PHYSWONDER: BUILDING SIMULATION-READY WORLDS WHERE DEFORMATION DYNAMICS REVEAL CONSTITUTIVE RESPONSE

Anonymous authors
Paper under double-blind review

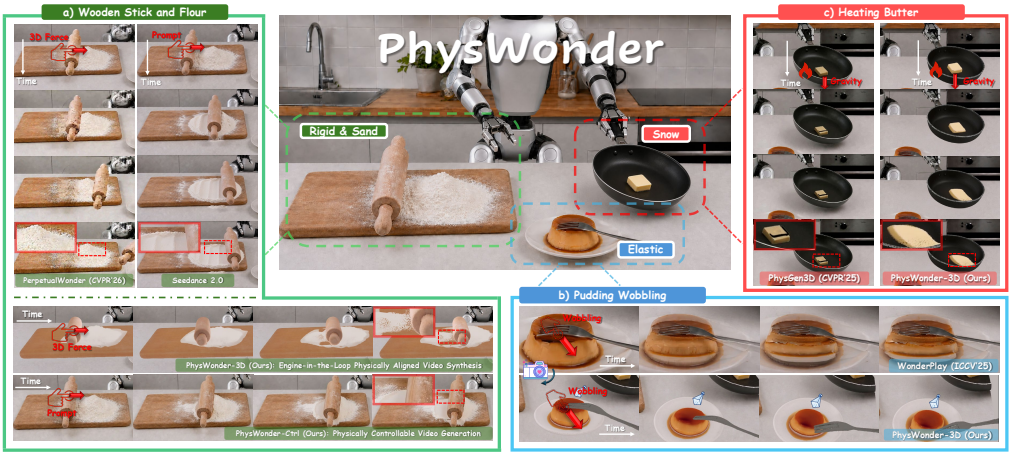


Figure 1: **Material properties are revealed not only by appearance, but also by how objects move and deform under physical interactions.** (a) PhysWonder-3D constructs simulation-ready particle worlds for physically aligned rigid and sand simulation under 3D force control, while PhysWonder-Ctrl internalizes constitutive-response knowledge for prompt-conditioned generation. Perpetual-Wonder (Zhan et al., 2026) leaves flour unresponsive to the rolling pin. Seedance 2.0 (Seedance et al., 2026) depicts the rolling interaction but unnaturally alters flour quantity and shape. (b) PhysWonder-3D renders 3D particle worlds for scalable synthesis, preserving object structure and surface appearance across camera views. Generative refinement in WonderPlay (Li et al., 2025) can hallucinate appearance and motion, making videos deviate from trajectories simulated with physical parameters. (c) PhysWonder-3D models background geometry and aligns objects with support surfaces, avoiding the interpenetration observed in PhysGen3D (Chen et al., 2025).

ABSTRACT

Recent video generation models show growing potential as world simulators, but physically controlling material-dependent dynamics remains difficult. The same interaction can produce different dynamics depending on how materials deform and whether they recover, reflecting their distinct constitutive responses. Existing approaches guide video generation with simulated motion, forces, or material attributes, but leave this constitutive response implicit. Our key insight is to reveal constitutive response by jointly modeling video generation and physically aligned, appearance-decoupled deformation dynamics. We present ***PhysWonder***, combining two novel modules: ***PhysWonder-3D*** for scalable, physically aligned synthesis and ***PhysWonder-Ctrl*** for learning material control. PhysWonder-3D builds simulation-ready particle worlds from single images and uses a hybrid Material Point Method (MPM) to vary constitutive settings under fixed world and interaction conditions. This produces matched rigid, elastic, sand, and snow dynamics, with RGB videos and deformation-response supervision co-rasterized from the same simulator states. PhysWonder-Ctrl jointly trains a Control Latent Predictor and video generator with response-latent and video supervision. We condition the video generator with the predicted latents, enabling material-controllable generation from a reference image

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.

and a physical prompt without simulator rollouts at inference. PhysWonder-3D preserves simulated trajectories while achieving a $2.46\times$ end-to-end speedup over RealWonder in single-image video synthesis. PhysWonder-Ctrl improves material controllability and motion similarity under both explicit and predicted guidance.

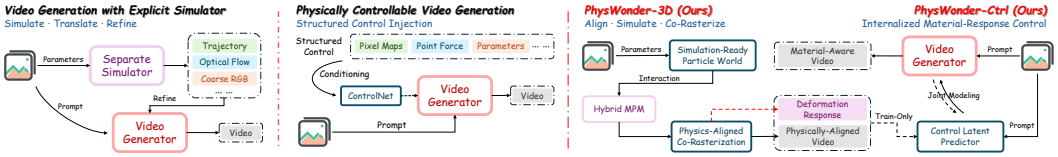


Figure 2: **Comparison of video generation pipelines.** Left: video generators refine simulator outputs, potentially hallucinating appearance or motion and weakening physical-visual alignment. Middle: structured controls specify forces, physical parameters, or trajectories, but leave implicit how material properties shape deformation under interaction. Right: PhysWonder-3D produces aligned videos and deformation responses from the same simulated states. PhysWonder-Ctrl jointly trains a Control Latent Predictor and video generator, enabling simulator-free material control at inference.

1 Introduction

Recent video generation foundation models have shown remarkable progress in visual quality, temporal consistency, and dynamic scene generation (Brooks et al., 2024; Agarwal et al., 2025; Wu et al., 2025). Large-scale pretraining has also enabled these models to generate increasingly plausible physical phenomena and interactions. However, these capabilities are learned implicitly from visual data, and the relation among physical parameters, material properties, and generated dynamics remains difficult to control explicitly. Therefore, using video generators as physically controllable world simulators remains challenging, especially for material-dependent dynamics. This raises a key question: how can video generation models learn to control material-dependent dynamics?

Existing approaches mainly follow two directions, as illustrated in Fig. 2. *Video Generation with Explicit Simulator* methods such as WonderPlay and PerpetualWonder (Li et al., 2025; Zhan et al., 2026) rely on different physics simulators to obtain trajectories or coarse 3D dynamics, and then refine these simulation results with pretrained video generators, including flow-conditioned models such as Go-With-The-Flow (Burgert et al., 2025). Such refinement can produce results that deviate from the simulated trajectories, weakening parameter-to-video alignment. *Physically Controllable Video Generation* methods instead condition video models with point forces (Gillman et al., 2026), physical parameter maps (Narayanan et al., 2026), or generated trajectories (Wang et al., 2026). These controls specify physical conditions or motion, but do not explicitly reveal how material properties translate into dynamic responses under interaction.

As illustrated in Fig. 1, materials can exhibit different behaviors under interaction. A pudding deforms and returns toward its original shape, while flour spreads and keeps its changed shape. These differences arise from constitutive behavior, which describes how a material deforms and whether it recovers after interaction. This observation suggests that physical conditions alone do not explicitly reveal material response. Instead, material properties become visible through how deformation evolves over time. These deformation dynamics reveal the material's constitutive response.

Our key insight is that constitutive-response capabilities can emerge in video generators by learning from physically aligned deformation dynamics. By pairing prescribed material conditions with deformation-response supervision and corresponding videos, the generator can learn how material properties shape visible dynamics. However, such supervision is difficult to obtain from real videos, where material parameters, applied interactions, and internal deformation cannot be reliably inferred. We realize this idea in ***PhysWonder***, which consists of two complementary components. ***PhysWonder-3D*** produces physically aligned videos and deformation-response supervision from the same simulated states, while ***PhysWonder-Ctrl*** internalizes this supervision for material-controllable video generation.

To obtain such supervision, we revisit simulator-coupled video synthesis. Physics simulators provide dynamics under prescribed material parameters, but their outputs often require generative refinement

for realistic rendering, which can alter the simulated trajectories. PhysWonder-3D instead samples appearance-carrying particles from aligned textured meshes to build a simulation-ready particle world. This allows RGB videos and deformation-response maps to be co-rasterized directly from simulated particle states without generative refinement. To isolate constitutive response, we further fix the particle world and interaction while varying only the material behavior. This motivates an interaction-shared, response-specific hybrid Material Point Method (MPM) for rigid, elastic, sand, and snow materials under the same interaction. The resulting videos and deformation responses therefore remain aligned with the prescribed material conditions.

PhysWonder-Ctrl turns this supervision into an inference-time constitutive-response capability. A Control Latent Predictor maps the reference image and physical condition to a deformation-response latent. Because latent matching alone does not ensure the same effect on dynamics, we condition the generator with the predicted latents during joint training with response alignment and video supervision. This enables material-controllable generation without simulator rollouts.

Our main contributions are:

1) We present PhysWonder, where learning from deformation dynamics enables constitutive-response capabilities to emerge within video generators for material-controllable generation.

2) We develop PhysWonder-3D for physically aligned synthesis through simulation-ready world construction, hybrid MPM coupling, and direct co-rasterization. It produces visually faithful counterfactual videos and aligned response supervision from shared simulator states.

3) We introduce PhysWonder-Ctrl for generator-in-the-loop response learning. Predicted latents condition the generator under joint response-latent and video supervision, enabling material control without simulated responses at inference. The learned controls also transfer across video backbones.

4) PhysWonder demonstrates physical fidelity, efficiency, and material controllability. For single-image world building under prescribed physical parameters, PhysWonder-3D preserves parameter-aligned dynamics across multiple constitutive settings with visually faithful rendering, while achieving speedups over RealWonder ($2.46\times$) and PhysGen3D ($2.12\times$). We construct a benchmark with parameter-aligned videos, where PhysWonder-Ctrl achieves 33.0% higher motion similarity than PhysCtrl and transfers across video backbones.

2 Related Work

Physical Simulation and Generative Dynamics. Classical and neural 3D simulators model material dynamics with high physical fidelity, but typically require accurate reconstruction, specified parameters, and scene-specific optimization (Stomakhin et al., 2013; Klár et al., 2016; Hu et al., 2018; Müller et al., 2005; Zhong et al., 2024; Cai et al., 2024; Xie et al., 2024; Hong et al., 2026). Conversely, video models learn plausible motion from data but do not expose the evolving material response (Brooks et al., 2024; Gao et al., 2025; Yang et al., 2024; Wan et al., 2025; Chefer et al., 2025; Zhang et al., 2026; Tan et al., 2026). PhysWonder-3D bridges this gap by constructing simulation-ready particle worlds and producing aligned deformation-response supervision at scale.

Video Generation with Explicit Simulator Coupling. Prior methods incorporate rigid priors or MPM-based dynamics into video synthesis (Liu et al., 2024; Zhang et al., 2024; Huang et al., 2025; Chen et al., 2025). WonderPlay, RealWonder, and PerpetualWonder obtain coarse visual or trajectory cues from a separate simulator and refine them with a video generator (Li et al., 2025; Liu et al., 2026; Zhan et al., 2026), which may hallucinate appearance or motion and weaken physical-visual alignment. PhysWonder-3D avoids generative post-processing by directly co-rasterizing video and deformation response from the same simulated state.

Physically Controllable Video Generation. Existing methods condition video generation on trajectories, forces, or pixel-aligned physical parameters (Le et al., 2025b; Yuan et al., 2025; Gillman et al., 2026; Narayanan et al., 2026). PhysCtrl (Wang et al., 2026) learns trajectories with deformation-gradient constraints and guides video synthesis through projected motion. These control interfaces leave local material stretch implicit. PhysWonder-Ctrl uses simulation-aligned stretch supervision and generator-in-the-loop learning to predict response latents directly in the video generator's control space, enabling material-conditioned generation without response-control videos at inference.

3 PRELIMINARIES

3.1 THE MATERIAL POINT METHOD

The Material Point Method (MPM) (Jiang et al., 2016) alternates between Lagrangian particles and an Eulerian grid through particle-to-grid (P2G) and grid-to-particle (G2P) transfers. Internal forces are determined by deformation gradients and constitutive responses, enabling diverse material dynamics. For elastic materials, the Lamé coefficients are (Boresi & Schmidt, 2002)

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

where E and ν denote Young's modulus and Poisson's ratio. The deformation gradient is updated as $\mathbf{F}_i^{n+1} = (\mathbf{I} + \Delta t\, \mathbf{C}_i^n)\, \mathbf{F}_i^n$, where $\mathbf{C}_i^n$ is the local affine velocity field obtained during G2P. Material-specific constitutive updates are then applied to $\mathbf{F}_i^{n+1}$.

3.2 VIDEO GENERATION BACKBONE

We build on Wan (Wan et al., 2025), a DiT-based video foundation model (Peebles & Xie, 2023) trained with Flow Matching (Lipman et al., 2022) in latent space. Given noise latent $\mathbf{z}_0$, target video latent $\mathbf{z}_1$, and condition c, it learns the velocity field along

$$\begin{aligned} \mathbf{z}_t &= (1-t)\mathbf{z}_0 + t\mathbf{z}_1, \qquad t \sim \mathcal{U}(0,1), \\ \mathcal{L}_{\text{FlowMatchSFT}} &= \mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1,t}\left[\|v_\theta(\mathbf{z}_t, t, c) - (\mathbf{z}_1 - \mathbf{z}_0)\|_2^2\right]. \end{aligned} \tag{2}$$

Since Wan does not explicitly model physical laws or controllable material dynamics, we augment it with deformation-response supervision and physics-guided control.

