PHYSVERSE: DEFORMABLE SIMULATION-READY 4D WORLD MODELING FROM MONOCULAR VIDEOS.

Anonymous authors
Paper under double-blind review

ABSTRACT

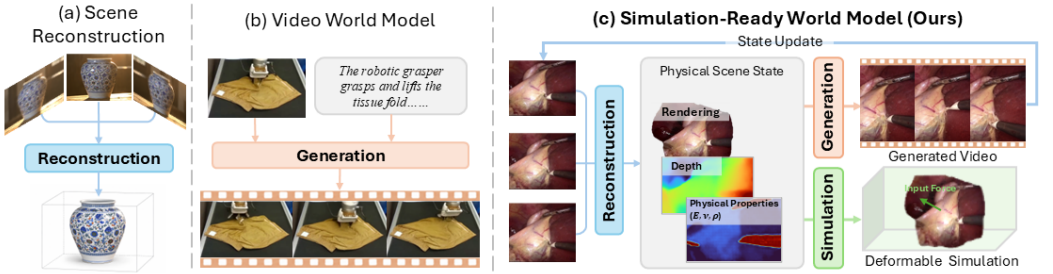
World modeling aims to predict future observations while retaining a scene state for downstream use. Existing reconstruction and generation-centric methods mainly capture geometry and visual dynamics, and rarely represent the spatially varying material parameters a soft-body simulator requires, so the retained state cannot be used directly for deformable simulation. We present PhysVerse, a deformable simulation-ready 4D world model that links reconstruction, video generation, and deformable simulation through a shared scene state. Given monocular video, a feed-forward reconstructor jointly predicts depth, dynamic Gaussian primitives, and dense maps of Young's modulus, Poisson's ratio, and density in a single pass. The reconstructed geometry and material maps initialize MPM-based deformable simulation, while the same depth and material maps condition future video prediction. Generated frames can be passed once again through the reconstructor to recover geometry and material maps. We evaluate PhysVerse on both surgical and general robotic video benchmarks. Experiments show that the reconstructed state supports deformable rollout under MPM, and that depth and material conditioning improve video fidelity and temporal consistency. Our demo is available at: https://anonymous.4open.science/w/doaneqzp-0DBD/.

1 INTRODUCTION

World modeling methods offer a scalable way to learn and predict scene dynamics from visual observations and other conditions. For soft and deformable scenes, that prediction is more useful when the retained scene state is one a soft-body simulator can advance. Such a state matters in general interaction, and becomes especially important in surgical operations and robotic manipulation of non-rigid objects, where spatially varying material parameters such as stiffness and density shape how contact unfolds (Dagli et al., 2025). Existing world modeling methods, however, primarily represent scene geometry, appearance, and visual dynamics. This leaves a gap in soft and deformable scenes, such as surgical environments, where rolling the scene forward in a simulator depends on material parameters that these models do not provide. In surgical settings, Endora and SurgSora improve the realism and controllability of video synthesis, while HieraSurg and Surgical Vision World Model introduce structured future prediction through semantic scene evolution and latent action modeling (Li et al., 2024; Chen et al., 2025; Biagini et al., 2025; Koju et al., 2025). More recently, Cosmos-H-Surgical formulates action-conditioned video generation for robotic manipulation (He et al., 2025), and simulation platforms like ORBIT-Surgical (Yu et al., 2024) benchmark dexterous surgical manipulation in physics-based soft-tissue environments. However, these conditions mainly characterize observable scene appearance, geometry, and motion, and leave the scene without the material parameters a soft-body solver needs, so the world model cannot by itself run deformable simulation, limiting the downstream application for deformable interaction.

Meanwhile, a separate line of work recovers material parameters for simulation from visual observations. NeRF2Physics optimizes language-embedded features over a reconstructed NeRF, whereas PhysGen and PhysFlow combine foundation-model reasoning with physics-based motion modeling or simulation (Zhai et al., 2024; Liu et al., 2024; 2025b). Earlier differentiable physical frameworks such as DiSECt (Heiden et al., 2021) demonstrate parameter identification for soft and cutting interactions, but require dedicated contact setups. More recent feed-forward methods predict material fields from reconstructed geometry or visual observations. VoMP aggregates multi-view features over a voxelized asset, while ReconPhys and PhysGM recover simulatable Gaussian representations together with

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.



Category	Method	Feed-forward	Input	Output			
			Type	Geometry	Video	Material Maps	Simulation
Scene Reconstruction	Endo-4DGS (Huang et al., 2024)	×	Video	✓	×	×	×
	SurgCUT3R (Xu et al., 2026)	✓	Video	✓	×	×	×
Video Generation	SurgSora (Chen et al., 2025)	–	Image + Control	×	✓	×	×
	Cosmos-H-Surgical (He et al., 2025)	–	Image	×	✓	×	×
Geometry-Guided World Modeling	NeoVerse (Yang et al., 2026)	✓	Video	✓	✓	×	×
Object-Centric Physical Modeling	VoMP (Dagli et al., 2025)	✓	Mesh / 3DGS	×	×	✓	×
	PhysGS (Chopra et al., 2026)	×	Image	✓	×	✓	✓
	PhysGM (Lv et al., 2026)	✓	Image	✓	×	∼	✓
Simulation-Ready World Modeling	**PhysVerse (Ours)**	✓	Image / Video	✓	✓	✓	✓

Figure 1: Comparison with representative world-modeling paradigms. PhysVerse combines feed-forward reconstruction, future video prediction, dense material maps, and deformable simulation. ✓/× denote supported/unsupported, while ∼ denotes partial support.

material parameters from monocular videos or images (Dagli et al., 2025; Wang et al., 2026; Lv et al., 2026). These methods primarily focus on object-centric assets and simulation, and do not maintain a scene-level 4D state that jointly supports future video prediction and deformable simulation.

To address these limitations, we propose PhysVerse, a deformable simulation-ready 4D world model built upon an explicit scene state that a soft-body simulator can consume. Given monocular video, PhysVerse reconstructs scene geometry, dynamic Gaussian primitives, and dense material maps of Young's modulus E, Poisson's ratio ν, and density ρ, which together form this scene state. A video generator conditions a frozen Wan2.2-TI2V-5B video diffusion backbone on the reconstructed material maps and depth for future video prediction. The reconstructed geometry and material maps are assigned to simulation particles, providing a scene representation for Material Point Method (MPM)-based deformable simulation (Hu et al., 2018; Xie et al., 2024). Generated frames can be passed again through the reconstructor to recover geometry and material maps, which we illustrate qualitatively. On surgical and general-domain deformable videos, depth and material conditioning improve generation quality and temporal consistency, while the reconstructed state supports MPM rollout. In summary, our contributions are:

1. **Deformable simulation-ready 4D world modeling.** We introduce a 4D world model that extends conventional visual and geometric scene representations with dense material maps, so the same state supports future video prediction and initializes deformable simulation.
2. **Feed-forward material prediction for simulation.** Our Physics DPT head predicts maps of Young's modulus, Poisson's ratio, and density alongside geometry and dynamic Gaussian primitives in a single forward pass from monocular video, yielding a scene representation that an MPM simulator can consume.
3. **Geometry and material-conditioned generation.** We condition a frozen video diffusion transformer on the reconstructed depth and material maps, enabling temporally coherent future frame prediction from the same scene state used for deformable simulation.
4. **Evaluation across reconstruction, generation, and simulation.** Experiments on surgical and general-domain datasets evaluate the predicted material maps, future video generation quality, temporal consistency, and the ability to support the downstream MPM simulation.

2 Related Work

Object-Centric Physical Modeling. Benefiting from advances in geometric and rendering reconstruction (Kerbl et al., 2023; Wang et al., 2025), a line of work recovers material parameters that can be consumed by a simulator, rather than modeling appearance alone. Early methods rely on iterative optimization to obtain these parameters. NeRF2Physics (Zhai et al., 2024) combines language-embedded 3D representations with LLM material priors and CLIP features to infer physical properties, while PhysGS (Chopra et al., 2026) estimates dense physical property fields over 3D Gaussian representations but requires per-scene optimization. PhysGen (Liu et al., 2024) instead combines vision-language reasoning, physics simulation, and video diffusion to generate object motion. More recent studies adopt feed-forward, optimization-free pipelines to improve reconstruction efficiency. VoMP (Dagli et al., 2025) predicts per-voxel physical properties (E, ν, ρ) in a single forward pass using DINOv2 features, but requires mesh or Gaussian splat inputs rather than multi-view RGB images. ReconPhys (Wang et al., 2026) recovers simulatable non-rigid 3D Gaussian Splatting (3DGS) assets and physical parameters from monocular videos in about one second, eliminating time-consuming per-scene optimization. PhysGM (Lv et al., 2026) takes a single RGB image and jointly predicts 3D Gaussian Splatting assets and physical parameters in a feed-forward manner, enabling fast 4D dynamic simulation without per-scene tuning. These methods primarily target individual objects. They provide object-level assets for simulation, but do not maintain a scene-level 4D state shared by future video prediction and deformable simulation.

Geometry-Guided World Models and Video Generation. Earlier visual dynamics architectures such as SlotFormer (Wu et al., 2023) model future rollouts via latent object-centric representations, but lack explicit continuum-mechanical states (e.g., stiffness and constitutive laws) necessary for forward physical simulation. Recent video-generation methods use reconstructed 3D/4D scene representations as geometric guidance, improving spatial consistency and controllability (Ren et al., 2025; Yang et al., 2026). NeoVerse couples feed-forward reconstruction with future video generation from monocular video, yet the retained state remains geometric and visual. Geometric guidance alone does not supply the spatially varying material parameters a soft-body simulator requires, so these world models cannot by themselves run deformable simulation. Surgical scenes make this gap concrete: soft-tissue deformation, occlusion, and tool–tissue interaction all depend on how the scene state is advanced under contact. Early approaches include TPG-VAE for future frame prediction (Gao et al., 2021), Endora for endoscopic video generation (Li et al., 2024), and text-guided SurGen and Ophora for laparoscopic and ophthalmic scenarios (Cho et al., 2024; Li et al., 2025). Recent methods introduce richer control signals. SurgSora conditions generation on RGB, depth, segmentation, optical flow, and trajectories (Chen et al., 2025), while HieraSurg predicts semantic maps before rendering videos with action and panoptic information (Biagini et al., 2025). VISAGE models tool-tissue interactions using action scene graphs for future video prediction (Yeganeh et al., 2024). Beyond standalone video synthesis, SurgWM learns controllable generation and dynamics prediction from visual inputs (Koju et al., 2025), while Cosmos-H-Surgical generates manipulation videos from da Vinci control signals (He et al., 2025). These conditions describe visible scene content, geometry, and motion. They leave the scene without a material field that a soft-body solver can integrate, so the predicted future remains a video rather than a deformable simulation.

