167	3.9500	3.5444	3.6833	4.4222	**1.9056**



(a) Ablation study for PHYSFIT.



(b) Ablation study for PHYSCOMP.

Figure 5: Ablation study for PHYSFORGE.

▪ PHYSFIT. We remove PHYSFIT, fitting physical parameters to the entire generated video. We compare quantitative fitting error and convergence speed using loss-versus-optimization-step curves in Figure 5a. Removing PHYSFIT prevents the fitting process from focusing on the windows that reveal physical properties. Over the longer sequence, the loss function struggles to identify the intervals that need to be fitted, which reduces fitting performance and increases fitting time. Therefore, PHYSFIT enables better physical-property fitting for 3D assets.

▪ PHYSCOMP. Removing PHYSCOMP merges semantically distinct components into a single simulated body, as shown in Figure 5b. Viewed as one connected object, the relative motion and interactions among components disappear, making the dynamics physically implausible. With PHYSCOMP, table, bottle and glass retain their own physical properties and motion state. These components that appear connected yet have distinct semantic identities can then collide and wobble upon contact, revealing their individual physical behavior and yielding more plausible interactions.

5 CONCLUSION

In this paper, we analyzed existing methods and found that insufficient material controllability and semantic component separability in reconstructed 3D assets lead to unrealistic physical simulations. To address these limitations, we proposed PHYSFORGE with two key modules: PHYSFIT and PHYSCOMP. PHYSFIT improved the efficiency of physical-property fitting through response-window selection and handling of Hallucinated Forcing, while PHYSCOMP used dynamic and semantic information to decompose reconstructed 3D assets so that individual components retain distinct material properties. Together, these modules enable the reconstruction of different material-conditioned 3D assets from the same input object under different material prompts, while keeping their semantic components extractable. Quantitative and qualitative evaluations show that material controllability and component separability yield more realistic simulations of 3D assets than existing methods. Ablation studies further confirm the effectiveness of both modules.

However, PHYSFORGE still has some limitations. PHYSCOMP uses generated videos to segment objects based on their dynamic and semantic information. Changing the material condition requires additional video generation and physical fitting for each material-condition change. The performance of PHYSFORGE also depends on how well the generated video represents physical behavior, although PHYSFIT can mitigate hallucinations from video generation to some extent. In future works, more advanced non-physical motion separation techniques can be employed to provide a cleaner window with less nonphysical motion for physical parameter fitting.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist with literature retrieval, provide feedback on experiments, assist with method implementation, translation, and support the interpretation of experimental results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. *Advances in Neural Information Processing Systems*, 38:93771–93784, 2026a.

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-anything: Simulation-ready physical 3d assets from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5839–5848, 2026b.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5881–5892. IEEE, 2025.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 international conference on robotics and automation (ICRA)*, pp. 2553–2560. Ieee, 2022.

Yuxiang Feng, Juncheng Wang, Chao Xu, Yijie Qian, Huihan Wang, Wenlong Hou, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. Newton: Agentic planning for physically grounded video generation. *arXiv preprint arXiv:2605.18396*, 2026.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4301–4311, 2026.

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

Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019.

Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson WH Lau. Dreamphysics: Learning physics-based 3d dynamics with video diffusion priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3733–3741, 2025.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7219–7230. IEEE, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *International Conference on Learning Representations*, 2023.

Yunyi Li, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Deform: Reasoning-guided physics-aware video generation via spatial-temporal masking. In *European Conference on Computer Vision*, pp. 455–471. Springer, 2026.

Xinyue Liang, Zhiyuan Ma, Lingchen Sun, Yanjun Guo, and Lei Zhang. Photo3d: Advancing photorealistic 3d generation through structure-aligned detail enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34237–34247, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. Omniphysgs: 3d constitutive gaussians for general physics-based dynamics generation. In *International Conference on Learning Representations*, volume 2025, pp. 70160–70187, 2025b.

Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Feng Yiqiang, Yadong Mu, and Katerina Fragkiadaki. Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers. *Advances in Neural Information Processing Systems*, 38:35387–35415, 2026.

Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024a.

Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9704–9715. IEEE, 2025.

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

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö Arik. Vista: A test-time self-improving video generation agent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6021–6032, 2026.

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, 2026.

Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-sam: Native 3d part segmentation. *arXiv preprint arXiv:2509.06784*, 2025.

Si-Wei Meng, Ya-Wei Luo, and Ping Liu. Phymagic: Physical motion-aware generative inference with confidence-guided llm. In *European Conference on Computer Vision*. Springer, 2026.

Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.

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.

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

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

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480. IEEE, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

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

Qingshan Xu, Jiao Liu, Shangshu Yu, Yuxuan Wang, Yuan Zhou, Junbao Zhou, Jiequan Cui, Yew-Soon Ong, and Hanwang Zhang. Neuspring: Neural spring fields for reconstruction and simulation of deformable objects from videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11361–11369, 2026.

Xinli Xu, Wenhang Ge, Dicong Qiu, ZhiFei Chen, Dongyu Yan, Zhuoyun Liu, Haoyu Zhao, Hanfeng Zhao, Shunsi Zhang, Junwei Liang, et al. Gaussianproperty: Integrating physical properties to 3d gaussians with lmms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7231–7240. IEEE, 2025.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18826–18836. IEEE, 2025.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. Physrvg: Physics-aware unified reinforcement learning for video generative models. *arXiv preprint arXiv:2601.11087*, 2026.

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

Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, pp. 407–423. Springer, 2024.

A Appendix

A.1 Implementation Details

We use Qwen3.5-4B for object and component analysis and initial material estimation, with SAM3 providing framewise masks. Depth Anything 3 processes at most 24 images per inference call, including shared reference frames spanning the video. Each chunk is aligned to these references using high-confidence depth pixels, and overlapping predictions are blended. Physical simulation of MPM uses a shared grid with 100–300 cells per axis, selected according to scene extent, object size, particle spacing, and motion constraints. Each video frame is integrated with eight substeps, giving $\Delta t = 1/(8f)$ for frame rate f. Material fitting runs for up to 20 rounds over the detected collision events, initialized from the positions and velocities obtained by video-force fitting. Young's modulus is updated from deformation errors, while candidate selection also accounts for contact duration, penetration, and a material-prior regularizer weighted by 0.01. Particle volumes are assigned by dividing each body's particle-cloud convex-hull volume by its particle count, and particle masses follow from the assigned density. Ground contact uses a planar boundary with restitution 0.5 and Coulomb friction coefficient 0.2.

A.2 More Results on Material-Controllable

We present further image results for our method on material-controllable in Figure 6–11. Figure 6–9 demonstrate the results produced by the same object under prompts specifying different materials. Figure 10 and 11 shows that our method enables material control for liquid and sand. You can refer to the supplementary materials for video and additional examples.