4 METHODOLOGY

Existing controls such as trajectories (Wang et al., 2026), particles (Li et al., 2025), and optical flow (Chefer et al., 2025) primarily describe observable motion. To learn constitutive-response capabilities from deformation dynamics, we construct aligned training examples by varying constitutive settings under fixed world and interaction conditions:

$$(\mathcal{W}, \boldsymbol{\xi}, \boldsymbol{\psi}^{(m)}) \xrightarrow{\mathcal{T}} \mathcal{S}_{1:T}^{(m)}, \qquad (\mathcal{W}, \mathcal{S}_{1:T}^{(m)}) \xrightarrow{\mathcal{R}} (\mathbf{Y}_{1:T}^{(m)}, \mathbf{G}_{1:T}^{(m)}), \tag{3}$$

where branch m shares the reconstructed world $\mathcal{W}$ and interaction $\boldsymbol{\xi}$, while $\boldsymbol{\psi}^{(m)}$ specifies the constitutive family and its calibrated response parameters. Here, $\mathcal{W}$ includes geometry, appearance, and camera, while $\boldsymbol{\xi}$ includes the initial state and external interaction. At frame k, $\mathcal{S}_k^{(m)} = \left\{\mathbf{x}_i^{k,(m)}, \mathbf{v}_i^{k,(m)}, \mathbf{F}_i^{k,(m)}, \mathbf{C}_i^{k,(m)}, J_{p,i}^{k,(m)}\right\}_{i=1}^{N}$ is the full simulator state, whereas $\mathbf{G}_{1:T}^{(m)}$ is its particle-supported deformation-response observation. Thus, constitutive attribution arises from cross-branch variation under fixed $(\mathcal{W}, \boldsymbol{\xi})$. This factorization motivates our response-specific hybrid MPM (Sec. 4.1), matched PhysWonder-3D synthesis (Sec. 4.2), and inference-time response internalization in PhysWonder-Ctrl (Sec. 4.3).

4.1 INTERACTION-SHARED, RESPONSE-SPECIFIC HYBRID MPM

MLS-MPM (Hu et al., 2018) naturally couples materials through shared particle-grid transport, but matched counterfactual supervision additionally requires distinct material responses to remain comparable under the same interaction. We organize the solver around shared particle-grid transport and material-specific state updates. Elastic and elastoplastic materials follow their constitutive updates, while rigid objects use native rigid-body evolution. All branches expose a common particle-state interface for aligned rendering and deformation readout. For particle i, the shared trial deformation is

$$\mathbf{F}_i^{\text{tr}} = (\mathbf{I} + \Delta t\, \mathbf{C}_i^n)\, \mathbf{F}_i^n, \tag{4}$$

followed by the material-specific update

$$\left(\bar{\mathbf{F}}_i^{n+1}, J_{p,i}^{n+1}, \boldsymbol{\tau}_i^{\text{K},n+1}\right) = \mathcal{P}_{\chi_i}\left(\mathbf{F}_i^{\text{tr}}, J_{p,i}^n; \boldsymbol{\psi}_i\right), \tag{5}$$

where χ_i and $\boldsymbol{\psi}_i$ denote the material type and its parameters. The returned Kirchhoff stress enters the shared MLS-MPM P2G update.

Elastic materials preserve recoverable stretch, while sand and snow project the singular values of $\mathbf{F}_i^{\mathrm{tr}} = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^\top$ according to their elastoplastic yield rules. Rigid particles instead retain only the proper rotation $\bar{\mathbf{F}}_i = \mathbf{U}_i \mathbf{D}_i \mathbf{V}_i^\top \in \mathrm{SO}(3)$, where $\mathbf{D}_i = \mathrm{diag}(1, 1, \det(\mathbf{U}_i \mathbf{V}_i^\top))$, and contribute no deformable stress. Each rigid cluster additionally maintains

$$\mathbf{b}_r = (\mathbf{c}_r, \mathbf{Q}_r, \mathbf{V}_r, \boldsymbol{\omega}_r)\,. \tag{6}$$

After shared particle-grid transport, intermediate velocities are projected onto the rigid velocity manifold. The body state is advanced, SDF-based contact is resolved, and rigid motion is written back to the common particle interface:

$$\mathbf{x}_i^{n+1} = \mathbf{c}_r^{n+1} + \mathbf{Q}_r^{n+1}\mathbf{r}_i^0, \qquad \mathbf{v}_i^{n+1} = \mathbf{V}_r^{n+1} + \boldsymbol{\omega}_r^{n+1} \times \mathbf{Q}_r^{n+1}\mathbf{r}_i^0, \qquad \mathbf{F}_i^{n+1} = \mathbf{Q}_r^{n+1}. \tag{7}$$

The solver is thus interaction-shared but response-specific: rigid, elastic, sand, and snow preserve distinct dynamics while exposing comparable motion and deformation through a common particle-state interface. Detailed projections and momentum derivations are provided in Supp. Sec. D.

4.2 PhysWonder-3D: Scalable Physically Aligned Synthesis

PhysWonder-3D produces visually faithful videos and explicit deformation-response supervision from the same simulated evolution. We construct a simulation-ready particle world that supports both physical simulation and direct rendering, then vary constitutive settings under fixed world and interaction conditions to obtain matched counterfactuals.

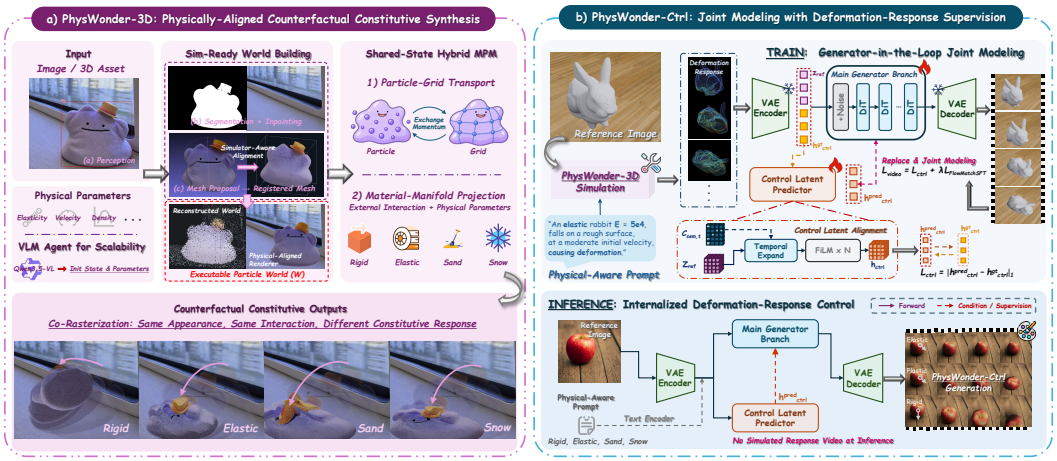


Figure 3: **Overview of PhysWonder.** Left: PhysWonder-3D builds a simulation-ready particle world and uses an interaction-shared, response-specific hybrid MPM to synthesize matched rigid, elastic, sand, and snow responses. The bottom row overlays RGB renderings with deformation-response maps from five simulation stages, compactly visualizing their evolution under the same appearance and interaction. Right: PhysWonder-Ctrl encodes the simulation-derived response video as privileged supervision and jointly trains the Control Latent Predictor with the video generator. At inference, the predicted response latent enables physically controllable generation from a reference image and physics-aware prompt without requiring a simulated response video.

Simulation-Ready World Building. A visually plausible reconstruction is not yet physically executable, while a physically valid simulation does not necessarily yield a faithful video. Prior pipelines often simulate coarse 3D dynamics and rely on video generation for refinement, which may hallucinate appearance or deviate from the simulated motion (Li et al., 2025); PhysGen3D (Chen et al., 2025) further struggles with rigid and high-stiffness objects, while directly rendering deformed simulation geometry can limit appearance fidelity and scene consistency.

We therefore build a simulation-ready world that preserves both physical executability and visual fidelity. For image inputs, SAM3 (Carion et al., 2026) extracts object and support-plane masks, ObjectClear completes the background (Zhao et al., 2026), and MoGe (Wang et al., 2025) recovers the scene point map and camera intrinsics. Conditioned on the object mask and point map, SAM3D (Chen et al., 2026) provides a textured mesh proposal.

Rather than directly sending this proposal to simulation, our simulator-aware alignment closes the gap between visual reconstruction and physical execution. As shown in Fig. 3, we first transform the mesh into the recovered camera frame and optimize its scale and translation by minimizing mask reprojection error. Grounded objects are further aligned with the estimated support plane, while visible texture regions are corrected using the input image. Compared with object-only reconstruction, this scene-aware alignment reduces object-ground interpenetration and background overlap artifacts, while yielding more faithful and scene-consistent rendering.

We then remove degenerate faces, sample uniformly spaced surface particles, and optionally fill the interior for volumetric objects. Appearance from the corrected textured mesh is transferred to these particles, allowing their simulated states to be rendered directly. Finally, the resulting particles are normalized into the metric MPM domain. This process aligns observation, contact, and discretization, yielding a shared executable world $\mathcal{W}$. The interaction configuration $\boldsymbol{\xi}$ and constitutive parameters $\boldsymbol{\psi}$ in Eq. 3 can be specified directly by users.

Scalable Matched Counterfactual Synthesis. Our goal is to expose how changes in material properties alter deformation dynamics. Independently configured simulations can confound material effects with differences in geometry, initialization, or external interaction. Our key design is therefore to construct matched counterfactuals: all branches share the executable world $\mathcal{W}$ and interaction configuration $\boldsymbol{\xi}$, while only the constitutive family $\chi^{(m)}$ and parameters $\boldsymbol{\psi}^{(m)}$ vary across rigid, elastic, sand, and snow. Qwen3.5-VL (Qwen Team, 2026) helps select material parameters within suitable simulation ranges for scalable, physically aligned video synthesis.

After material-specific updates and rigid-body synchronization, we extract a rotation-invariant local stretch descriptor from each particle's deformation-gradient state:

$$r_i^{k,(m)} = \left\| \log \boldsymbol{\sigma}\left(\mathbf{F}_i^{k,(m)}\right) \right\|_2, \tag{8}$$

where $\boldsymbol{\sigma}$ denotes the singular values. We rasterize this descriptor on the evolving particle support, using a shared camera projection, normalization $\mathcal{N}$, and color mapping $\mathcal{C}$ across matched branches:

$$G_k^{(m)} = \mathcal{C}\left(\mathcal{R}_{\pi_{\mathrm{cam}}}\left(\left\{\left(\mathbf{x}_i^{k,(m)}, \mathcal{N}(r_i^{k,(m)})\right)\right\}_{i=1}^{N}\right)\right). \tag{9}$$

The resulting sequence $G_{1:T}^{(m)}$ pairs particle transport with local stretch, capturing both the spatial distribution and intensity of deformation over time. Under matched interactions, these maps reveal how material properties shape deformation dynamics. RGB videos and deformation-response maps are co-rasterized from the same particle states, providing physically aligned supervision.

4.3 PhysWonder-Ctrl: Internalizing Constitutive Response

PhysWonder-3D exposes constitutive behavior through matched deformation dynamics and provides explicit deformation-response supervision from the same simulated states. We treat this supervision as privileged information during training, using PhysWonder-3D as a bridge from prescribed material conditions to material-dependent video generation. Following Wan's variational autoencoder (VAE) architecture (Wan et al., 2025; Kingma & Welling, 2013), we use a shared VAE to encode the reference image and deformation-response sequence as

$$z_{\mathrm{ref}} = \mathcal{E}_{\mathrm{VAE}}(I_{\mathrm{ref}}), \qquad h_{\mathrm{ctrl}}^{\mathrm{gt}} = \mathcal{E}_{\mathrm{VAE}}(\mathbf{G}_{1:T}), \tag{10}$$

where z_{ref} anchors appearance and layout, while $h_{\mathrm{ctrl}}^{\mathrm{gt}}$ provides privileged response supervision.

