LP^{2}SLAT: LEARNING PHYSICAL PRIORS ON STRUCTURED LATENTS FOR SINGLE-IMAGE GENERATED 3D ASSETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Turning a single-image generated asset into a deformable object requires physical properties that vary across its parts. We introduce LP^{2}SLAT, a feed-forward network that predicts material type, density, Young's modulus, and Poisson ratio from the structured latent (SLat) of a 3D generator and the source image. The SLat branch combines local voxel features with object context to predict a base physical field. The image branch supplies appearance features at projected voxel locations, and a learned gate weights their residual corrections. To supervise generated geometry using reference asset annotations, we establish correspondence in the shared source view and filter assignments at part boundaries or with mixed label support. On the PhysX–Pixie split, LP^{2}SLAT raises material macro-F1 from 0.216 to 0.585 over a SLAT-Phys-style baseline and reduces log-density and log-modulus MAE by 36.5% and 32.5%. Source-view supervision raises F1 by 0.237 over ICP-based transfer. We also construct PhysSLat-227, a collection of 227 textured assets with part-level physical priors, and evaluate predicted fields in MPM simulations.

1 INTRODUCTION

An image-to-3D model can generate the textured leaves, stems, and container of a potted plant. Simulating that plant under a load requires a further distinction: the leaves should bend, the stems should support them, and the container should remain comparatively stiff. These responses depend on physical properties that vary within the object. Models such as TRELLIS (Xiang et al., 2025) supply geometry and appearance from one image; spatial material and mechanical fields connect these assets to downstream simulation and to physics-grounded generation for embodied environments and interactive content (Xie et al., 2023; Yang et al., 2024; 2026; Cao et al., 2026a).

Existing predictors use different observations to construct these fields. NeRF2Physics estimates properties on a feature field reconstructed from multiple views (Zhai et al., 2024); PIXIE learns a supervised mapping from distilled 3D visual features (Le et al., 2025). SLAT-Phys reads physical properties directly from the structured latent (SLat) generated by TRELLIS from one image (Das & Manocha, 2026). In this single-image setting, the generator's SLat and the original image offer two routes to prediction: SLat places shape and appearance features on the generated surface, and the source image retains local texture and material cues at the observed view. We use the former to define the physical field and the latter to refine it.

We introduce LP^{2}SLAT, a two-branch network for material type, density, Young's modulus, and Poisson ratio prediction. Its SLat branch combines local voxel features with pooled object context to predict a base field at every active voxel. The source-image branch samples appearance features at projected voxel locations. A fusion module uses both local features to predict a residual correction and a gate that controls its contribution. We train the predictor jointly with the generator and image encoder frozen, then attach the field to the mesh decoded from the same SLat.

Training requires labels on generated voxels, whereas available annotations belong to reference assets. Differences in shape and scale complicate direct 3D correspondence. We use the shared source view to connect the two: silhouette fitting aligns the generated asset with the image, and

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.



(a) Representative construction assets **(b) Construction metrics**

Figure 1: **PhysSLat-227 dataset overview.** Representative texture-preserving assets (left), category coverage, and source composition of 227 curated objects (right).

depth-aware projection associates its visible voxels with reference parts. Boundary erosion and label-purity filtering remove ambiguous assignments before these correspondences become training targets.

We evaluate the learned predictors on an object-disjoint PhysX–Pixie split. Compared with a SLAT-Phys-style baseline, LP^{2}SLAT raises held-out material macro-F1 from 0.216 to 0.585 and reduces log-density and log-modulus MAE by 36.5% and 32.5%, respectively. We further compare MPM deformation responses driven by predicted physical fields and reference annotations under the same loads; trajectory RMSE decreases by at least 25% on 20 of 25 tested objects. We also construct PHYSSLAT-227, with 227 textured objects across 18 categories and part-level physical priors.

Our contributions are:

- **Gated latent–image physics prediction.** We introduce LP^{2}SLAT, a network that predicts material type and mechanical properties for single-image generated 3D assets. It learns a base physical field from structured 3D latents and uses gated source-image residuals to refine it with local appearance.
- **Cross-asset supervision through source-view correspondence.** We transfer part-level physical annotations from reference assets to generated 3D latents through shared-view correspondence, accommodating geometric differences without direct 3D registration. Boundary erosion and label-purity filtering exclude ambiguous assignments.
- **PhysSLat-227 dataset.** We construct a dataset of 227 textured objects spanning 18 categories with part-level material, density, Young's-modulus, and Poisson-ratio priors.