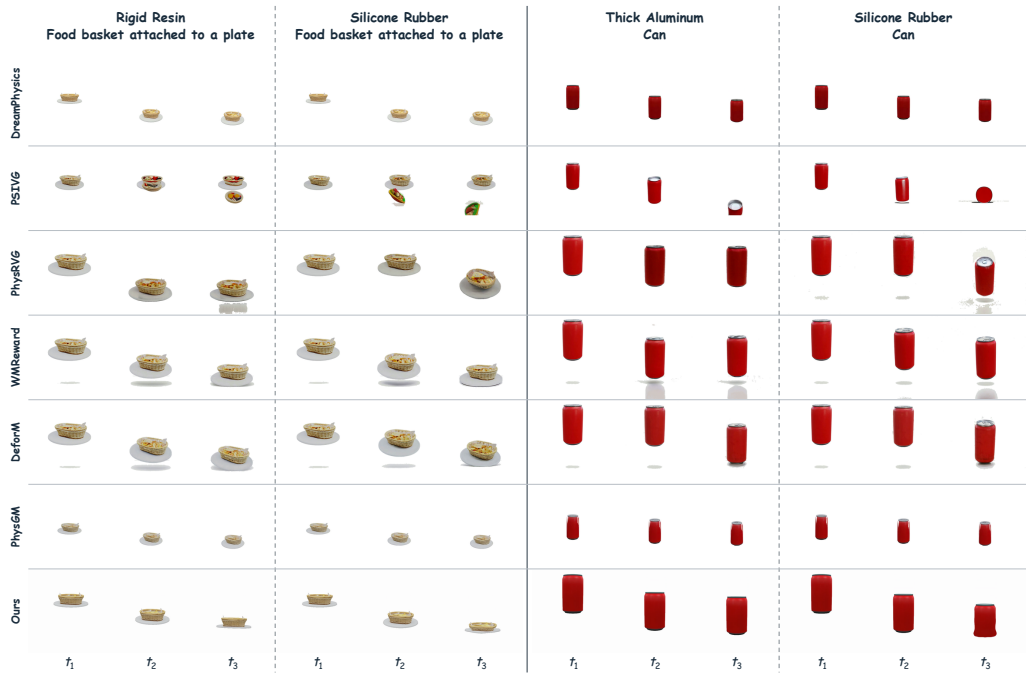


Figure 6: Additional material-controllable comparisons.

A.3 More Results on Components Separable

Figure 12 presents further examples of separable components. We can observe relative motion occurring between a cup and its saucer, a pot and its lid, and two books. Please refer to the supplementary materials for clearer video and additional examples.

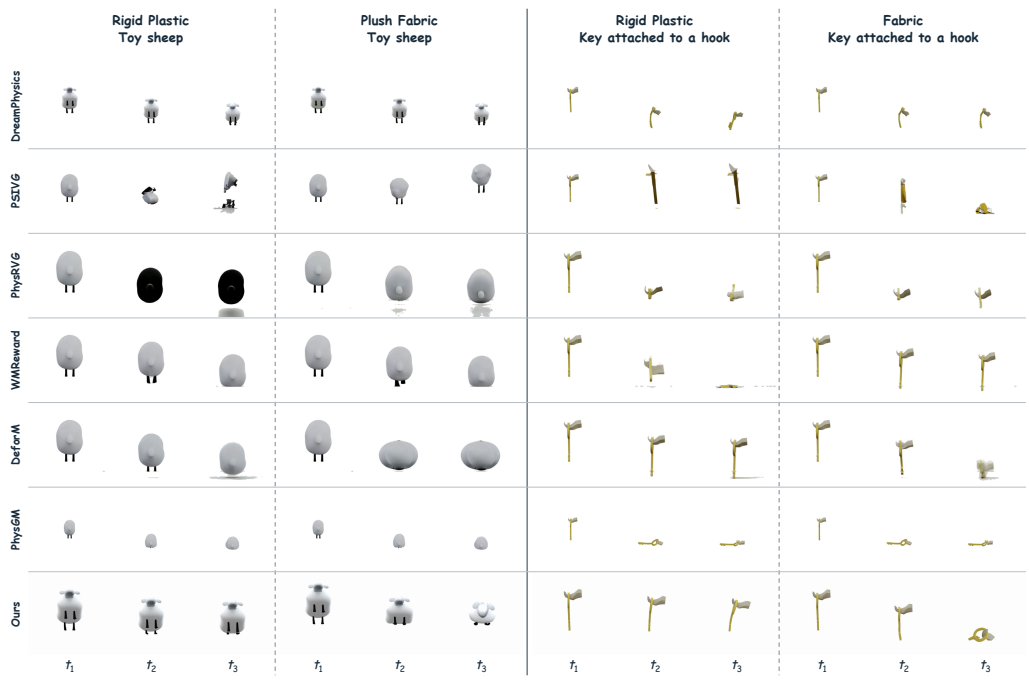


Figure 7: Additional material-controllable comparisons.