Directly predicting the complete deformation-response sequence for an unseen input would introduce another spatiotemporal generation problem. We instead design a Control Latent Predictor (CLP) to amortize the parameter-to-response relation into an inference-available latent:

$$h_{\mathrm{ctrl}} = \mathcal{P}_\phi(z_{\mathrm{ref}}, t, c_{\mathrm{sem}}), \tag{11}$$

where c_{sem} fuses pooled text and image features. $\mathcal{P}_\phi$ temporally expands z_{ref} as a spatiotemporal scaffold and Feature-wise Linear Modulation (FiLM)-modulates stacked 3D residual blocks with semantic and timestep conditions, thereby predicting the deformation-response latent associated with the reference and physical condition. Crucially, matching the privileged latent alone does not ensure the same effect on generated dynamics. We therefore replace the privileged latent with the predicted one inside the video generator and jointly optimize the predictor and generator. With $c_{\mathrm{gen}} = (z_{\mathrm{ref}}, c_{\mathrm{sem}}, h_{\mathrm{ctrl}})$, we use

$$\mathcal{L}_{\mathrm{Video}} = \mathcal{L}_{\mathrm{ctrl}} + \lambda \mathcal{L}_{\mathrm{FlowMatchSFT}} = \left\| h_{\mathrm{ctrl}} - h_{\mathrm{ctrl}}^{\mathrm{gt}} \right\|_1 + \lambda \mathbb{E}_{z_0, z_1, t}\left[\left\| v_\theta\left(z_t, t, c_{\mathrm{gen}}\right) - (z_1 - z_0) \right\|_2^2 \right]. \tag{12}$$

Here, z_t follows the flow path in Sec. 3.2 and λ balances response alignment and generator-side supervision. The first term transfers the simulation-derived deformation-response structure, while the second preserves its effect on the generated dynamics. Thus, CLP learns not merely to recover a missing control modality, but to internalize the physical-parameter-to-response relationship as a generator-adapted condition. At inference, PhysWonder-Ctrl predicts h_{ctrl} from the reference image, physical-semantic condition, and timestep without requiring a simulated response video. PhysWonder-3D therefore provides scalable, physically aligned deformation-response supervision during training, while PhysWonder-Ctrl internalizes its generative effect into an inference-time constitutive-response capability. The detailed CLP architecture is provided in Supp. Sec. E.

5 Experiments

5.1 Experimental Setup

Benchmark and Evaluation. We construct the PhysWonder benchmark with PhysWonder-3D using 1,000 assets from Objaverse (Deitke et al., 2023) and Google Scanned Objects (Downs et al., 2022), synthesizing matched rigid, elastic, sand, and snow responses and holding out 200 assets for evaluation. PhysWonder-Ctrl uses a Wan2.1-based 1.3B control backbone; fine-tuned baselines share training data and LoRA settings. Given the reconstruction sensitivity and case-specific initialization of simulator-coupled baselines, we evaluate 12 representative assets across four constitutive settings (48 rollouts), measuring runtime, success rate, and human 2AFC. For PhysWonder-Ctrl, PSNR, SSIM, and LPIPS assess visual fidelity, while Sim_{Flow} measures alignment with parameter-matched simulation dynamics via cosine similarity of extracted optical flows. Fixed-rubric Vision-Language Model (VLM) judging and human 2AFC additionally assess material-response plausibility against representative strong-performing and closed-source models. Full details appear in Appendix Sec. B.

5.2 Evaluation of PhysWonder-3D

We compare PhysWonder-3D with recent simulator-coupled pipelines (Chen et al., 2025; Li et al., 2025; Zhan et al., 2026; Liu et al., 2026). As shown in Fig. 4, prior methods can suffer from reconstruction–simulation mismatch or refinement-induced trajectory drift and hallucination. PhysGen3D further struggles with rigid dynamics and scene-consistent rendering, while PhysWonder-3D preserves appearance and simulated trajectories across rigid, elastic, sand, and snow materials without generative refinement. Table 1 reports 48/48 successful rollouts, compared with RealWonder's 33/48. Mean runtime over successful runs decreases from 411.9 to 167.2 s (2.46× faster). These results support PhysWonder-3D as a scalable, physically aligned synthesis pipeline. Appendix Sec. C Fig. 9 further shows that local stretch decreases with increasing stiffness under matched loading.

Table 1: **Efficiency, robustness, and perceived physical plausibility of PhysWonder-3D.** We average end-to-end image-to-video runtimes over successful runs (no crashes or visible output corruption), using identical single-image inputs and each method's own complete pipeline, and 2AFC denotes human preference. **First**, **Second**, and **Third** denote the top three results.

Method	Venue	Elastic↓	Rigid↓	Sand↓	Snow↓	Avg.↓	Succ.↑	2AFC(%)↑
PhysGen3D (Chen et al., 2025)	CVPR'25	**362.7**	–	**345.8**	**353.4**	**354.0**	**31/48**	**61.11**
WonderPlay (Li et al., 2025)	ICCV'25	934.5	**890.3**	906.0	893.4	906.1	28/48	–
PerpetualWonder (Zhan et al., 2026)	CVPR'26	6855.9	8225.7	6767.8	6654.2	7125.9	14/48	–
RealWonder (Liu et al., 2026)	arXiv'26	**398.4**	**449.9**	**403.2**	**396.2**	**411.9**	**33/48**	**10.00**
PhysWonder-3D	Ours	**166.2**	**181.3**	**160.2**	**161.2**	**167.2**	**48/48**	**78.89**

5.3 Evaluation of PhysWonder-Ctrl

Table 2 shows that deformation-response guidance improves motion alignment across tested video backbones, while larger prompt-only models remain weakly aligned with matched physical rollouts. PhysWonder-Ctrl achieves the strongest explicit-control scores, and predicted guidance maintains strong alignment using only an image and physical prompt. Fig. 5 shows clearer deformation changes across stiffness settings, while Fig. 6 illustrates distinct counterfactual material dynamics. Together, these qualitative comparisons show response variation within and across constitutive families. Human and VLM judgments in Table 3 provide complementary evidence of material-response plausibility.



Figure 4: **Qualitative comparison of PhysWonder-3D.** Prior methods exhibit reconstruction-simulation mismatch, trajectory drift, or generative artifacts, while PhysWonder-3D preserves appearance and parameter-aligned dynamics across four materials.

Table 2: **Quantitative comparison on engine-in-the-loop scenes.** All methods use identical reference images and prompts; models requiring post-training use the same training scale.

Method	PSNR↑	SSIM↑	LPIPS↓	$\mathrm{Sim}_{\mathrm{Flow}}$ ↑
Physically Controllable Video Generation				
Force-Prompting (NeurIPS'25) (Gillman et al., 2026) / Prompt	21.55	0.6288	0.2741	0.0961
Force-Prompting (NeurIPS'25) / Prompt + Force Control	23.21	0.6180	0.2648	0.2366
VideoREPA-5B (NeurIPS'25) (Zhang et al., 2026) / Prompt	23.38	0.6345	0.2626	0.0675
PhysCtrl-5B (NeurIPS'25) (Wang et al., 2026) / Predicted Trajectory Control / Prompt	23.25	0.6356	0.2776	0.2564
PhyCo-2B (CVPR'26) (Narayanan et al., 2026) / Prompt + Physical Property Maps	23.49	0.6284	0.2490	0.1635
Video Foundation Models (w/o Task-Specific Fine-Tuning)				
Wan2.1-TI2V-1.3B / Engine-in-the-Loop Deformation Control	**23.82**	0.6167	0.2659	0.3165
Wan2.1-TI2V-1.3B (Wan et al., 2025) / Prompt	20.41	0.5345	0.5642	0.0432
Wan2.2-TI2V-14B (Wan et al., 2025) / Prompt	20.42	0.5351	0.5645	0.0423
Wan2.2-TI2V-14B / Engine-in-the-Loop Deformation Control	23.66	0.6179	0.3124	0.2985
MiniMax-H3-33B (MiniMax AI, 2026) / Prompt	16.11	0.5449	0.5867	0.0694
MiniMax-H3-33B / Engine-in-the-Loop Deformation Control	21.66	0.5977	0.3240	0.2822
Ablation: Control Modality under Matched Fine-Tuning				
Wan2.1-TI2V-1.3B / Prompt	20.35	0.5274	0.6088	0.0182
Wan2.1-TI2V-1.3B / Engine-in-the-Loop Trajectory Control	23.59	0.6303	**0.2484**	0.2756
Wan2.1-TI2V-1.3B / Optical-Flow Control	20.56	0.4795	0.6645	0.2307
Wan2.1-TI2V-1.3B / Engine-in-the-Loop Deformation Control	**24.24**	**0.6463**	**0.2314**	**0.3398**
Ablation: Control-Latent Alignment Only ($\lambda = 0$)				
PhysWonder-Ctrl-1.3B / Engine-in-the-Loop Deformation Control	23.20	0.6232	0.2609	0.3273
PhysWonder-Ctrl-1.3B / Predicted Control Latent / Prompt	20.28	0.5035	0.6273	0.1021
Full Model: Generator-in-the-Loop Joint Modeling ($\mathcal{L}_{\mathrm{Video}} = \mathcal{L}_{\mathrm{ctrl}} + \lambda \mathcal{L}_{\mathrm{FlowMatchSFT}}$)				
PhysWonder-Ctrl-1.3B / Engine-in-the-Loop Deformation Control	**24.50**	**0.6513**	**0.2301**	**0.3575**
PhysWonder-Ctrl-1.3B / Predicted Control Latent / Prompt	23.77	**0.6359**	0.2485	**0.3410**

5.4 Ablation and Transfer Analysis

Control Representation. Under matched training, deformation-response guidance achieves the best fidelity and motion alignment (Table 2), improving $\mathrm{Sim}_{\mathrm{Flow}}$ by approximately 23% over trajectories and 47% over optical flow. With deformation LoRA, the stretch readout yields lower LPIPS than transport-only and volume-only cues at both tested stiffnesses (Appendix Sec. C Fig. 10).

Generator-in-the-Loop Internalization. With latent alignment alone, predicted guidance retains only 31% of the simulator-guided $\mathrm{Sim}_{\mathrm{Flow}}$ score. Joint training raises it to over 95% and improves predicted-control alignment by $3.3\times$. Closing this gap supports learning response latents through their effect on generated dynamics, with video supervision complementing latent matching.

Cross-Backbone Transfer. We reuse the same frozen Control Latent Predictor across heterogeneous generators, adapting only lightweight control interfaces. Table 4 shows motion-alignment gains for CogVideoX, PhyCo, and Force Prompting, supporting transfer across general-purpose and physics-conditioned generators. CLP injection also improves PhyCo on most evaluated deformation-sensitive and temporal-causality cases in Physics-IQ (Appendix Sec. C Table 6).

Table 3: **Visual evaluation of material-response plausibility.** VLM Win reports anonymized pairwise preference rates under a fixed rubric. Human 2AFC evaluates top VLM performers and the Wan backbone as an ablation baseline, separately from Table 1. Baseline scores report win rates against PhysWonder-Ctrl, while Ours reports its mean win rate across baselines.

Method	VLM Win (%)↑	Human 2AFC (%)↑
Physics-Aligned Synthesis		
PhysWonder-3D (Ours)	**85.7**	–
Video Foundation Models		
Seedance 2.0	60.7	22.2
MiniMax-H3-33B	53.6	24.4
Wan2.1-TI2V-1.3B	14.3	8.9
Physically Controllable Video Generation		
Force-Prompting-5B	17.9	–
VideoREPA-5B	25.0	–
PhyCo-2B	**64.3**	26.7
PhysWonder-Ctrl (Ours)	**78.6**	**79.5**

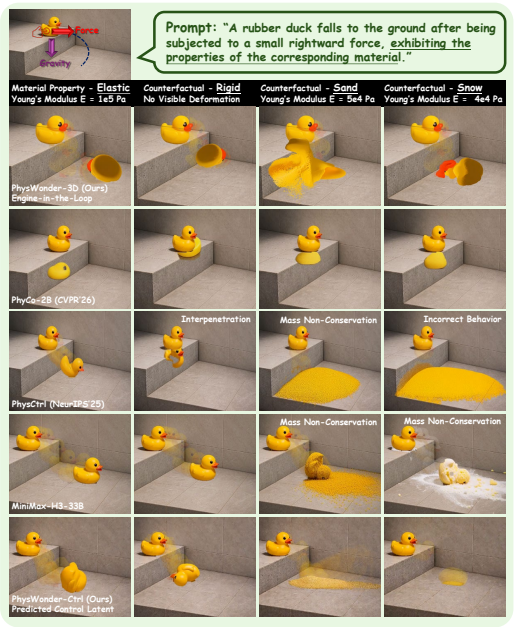
Figure 5: **Stiffness Sweep: $E = 10^3$–10^5 Pa and Rigid.** PhysWonder-3D's engine-in-the-loop synthesis yields stiffness-consistent deformation under matched interactions. PhysCtrl shows limited elastic variation and artifacts in the rigid case, while MiniMax-H3-33B changes little. PhysWonder-Ctrl provides more coherent prompt-only material responses. Two translucent earlier frames overlay the opaque final frame.



Table 4: **Cross-backbone transfer and capability injection on the PhysWonder benchmark.** The same frozen CLP improves motion alignment across general-purpose and physics-conditioned generators via lightweight interface adaptation.