3 Method

PhysVerse reconstructs a scene state from monocular video and uses it for both future video prediction and deformable simulation. The state consists of depth, dynamic Gaussian primitives, and dense material maps of Young's modulus, Poisson's ratio, and density. Depth and material maps condition the video generator, while the geometry and material maps are assigned to particles for MPM. Generated frames can be passed once through the reconstructor to recover geometry and material maps.

3.1 Feed-Forward Material Prediction

In this work, we extend geometry and appearance reconstruction with dense material maps, so the reconstructed scene state can be passed to a soft-body simulator.

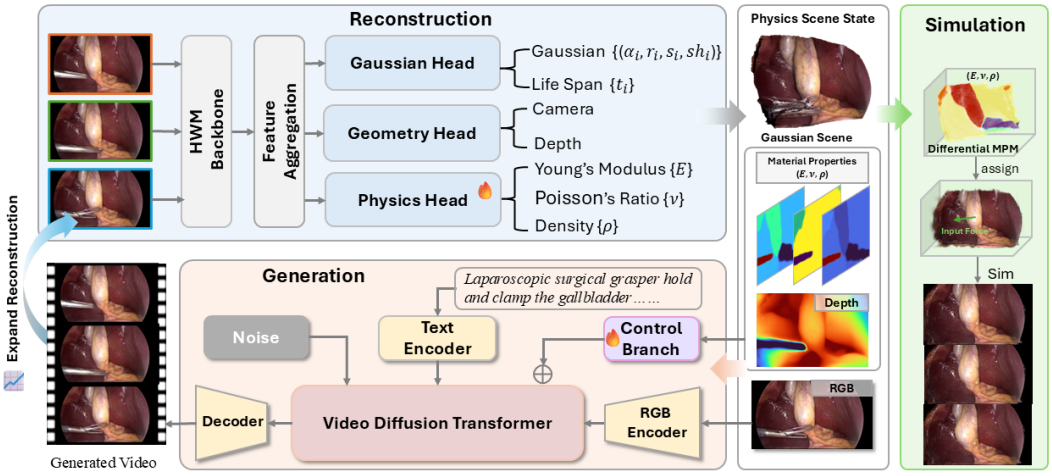


Figure 2: PhysVerse Overview. PhysVerse reconstructs geometry and dense material maps (E, ν, ρ) in a single forward pass from the monocular video input. The reconstructed geometry and material maps condition future video prediction and initialize deformable simulation.

Building upon HunyuanWorld-Mirror (Liu et al., 2025a), we retain its reconstruction backbone $\mathcal{R}_{\phi^*}$ with frozen parameters and introduce a trainable Physics Head $\mathcal{H}_\theta^m$, where the superscript m denotes the physical-property prediction branch. This architecture forms our reconstructor $\mathcal{F}_{\phi^*,\theta}$. Given an input image sequence $\mathbf{I}_{0:N-1}$, it predicts geometry, appearance, and material maps in a single forward pass:

$$\left\{\hat{\mathbf{D}}_i,\ \hat{\mathbf{C}}_i,\ \hat{\mathbf{M}}_i,\ \hat{\mathcal{G}}_i^{4\mathrm{D}}\right\}_{i=0}^{N-1} = \mathcal{F}_{\phi^*,\theta}(\mathbf{I}_{0:N-1}), \tag{1}$$

Here, $i \in \{0, \ldots, N-1\}$ indexes the input frames, while $\hat{\mathbf{D}}_i$, $\hat{\mathbf{C}}_i$, $\hat{\mathbf{M}}_i$, and $\hat{\mathcal{G}}_i^{4\mathrm{D}}$ denote the predicted depth map, camera parameters, material map, and dynamic Gaussian primitives for frame i, respectively. The parameters ϕ^* are frozen, whereas θ parameterizes the trainable Physics Head $\mathcal{H}_\theta^m$. For each input frame i, the frozen reconstruction backbone produces output tokens $\hat{\mathbf{T}}_i \in \mathbb{R}^{L\times D}$, where L is the number of tokens and D is their channel dimension. The Physics Head $\mathcal{H}_\theta^m$ decodes these tokens into dense material maps:

$$\hat{\mathbf{M}}_i = \left[\hat{\mathbf{M}}_i^E,\ \hat{\mathbf{M}}_i^\nu,\ \hat{\mathbf{M}}_i^\rho\right] = \mathcal{H}_\theta^m\left(\hat{\mathbf{T}}_i\right), \qquad \hat{\mathbf{M}}_i \in [0,1]^{H\times W\times 3} \tag{2}$$

Here, $\hat{\mathbf{M}}_i^E$, $\hat{\mathbf{M}}_i^\nu$, and $\hat{\mathbf{M}}_i^\rho$ denote the normalized dense maps of Young's modulus, Poisson's ratio, and density, respectively, and H and W are the height and width of each map. The predicted material maps are lifted into 3D with the reconstructed geometry and assigned to simulation particles. Together with the Gaussian primitives, they constitute a unified physical scene representation, enabling neural rendering alongside Material Point Method (MPM) simulation (Liu et al., 2026).

3.2 Geometry and Material Conditioned Generation

For video generation, we pass the initial frame through the reconstructor to obtain depth and material maps. We keep Wan2.2-TI2V-5B (Team et al., 2025) frozen and use its visual and temporal prior as the video backbone. A lightweight control branch injects the predicted depth and material latents into its DiT as residual guidance. Given an initial frame $\mathbf{I}_0$, a text condition c_{text}, and its reconstructed conditions $(\hat{\mathbf{D}}_0, \hat{\mathbf{M}}_0)$, the generator predicts the T future frames $\hat{\mathbf{V}}_{1:T}$ as:

$$\hat{\mathbf{V}}_{1:T} = G_\psi(\mathbf{I}_0, c_{\text{text}}, \hat{\mathbf{D}}_0, \hat{\mathbf{M}}_0). \tag{3}$$

Here, ψ denotes the trainable parameters of the control branch, with the Wan2.2-TI2V-5B DiT backbone kept frozen. Depth provides geometric guidance, while the material map provides spatially varying physical cues for future video generation. The RGB input is processed by the backbone, and the normalized depth and material maps are encoded by the frozen Wan VAE into $\mathbf{z}_d = \mathrm{Enc}_{\mathrm{VAE}}(\widetilde{\hat{\mathbf{D}}}_0)$

and $\mathbf{z}_m = \mathrm{Enc}_{\mathrm{VAE}}(\widetilde{\mathbf{M}}_0)$. The depth and material inputs are first mapped by modality-specific input projections. They are then processed separately by a shared set of control blocks, producing residuals $\mathbf{s}_d^{(i)}$ and $\mathbf{s}_m^{(i)}$ at stage i. The two residuals are fused using learnable sigmoid gates:

$$\mathbf{s}^{(i)} = \sigma\left(\gamma_d^{(i)}\right)\mathbf{s}_d^{(i)} + \sigma\left(\gamma_m^{(i)}\right)\mathbf{s}_m^{(i)}. \tag{4}$$

Here, $\gamma_d^{(i)}$ and $\gamma_m^{(i)}$ are learned per-stage scalar gates and σ denotes the sigmoid function. This fusion balances the depth and material conditions at each control stage. The fused residual is added to the hidden state of the corresponding DiT block on the first frame latent. Later frames are not given a separate depth or material map. Writing $\mathbf{x}^{(i)}$ for the hidden state entering block i, the update is:

$$\mathbf{x}^{(i+1)} = \mathrm{DiTBlock}_i\left(\mathbf{x}^{(i)}\right) + \mathbf{s}^{(i)}. \tag{5}$$

Further implementation details are provided in Appendix E.2.

3.3 Training Dataset Curation

Material Property Datasets. We developed an automatic procedure to construct triplets (E, ν, ρ) from multiple datasets, as shown in Figure 3. For each raw dataset, we first use a vision-language model (VLM) to detect object instances and obtain their bounding boxes and labels, which are then passed to SAM to generate pixel-wise semantic masks. For each unified semantic category, we obtain reference values (E, ν, ρ) from multiple online databases (MatWeb, LLC, 2025; The Engineering ToolBox, 2024; Department of Engineering, University of Cambridge, 2011; Baumgartner et al., 2025) and assign that category-level triplet to the corresponding semantic pixels as dense supervision. The categories include both rigid materials, such as metal instruments and other man-made objects, and deformable materials, such as soft tissue. We apply a $\log_{10}$ transformation to E and ρ, while ν is directly normalized to $[0, 1]$. For datasets that already provide semantic segmentation annotations, we skip instance detection and begin directly from the semantic-mask processing step. Dataset details are provided in Appendix C.1.

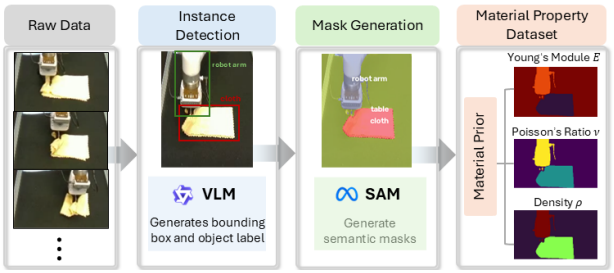


Figure 3: Material Property Dataset Curation.

Video Generation Datasets. For video generation, we construct our training and evaluation data from the public Cholec80 (Twinanda et al., 2016) and VLA-OS (Gao et al., 2026) datasets. A vision-language model (Qwen3-VL-8B-Instruct (Team, 2025)) is utilized to generate descriptions of the corresponding interactions. For Cholec80, we utilize 5,000 laparoscopic clips of 49 frames each, while for VLA-OS, we use 1000 robotic deformable-manipulation clips of 49 frames each. Each clip is paired with aligned text, depth, and material maps, as detailed in Appendix C.2.