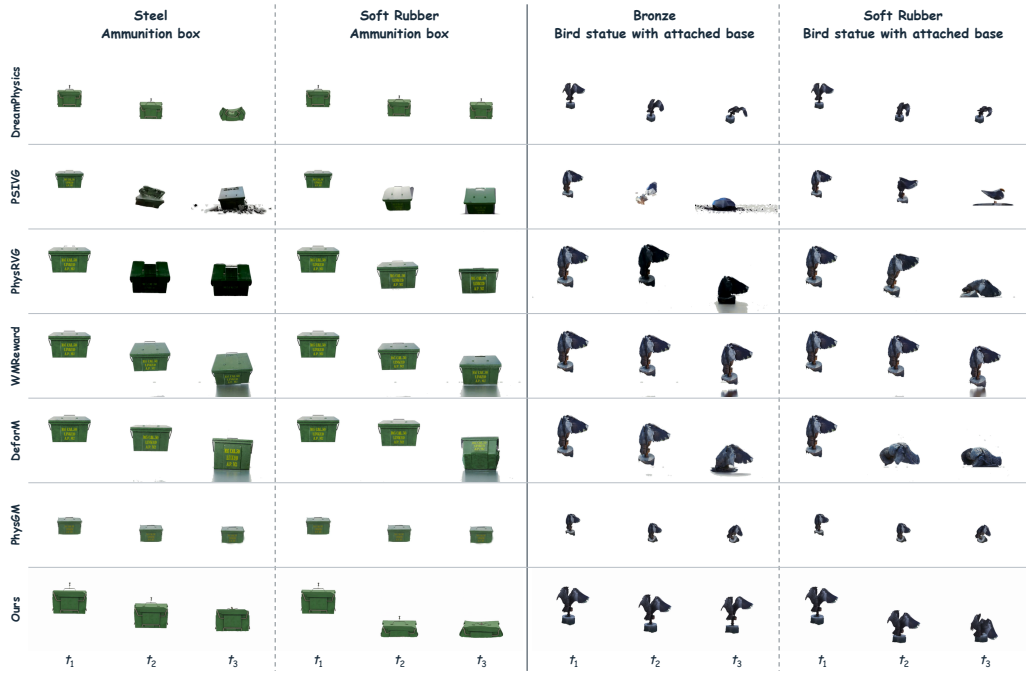


Figure 8: Additional material-controllable comparisons.