Setting	PSNR↑	SSIM↑	LPIPS↓	Sim_{Flow}↑
CogVideoX-5B *Capability injection*				
Prompt-Only Fine-Tuning	21.585	0.6209	0.2801	0.0892
+ Control Latent Predictor	**22.454**	**0.6503**	**0.2747**	**0.2510**
PhyCo-2B *Cosmos-Predict-2B*				
Physical Property Maps	**23.485**	0.6284	0.2490	0.1635
+ Control Latent Predictor	23.404	**0.6495**	**0.2398**	**0.1738**
Force-Prompting-5B *CogVideoX-5B*				
Force Control	23.209	0.6179	0.2648	0.2366
+ Control Latent Predictor	**24.000**	**0.6530**	**0.2359**	**0.2937**

Figure 6: **Observed-State and Counterfactual Material Responses.** PhysWonder-3D generates prior-guided and counterfactual responses under matched interactions. Pretrained models exhibit inconsistent deformation and motion, while PhysWonder-Ctrl internalises constitutive responses to produce more physically coherent material dynamics. We show the first and final original frames, with eight uniformly sampled intermediate frames overlaid at low opacity.



6 Conclusion and Future Work

We presented PhysWonder, where deformation dynamics reveal constitutive response and provide supervision for material-controllable video generation. Experiments demonstrate improvements in synthesis efficiency, robustness, material controllability, and cross-backbone transfer. Future work will explore closing the real-to-sim gap through agentic video generation, learning to map real videos into executable simulation-ready twin spaces where geometry, interactions, and material properties can be iteratively inferred and refined to reproduce observed dynamics.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 118456–118470, 2026.

Arthur P Boresi and Richard J Schmidt. *Advanced mechanics of materials*. John Wiley & Sons, 2002.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:75035–75063, 2024.

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 (ICLR)*, volume 2026, pp. 138846–138923, 2026.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

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 (CVPR)*, pp. 6178–6189, 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 (CVPR)*, pp. 7220–7232, 2026.

Rishit Dagli, Donglai Xiang, Vismay Modi, Charles Loop, Clement Fuji Tsang, Anka He Chen, Anita Hu, Gavriel State, David IW Levin, and Maria Shugrina. Vomp: Predicting volumetric mechanical property fields. *arXiv preprint arXiv:2510.22975*, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13142–13153, 2023.

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 *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:103174–103201, 2026.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, et al. $phyworldbench$: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 75130–75164, 2026.

Haoqin Hong, Ding Fan, Fubin Dou, Zhi-Li Zhou, Haoran Sun, Congcong Zhu, and Jingrun Chen. Physics-informed deformable gaussian splatting: Towards unified constitutive laws for time-evolving material field. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4725–4733, 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.

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 (AAAI)*, pp. 3733–3741, 2025.

Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *ACM SIGGRAPH 2016 Courses*, pp. 1–52, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Gergely Klár, Theodore Gast, Andre Pradhana, Chuyuan Fu, Craig Schroeder, Chenfanfu Jiang, and Joseph Teran. Drucker-prager elastoplasticity for sand animation. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

Long Le, Ryan Lucas, Chen Wang, Chuhao Chen, Dinesh Jayaraman, Eric Eaton, and Lingjie Liu. Pixie: Fast and generalizable supervised learning of 3d physics from pixels. *arXiv preprint arXiv:2508.17437*, 2025a.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025b.

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

Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. Omniphysgs: 3d constitutive gaussians for general physics-based dynamics generation. *arXiv preprint arXiv:2501.18982*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

MiniMax AI. MiniMax H3. `https://github.com/MiniMax-AI/MiniMax-H3`, 2026. GitHub repository, accessed August 8, 2026.

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

Matthias Müller, Bruno Heidelberger, Matthias Teschner, and Markus Gross. Meshless deformations based on shape matching. *ACM Transactions on Graphics (TOG)*, 24(3):471–478, 2005.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. Phyco: Learning controllable physical priors for generative motion. *arXiv preprint arXiv:2604.28169*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. Phantom: Physics-infused video generation via joint modeling of visual and latent physical dynamics. *arXiv preprint arXiv:2604.08503*, 2026.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

Boming Tan, Xiangdong Zhang, Ning Liao, Yuqing Zhang, Shaofeng Zhang, Xue Yang, Qi Fan, and Yanyong Zhang. Dreamworld: Unified world modeling in video generation. *arXiv preprint arXiv:2603.00466*, 2026.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 402–419. Springer, 2020.

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.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:167907–167932, 2026.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5261–5271. IEEE, 2025.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

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 (CVPR)*, pp. 4389–4398, 2024.

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

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *arXiv preprint arXiv:2509.21309*, 2025.

Jiahao Zhan, Zizhang Li, Hong-Xing Yu, and Jiajun Wu. Perpetualwonder: Long-horizon action-conditioned 4d scene generation. *arXiv preprint arXiv:2602.04876*, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 388–406. Springer, 2024.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:122647–122676, 2026.

Jixin Zhao, Zhouxia Wang, Peiqing Yang, and Shangchen Zhou. Precise object and effect removal with adaptive target-aware attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19370–19379, 2026.

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

APPENDIX

To clarify the motivation behind PhysWonder, we first compare representative approaches through their *input - intermediate processing - output* pathways. We distinguish methods that retain explicit physics simulation at inference from those that learn physics-conditioned generation.

Video synthesis with explicit physics simulation. Given an image and 3D actions, WonderPlay (Li et al., 2025) reconstructs a scene, simulates coarse dynamics, and uses rendered RGB and optical flow to guide video generation and subsequent dynamic-scene refinement. PerpetualWonder (Zhan et al., 2026) extends this pathway through coupled physical and visual representations and multi-view refinement, allowing corrected states to support subsequent actions. RealWonder (Liu et al., 2026) reconstructs a lightweight point-cloud scene and converts online simulation into flow and coarse RGB conditions for a distilled video generator. These approaches depend on accurate scene reconstruction and simulation-ready initialization, while their generative stages may alter appearance or motion and weaken correspondence between prescribed physical parameters and the resulting dynamics.

A related direct-rendering approach, PhysGen3D (Chen et al., 2025), reconstructs textured meshes, simulates particle dynamics with MPM, and renders the updated geometry. Its elastic MPM formulation approximates rigid behavior through finite stiffness, making high-modulus settings more demanding numerically, while its background collision model is less suited to complex scene geometry. Another line of work performs physics simulation on reconstructed 3D representations. PhysGaussian and related Gaussian-based methods (Xie et al., 2024; Zhang et al., 2024) couple explicit 3D Gaussian assets with physics solvers to produce physically driven dynamics. OmniPhysGS (Lin et al., 2025) similarly builds physical simulation on top of an explicit reconstructed 3D representation and optimized material properties. These methods are effective for scene-specific physical modeling, but typically require an existing multi-view 3D reconstruction together with additional per-scene physical optimization or simulation preparation. They therefore do not directly

initialize simulation from a single image, and their scene-specific processing cost makes them less suitable as the backend of a scalable data-synthesis pipeline.

Complementary material-estimation approaches such as VoMP (Dagli et al., 2025) and Pixie (Le et al., 2025a) predict material-property fields from static appearance and geometry. Their supervision targets physical attributes that are subsequently used by a simulator to produce dynamics.

PhysWonder-3D instead targets scalable, single-image physically aligned synthesis. It constructs a simulation-ready particle world and renders RGB videos and deformation-response supervision from the same simulator states. Under a fixed scene and interaction, only the constitutive family and calibrated parameters are varied, producing matched counterfactuals whose parameter changes, simulator trajectories, and rendered dynamics remain explicitly aligned. This design avoids both scene-specific physical optimization and generative refinement, making PhysWonder-3D suitable for scalable video-response data construction.

Learned physics-conditioned video generation. Recent works study how physical structure can be learned or evaluated within video generators. NewtonGen (Yuan et al., 2025) introduces learnable Neural Newtonian Dynamics for object-level dynamical modeling and controllable video generation. PHANTOM (Shen et al., 2026) jointly models video content and V-JEPA2 (Assran et al., 2025) embeddings to improve physical plausibility, but its latent supervision derives from video features rather than constitutive responses aligned with prescribed material parameters. Complementary benchmarks such as VideoPhy-2 (Bansal et al., 2026) and PhyWorldBench (Gu et al., 2026) assess broad physical commonsense and realism in generated videos, without specifically emphasizing material-dependent dynamic responses. Physics-IQ (Motamed et al., 2026) provides a complementary evaluation whose solid-mechanics, material-property, and temporal-causality cases are more closely aligned with our focus on physically grounded dynamics.

PhyCo (Narayanan et al., 2026) conditions video generation on pixel-aligned physical property maps, while Force Prompting (Gillman et al., 2026) encodes global wind or localized forces as external control signals. These approaches specify physical attributes or interventions, but do not explicitly expose how prescribed material parameters translate into local deformation under interaction. Consequently, material response must be inferred implicitly from video supervision, where appearance, motion, external forces, and deformation remain entangled. PhysWonder instead isolates this parameter-to-response relationship through matched constitutive counterfactuals, directly aligning controlled material changes with the local deformation they induce.

PhysCtrl (Wang et al., 2026) reconstructs object geometry and uses a separate trajectory model to predict physics-conditioned 3D point trajectories, which are then projected as motion controls for video synthesis. Its deformation-gradient objective regularizes trajectory prediction, but the trajectory model and video generator are not jointly optimized. Moreover, the resulting trajectories reflect the combined effects of material properties, external interactions, and scene constraints, rather than explicitly isolating the deformation induced by a prescribed change in constitutive behavior.

PhysWonder-Ctrl follows our central insight that constitutive-response capabilities can be learned from physically aligned deformation dynamics. PhysWonder-3D fixes the world and interaction while varying constitutive settings, producing matched videos and deformation-response sequences from the same simulator states. These parameter-aligned responses serve as privileged targets for a Control Latent Predictor operating in the generator's control space. Joint response-latent and video supervision then preserves both the deformation-response structure and its effect on visible dynamics. At inference, an image and physical prompt suffice without simulator rollouts or an explicit trajectory-generation stage.

Questions addressed by PhysWonder. PhysWonder studies how deformation dynamics can provide explicit supervision for learning constitutive-response capabilities in video generators. This leads to three questions:

(1) What additional supervision does local deformation provide beyond observable motion under matched interactions?

(2) How can PhysWonder-3D produce scalable, physically aligned videos and deformation-response supervision while fixing the world and interaction?

(3) How can this training-time supervision be internalized as an inference-time control predicted from an image and physical prompt?

Accordingly, the Appendix provides an AI use statement (Sec. A), evaluation metrics and protocols (Sec. B), extended experimental results (Sec. C), the hybrid MPM solver formulation and update procedure (Sec. D), the Control Latent Predictor architecture and training-objective analysis (Sec. E), and the large-scale physically aligned synthesis pipeline (Sec. F). Table 5 summarizes the accompanying video demonstrations and the capability illustrated by each.

Table 5: **Supplementary video demonstrations and their purposes.**

Video	Purpose
Teaser Demo	Full video demonstrations corresponding to Fig. 1, showing physically aligned simulation and material-dependent interactions with PhysWonder.
Figure 5 Demo	Shows stiffness-dependent dynamics under different Young's moduli, highlighting improved deformation and motion control as material stiffness changes.
Figure 6 Demo	Shows both the observed constitutive state and counterfactual material states under matched interactions, while demonstrating that PhysWonder-Ctrl produces more physically coherent material responses.
Figure 7 Demo	Compares PhysWonder-3D with simulator-coupled synthesis baselines.

A AI Use Statement

For manuscript preparation, we used OpenAI GPT-5.6 Sol to improve the grammar, wording, readability, and internal consistency of text written by the authors. The scientific ideas, methodology, implementation, experimental design, execution, result analysis, figure construction, and conclusions were developed and verified by the authors. All AI-assisted edits were manually reviewed and revised by the authors. Generative AI systems also serve as explicitly reported components and objects of study in our research. A vision-language-model agent is used to ground simulator configurations; model-generated image and video samples are used as experimental cases for evaluating generation quality and generalization; and AI agents are used for automatic method evaluation under the fixed prompts and criteria documented in the paper. These outputs are treated as experimental inputs or automated measurements rather than substitutes for scientific judgment by the authors. All figure layouts, annotations, quantitative summaries, and interpretations were prepared by the authors, although some figures display model-generated outputs obtained during the experiments.

The authors reviewed all AI-assisted content and take full responsibility for the text, claims, results, and artifacts presented in this work.

B Additional Evaluation Details

B.1 Evaluation Principle

Evaluating physically controllable video generation requires separating visual fidelity from physical-response alignment. Conventional image metrics such as PSNR, SSIM, and LPIPS measure whether generated frames resemble a reference, but do not determine whether the visible dynamics follow those induced by prescribed physical parameters and interactions. Conversely, recovering numerical material parameters directly from a generated video is ill-posed and does not provide a reliable measure of dynamic correctness.

We therefore use two complementary evaluation settings. When a matched simulator rollout is available, we evaluate visual fidelity and whether the generated motion follows the dynamics induced by the same physical condition. When no unique trajectory should be treated as ground truth, we assess relative material-response plausibility through anonymized pairwise evaluation. Human 2AFC provides intuitive perceptual validation, while a fixed-rubric VLM judge applies structured criteria for temporal evolution, material response, and state consistency.

B.2 Engine-in-the-Loop Alignment with Matched Physical References

For each engine-in-the-loop scene, the simulator target and generated video share the same scene configuration, constitutive parameters, camera, and external interaction. This pairing evaluates not whether a model numerically recovers a material parameter, but whether it reproduces the *visible dynamics induced by that parameterized physical condition*.