3.4 Training Scheme

Reconstruction For the material maps, we adopt a dense regression loss with confidence weighting, similar to that used for dense depth prediction. For each frame i and pixel $u \in \Omega$, the Physics Head predicts a material vector $\hat{\mathbf{M}}_{i,u} = [\hat{M}^E_{i,u}, \hat{M}^\nu_{i,u}, \hat{M}^\rho_{i,u}]$, supervised by the normalized reference target $\mathbf{M}_{i,u}$, where Ω denotes the image-pixel domain. The material-map objective combines a pixel-wise regression term $\mathcal{L}_{\mathrm{reg}}$, a confidence-weighted term $\mathcal{L}_{\mathrm{conf}}$, and a multi-scale gradient consistency term $\mathcal{L}_{\mathrm{grad}}$:

$$\mathcal{L}_{\mathrm{mat}} = \mathcal{L}_{\mathrm{reg}} + \mathcal{L}_{\mathrm{conf}} + \mathcal{L}_{\mathrm{grad}}. \tag{6}$$

The three components are given explicitly in Appendix E.3.

Generation To train the geometry and material conditioned generator, we adopt Wan2.2-TI2V-5B as the frozen DiT backbone and follow its latent flow-matching (Liu et al., 2022) objective to learn

Table 1: Material Property Prediction for different reconstruction backbones. Best is highlighted.

Backbone	CholecSeg8k			EndoVis2018			CaDISv2		
	$\text{ALDE}_E\downarrow$	$\text{ADE}_\nu\downarrow$	$\text{ARE}_\rho\downarrow$	$\text{ALDE}_E\downarrow$	$\text{ADE}_\nu\downarrow$	$\text{ARE}_\rho\downarrow$	$\text{ALDE}_E\downarrow$	$\text{ADE}_\nu\downarrow$	$\text{ARE}_\rho\downarrow$
HWM	0.0836	0.0036	0.0534	0.0720	0.0035	0.0741	0.0987	0.0066	0.0441
VGGT	0.0850	0.0041	0.0588	0.0693	0.0036	0.0740	0.0852	0.0058	0.0539
DA3	0.1196	0.0058	0.0520	0.1006	0.0053	0.1150	0.1343	0.0105	0.0582

Backbone	VLA-OS			VKitti			UrbanSyn		
	$\text{ALDE}_E\downarrow$	$\text{ADE}_\nu\downarrow$	$\text{ARE}_\rho\downarrow$	$\text{ALDE}_E\downarrow$	$\text{ADE}_\nu\downarrow$	$\text{ARE}_\rho\downarrow$	$\text{ALDE}_E\downarrow$	$\text{ADE}_\nu\downarrow$	$\text{ARE}_\rho\downarrow$
HWM	0.0845	0.0043	0.0425	0.2487	0.0149	0.1557	0.1446	0.0084	0.0761
VGGT	0.0902	0.0044	0.0364	0.2747	0.0172	0.1671	0.1424	0.0087	0.0779
DA3	0.1110	0.0056	0.0377	0.3049	0.0175	0.1852	0.1765	0.0107	0.1020

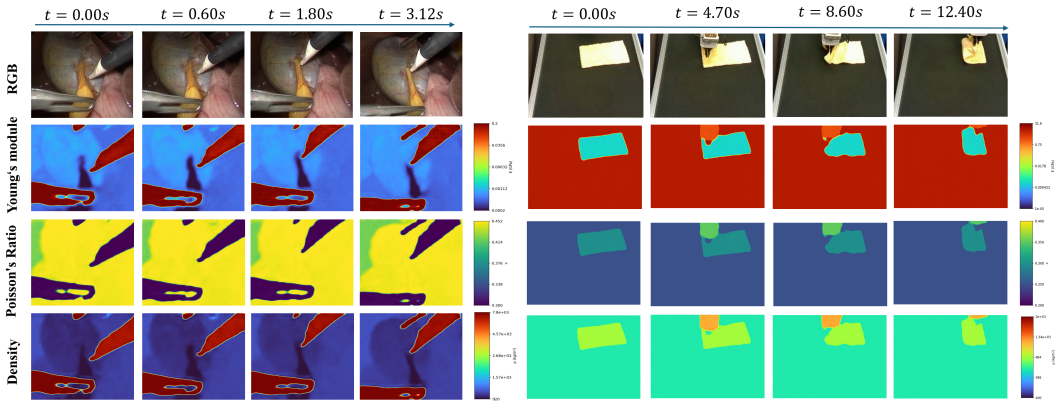


Figure 4: Qualitative Results of Material Maps. Predicted Young's modulus, Poisson's ratio, and density stay aligned with object boundaries across time.

the latent-space velocity field. Let x_1 denote the latent embedding of ground-truth clean video, and $x_0 \sim \mathcal{N}(0, I)$ be sampled standard Gaussian noise. The training objective of the control-augmented flow-based generation network f_ψ is formulated as:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{x_0,x_1,t}\left\| f_\psi(x_t, t; \mathbf{I}_0, \hat{\mathbf{D}}_0, \hat{\mathbf{M}}_0, c_{\text{text}}) - v_t \right\|_2^2, \tag{7}$$

where $x_t = (1-t)x_0 + tx_1$ stands for the linear interpolation between clean latent x_1 and noise x_0, and $v_t = x_1 - x_0$ represents the corresponding target flow velocity. c_{text} denotes text condition embeddings extracted from a pre-trained language model umT5 (Chung et al., 2023).

4 Experiments and Results

4.1 Experimental Setup

For material-map prediction, we compare three frozen backbones with the same Physics Head: HunyuanWorld-Mirror (HWM) (Liu et al., 2025a), VGGT (Wang et al., 2025), and Depth-Anything-3 (DA3) (Lin et al., 2025), and report ALDE_E, ADE_ν, and ARE_ρ. For video generation, we compare against general-purpose and domain-specific baselines, including Wan2.2-TI2V-5B (Team et al., 2025), Cosmos-Predict2.5 (Ali et al., 2025), Cosmos-H-Surgical (He et al., 2025), and SurgSora (Chen et al., 2025), and report FVD, FID, LPIPS, and VBench metrics. We further evaluate these maps in deformable simulation, reporting nRMSE and EPE against a reference MPM rollout that uses the annotated material maps and a prescribed interaction force, under identical geometry and simulation settings. For fair comparisons, we vary only material assignments in simulation. Detailed metric definitions and implementation settings are provided in Appendix B. For implementation, we freeze the HunyuanWorld-Mirror and Wan2.2-TI2V-5B backbones and train only the Physics DPT head and control modules for 10,000 and 5,000 steps, respectively, on an NVIDIA A6000 GPU. Video generation produces 49 frames at 384×640 resolution. Full training, sampling, simulation configurations, and runtime details are provided in Appendix D.

Table 2: Material Maps in Simulation. All variants use the same geometry and simulation settings, differing only in the assigned material maps. Best and second-best are highlighted.

Method	Material Map Metrics			Simulation Metrics	
	$\text{ALDE}_E \downarrow$	$\text{ADE}_\nu \downarrow$	$\text{ARE}_\rho \downarrow$	nRMSE $\downarrow$	EPE $\downarrow$
Grok4.6 (Global)	1.3676	0.0316	0.1153	21.443	32.6
ChatGPT-5.6 (Global)	1.2108	0.0324	0.1090	13.686	31.2
Ours (Global)	0.2807	0.0150	0.0852	1.199	17.5
Ours (Dense)	0.1056	0.0062	0.1325	0.260	12.7

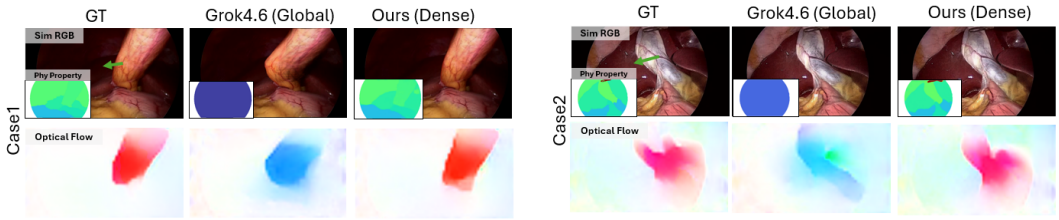


Figure 5: Simulation Comparison. Simulated deformation and optical flow when the assigned material maps differ, under the same geometry and applied force.

4.2 Main Results

4.2.1 Material Property Prediction

Given the limited prior work on dense physical property prediction, we compare three frozen reconstruction backbones for material-map prediction: HunyuanWorld-Mirror (HWM) Liu et al. (2025a), VGGT Wang et al. (2025), and Depth-Anything-3 (DA3) Lin et al. (2025). We attach the same Physics Head to each frozen backbone and train only the head under the same settings. As shown in Table 1, HWM achieves the lowest ALDE_E and ADE_ν on both CholecSeg8k and VLA-OS, and obtains the best results on all three metrics for VKitti. On UrbanSyn, it achieves the lowest ADE_ν and ARE_ρ, while VGGT attains a marginally lower ALDE_E. On EndoVis2018 and CaDISv2, VGGT attains a lower ALDE_E, and also a lower ARE_ρ and ADE_ν, respectively. HWM remains competitive on these sets, and it also provides the depth, camera parameters, and dynamic Gaussian primitives used by generation and MPM. We therefore use HWM as the reconstruction backbone of PhysVerse.

Figure 4 shows predicted material maps for temporally continuous surgical and general-domain robotic manipulation sequences. Across frames, the predictions remain spatially aligned with moving objects and preserve sharp boundaries between materials. In the surgical sequences, metallic instruments receive higher Young's modulus and density, together with lower Poisson's ratio, than the surrounding soft tissue. Gallbladder and fat receive lower Young's modulus than muscle, following the category-level reference values used as supervision. The robotic examples show the same spatial alignment under deformable object manipulation. Additional qualitative results on other datasets are provided in Appendix A.1.