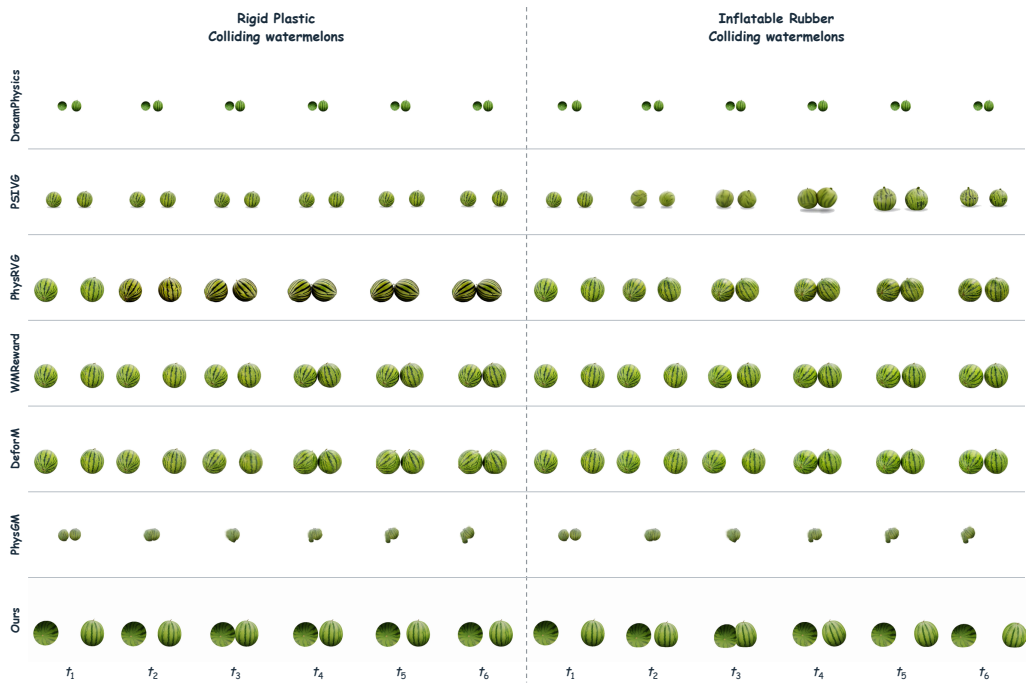


Figure 9: Additional material-controllable comparisons.

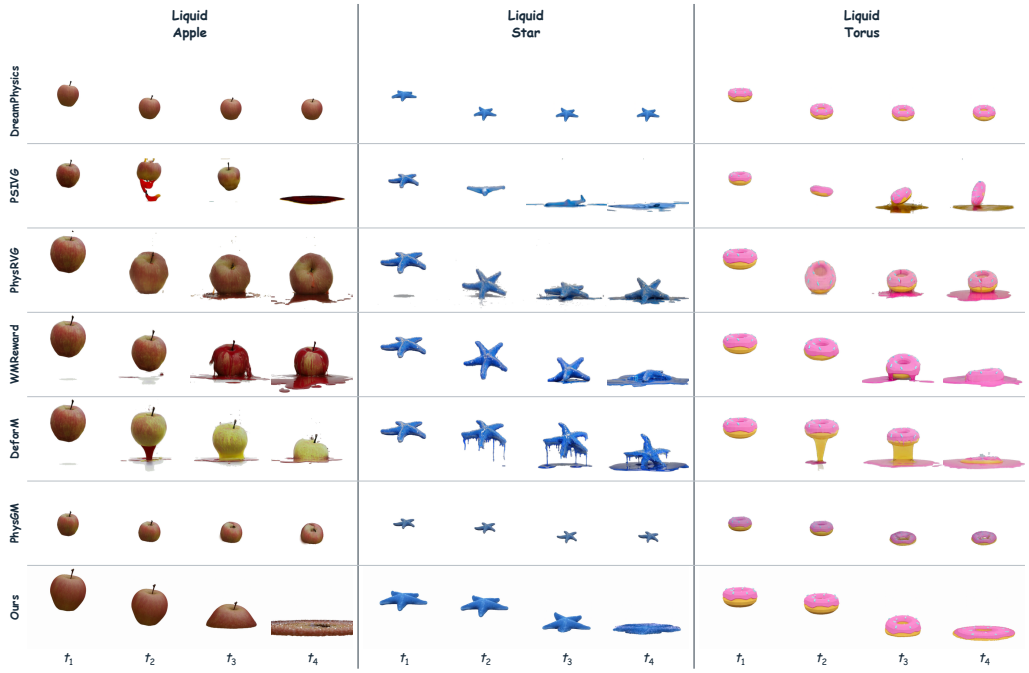


Figure 10: Additional material-controllable comparisons on liquid.