(a) Video Fidelity. We report PSNR, SSIM, and LPIPS between generated and simulator-rendered target videos, averaged over all frames. PSNR and SSIM measure pixel-level and structural agreement, while LPIPS measures perceptual similarity using pretrained deep features (Simonyan & Zisserman, 2014; Zhang et al., 2018). These metrics quantify appearance-level fidelity but cannot determine whether two videos follow the same physical evolution.

(b) Parameter-Conditioned Motion Alignment. Two videos may contain visually similar frames while following different motion trajectories, or differ in local appearance while preserving similar dynamics. We therefore use $\mathrm{Sim}_{\mathrm{Flow}}$ to measure alignment between generated motion and the parameter-matched simulator rollout.

Using the same pretrained optical-flow estimator $\mathcal{F}$(Teed & Deng, 2020), we extract flow between consecutive generated and target frames:

$$\hat{\mathbf{U}}_n = \mathcal{F}(\hat{I}_n, \hat{I}_{n+1}), \qquad \mathbf{U}_n^* = \mathcal{F}(I_n^*, I_{n+1}^*), \tag{13}$$

where $\hat{\mathbf{U}}_n$ and $\mathbf{U}_n^*$ denote the generated and engine-in-the-loop flow fields, respectively. We compute

$$\mathrm{Sim}_{\mathrm{Flow}} = \frac{1}{N-1} \sum_{n=1}^{N-1} \frac{\left\langle \phi(\hat{\mathbf{U}}_n), \phi(\mathbf{U}_n^*) \right\rangle}{\left\| \phi(\hat{\mathbf{U}}_n) \right\|_2 \left\| \phi(\mathbf{U}_n^*) \right\|_2 + \varepsilon}, \tag{14}$$

where $\phi(\cdot)$ vectorizes the flow field and $\varepsilon > 0$ ensures numerical stability.

Optical flow provides a dense image-space description of visible motion. Because the generated and simulator-rendered videos are temporally aligned and share the same camera, constitutive parameters, and interaction, their flow similarity reflects whether the generated dynamics follow the direction and spatial structure of the corresponding physical rollout. The cosine formulation is less sensitive than direct flow regression to moderate magnitude differences caused by rendering or generation noise. We therefore interpret $\mathrm{Sim}_{\mathrm{Flow}}$ as a measure of *parameter-conditioned motion alignment*.

B.3 Reference-Free Material-Response Evaluation

For open-ended generation, the same high-level material and interaction condition may admit multiple physically plausible trajectories. Treating a single simulator rollout as the unique ground truth may therefore penalize valid variations. We instead evaluate relative material-response plausibility under a shared initial condition and physical prompt.

Pair Construction and Temporal Sampling. All compared methods receive the same reference image, physical prompt, and interaction specification. Candidate identities are anonymized and their presentation order is randomized for each pairwise comparison. For automatic evaluation, we uniformly sample the same timestamps from both candidate videos and present these temporal samples together with the shared initial frame. This prevents the judgment from depending on a single favorable frame and ensures that both methods are compared over the same temporal evolution.

(a) Fixed-Rubric VLM Pairwise Evaluation. We use GPT-5.6-Sol-High as a fixed automatic judge. For each comparison, the VLM receives the shared initial frame, physical prompt, and temporally aligned samples from two anonymized candidate videos. Unlike conventional perceptual metrics, the fixed rubric explicitly examines several aspects of the visible physical evolution:

- *consistency of motion direction and temporal progression,*
- *material-consistent deformation, recovery, or irreversible change,*
- *preservation of object identity and physical-state coherence,*
- *agreement with the prescribed interaction and material condition.*

Visual quality affects the decision only when artifacts prevent reliable physical comparison. The evaluator is explicitly instructed not to infer the exact numerical material parameter or require frame-by-frame agreement with a particular simulator trajectory. This design keeps the judgment focused on

observable physical evolution rather than appearance quality or unverifiable parameter recovery. We report the VLM win rate as the fraction of randomized pairwise comparisons favoring each method. The score is therefore interpreted as structured evidence of relative material-response plausibility rather than an absolute physical-accuracy metric.

(b) Human 2AFC Study. We additionally conduct an anonymized two-alternative forced-choice (2AFC) study as complementary perceptual evaluation. Following human-evaluation practice in Force-Prompting (Gillman et al., 2026) and WonderPlay (Li et al., 2025), we ask participants to assess observable motion and deformation rather than the correspondence between numerical material parameters, constitutive behavior, and generated dynamics. Such parameter-level judgments require specialized knowledge and are difficult to assess reliably from videos alone.

To keep each questionnaire within a practical completion time, we select representative strong-performing baselines from the evaluated methods and compare each of them directly against PhysWonder-Ctrl in randomized 1-vs.-1 trials across multiple scenes. For each comparator, we record the fraction of pairwise judgments won by PhysWonder-Ctrl and aggregate the complementary preference rates under a common normalization, yielding the reported Human 2AFC score. Through an anonymized evaluation website, we collect 45 completed questionnaires with randomized candidate ordering. For each comparison, participants receive the same initial condition and physical prompt together with two anonymized candidate videos.

Participants select the result that appears more visually coherent and whose motion and deformation are more physically plausible under the specified material condition. They are not asked to infer an exact Young's modulus or numerically verify the complete trajectory. Human 2AFC therefore serves primarily as an intuitive check of visual coherence and perceived material-response plausibility, while $\mathrm{Sim}_{\mathrm{Flow}}$ provides the complementary quantitative measure of alignment with the parameter-matched physical trajectory. Together with the structured VLM rubric, these evaluations cover perceptual plausibility, structured physical judgment, and parameter-conditioned motion alignment.

Automatic Pairwise Evaluation Prompt

Task. You are an expert evaluator for physically grounded video generation.
Physical prompt. *"A Sand Bagel with Young's modulus $E = 10^3$ starts on the ground, initially moving to the right, and may slide briefly on a rough surface, causing deformation."*
You are given a shared initial frame and temporally sampled results from two anonymized candidate videos. Candidate A and Candidate B are sampled at the same timestamps and are generated under the same initial condition and physical prompt.
Choose the candidate whose temporal evolution better matches the intended physical behavior.
Evaluation criteria. Focus primarily on:

- whether motion follows the intended interaction,
- whether deformation and recovery are consistent with the material,
- whether object identity and physical state remain coherent,
- whether temporal evolution matches the specified physical condition.

Visual quality should affect the decision only when artifacts prevent reliable physical comparison. Do not attempt to recover the exact numerical material parameter or demand an exact frame-by-frame trajectory.
Required output format.

```
Choice: A or Choice: B
Reason: one-sentence explanation
```

Human 2AFC Evaluation Instruction

Given the same initial condition and physical prompt, compare Candidate A and Candidate B. Choose the candidate whose temporal motion and deformation better match the specified physical condition.
Please consider:

- whether the object moves in the intended direction,
- whether its deformation or recovery is appropriate for the material,
- whether the object's physical state remains coherent over time.

You do not need to determine the exact numerical material parameters or verify the complete trajectory.
Choice: Candidate A or Candidate B

B.4 Scope of Physical Evaluation

No single metric establishes absolute physical correctness for generated videos, because visual fidelity, trajectory alignment, and material-response plausibility capture different aspects of physical behavior. With matched simulator references, PSNR, SSIM, and LPIPS assess visual fidelity, while $\mathrm{Sim}_{\mathrm{Flow}}$ measures whether the generated motion follows the dynamics induced by the same prescribed physical condition. When no unique trajectory exists, VLM and Human 2AFC instead assess whether the visible deformation and state evolution are plausible under the specified material and interaction. This layered protocol separates what can be measured against a matched physical reference from what must be judged relatively under open-ended generation, providing complementary rather than interchangeable evidence of physical consistency.

C Additional Experimental Results

C.1 Implementation and Paired-Data Construction

Training and Implementation. We construct the PhysWonder benchmark from 1,000 assets from Objaverse (Deitke et al., 2023) and Google Scanned Objects (Downs et al., 2022). For each asset, PhysWonder-3D synthesizes matched rigid, elastic, sand, and snow responses, while 200 assets are held out for evaluation. Each training example contains a reference image, physical prompt, simulator-rendered RGB target, and particle-supported deformation-response sequence obtained from the same simulator states. PhysWonder-Ctrl uses a Wan2.1-based 1.3B control backbone. All fine-tuned baselines use the same training data and LoRA setup (480×832, learning rate 1×10^{-5}, rank 32). We first fine-tune the video generator for 5 epochs on four NVIDIA A800 GPUs ($\sim$25 hours), followed by 2 epochs of generator-in-the-loop joint training of the Control Latent Predictor and video generator LoRA ($\sim$6.5 hours). Generating a 101-frame video takes approximately 2 minutes on one NVIDIA RTX 4090, with comparable inference time between CLP-predicted control and explicit deformation-response guidance encoded by the VAE.

Training Domain and Generalization. PhysWonder-Ctrl is trained only on scalable single-image data rendered from 3D assets; no real-captured or AI-generated images are used for post-training. Evaluation additionally includes real-captured and AI-generated single images, testing whether the learned constitutive-response capability generalizes beyond the rendered-asset training distribution.

Matched Paired Construction. Within each counterfactual group, the executable world, appearance, camera, initial state, and external interaction are fixed, while only the constitutive family and calibrated parameters vary. RGB videos and deformation-response supervision are co-rasterized from the same simulator states. For control-modality ablations, trajectory, optical-flow, and deformation-response controls are derived from the same engine-in-the-loop rollouts and trained at the same scale. This isolates the control representation rather than differences in training data or target dynamics. At predicted-control inference, PhysWonder-Ctrl uses only the reference image and physical prompt, without a simulator-derived deformation-response sequence.

Large-Scale Synthesis Robustness. All PhysWonder benchmark scenes, rendered from individual textured 3D assets, remain below the 4.2×10^6 particle threshold during batch synthesis. Tests on diverse Internet and AI-generated tabletop images further show that even scenes with multiple simultaneously simulated and interacting objects can generally be processed reliably within the supported particle budget. Only unusually cluttered scenes with extensive backgrounds or many objects may exceed this threshold and require manual adjustment of particle sampling.

C.2 Extended PhysWonder-3D Results

Fig. 7 extends the main-paper comparison to additional constitutive settings. Across these examples, PhysWonder-3D preserves both visual appearance and the trajectories generated under prescribed physical parameters. Elastic responses vary consistently with stiffness, sand and snow retain irreversible deformation, and rigid objects follow native rigid-body dynamics rather than high-stiffness elastic approximations. Together with the runtime and success-rate results in the main paper, these results support Ours as a scalable source of physically aligned deformation-response supervision.

C.3 Generalization to Interaction-Rich Generated Inputs

PhysWonder-Ctrl is trained only on rendered 3D assets, without real-captured or AI-generated images for post-training. To further examine out-of-domain applicability, Fig. 8 shows three interaction-rich scenes initialized from AI-generated single images. Despite the domain shift and multi-object interactions, PhysWonder-3D constructs executable particle worlds and produces material-dependent responses directly from simulation. These examples illustrate that the synthesis pipeline is not restricted to curated 3D assets and can also operate on generated visual content. PhysWonder-Ctrl further transfers the learned constitutive-response capability to these inputs without requiring simulator-derived deformation-response sequences at inference.

C.4 Analysis of the Deformation-Response Representation

We analyze whether the deformation-response representation captures stiffness-dependent deformation and provides useful generation guidance beyond particle transport and volumetric change. We select three assets from the PhysWonder benchmark: Bagel and Baguette for the stiffness analysis, and an apple for the response-readout ablation. For each asset, PhysWonder-3D synthesizes the analyzed sequences from a single rendered image.

Stiffness-dependent response. We test whether the response readout captures stiffness beyond global motion using Bagel and Baguette under matched loading, with $E = 10^4$–10^7 Pa and a rigid baseline. As shown in Fig. 9, increasing stiffness reduces both peak mean response $r(F)$ and normalized shape change; the response maps weaken accordingly, while the rigid branch remains near zero. This supports local stretch as a stiffness-sensitive cue complementary to particle transport.

Response-readout ablation. To isolate the contribution of local stretch beyond moving geometry, we compare three controls derived from the same physical trajectory: particle transport T, volumetric change $v = |\log J|/\sqrt{3}$, and stretch response $r = \|\log \sigma(F)\|_2$. We evaluate the same reconstructed apple at $E \in \{3, 10\}$ kPa using three paired seeds, keeping the physical trajectory and rendering setup fixed within each stiffness.

As shown in Fig. 10, transport preserves motion and silhouette, while the volumetric cue captures expansion and compression. The stretch response additionally captures anisotropic shape-changing deformation. With deformation LoRA, r achieves the lowest mean LPIPS at both tested stiffnesses and improves foreground motion alignment over transport-only control. This supports particle-supported local stretch as complementary physical supervision beyond motion or volume alone.