We further evaluate the deformable simulation on the test sets of Cholec80-5k. Each scene uses the same reconstructed 3D Gaussian geometry and external force, while the material parameters assigned to particles come from different maps. Grok4.6 (Global) and ChatGPT-5.6 (Global) infer a material map from the full RGB frame with the prompt in Appendix D.4. Grok4.6 (Global) reaches an nRMSE of 21.443, about 21 times the mean reference displacement, which indicates an unstable rollout under the prescribed impulse. Ours (Global) uniformly assigns the spatial median of each predicted channel to all particles, whereas Ours (Dense) uses the predicted spatial field. As shown in Table 2, Ours (Dense) reduces nRMSE by 78.3% and EPE by 27.4% compared with Ours (Global), because the spatially varying map stays closer to the annotated-map reference rollout than a single global assignment. This gap is why predicting a segmentation and then filling a stored table is not a substitute for the trained dense head. A lookup assigns one database triplet to each predicted class and still needs a class inventory and a database query at test time. The head is instead trained to regress the spatial material map in one forward pass, and the MPM rollout uses that map. However,

Table 3: Quantitative Comparison of Video Generation. Results on the surgical Cholec80-5k and general-domain VLA-OS datasets. ↑ indicates higher is better, while ↓ indicates lower is better. Top-3 results are highlighted as first, second, and third.

Method	FVD ↓	FID ↓	LPIPS ↓	IQ ↑	Subj-Cons ↑	Mot-Smooth ↑	Time (s/clip) ↓
Cholec80-5k							
Wan2.2-TI2V-5B (Team et al., 2025)	83.60	95.88	0.4299	57.6	88.5	97.2	∼30
SurgSora (Chen et al., 2025)	75.85	80.62	0.3855	60.1	90.7	97.2	∼70
Cosmos-H-Surgical (He et al., 2025)	63.13	69.93	0.372	63.9	93.3	98.8	∼150
PhysVerse (Ours)	55.09	66.89	0.348	60.9	93.72	98.98	∼31
VLA-OS							
Wan2.2-TI2V-5B (Team et al., 2025)	117.68	131.65	0.514	43.5	80.1	95.4	∼30
SurgSora (Chen et al., 2025)	128.95	173.11	0.526	43.9	86.5	97.9	∼70
Cosmos-H-Surgical (He et al., 2025)	174.72	228.35	0.665	48.9	69.2	94.8	∼31
Cosmos-Predict2.5 (Ali et al., 2025)	196.27	214.00	0.679	45.3	68.9	93.7	∼150
PhysVerse (Ours)	32.31	34.95	0.189	48.2	95.0	99.4	∼31

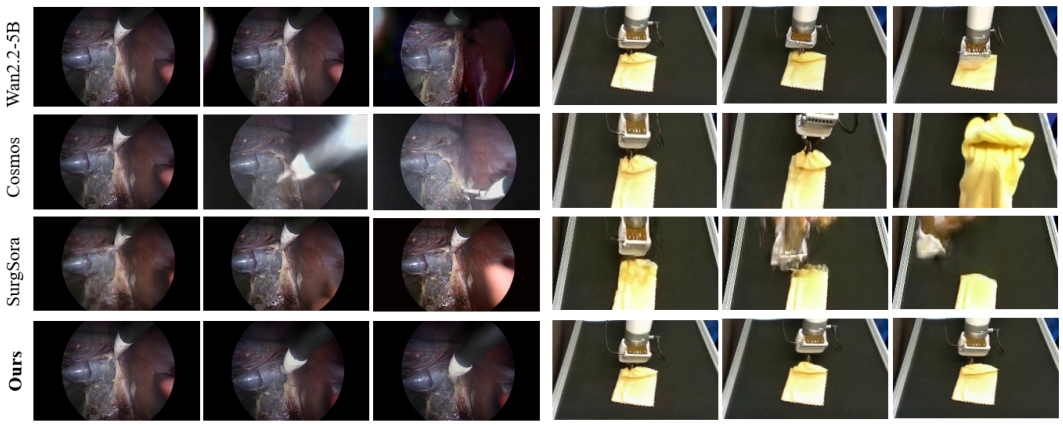


Figure 6: Qualitative Comparison of Video Generation. Each triplet shows three consecutive frames from representative Cholec80-5k and VLA-OS sequences. PhysVerse better preserves object geometry, appearance, and temporal consistency across both surgical and general-domain scenes.

Ours (Dense) has a higher ARE_ρ because the spatial median matches the dominant density, whereas local deviations in the dense ρ map increase the mean relative error. The nRMSE and EPE gains follow the more accurate spatial field of E. The remaining limits of this category-level supervision are discussed in Appendix F. Figure 5 presents the qualitative results of Table 2, showing that the predicted material property improves deformation accuracy for the simulation. Further details on the simulation setup and evaluation metrics are in Appendices D.4 and B.3.

4.2.2 Geometry and Material Conditioned Generation

We compare PhysVerse with existing video generation methods with average results on the test sets of surgical Cholec80-5k and general-domain VLA-OS datasets (Table 3). All methods are evaluated on the same test clips with matched evaluation resolution, using the same metric implementations. On Cholec80-5k, PhysVerse achieves the best performance across most reported metrics. Compared with the second-best baseline, Cosmos-H-Surgical, PhysVerse reduces FVD, FID, and LPIPS by 12.7%, 4.3%, and 6.5%, respectively. However, Cosmos-H-Surgical scores higher on IQ, gaining from its video-foundation-model pretraining, while PhysVerse still raises IQ over the original Wan2.2 backbone. On VLA-OS, PhysVerse achieves the best performance on most metrics, with consistent results in general-domain robotic manipulation. Overall, depth and material conditioning improve video fidelity and temporal consistency on both surgical and general-domain scenes. Figure 6 presents qualitative comparisons on representative surgical and general-domain robotic manipulation sequences. PhysVerse better preserves object geometry, appearance, and spatial consistency across frames, while the baselines exhibit shape drift, appearance changes, and motion artifacts.

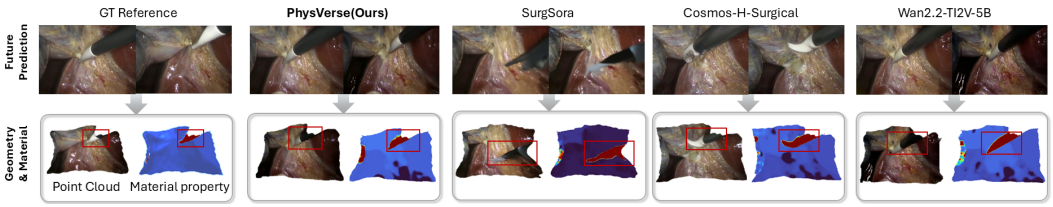


Figure 7: Generation Feedback to Reconstruction. Future frames generated by each method are fed back into the same frozen PhysVerse reconstructor.

Table 4: Ablation of depth (D) and material-map (M) conditioning on the test set. Best and second-best results are highlighted as first and second.

Method	w/Depth	w/Material	FVD↓	FID↓	SSIM↑	PSNR↑	LPIPS↓	Subj-Cons↑	Mot-Smooth↑	Temp-Flicker↑	Bg-Cons↑
PhysVerse	×	×	61.04	67.51	0.587	17.32	0.365	92.63	98.73	98.01	96.55
	✓	×	58.55	66.93	0.582	17.25	0.361	92.97	98.75	98.06	96.67
	×	✓	56.74	64.96	0.586	17.38	0.356	93.47	98.86	98.18	96.79
	✓	✓	55.09	66.89	0.589	17.60	0.348	93.72	98.98	98.37	96.95

We further check whether generated frames retain enough geometry and material information for reconstruction. We feed the generated frames once again into the frozen PhysVerse reconstructor to recover the Gaussian scene and material maps. As shown in Figure 7, PhysVerse better preserves scene geometry and instrument structure, while maintaining material boundaries aligned with the reconstructed objects. In contrast, competing methods exhibit geometric distortions, tool-shape drift, and inconsistent material regions. These results show that PhysVerse's future rollout keep geometric structure and material boundaries through the later reconstruction pass results.

4.3 Ablation Studies

Table 4 is evaluated on the Cholec80-5k test set, and each reported number is the average over all test clips. Table 4 compares four separately trained conditioning configurations. The unconditioned setting disables both depth and material inputs, whereas D-only, M-only, and $D{+}M$ use depth, material maps, and both conditions, respectively. Both single-condition variants improve FVD over the unconditioned model (58.55 for D-only and 56.74 for M-only versus 61.04), while their joint use further reduces FVD to 55.09. The $D{+}M$ model also achieves the best SSIM, PSNR, LPIPS, subject consistency, motion smoothness, temporal flickering, and background consistency. Although M-only attains the lowest FID, it does not yield the strongest overall video fidelity, whereas the joint condition improves both frame-level quality and temporal behavior. These results show that depth and material maps help generated frames preserve the geometry and material layout used to reconstruct the scene state. The qualitative ablation results are shown in Appendix A.3.

5 Conclusion

We presented PhysVerse, a deformable simulation-ready 4D world model that links feed-forward reconstruction, future video generation, and MPM simulation through a shared scene state. Given monocular video, PhysVerse reconstructs depth, camera parameters, dynamic Gaussian primitives, and dense material maps of Young's modulus, Poisson's ratio, and density. The depth and material maps condition a frozen video diffusion backbone for future prediction, while the reconstructed geometry and material maps initialize MPM-based deformable simulation. Experiments on surgical and general deformable datasets show that this state supports MPM rollout, and that depth and material conditioning improves video fidelity and temporal consistency. We hope this scene state helps world models support deformable simulation together with future video prediction.

AI Use Disclosure

AI tools were used to assist with language editing and improving the clarity and readability of the manuscript. All AI-assisted content was reviewed and revised by the authors.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

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*, volume 2026, pp. 118456–118470, 2026.

C. Baumgartner, P. A. Hasgall, F. Di Gennaro, E. Neufeld, B. Lloyd, M. C. Gosselin, D. Payne, A. Klingenböck, and N. Kuster. IT'IS Database for Thermal and Electromagnetic Parameters of Biological Tissues, 2025. Version 5.0.