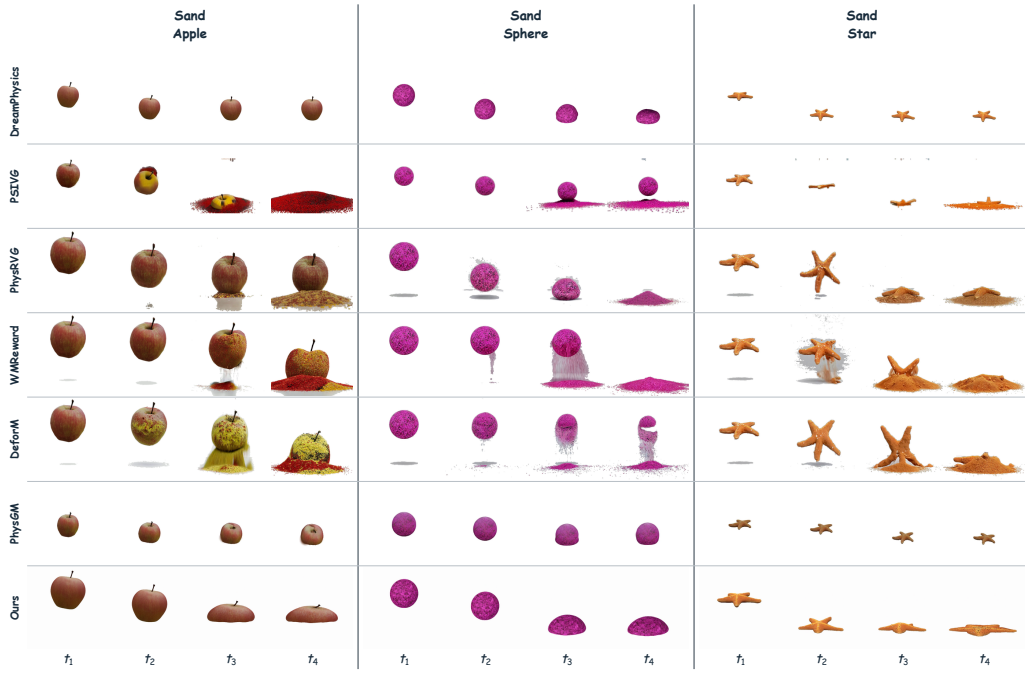


Figure 11: Additional material-controllable comparisons on sand.

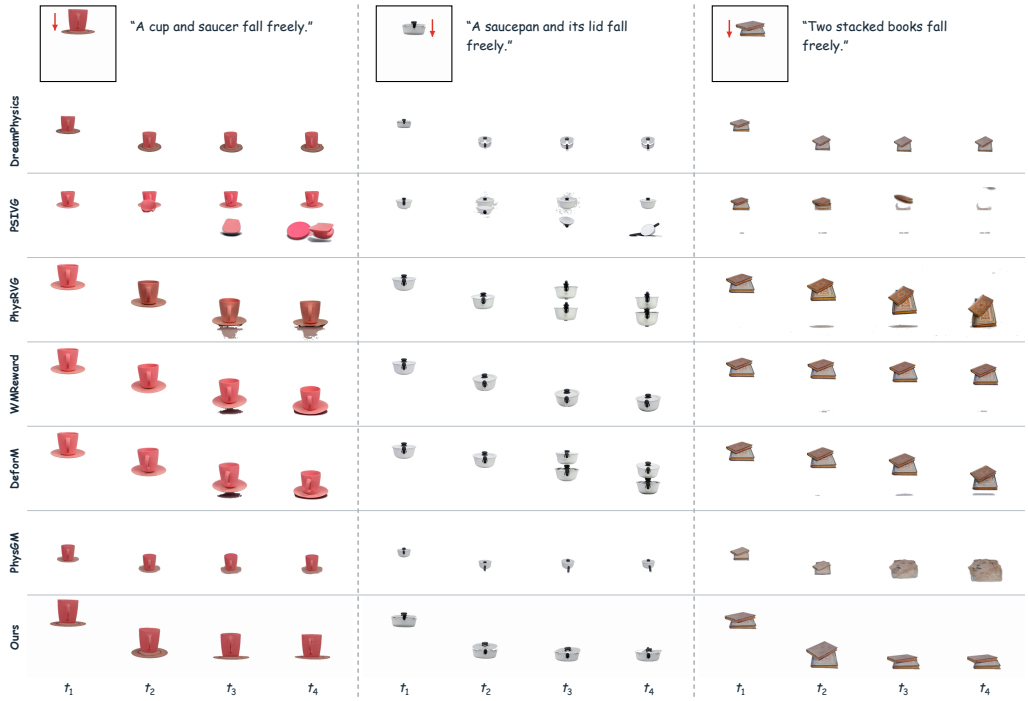


Figure 12: Additional component separable comparisons.