Figure 7: **Scalable physically aligned synthesis with PhysWonder-3D.** Across elastic, rigid, sand, and snow settings, PhysWonder-3D directly renders matched simulator states while preserving appearance and the trajectory induced by prescribed physical parameters. Lower-stiffness elastic cases deform more, sand and snow exhibit irreversible responses, and the native rigid branch avoids artifacts from stiff-elastic approximations. These results highlight PhysWonder-3D as a scalable synthesis pipeline for producing videos whose physical parameters, simulator trajectories, and rendered dynamics remain explicitly aligned. Prior pipelines can instead suffer from simulation failure or refinement-induced appearance and motion drift.

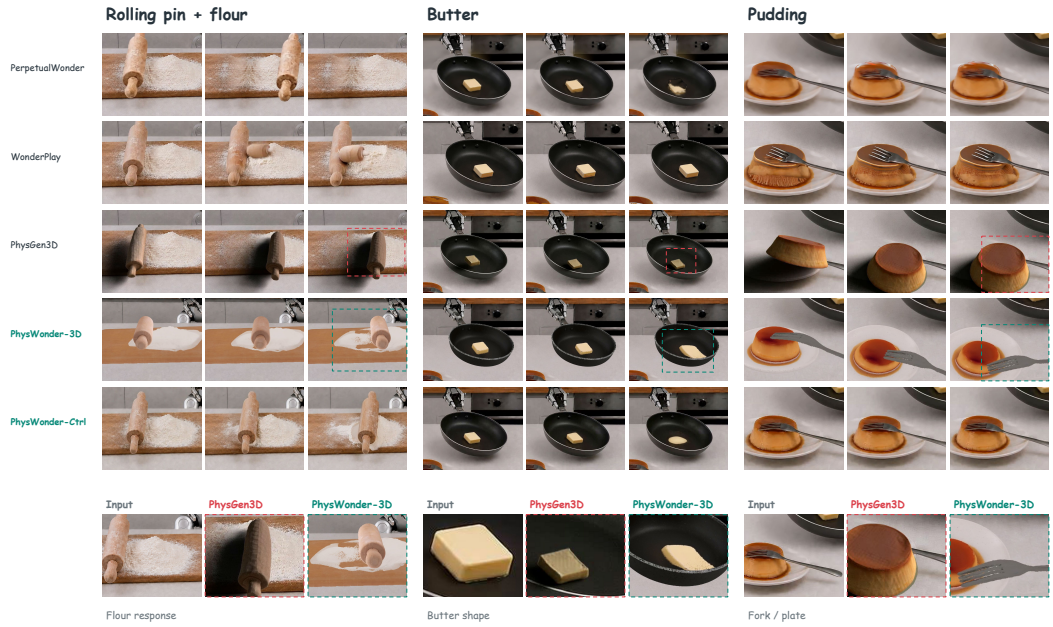


Figure 8: **Qualitative comparison on interaction-rich AI-generated scenes.** All three scenes are initialized from AI-generated single images, providing out-of-domain inputs beyond the rendered 3D assets used for PhysWonder-Ctrl training. For the rolling pin and flour, PerpetualWonder leaves the flour largely unchanged as the pin moves out of view, WonderPlay introduces an additional pin-like object, and PhysGen3D produces little visible flour response. PhysWonder-3D instead produces a clear displacement trail and granular redistribution. For butter, WonderPlay and PhysGen3D largely preserve a block-like shape, whereas PhysWonder-3D and PhysWonder-Ctrl show more pronounced flattening and spreading. For pudding, PhysGen3D exhibits a mismatch between the reconstructed object and the fork–plate configuration, while PhysWonder-3D produces localized deformation near the fork. PhysWonder-Ctrl better preserves the reference appearance, although its deformation is less pronounced. Red dashed boxes highlight baseline regions enlarged below; teal boxes show the corresponding PhysWonder-3D regions alongside input-image crops. Frames show successive stages within each video and are not synchronized across methods in physical time.

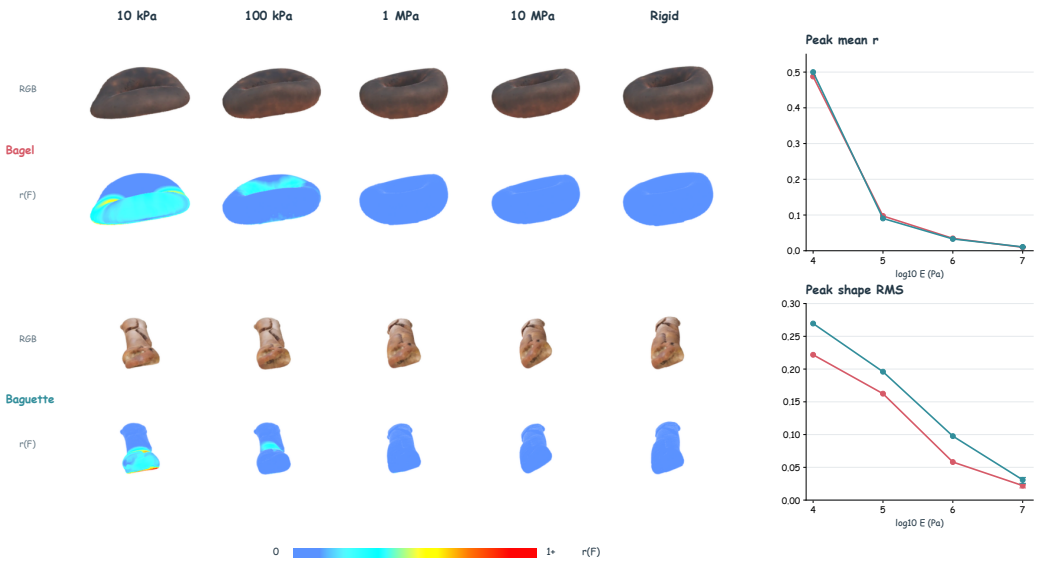


Figure 9: **Stiffness-dependent deformation response.** Left: RGB and $r(F) = \|\log \boldsymbol{\sigma}(F)\|_2$ at $t = 0.4\,\mathrm{s}$ under a shared response scale. Right: peak mean response and rigid-aligned shape RMS over the trajectory. Both decrease consistently with increasing stiffness across the tested objects, while the rigid branch remains near numerical zero.

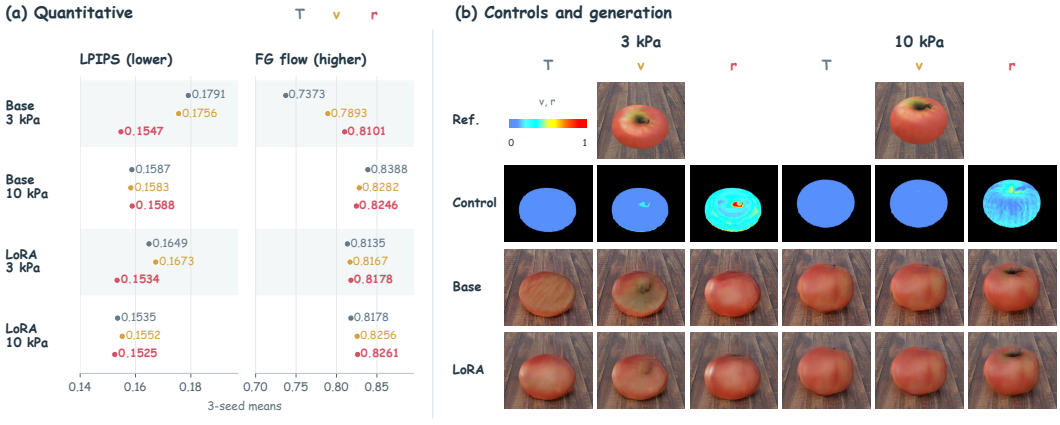


Figure 10: **Response-readout ablation.** (a) Three-seed mean LPIPS and foreground-flow cosine at two stiffnesses. (b) Synchronized simulator references, controls, and generated frames for particle transport T, volumetric change $v = |\log J|/\sqrt{3}$, and local stretch $r = \|\log \sigma(F)\|_2$. With deformation LoRA, r yields the lowest LPIPS at both stiffnesses and improves foreground motion alignment over transport alone, supporting its complementary deformation-aware conditioning.

Table 6: **Capability injection on Physics-IQ.** We inject the same frozen CLP into PhyCo and evaluate cases involving deformation-sensitive response or temporal physical causality. The additional "Phenomenon" column highlights the physical behavior probed by each case. CLP improves 6 of 8 selected cases and also increases the original Physics-IQ score from 59.24 to 60.22.

Scene	Physical Phenomenon	PhyCo	+CLP	Δ
Deformation-Sensitive Events				
Mug Breaks	Fracture / deformation response	96.05	**96.39**	+0.33
Weight on Ceramic	Load-bearing stiffness	**16.04**	15.87	−0.17
Weight on Paper	Compliance under loading	93.27	**97.28**	**+4.00**
Weight on Pillow	Soft-body compression	40.09	**43.80**	**+3.71**
Temporal-Causality Effects				
Stable vs. Unstable Support	Support-state transition	92.97	**97.13**	**+4.16**
Unstable Block Stack	Instability propagation	22.91	**30.82**	**+7.91**
Single Cradle	Contact-driven momentum transfer	18.61	**19.85**	+1.24
Ball Hits Duck	Collision-induced state change	**78.99**	78.23	−0.76
Physics-IQ Original	*Overall benchmark*	59.24	**60.22**	**+0.97**

C.5 Cross-Backbone Capability Transfer

The main paper shows that the same frozen Control Latent Predictor improves motion alignment when injected into CogVideoX, PhyCo, and Force Prompting through lightweight interface adaptation. Table 6 further examines this transfer on Physics-IQ (Motamed et al., 2026), covering deformation-sensitive events and temporal physical causality.

Our transfer design separates learned material-response prediction from backbone-specific conditioning. Both the Control Latent Predictor and target video backbone remain frozen, while decoded Control Latent Predictor responses are adapted to each model's control interface. For CogVideoX and Force Prompting, we train only the ControlNet input encoders, patch projection, and output projections, with 19.69M and 57.45M trainable parameters, respectively. For PhyCo, decoded responses provide temporal soft masks that update the deformation-property maps, requiring only 1.31M trainable parameters in temporal output gates. This allows the learned response information to be reused across different conditioning architectures without retraining the predictor or video backbone. The PhyCo adapter is further reused on Physics-IQ without additional training.

Control Latent Predictor injection improves PhyCo on six of eight selected cases. *Weight on Paper* and *Weight on Pillow* probe how materials deform under loading, whereas *Stable vs. Unstable Support* and *Unstable Block Stack* probe how changes in support affect subsequent motion. The gains across these cases suggest that deformation-response supervision provides useful cues for both local material response and the temporal evolution of physical interactions. The overall Physics-IQ score also increases from 59.24 to 60.22, although the two small regressions indicate that the benefits remain task-dependent. Across the full Physics-IQ benchmark, performance remains largely unchanged on scenarios that do not primarily depend on intrinsic material properties. Together, these observations suggest that Control Latent Predictor injects transferable material-response understanding into an existing video generator while largely preserving its broader capabilities. This extends the role of the learned response latent beyond the original Wan conditioning interface, without implying uniform improvement across all physical-reasoning tasks.

C.6 Failure Cases and Limitations

PhysWonder-3D is currently designed primarily for tabletop-scale scenes. Its simulation-ready world construction samples the reconstructed object, background geometry, and support surface into a particle representation, which enables explicit contact handling and reduces object-ground interpenetration. This design works well for compact scenes with relatively simple support geometry, but scaling to large, cluttered, or geometrically complex environments may require denser or spatially adaptive sampling and, in some cases, manual adjustment of the particle sampling frequency.



Figure 11: **Failure cases of PhysWonder-Ctrl.** Across rigid, elastic, sand, and snow conditions on the same reference object, predicted control generally follows the corresponding engine-in-the-loop dynamics. Remaining failures mainly arise from appearance drift and semantic entanglement: in the sand case, the generator can introduce granular texture or geometry in addition to the intended deformation response, revealing incomplete disentanglement of appearance and physics.

PhysWonder-Ctrl is additionally limited by the capacity and visual priors of its video backbone. Although predicted control improves material-dependent motion and deformation, appearance and object identity may drift, and material labels can remain partially entangled with visual semantics. For example, a sand-like response may introduce granular texture or local geometry changes in addition to the intended deformation. Future work will extend simulation-ready world construction to larger and more complex scenes, while stronger video backbones and better disentanglement between appearance and material response may further improve generation fidelity.

D Interaction-Shared, Response-Specific Hybrid Solver

This appendix provides the solver details needed to instantiate the simulation operator $\mathcal{T}$ in Eq. (3). MLS-MPM (Hu et al., 2018) supplies the shared particle-grid transport, while material dynamics are determined by response-specific constitutive updates and native rigid-body evolution. The resulting branches expose a common particle-state interface without collapsing rigid, elastic, sand, and snow into a single constitutive model.