Diego Biagini, Nassir Navab, and Azade Farshad. Hierasurg: Hierarchy-aware diffusion model for surgical video generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 310–319. Springer, 2025.

S Bodenstedt, S Speidel, M Wagner, J Chen, A Kisilenko, BP Müller, L Maier-Hein, B Oliveira, S Hong, J Zamora-Anaya, et al. Heichole surgical workflow analysis and full scene segmentation (heisurf). In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, pp. 1–12, 2021.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Tong Chen, Shuya Yang, Junyi Wang, Long Bai, Hongliang Ren, and Luping Zhou. Surgsora: Object-aware diffusion model for controllable surgical video generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 521–531. Springer, 2025.

Joseph Cho, Samuel Schmidgall, Cyril Zakka, Mrudang Mathur, Dhamanpreet Kaur, Rohan Shad, and William Hiesinger. Surgen: Text-guided diffusion model for surgical video generation. *arXiv preprint arXiv:2408.14028*, 2024.

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

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023.

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.

Alban De Vaucorbeil, Vinh Phu Nguyen, Sina Sinaie, and Jian Ying Wu. Material point method after 25 years: Theory, implementation, and applications. *Advances in applied mechanics*, 53:185–398, 2020.

Department of Engineering, University of Cambridge. *Materials Data Book*. University of Cambridge, 2011.

Chongkai Gao, Zixuan Liu, Zhenghao Chi, Junshan Huang, Xin Fei, Yiwen Hou, Yuxuan Zhang, Yudi Lin, Zhirui Fang, and Lin Shao. Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models. *Advances in Neural Information Processing Systems*, 38:136705–136736, 2026.

Xiaojie Gao, Yueming Jin, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Future frame prediction for robot-assisted surgery. In *International Conference on Information Processing in Medical Imaging*, pp. 533–544. Springer, 2021.

Jose L Gómez, Manuel Silva, Antonio Seoane, Agnes Borrás, Mario Noriega, Germán Ros, Jose A Iglesias-Guitian, and Antonio M López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *Neurocomputing*, 637:130038, 2025.

Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenolé Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. Cadis: Cataract dataset for surgical rgb-image segmentation. *Medical Image Analysis*, 71:102053, 2021.

Yufan He, Pengfei Guo, Mengya Xu, Zhaoshuo Li, Andriy Myronenko, Dillan Imans, Bingjie Liu, Dongren Yang, Mingxue Gu, Yongnan Ji, et al. Cosmos-h-surgical: Learning surgical robot policies from videos via world modeling. *arXiv preprint arXiv:2512.23162*, 2025.

Eric Heiden, Miles Macklin, Yashraj Narang, Dieter Fox, Animesh Garg, and Fabio Ramos. DiSECt: A differentiable simulation engine for autonomous robotic cutting. In *Robotics: Science and Systems (RSS)*, 2021.

W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

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*, 37(4):1–14, 2018.

Yiming Huang, Beilei Cui, Long Bai, Ziqi Guo, Mengya Xu, Mobarakol Islam, and Hongliang Ren. Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 197–207. Springer, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

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

Saurabh Koju, Saurav Bastola, Prashant Shrestha, Sanskar Amgain, Yash Raj Shrestha, Rudra PK Poudel, and Binod Bhattarai. Surgical vision world model. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pp. 1–10. Springer, 2025.

Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *International conference on medical image computing and computer-assisted intervention*, pp. 230–240. Springer, 2024.

Wei Li, Ming Hu, Guoan Wang, Lihao Liu, Kaijing Zhou, Junzhi Ning, Xin Guo, Zongyuan Ge, Lixu Gu, and Junjun He. Ophora: a large-scale data-driven text-guided ophthalmic surgical video generation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 425–435. Springer, 2025.

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

Changjing Liu, Yiming Huang, Long Bai, Beilei Cui, and Hongliang Ren. Endogsim: Physics-aware 4d dynamic endoscopic scene simulations via mllm-guided gaussian splatting. *arXiv preprint arXiv:2605.16022*, 2026.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025a.

Zhuoman Liu, Weicai Ye, Yan Luximon, Pengfei Wan, and Di Zhang. Unleashing the potential of multi-modal foundation models and video diffusion for 4d dynamic physical scene simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11016–11025, 2025b.

Chunji Lv, Zequn Chen, Donglin Di, Weinan Zhang, Hao Li, Chen Wei, Yinjie Lei, and Changsheng Li. Physgm: Large physical gaussian model for feed-forward 4d synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29855–29865, 2026.

MatWeb, LLC. Matweb material property data. https://www.matweb.com/, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

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

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2(3):4, 2025.

The Engineering ToolBox. Material properties. https://www.engineeringtoolbox.com/, 2024.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

Boyuan Wang, Xiaofeng Wang, Yongkang Li, Zheng Zhu, Yifan Chang, Angen Ye, Guosheng Zhao, Chaojun Ni, Guan Huang, Yijie Ren, et al. Reconphys: Reconstruct appearance and physical attributes from single video. *arXiv preprint arXiv:2604.07882*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International conference on medical image computing and computer-assisted intervention*, pp. 431–441. Springer, 2022.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations (ICLR)*, 2023.

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 Computer Vision and Pattern Recognition (CVPR)*, pp. 4389–4398, 2024.

Kaiyuan Xu, Fangzhou Hong, Daniel Elson, and Baoru Huang. Surgcut3r: Surgical scene-aware continuous understanding of temporal 3d representation. *arXiv preprint arXiv:2603.06971*, 2026.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Yousef Yeganeh, Rachmadio Lazuardi, Amir Shamseddin, Emine Dari, Yash Thirani, Nassir Navab, and Azade Farshad. Visage: Video synthesis using action graphs for surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 146–156. Springer, 2024.

Qinxi Yu, Masoud Moghani, Karthik Dharmarajan, Vincent Schorp, William Chung-Ho Panitch, Jingzhou Liu, Kush Hari, Huang Huang, Mayank Mittal, Ken Goldberg, and Animesh Garg. ORBIT-Surgical: An open-simulation framework for learning surgical augmented dexterity. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Albert J Zhai, Yuan Shen, Emily Y Chen, Gloria X Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang. Physical property understanding from language-embedded feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28296–28305, 2024.

Supplementary Contents

A Additional Results **15**

A.1 Material Map Visualizations 15

A.2 Zero-Shot Material Prediction on Unseen Deformable Scenes 15

A.3 Ablation Studies on Geometry and Material Conditioned Generation 16

B Evaluation Metrics **17**

B.1 Material Map Metrics. 17

B.2 Video Generation Metrics 18

B.3 Simulation Metrics 18

C Dataset Details **19**

C.1 Material Property Datasets 19

C.2 Video Generation Datasets 19

D Implementation Details **19**

D.1 Feed-Forward Material Prediction 19

D.2 Geometry and Material Conditioned Generation 20

D.3 Runtime Analysis 20

D.4 Simulation Setup 20

E Additional Method Details **21**

E.1 Geometry, Appearance, and Material Map Details 21

E.2 Architectural Details of the Control Branch 22

E.3 Material-Map Loss Decomposition 22

E.4 Gaussian-to-Particle Mapping and MPM Simulation 22

F Limitations **23**

A Additional Results

A.1 Material Map Visualizations

Figure 8 presents material maps on general-domain scenes from Virtual KITTI, VideoPhy2, and VLA-OS. Across driving, object-centric, and robotic manipulation scenes, the predictions remain aligned with object and scene boundaries, distinguishing foreground objects from their surrounding environments despite substantial appearance and domain changes. Figure 9 provides complementary results on surgical scenes from CholecSeg8k, EndoVis2018, and CaDISv2.

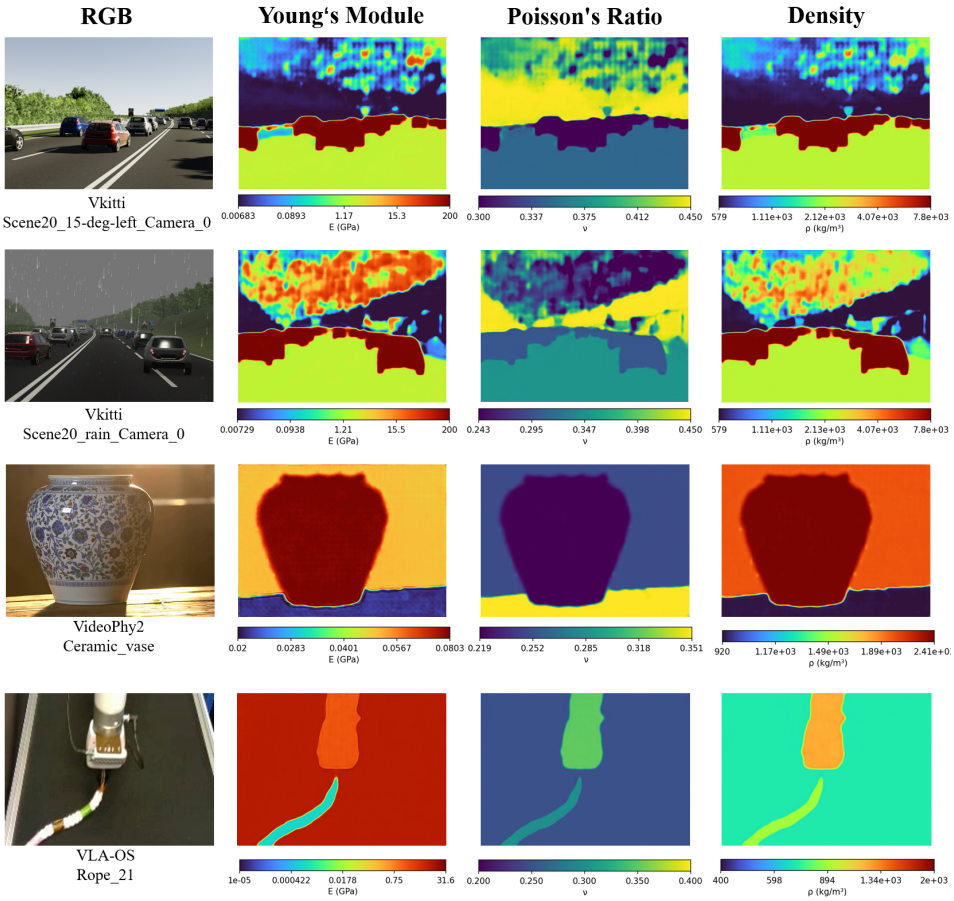


Figure 8: Material Maps on General Scenes. Predicted Young's modulus E, Poisson's ratio ν, and density ρ on representative Virtual KITTI, VideoPhy2, and VLA-OS examples.

A.2 Zero-Shot Material Prediction on Unseen Deformable Scenes

We evaluate the frozen PhysVerse reconstructor on previously unseen deformable scenes from the EndoNeRF (Wang et al., 2022) dataset without test-time optimization. The four scenes cover tool–tissue pulling and cutting, together with gallbladder and stomach views, so the test changes both the interaction and the anatomy. Table 5 and Figure 10 summarize this transfer.

The predicted material maps stay spatially aligned with the reconstructed instruments and soft tissue, and the two regions keep distinct (E, ν, ρ). Poisson's ratio is the most stable channel: ADE_ν lies in 0.007–0.028 and is lower than VGGT on every scene. Density transfers most cleanly under tool–tissue contact, with ARE_ρ of 0.002 on Pulling and 0.007 on Cutting, and rises to 0.047 and 0.044 on Gallbladder and Stomach. Young's modulus is more sensitive to the held-out contact: ALDE_E is 0.861 and 0.822 on Pulling and Cutting, compared with 0.291 and 0.568 on Gallbladder and Stomach. Ours remains lower than VGGT on ALDE_E for Pulling, Gallbladder, and Stomach,

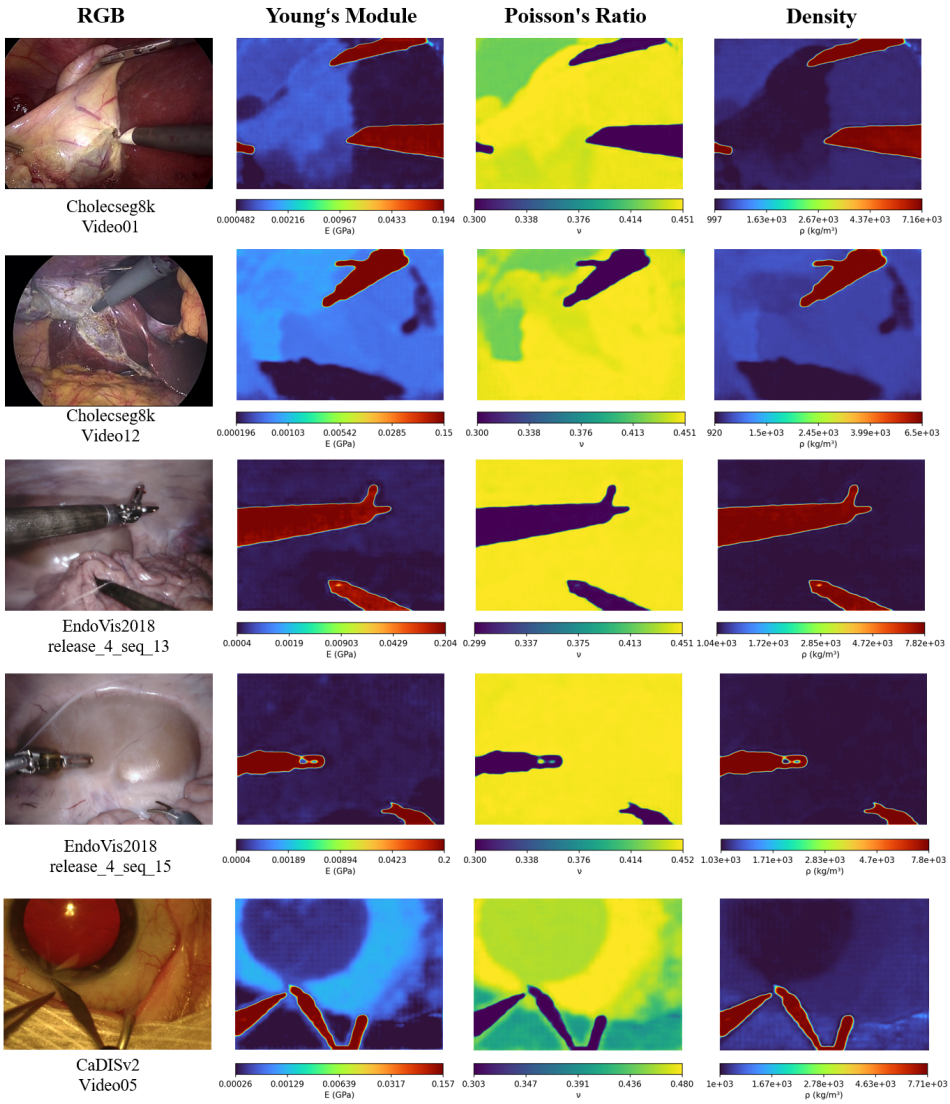


Figure 9: Material Maps on Surgical Scenes. Predicted Young's modulus, Poisson's ratio, and density on representative surgical scenes from multiple datasets.

while Cutting is the exception. The feed-forward map therefore keeps the instrument–tissue material layout without per-scene adaptation, with ν stable across scenes and E the main source of variation under large contact.

A.3 Ablation Studies on Geometry and Material Conditioned Generation

Figure 11 qualitatively compares different generation conditions on three test sequences. Depth-only and material-map-only conditioning exhibit appearance or geometric inconsistencies (highlighted in red), whereas joint conditioning better preserves instrument shape, tissue appearance, and temporal consistency.

Table 5: Zero-Shot Material Prediction. Lower is better for all metrics. Best is highlighted.

Scene	Method	ALDE$_E$ ↓	ADE$_\nu$ ↓	ARE$_\rho$ ↓
Pulling	VGGT	0.874	0.034	0.065
	Ours	0.861	0.028	0.002
Cutting	VGGT	0.792	0.031	0.061
	Ours	0.822	0.026	0.007
Gallbladder	VGGT	0.301	0.015	0.016
	Ours	0.291	0.007	0.047
Stomach	VGGT	0.648	0.021	0.022
	Ours	0.568	0.012	0.044

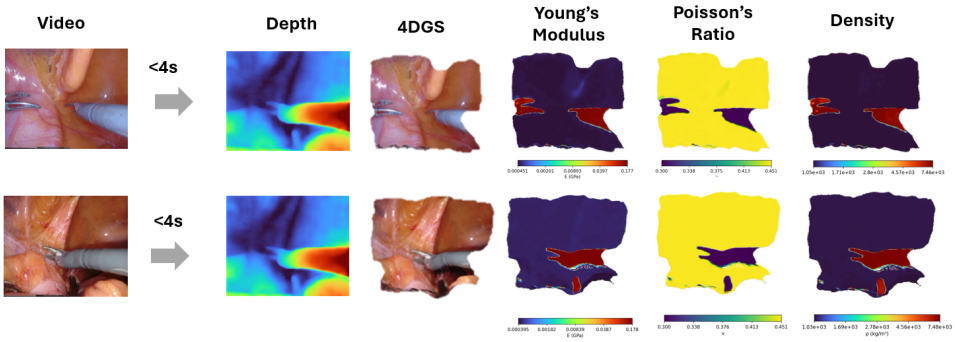


Figure 10: Zero-Shot Material Maps. Predicted Young's modulus E, Poisson's ratio ν, and density ρ on EndoNeRF (Wang et al., 2022).

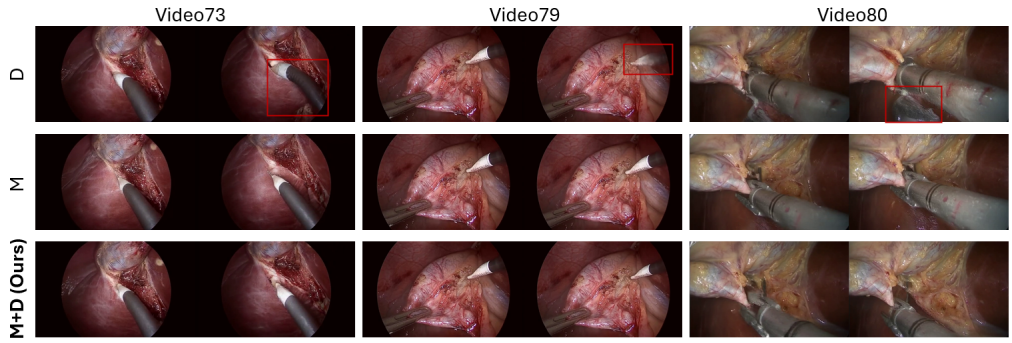


Figure 11: Qualitative Ablation of Generation Conditions. We compare depth-only (D), material-map-only (M), and joint depth–material conditioning (D+M, ours) on three representative test videos.

B Evaluation Metrics

B.1 Material Map Metrics.

To evaluate the predicted material maps, we use standard error metrics for scalar fields. Let y denote a material target at a valid pixel (e.g., Young's modulus, Poisson's ratio, or density), and let $\hat{y}$ denote its prediction.

Absolute Difference Error (ADE). ADE is the mean absolute error in the original linear space:

$$\mathrm{ADE}(y, \hat{y}) = \frac{1}{N} \sum_{p=1}^{N} |\hat{y}_p - y_p| \,. \tag{B.1}$$

It is appropriate for bounded quantities that do not span multiple orders of magnitude.

Average Log Difference Error (ALDE). ALDE measures the mean absolute error in base-10 logarithmic space:

$$\mathrm{ALDE}(y, \hat{y}) = \frac{1}{N} \sum_{p=1}^{N} \left| \log_{10}(\hat{y}_p + \varepsilon) - \log_{10}(y_p + \varepsilon) \right|. \tag{B.2}$$

It captures multiplicative error and is useful for positive quantities with a large dynamic range.

Average Relative Error (ARE). ARE measures the mean relative deviation:

$$\mathrm{ARE}(y, \hat{y}) = \frac{1}{N} \sum_{p=1}^{N} \frac{|\hat{y}_p - y_p|}{y_p + \varepsilon}. \tag{B.3}$$

This dimensionless metric compares errors proportionally across regions with different target scales.