D.1 Response-Specific Hybrid Update

For particle i, the shared trial deformation gradient is

$$\mathbf{F}_i^{\mathrm{tr}} = \left(\mathbf{I}_3 + \Delta t\, \mathbf{C}_i^n\right) \mathbf{F}_i^n. \tag{15}$$

The material-dependent update is

$$\left(\bar{\mathbf{F}}_i, J_{p,i}^{n+1}, \boldsymbol{\tau}_i^{\mathrm{K}}\right) = \mathcal{P}_{\chi_i}\left(\mathbf{F}_i^{\mathrm{tr}}, J_{p,i}^n; \boldsymbol{\psi}_i\right), \qquad \mathbf{F}_i^{n+1} = \bar{\mathbf{F}}_i, \tag{16}$$

where χ_i and $\boldsymbol{\psi}_i$ denote the material family and its parameters. Elastic particles retain recoverable stretch using a fixed-corotated stress; sand applies a Drucker–Prager return mapping; snow uses capped principal stretches with a plastic-volume update (Stomakhin et al., 2013; Klár et al., 2016). Rigid particles retain only the proper rotational factor and contribute no deformable stress:

$$\mathcal{P}_{\mathrm{rigid}}\left(\mathbf{F}^{\mathrm{tr}}, J_p\right) = \left(\mathrm{Rot}\left(\mathbf{F}^{\mathrm{tr}}\right), J_p, \mathbf{0}\right). \tag{17}$$

The returned Kirchhoff stress enters the shared affine-momentum update

$$\mathbf{A}_i = m_i \mathbf{C}_i^n - \frac{4\Delta t\, V_i^0}{\Delta x^2} \boldsymbol{\tau}_i^{\mathrm{K}}. \tag{18}$$

P2G scatters particle mass and affine momentum using quadratic B-spline weights. A global grid contains all particles and updates non-rigid particles, while the body-conditioned grid for rigid cluster r contains that cluster together with all non-rigid particles:

$$\mathcal{I}_{\text{all}} = \mathcal{P}, \qquad \mathcal{I}_r = \mathcal{P}_{\text{nr}} \cup \mathcal{P}_r. \tag{19}$$

Rigid cluster r reads from its body-conditioned grid, whereas non-rigid particles read from the global grid. Direct rigid–rigid interaction is therefore left to the native contact stage.

D.2 Native Rigid Coupling

Each rigid cluster maintains the authoritative body state

$$\mathbf{b}_r^n = (\mathbf{c}_r^n, \mathbf{Q}_r^n, \mathbf{V}_r^n, \boldsymbol{\omega}_r^n)\,. \tag{20}$$

After body-conditioned G2P, its provisional particle velocities are projected onto the rigid velocity manifold:

$$\left(\widetilde{\mathbf{V}}_r, \widetilde{\boldsymbol{\omega}}_r\right) = \arg\min_{\mathbf{V},\boldsymbol{\omega}} \sum_{i\in\mathcal{P}_r} m_i \left\|\widetilde{\mathbf{v}}_i - \mathbf{V} - \boldsymbol{\omega} \times (\mathbf{x}_i^n - \mathbf{c}_r^n)\right\|_2^2 + \varepsilon_r \left\|\boldsymbol{\omega}\right\|_2^2. \tag{21}$$

The reduced twist predicts a pre-contact state, after which native SDF contact produces the final state:

$$\mathbf{b}_r^- = \text{Advance}\left(\mathbf{b}_r^n, \widetilde{\mathbf{V}}_r, \widetilde{\boldsymbol{\omega}}_r\right), \qquad \left\{\mathbf{b}_r^{n+1}\right\}_r = \text{ContactSolve}\left(\left\{\mathbf{b}_r^-\right\}_r; \boldsymbol{\xi}\right). \tag{22}$$

We retain at most eight contact samples per oriented body pair, apply eight Gauss–Seidel normal and friction sweeps, and use two contact-redetection and position-correction passes.

Let $\mathbf{r}_i^0 = (\mathbf{Q}_r^0)^\top(\mathbf{x}_i^0 - \mathbf{c}_r^0)$ be the fixed body-frame coordinate. The final body state is written back to the common particle interface:

$$\begin{aligned} \mathbf{x}_i^{n+1} &= \mathbf{c}_r^{n+1} + \mathbf{Q}_r^{n+1}\mathbf{r}_i^0, \\ \mathbf{v}_i^{n+1} &= \mathbf{V}_r^{n+1} + \boldsymbol{\omega}_r^{n+1} \times \left(\mathbf{x}_i^{n+1} - \mathbf{c}_r^{n+1}\right), \\ \mathbf{C}_i^{n+1} &= \left[\boldsymbol{\omega}_r^{n+1}\right]_\times, \qquad \mathbf{F}_i^{n+1} = \mathbf{Q}_r^{n+1}. \end{aligned} \tag{23}$$

D.3 Deformation-Response Readout

The synchronized particle state provides the response observation used by PhysWonder-3D. For material branch m at frame k, we compute

$$r_i^{k,(m)} = \left\|\log \boldsymbol{\sigma}\left(\mathbf{F}_i^{k,(m)}\right)\right\|_2, \qquad G_k^{(m)} = \mathcal{C}\left[\mathcal{R}_{\pi_{\text{cam}}}\left(\left\{\left(\mathbf{x}_i^{k,(m)}, \mathcal{N}(r_i^{k,(m)})\right)\right\}_{i=1}^N\right)\right]. \tag{24}$$

The moving support records particle transport, while the attached deformation-gradient-derived stretch records local deformation. Their joint evolution provides aligned, particle-supported deformation-response supervision rather than a complete stress or constitutive-state representation.

E Details of Control Latent Predictor and Generator-in-the-Loop Response Learning

PhysWonder-Ctrl implements our core insight: constitutive-response capabilities can emerge within video generators through learning deformation dynamics. We realize this idea by jointly learning a response representation and its effect on generated dynamics. Aligned deformation-response supervision grounds the representation in simulated material behavior, while video supervision teaches the generator to express this behavior.

Control Latent Predictor. As introduced in Sec. 4.3, the Control Latent Predictor maps the reference scene and physical condition to a deformation-response latent:

$$\mathbf{h}_{\text{ctrl}} = \mathcal{P}_\phi\left(\mathbf{z}_{\text{ref}}, t, \mathbf{c}_{\text{sem}}\right), \tag{25}$$

Algorithm 1: One substep of the interaction-shared, response-specific hybrid solver

Input: Particle and rigid states; material families and parameters $\{\chi_i, \boldsymbol{\psi}_i\}_i$; interaction configuration $\boldsymbol{\xi}$; time step Δt and grid spacing Δx
Output: Updated particle and rigid states
Clear the global grid $\mathcal{G}^{\mathrm{all}}$ and body-conditioned grids $\left\{\mathcal{G}^{(r)}\right\}_r$;
foreach *particle* i **do**
  $\mathbf{F}_i^{\mathrm{tr}} \leftarrow (\mathbf{I}_3 + \Delta t\, \mathbf{C}_i)\, \mathbf{F}_i$;
  $\left(\bar{\mathbf{F}}_i, \bar{J}_{p,i}, \boldsymbol{\tau}_i^{\mathrm{K}}\right) \leftarrow \mathcal{P}_{\chi_i}\left(\mathbf{F}_i^{\mathrm{tr}}, J_{p,i}; \boldsymbol{\psi}_i\right)$;
  $\left(\mathbf{F}_i, J_{p,i}\right) \leftarrow \left(\bar{\mathbf{F}}_i, \bar{J}_{p,i}\right)$;
  Scatter particle mass and affine momentum to $\mathcal{G}^{\mathrm{all}}$ and its applicable body-conditioned grids;
Normalize the grids and apply gravity and simulation-domain boundary conditions;
foreach *particle* i **do**
  Select $\mathcal{G}^{(r)}$ if $i \in \mathcal{P}_r$; otherwise select $\mathcal{G}^{\mathrm{all}}$;
  Reconstruct $\left(\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{C}}_i\right)$ by MLS G2P;
Update non-rigid particle positions, velocities, and affine fields;
foreach *rigid cluster* r **do**
  Reduce $\{\widetilde{\mathbf{v}}_i : i \in \mathcal{P}_r\}$ to $\left(\widetilde{\mathbf{V}}_r, \widetilde{\boldsymbol{\omega}}_r\right)$;
  Predict the pre-contact body state $\mathbf{b}_r^-$;
Resolve native rigid-ground and rigid–rigid contact to obtain $\left\{\mathbf{b}_r^{n+1}\right\}_r$;
foreach *rigid cluster* r **do**
  Write the final body state back to its particles using Eq. (23);

where $\mathbf{z}_{\mathrm{ref}}$ is the reference-image latent, t is the flow-matching timestep, and $\mathbf{c}_{\mathrm{sem}}$ fuses pooled text and image features. The reference latent provides the spatial scaffold, while conditional modulation allows the predictor to represent material-dependent responses within that scene.

Specifically, the reference latent $\mathbf{z}_{\mathrm{ref}} \in \mathbb{R}^{B\times C\times 1\times H\times W}$ is expanded along the temporal dimension and projected by a $3\times 3\times 3$ convolution into a spatiotemporal feature tensor $\mathbf{X}$. The predictor applies N stacked feature-wise linear modulation (FiLM)-conditioned 3D residual blocks. In each block, semantic and timestep conditions are mapped to modulation parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$:

$$\hat{\mathbf{X}} = \mathrm{GroupNorm}(\mathbf{X}) \odot (1 + \boldsymbol{\gamma}) + \boldsymbol{\beta}. \tag{26}$$

Sigmoid-Weighted Linear Unit (SiLU) activations, 3D convolutions, and residual connections process the modulated features, followed by a final projection to $\mathbf{h}_{\mathrm{ctrl}}$.

Learning Responses Through Their Generative Effect. During training, the simulation-derived deformation-response sequence provides the target latent

$$\mathbf{h}_{\mathrm{ctrl}}^{gt} = \mathcal{E}_{\mathrm{VAE}}(\mathbf{G}_{1:T}), \tag{27}$$

where $\mathcal{E}_{\mathrm{VAE}}$ is the VAE encoder. Latent alignment transfers the simulated response structure, but matching a latent alone does not ensure that the generator reproduces the corresponding deformation dynamics. We therefore use the predicted latent as the actual conditioning input during joint training. The video loss propagates through this latent to the predictor, optimizing both the predicted response and how the generator expresses it. At inference, the predictor uses only the reference image, physical prompt, and timestep, without simulator rollouts or deformation-response videos.

Flow-Matching View of Generator-in-the-Loop Response Learning. We examine how video supervision trains the predicted latent through its effect on generation. Following Sec. 3.2, let

$$\mathbf{y} = \mathbf{z}_1 - \mathbf{z}_0 \tag{28}$$

denote the FlowMatchSFT (Wan et al., 2025) regression target, and define the privileged conditioning variable

$$\mathbf{c}_{\mathrm{gen}}^{gt} = \left(\mathbf{z}_{\mathrm{ref}}, \mathbf{c}_{\mathrm{sem}}, \mathbf{h}_{\mathrm{ctrl}}^{gt}\right). \tag{29}$$

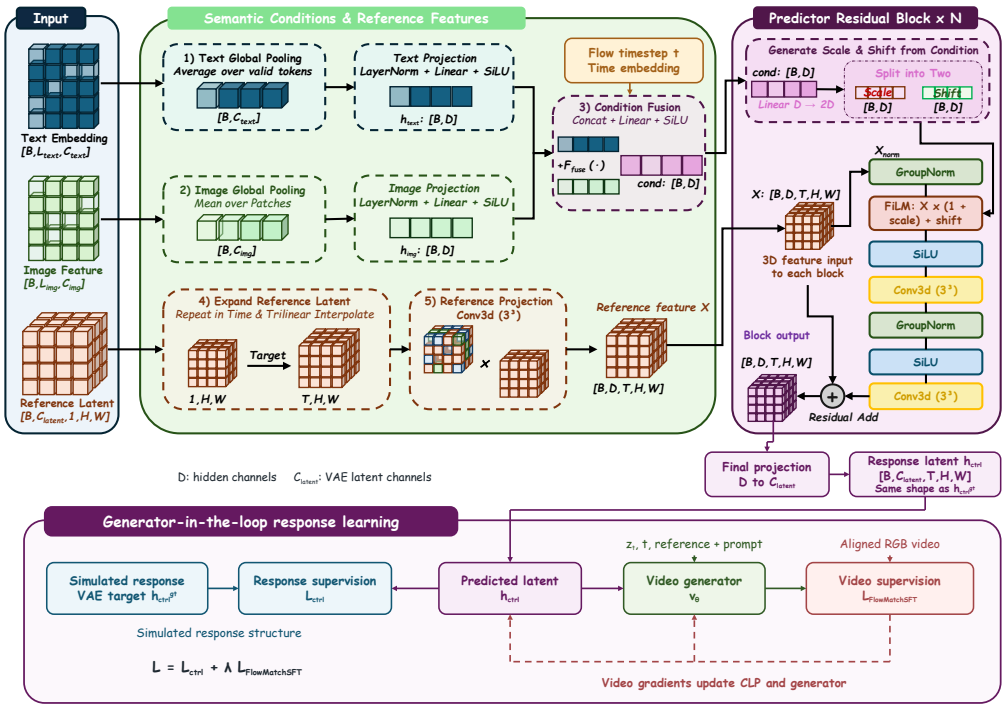


Figure 12: **Detailed architecture of the Control Latent Predictor in PhysWonder-Ctrl.** The reference latent is temporally expanded into a spatiotemporal scaffold. Pooled text and image features form the semantic condition, which combines with the timestep condition to FiLM-modulate stacked 3D residual blocks. The resulting response latent conditions video generation and is jointly optimized with response-latent and video supervision.