B.2 Video Generation Metrics

We evaluate paired video fidelity using Fréchet Video Distance (FVD), Fréchet Inception Distance (FID), Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) against the corresponding test videos. We further report the following VBench video-quality dimensions (Huang et al., 2023). Each VBench score is scaled to 0–100, and a higher score is better.

Dynamic Degree (DD). DD measures whether a generated video contains appreciable motion, rather than obtaining high temporal-consistency scores by remaining nearly static. VBench estimates motion using RAFT optical flow; a higher score indicates stronger video dynamics.

Imaging Quality (IQ). IQ assesses frame-level visual distortions, including blur, noise, and over- or under-exposure. Following VBench, it is estimated by an image-quality predictor, so it reflects perceptual image quality rather than agreement with the paired reference video.

Subject Consistency. This metric assesses whether the appearance of the main surgical subject, including anatomy and instruments, remains stable over the generated sequence. VBench computes cross-frame similarity of DINO features.

Background Consistency. Background consistency measures whether the surrounding scene remains stable across frames. It is computed from cross-frame CLIP feature similarity.

Motion Smoothness. Motion smoothness evaluates whether the generated motion evolves smoothly. It uses the motion prior of a video-frame interpolation model, and higher scores indicate smoother temporal motion.

Temporal Flickering. This metric targets high-frequency temporal changes in details that should be stable. It is based on frame differences in static content; the reported VBench score is higher when flickering is lower.

B.3 Simulation Metrics

We evaluate simulation accuracy against the reference data using nRMSE for 3D particle deformation and Flow Endpoint Error (EPE) for rendered motion.

Let $x_{T,i}^{\mathrm{pred}}, x_{T,i}^{\mathrm{ref}} \in \mathbb{R}^3$ be the predicted and reference particle positions at the final frame T, and let $u_{T,i}^{\mathrm{ref}} = x_{T,i}^{\mathrm{ref}} - x_{0,i}^{\mathrm{ref}}$ denote the displacement from the initial state. All reported numbers use the final frame.

nRMSE. Terminal position RMSE, normalized by the mean reference displacement:

$$\mathrm{nRMSE} = \frac{\sqrt{\frac{1}{N}\sum_i \left\|x_{T,i}^{\mathrm{pred}} - x_{T,i}^{\mathrm{ref}}\right\|^2}}{\frac{1}{N}\sum_i \|u_{T,i}^{\mathrm{ref}}\| + \varepsilon}. \tag{B.4}$$

Flow Endpoint Error (EPE). We evaluate rendered motion consistency using optical flow. For each consecutive frame pair $(t-1, t)$, we estimate dense optical flow from the predicted and reference renderings using RAFT-Large, denoted as $\mathbf{U}_t^{\mathrm{pred}}$ and $\mathbf{U}_t^{\mathrm{ref}}$, respectively. EPE is computed as the mean endpoint error over all pixels and frame pairs:

$$\mathrm{EPE} = \frac{1}{T|\Omega|}\sum_{t=1}^{T}\sum_{\mathbf{p}\in\Omega}\left\|\mathbf{U}_t^{\mathrm{pred}}(\mathbf{p}) - \mathbf{U}_t^{\mathrm{ref}}(\mathbf{p})\right\|_2. \tag{B.5}$$

where $t = 1, \ldots, T$ indexes consecutive frame pairs $(t-1, t)$ and Ω is the rendered image domain.

C Dataset Details

C.1 Material Property Datasets

We construct the material-map training set from eight medical and general-domain datasets, as summarized in Table 6. For every dataset, we split source videos or source cases into 80% training and 20% test sets, ensuring that frames from the same sequence never appear in both splits. Each frame is associated with a corresponding material map (E, ν, ρ), resulting in 662 video clips and 58,878 frames. For every dataset, training and test samples are separated at the source-video or source-case level, such that frames originating from the same sequence never appear in both splits.

Table 6: Material Property Datasets.

Dataset	Type	Total clips	Total frames
CholecSeg8k (Hong et al., 2020)	Medical	40	1,076
EndoVis2018 (Allan et al., 2020)	Medical	15	750
CaDISv2 (Grammatikopoulou et al., 2021)	Medical	22	1,369
HeiSurF (Bodenstedt et al., 2021)	Medical	24	466
VKitti (Cabon et al., 2020)	General	100	42,520
UrbanSyn (Gómez et al., 2025)	General	118	7,539
VideoPhy2 (Bansal et al., 2026)	General	43	1,558
VLA-OS deformable (Gao et al., 2026)	General	300	3,600
Total		662	58,878

C.2 Video Generation Datasets

For video generation, we use the surgical Cholec80-5k dataset and a deformable-interaction subset of VLA-OS. Each clip contains 49 consecutive frames and is paired with aligned text, depth, and material maps. Cholec80-5k is split by source video to prevent frames from the same procedure from appearing in both training and test sets. Table 7 summarizes both datasets.

D Implementation Details

D.1 Feed-Forward Material Prediction

We freeze the HunyuanWorld-Mirror backbone and optimize only the Physics DPT head on the mixed dataset summarized in Table 6. Training uses AdamW with $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, and a learning rate of 2×10^{-4}. We train for 10,000 steps in bf16 precision using a warmup-cosine learning-rate schedule with $\eta_{\min}/\eta_{\max} = 0.1$. All runs use a random seed of 42 and are evaluated on the test set every 1,000 steps on a single NVIDIA A6000 GPU. The data loader samples at least two views per example and processes up to eight images per GPU, with aspect ratios in $[0.5, 2.0]$ and image sizes ranging from 100,000 to 250,000 pixels.

Table 7: Video Generation Datasets.

Dataset	Split	Source units	Clips	Frames / clip	Total frames
Cholec80-5k (Twinanda et al., 2016)	Train	72 videos	4,504	49	220,696
	Test	8 videos	496	49	24,304
	Total	80 videos	5,000	49	245,000
VLA-OS (Gao et al., 2026)	Train	270 cases	900	49	44,100
	Test	30 cases	100	49	4,900
	Total	300 cases	1,000	49	49,000
All datasets	**Total**	–	6,000	49	294,000

Table 8: Runtime Breakdown of Reconstruction and Generation. Measurements use one NVIDIA A6000 GPU, bf16 precision, and batch size one. Reconstruction processes eight input frames at an input width of 518 pixels; generation produces 49 frames at 384×640 resolution with 50 Euler steps.

Stage	Component	Time(s)	Ratio
Reconstruction (8-frame clip)	Visual geometry backbone	1.847	53.1%
	Geometry Head (camera, depth)	0.984	28.3%
	Physics Head	0.191	5.5%
	Gaussian Head	0.456	13.1%
	Total	**3.478**	**100%**
Generation (49-frame rollout)	Text encoding (UMT5-XXL)	0.060	0.2%
	RGB VAE encoding	0.017	0.1%
	Control encoding (depth + material)	0.083	0.3%
	Denoising (DiT + control branch, 50 steps)	29.513	92.5%
	Tiled VAE decoding	2.213	6.9%
	Other	0.010	∼0%
	Total	**31.897**	**100%**

D.2 Geometry and Material Conditioned Generation

We freeze the pretrained Wan2.2-TI2V-5B backbone and optimize only the control modules for 5,000 steps on an NVIDIA A6000 GPU using AdamW with a learning rate of 1×10^{-5}. Training and inference use 49-frame clips at a resolution of 384×640. The initial RGB frame follows the native image-to-video pathway, whereas the corresponding depth and material maps are encoded by the frozen Wan VAE; the umT5 text encoder also remains frozen. The control branch contains $N_c = 15$ control stages, and inference uses 50 Euler sampling steps.

D.3 Runtime Analysis

We report the runtime breakdown of feed-forward reconstruction and geometry and material conditioned generation in Table 8.

D.4 Simulation Setup

Simulation Reference Data. For each scene, we derive the dense reference material map from the original semantic annotations and the database reference values (E, ν, ρ), and assign it to the reconstructed particles. Running the MPM solver with the prescribed interaction produces the reference particle trajectories and rendered motion.

Material assignment. Each scene is reconstructed once (HWM + 4DGS). We reuse this geometry for four MPM simulations with fixed particles, cameras, impulse, and boundaries, varying only the

material assignment. For the simulation evaluation, Ours (Global) assigns the spatial median of each predicted channel to all particles, $\bar{\mathbf{M}} = (\mathrm{median}_p\, \hat{M}_p^E, \mathrm{median}_p\, \hat{M}_p^\nu, \mathrm{median}_p\, \hat{M}_p^\rho)$. Ours (Dense) uses the predicted material map directly. Grok4.6 (Global) and ChatGPT-5.6 (Global) each infer a global material triplet from the full RGB frame with the following prompt, without receiving a mask, class list, or material table:

```
Input:  one RGB image.  No mask, class list, or material
table is provided.
Estimate one representative material triplet for the scene,
to be assigned uniformly to all simulation particles.
Return only a JSON object containing three scalar values,
with no markdown or extra text:
{"E": ..., "nu": ..., "rho": ...}.
E is Young's modulus in Pa, with 1 <= E <= 1e12; nu is
Poisson's ratio, with 0 <= nu <= 0.5; rho is density in
kg/m^3, with 1 <= rho <= 20000.
```

The predicted triplet is assigned uniformly to all simulation particles and broadcast over the image for material-map evaluation. All methods use the same valid-pixel mask and metric definitions.

MPM setup. We use the jelly constitutive model in MLS-MPM with a 25^3 grid. The simulation time step and frame interval are 2×10^{-4}s and 5×10^{-3}s, respectively. Each rollout contains 50 frames (≈ 0.25s). The scaling factors for Young's modulus, yield stress, and viscosity are 10^7, 10^5, and 10^4, respectively. These fixed factors put database-scale Young's modulus, together with the jelly model's yield stress and viscosity, into the non-metric units of the 25^3 grid, time step, and impulse. The reference rollout and every compared assignment use the same factors, so the simulation metrics compare material layout under this shared calibration.

Impulse. Pick the particle along the camera ray through the image center; apply a leftward screen drag of 0.2 (normalized width) with UI force 150, mapped to $\|\mathbf{F}\| = 150 \times 0.2 \times 25 = 750$ MPM units. Impulse radius 0.08; force direction follows the 3D ray difference at the pick depth.

E Additional Method Details

E.1 Geometry, Appearance, and Material Map Details

This section details the geometry-and-appearance component of the reconstructor $\mathcal{F}_{\phi^*,\theta}$. Given an input image sequence, the frozen backbone $\mathcal{R}_{\phi^*}$ produces per-frame output tokens $\hat{\mathbf{T}}_i \in \mathbb{R}^{L\times D}$. Its frozen geometry and Gaussian heads decode depth, camera, and dynamic Gaussian primitives, as summarized in Eq. equation 1. The Gaussian DPT head separately predicts a pixel-wise Gaussian depth map and a Gaussian feature map:

$$\hat{\mathbf{D}}_i^g,\, \mathbf{F}_i^g = \mathrm{DPT}_g\Big(\hat{\mathbf{T}}_i\Big).$$

The Gaussian centers are obtained by back-projecting the predicted Gaussian depth with the corresponding camera parameters,

$$\boldsymbol{\mu}_i^g = \Pi^{-1}\Big(\hat{\mathbf{D}}_i^g, \hat{\mathbf{C}}_i\Big),$$

while the remaining appearance attributes are decoded from Gaussian features and the input image:

$$\hat{\mathcal{A}}_i^g = \mathrm{Conv}(\mathbf{F}_i^g, \mathbf{I}_i) = \Big\{\hat{\alpha}_i^g,\, \hat{\mathbf{r}}_i^g,\, \hat{\mathbf{s}}_i^g,\, \Delta\hat{\mathbf{sh}}_i^g,\, \hat{w}_i^g\Big\}.$$

Here, $\hat{\alpha}_i^g$, $\hat{\mathbf{r}}_i^g$, $\hat{\mathbf{s}}_i^g$, $\Delta\hat{\mathbf{sh}}_i^g$, and $\hat{w}_i^g$ denote opacity, orientation, scale, residual spherical-harmonic color coefficients, and fusion weight, respectively. In the dynamic reconstruction setting, the Gaussian head additionally predicts a lifespan parameter $\hat{\tau}_i^g$, yielding the 4D Gaussian representation

$$\hat{\mathcal{G}}_{0:N-1}^{\mathrm{4D}} = \Big\{\boldsymbol{\mu}_i^g,\, \hat{\mathcal{A}}_i^g,\, \hat{\tau}_i^g\Big\}_{i=0}^{N-1}.$$

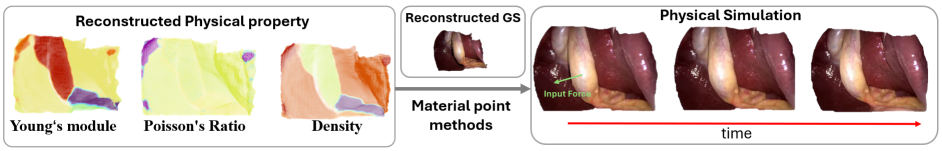


Figure 12: Simulation Pipeline. The reconstructed splatting and predicted material maps are used as input to the MPM simulation, which outputs the simulated particle positions and velocities.

E.2 Architectural Details of the Control Branch

The reconstructed first-frame depth and material maps are normalized and encoded by the frozen Wan VAE into $\mathbf{z}_d$ and $\mathbf{z}_m$, respectively. Each modality has a dedicated input projection, while both modality streams are processed independently using a shared stack of $N_c = 15$ control blocks. For $q \in \{d, m\}$, the first control block receives the noisy video token $\mathbf{x}$ and computes

$$\mathbf{u}_q^{(0)} = \mathrm{Block}_0(W_q \mathbf{z}_q + \mathbf{x}), \tag{E.1}$$

and the subsequent control blocks propagate the modality-specific state as

$$\mathbf{u}_q^{(i)} = \mathrm{Block}_i\left(\mathbf{u}_q^{(i-1)}\right), \qquad \mathbf{s}_q^{(i)} = W_i^{\mathrm{out}} \mathbf{u}_q^{(i)}. \tag{E.2}$$

Here, W_i^{out} denotes the zero-initialized residual projection at stage i. The two residual hints are fused and injected into the frozen Wan DiT exactly as specified by Eqs. 4 and 5. This design keeps the depth and material representations separate before their stage-wise gated fusion.

E.3 Material-Map Loss Decomposition

We provide the full definitions of the three sub-losses comprising $\mathcal{L}_{\mathrm{mat}}$ in Eq. equation 6. The Physics DPT head additionally predicts a training-only confidence map $\mathbf{c}_i$. For frame i and pixel u, the squared error between the predicted and normalized material vectors is

$$e_{i,u} = \left\|\hat{\mathbf{M}}_{i,u} - \mathbf{M}_{i,u}\right\|_2^2, \tag{E.3}$$

and the three sub-losses are then defined as

$$\mathcal{L}_{\mathrm{reg}} = \mathbb{E}_{i,u}\left[e_{i,u}\right], \tag{E.4a}$$

$$\mathcal{L}_{\mathrm{conf}} = \mathbb{E}_{i,u}\left[\gamma\, e_{i,u}\, c_{i,u} - \alpha \log c_{i,u}\right], \tag{E.4b}$$

$$\mathcal{L}_{\mathrm{grad}} = \sum_{s=0}^{3} \mathcal{L}_{\mathrm{grad}}^{(2^s)}. \tag{E.4c}$$

where $c_{i,u} \in \mathbb{R}_{>0}$ is the positive confidence at pixel u of frame i, $\mathcal{L}_{\mathrm{grad}}^{(2^s)}$ is the L_1 gradient consistency term on $\hat{\mathbf{M}}_i - \mathbf{M}_i$ at scale 2^s, and the hyper-parameters are fixed at $\gamma = 1.0$ and $\alpha = 0.2$.

E.4 Gaussian-to-Particle Mapping and MPM Simulation

Gaussian-to-particle assignment. Since the Gaussian primitives and material maps are decoded on the same image grid, each reconstructed Gaussian is directly associated with the material vector predicted at the corresponding pixel. We instantiate one MPM particle per Gaussian and assign

$$\mathbf{x}_p^0 = \boldsymbol{\mu}_p, \qquad (E_p, \nu_p, \rho_p) = \mathcal{T}^{-1}(\hat{\mathbf{M}}_p), \qquad m_p = \rho_p V_p, \tag{E.5}$$

where $\mathcal{T}^{-1}$ converts normalized predictions to physical units. Young's modulus and Poisson's ratio are further converted to the Lamé parameters used by the constitutive model.

MPM Formulation. The Material Point Method (MPM) (Stomakhin et al., 2013; De Vaucorbeil et al., 2020) is a particle-grid numerical framework for simulating continuum dynamics. Its particle-based representation also makes it well suited for coupling with 3D Gaussian scenes, as explored in

PhysGaussian (Xie et al., 2024). MPM alternates between three main stages: particle-to-grid (P2G) transfer, grid-based dynamics update, and grid-to-particle (G2P) transfer. During the P2G stage, particle mass and momentum are accumulated onto neighboring grid nodes as follows:

$$m_i^n = \sum_p w_{ip}^n m_p, \quad m_i^n \mathbf{v}_i^n = \sum_p w_{ip}^n m_p \big(\mathbf{v}_p^n + \mathbf{C}_p^n(\mathbf{x}_i - \mathbf{x}_p^n)\big), \tag{E.6}$$

Here, p and i index the Lagrangian particles and Eulerian grid nodes, respectively. At time t_n, each particle p is characterized by its volume V_p, mass m_p, position $\mathbf{x}_p^n$, velocity $\mathbf{v}_p^n$, deformation gradient $\mathbf{F}_p^n$, and affine momentum $\mathbf{C}_p^n$. The interpolation weight w_{ip}^n is given by the B-spline kernel between particle p and grid node i, evaluated according to the particle position $\mathbf{x}_p^n$.

After the P2G transfer, the velocities on the grid are updated by integrating both internal and external forces:

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^n - \frac{\Delta t}{m_i} \sum_p \tau_p^n \nabla w_{ip}^n V_p^0 + \Delta t \frac{\mathbf{f}^{ext}}{m_i}, \tag{E.7}$$

where τ_p^n is the Cauchy stress tensor, ∇w_{ip}^n is the gradient of the kernel function. In G2P, the velocities are transferred back to particles to update their states:

$$\mathbf{v}_p^{n+1} = \sum_i w_{ip} \mathbf{v}_i^{n+1}. \tag{E.8}$$

The deformation gradient $\mathbf{F}_p$ is updated at each step to capture local material deformation, with plastic corrections applied when necessary. We use MLS-MPM (Hu et al., 2018) to simulate the deformable dynamics from the reconstructed geometry and predicted material maps. As illustrated in Figure 12, the reconstructed 3DGS geometry and predicted material maps are assigned to MPM particles, whose positions and velocities are then evolved through the simulation.

F Limitations

Our proposed method is still limited in what the category-level supervision can represent. Each semantic category is assigned one database triplet (E, ν, ρ), so the reference maps do not contain fine fluctuations inside one material, and the predicted maps have limited ability to represent intra-material variation of (E, ν, ρ). Materials absent from the training categories are also uncovered, so generalization to unknown materials remains limited. These limits concern the training targets. At test time, the dense head regresses the spatial material map in one forward pass and does not predict a segmentation or query a stored table. Table 2 shows the value of that spatial map: Ours (Dense) reduces nRMSE by 78.3% and EPE by 27.4% relative to a single global assignment. This comparison does not remove the limits on intra-material variation and unknown materials. This work also presents two challenges for accumulated error in two settings. The first setting is multi-frame generation, where the generation model predicts a sequence of future frames and error can build up along that sequence. The second is the multi-round update, where after each round, the generated frames are sent back through the reconstructor and then generated again. Therefore, we only present a qualitative comparison of future prediction after a single feedback pass in Figure 7.

Reproducibility Statement

Our model architecture and training objectives are described in the main text, with additional methodological details in Appendix E. Dataset construction and splits, evaluation metrics, and training, inference, and simulation settings are provided in Appendices C.1, C.2, B, and D to facilitate reproducibility.