For the linear flow path, the optimal conditional velocity field under privileged deformation-response supervision is

$$\mathbf{u}_t^{gt}(\mathbf{z}) = \mathbb{E}\left[\mathbf{y} \mid \mathbf{z}_t = \mathbf{z},\ t,\ \mathbf{c}_{\text{gen}}^{gt}\right]. \tag{30}$$

During joint training and inference, the predicted latent from Eq. 25 supplies the conditioning variable

$$\mathbf{c}_{\text{gen}}^{pred} = (\mathbf{z}_{\text{ref}}, \mathbf{c}_{\text{sem}}, \mathbf{h}_{\text{ctrl}}). \tag{31}$$

Let

$$\mathbf{v}^{pred} = v_\theta\left(\mathbf{z}_t, t, \mathbf{c}_{\text{gen}}^{pred}\right). \tag{32}$$

Since $\mathbf{h}_{\text{ctrl}}$ is a deterministic function of $(\mathbf{z}_{\text{ref}}, t, \mathbf{c}_{\text{sem}})$, $\mathbf{v}^{pred}$ is measurable with respect to $(\mathbf{z}_t, t, \mathbf{c}_{\text{gen}}^{gt})$. Writing $\mathbf{u}_t^{gt}$ for $\mathbf{u}_t^{gt}(\mathbf{z}_t)$, the orthogonality property of conditional expectation gives

$$\begin{aligned}\mathcal{L}_{\text{FlowMatchSFT}} &= \mathbb{E}\left[\left\|\mathbf{y} - \mathbf{v}^{pred}\right\|_2^2\right] \\ &= \underbrace{\mathbb{E}\left[\left\|\mathbf{y} - \mathbf{u}_t^{gt}\right\|_2^2\right]}_{\text{irreducible conditional variance}} + \underbrace{\mathbb{E}\left[\left\|\mathbf{v}^{pred} - \mathbf{u}_t^{gt}\right\|_2^2\right]}_{\text{privileged conditional-flow mismatch}}.\end{aligned} \tag{33}$$

The cross term vanishes because

$$\mathbb{E}\left[\mathbf{y} - \mathbf{u}_t^{gt} \mid \mathbf{z}_t, t, \mathbf{c}_{\text{gen}}^{gt}\right] = \mathbf{0}. \tag{34}$$

The first term in Eq. 33 is independent of the predictor and generator. Minimizing FlowMatchSFT therefore minimizes the mismatch between the velocity field conditioned on the predicted response and the optimal field conditioned on the simulated response. Because the predicted latent lies on the generation path, this objective trains the predictor through the dynamics produced by the generator.

We combine this objective with response-latent alignment:

$$\mathcal{L}_{\text{ctrl}} = \left\| \mathbf{h}_{\text{ctrl}} - \mathbf{h}_{\text{ctrl}}^{gt} \right\|_1, \tag{35}$$

$$\mathcal{L}_{\text{Video}} = \mathcal{L}_{\text{ctrl}} + \lambda \mathcal{L}_{\text{FlowMatchSFT}}. \tag{36}$$

Together, these losses connect material conditions, simulated deformation responses, and generated dynamics. Learning this connection turns simulation-derived supervision into an internal material-response capability, enabling material-controllable generation without simulation at inference.

F Design of a Large-Scale Physically Aligned Synthesis Pipeline

PhysWonder-3D supplies physically aligned videos and deformation-response supervision for training PhysWonder-Ctrl. This appendix documents the practical steps that turn a single image into initialized geometry and appearance: instance-level configuration, geometric alignment, particle discretization, and appearance correction.

F.1 Instance-Level Analysis and Agent Prompt

We use Qwen3.5 (Qwen Team, 2026) to assign each foreground instance a simulation configuration. The model receives the parameter-selection prompt, the original image, and a copy annotated with SAM3 (Carion et al., 2026) bounding boxes and instance IDs. The IDs are ordered from left to right and link the response to the corresponding masks and reconstructed objects. The output is a JSON object indexed by these IDs, containing *E_default* (Young's modulus in Pa), *constitutive* (rigid, elastic, sand, or snow), and a numerical initial velocity. It also includes two initialization decisions: `on_ground` indicates contact with the visible support plane, and *filling* determines whether interior particles are added. The predicted stiffness is used as a bounded simulation setting, not a measured material constant. The implementation validates instance IDs and field types and bounds the stiffness before use; density and contact friction are supplied by the simulator configuration.

In *auto* mode, each foreground object uses its predicted constitutive type, allowing different materials within one scene. Counterfactual synthesis assigns one specified constitutive family to all foreground objects regardless of appearance, keeping geometry and initial conditions fixed to produce paired material responses (Appendix C).

Prompt Given to the Parameter Selection Agent

You are a physical parameter analyst. The foreground instance IDs, ordered left-to-right, are {`instance IDs`}. Estimate every listed instance's constitutive type, Young's modulus in Pa, whether it is initially on the visible ground plane, initial velocity, and whether it should be volume-filled. Detailed description:
– **velocity:** `right_mps` means image-right along the ground, `forward_mps` means away from the camera along the ground, and `up_mps` means away from the ground. Unit is m/s. Use 0.0 for a stationary component.
– **Motion requirement:** velocities are simulation initial conditions chosen to produce a physically plausible, visually informative motion, not merely an estimate of motion visible in this single image. At least one foreground instance must have a non-zero velocity with speed at least 0.3 m/s. Prefer an object that appears tilted, unsupported, precariously balanced, overhanging an edge, rolling, or likely to slide or fall. If no object appears unstable, choose the most suitable movable foreground object and assign a plausible horizontal push. Keep clearly fixed or well-supported objects stationary unless motion is physically plausible. Use realistic speeds: normally 0.2-1.0 m/s and never above 2.0 m/s.
– **E_default:** Young's modulus (elastic stiffness) of the object, in Pa.
– **constitutive:** exactly one of `rigid` (undeformable solid), `elastic` (recoverable deformable solid), `sand` (dry granular material), or `snow` (compressible snow-like material).
– **on_ground:** whether the object is initially on the ground plane.
– **filling:** true for a volumetrically solid object that should receive interior particles; false for a hollow object.
Return only a JSON object with top-level key `objects`. Its keys must be exactly the listed instance IDs. Each object value must contain `constitutive`, `E_default`, `on_ground`, `filling`, and `velocity`; velocity must contain `right_mps`, `forward_mps`, and `up_mps`.

The box reproduces the implemented instruction with typographic formatting; {`instance IDs`} is replaced with the detected IDs at runtime.

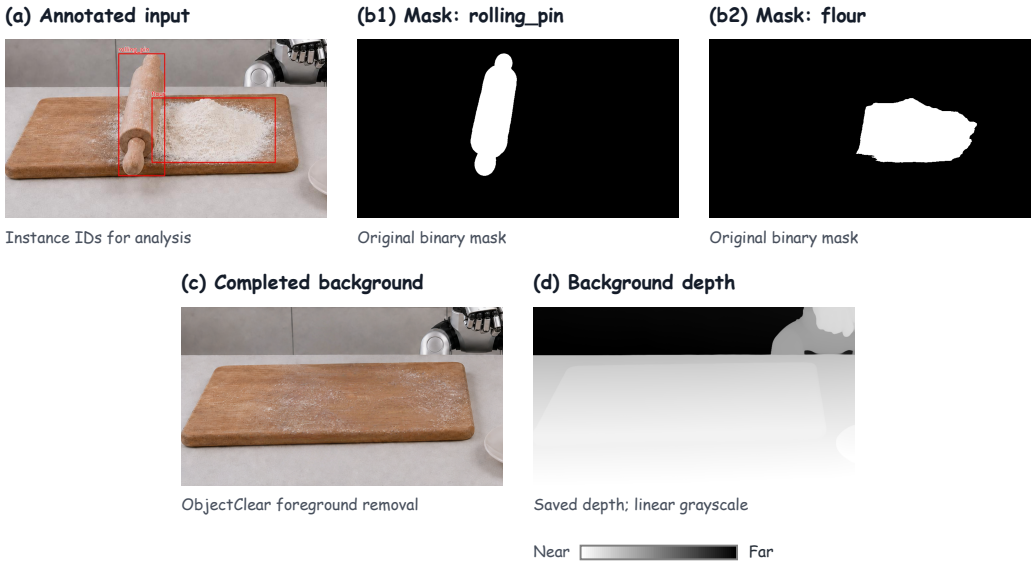


Figure 13: **Geometric preparation in a recorded PhysWonder-3D example.** (a) Input image with SAM3 annotations; (b1–b2) the original binary masks for the two foreground instances; (c) completed background; (d) saved background depth, displayed in continuous grayscale with nearer regions brighter. All panels derive from the same run. The depth display uses a linear mapping between the second and ninety-eighth percentiles of valid depth, clipping values outside this range; no contour overlay or smoothing is applied. Original-image depth guides object reconstruction after cross-image scale alignment, while background depth supplies the support-plane estimate.

F.2 Geometric Alignment from Image and Depth

Figure 13 illustrates the intermediate results of geometric preparation. SAM3 provides instance masks and a support-plane mask from the input image. The union of foreground masks guides ObjectClear (Zhao et al., 2026) to remove the foreground and complete the background, providing an unobstructed scene for background rendering and support estimation. MoGe (Wang et al., 2025) is applied separately to the original and completed images. Because their point maps can have different scales, we align the original-image point map to the completed-background point map using valid pixels outside the foreground masks. We reject outliers in the ratios of camera-to-point distances and fit a single scale on the remaining correspondences. This aligns the estimates without per-pixel depth optimization. The original RGB image, each instance mask, and the rescaled original-image point map are passed to SAM-3D Objects (Chen et al., 2026) to obtain textured meshes and initial placements. Separately, we fit a support plane by RANSAC to valid background points selected by the support mask. We then refine each mesh's scale and, for an instance marked *on_ground*, its translation along the support plane while enforcing contact. The selection criterion combines projected-center and bounding-box extent errors with the fraction of projected vertices outside the instance mask. This reduces gross size and placement mismatch without requiring exact silhouette agreement with an imperfect single-view reconstruction. Finally, a shared rotation, translation, and, when needed, uniform scene scaling place the geometry in the simulation domain with the support normal pointing upward; the inverse transform preserves the reference camera for rendering.

F.3 From Reconstructed Meshes to Material Particles

The reconstructed mesh specifies the object's initial shape and appearance. For simulation, we sample its surface at a shared spatial spacing and, when *filling* is true, add voxel-based interior samples while avoiding duplicate samples near the surface. The samples become material particles with a common reference volume determined by the spacing, and their subsequent response is governed by the selected material branch.

This representation is particularly useful for sand and snow. A fixed-connectivity mesh carried through the simulation would continue to connect vertices as material spreads or separates, potentially leaving stretched faces across regions that should have dispersed. Material particles can undergo large irreversible rearrangements without retaining the reconstruction mesh's face connectivity, allowing granular spreading and snow-like compaction or breakup to be represented by the hybrid MPM solver. The particle conversion alone does not prescribe these behaviors: they arise from the constitutive updates in Appendix D. It provides a shared geometric representation for material responses, aligned rendering, and supervision without separate remeshing or fracture procedures.

F.4 Appearance Correction and Interior Colors

The texture generated by SAM-3D Objects can differ from the input in brightness and color, including lighting baked into the reconstruction. These differences can weaken the visual correspondence between the reference image and the synthesized training video. We therefore correct the reconstructed UV texture using observations from the input image before transferring colors to material particles.

Mesh vertices are projected into the reference view, and the nearest vertex at each occupied pixel is retained as an approximate visibility test. Only projections inside an eroded instance mask are used for color comparison, reducing contamination from the background and uncertain silhouettes. We estimate a robust median color residual in CIELAB space between these image observations and their sampled texture colors, and apply the correction to the UV texture. This compensates for an overall appearance mismatch while preserving the reconstructed texture's spatial detail; it is not an inverse-lighting or intrinsic-reflectance estimate.

Both surface and interior particles receive colors from the corrected texture. For each particle, the closest point on the reconstructed surface identifies a triangle; barycentric interpolation of its UV coordinates supplies the texture lookup. Assigning colors to the interior avoids uncolored regions when deformation or granular spreading exposes previously hidden material. For these single-view assets, transferring nearby surface appearance is a practical prior for the unobserved interior, rather than a recovery of its true texture. The colors remain attached to the simulated particles, maintaining appearance continuity as their positions evolve.