2 Related Work

Spatial physical-property estimation. Spatial property prediction requires associating visual evidence with 3D locations. NeRF2Physics combines language-proposed material properties with CLIP-based kernel regression on a reconstructed point cloud (Zhai et al., 2024), whereas PIXIE learns a supervised mapping from distilled 3D visual features to material classes and mechanical parameters (Le et al., 2025). Both construct their spatial features from posed multi-view images. SLAT-Phys instead predicts material fields directly from TRELLIS structured latents generated from a single image (Xiang et al., 2025; Das & Manocha, 2026), making it the closest approach to ours. LP^{2}SLAT combines a local–global SLat base predictor with an image-conditioned residual decoder. The former defines predictions over the generated surface; the latter refines them using projected source appearance and a learned gate. Our training correspondence is established in the shared source view, linking reference-part annotations to generated voxels across geometric differences.

Physics-aware 3D asset generation. Another line of work predicts physical attributes while generating geometry and object structure. PhysGM jointly predicts a Gaussian representation and phys-

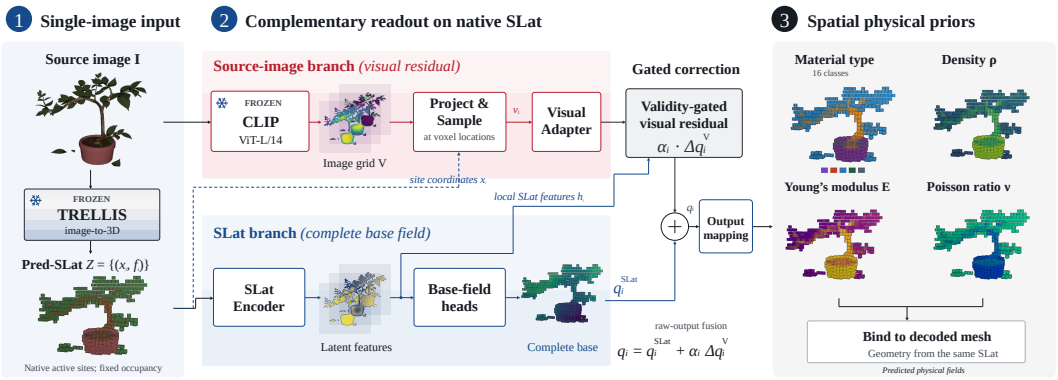

Figure 2: **Architecture of LP²SLAT.** Frozen TRELLIS and CLIP encode the generated asset and source image. The SLat branch predicts a base physical field; projected image features and local SLat features jointly determine gated residual corrections. The resulting material, density, modulus, and Poisson-ratio fields are attached to the decoded mesh.

ical parameters for downstream simulation, supporting both single- and multi-view inputs (Lv et al., 2025). PhysX-Anything and PhysX-Omni use vision-language models to generate structured geometry and physical attributes from a single image (Cao et al., 2026a;b); PhysX-Omni covers rigid, deformable, and articulated objects. PhysForge conditions geometry and kinematic generation on a hierarchical physical blueprint (Yang et al., 2026). Our setting keeps the geometry generator fixed and learns spatial physical properties on its native latent support.

Datasets for physical 3D learning. Training these predictors requires labels at an appropriate spatial scale. PIXIEVERSE pairs assets with spatial material annotations (Le et al., 2025), while PhysAssets also supplies reference simulation videos for learning geometry and physical behavior (Lv et al., 2025). PhysX-Mobility and PhysXVerse associate object parts with physical, functional, and kinematic attributes (Cao et al., 2026a;b); PhysDB organizes material, functional, and interaction annotations in a hierarchical scheme (Yang et al., 2026). Object-level physical parameters, as used in PhysAssets (Lv et al., 2025), do not resolve mechanical variation within an object. PHYSSLAT-227 emphasizes such variation in deformable and mixed-material assets, linking accepted face-level part assignments to part-specific density, Young's modulus, and Poisson-ratio priors.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

LP²SLAT assigns spatially varying physical properties to an asset generated from one RGB image. It predicts material type, density, Young's modulus, and Poisson ratio at each active voxel by combining the generator's 3D latent features with source-image appearance. The SLat branch supplies the base field, and the image branch supplies features for its refinement.

We obtain these two representations from the background-removed source image $I \in \mathbb{R}^{H\times W\times 3}$, where H and W are its height and width. A frozen TRELLIS generator $\mathcal{G}$ (Xiang et al., 2025) produces a structured latent representation (SLat) $\mathcal{Z}$, and a frozen CLIP encoder extracts a 2D patch-feature grid $\mathcal{V}$:

$$\mathcal{Z} = \mathcal{G}(I) = \{(\mathbf{x}_i, \mathbf{f}_i)\}_{i=1}^{N}, \qquad \mathcal{V} = \mathrm{CLIP}_{\mathrm{patch}}(I). \tag{1}$$

We call this image-generated SLat *Pred-SLat*. It contains N active voxels representing the object's surface, with grid locations $\mathbf{x}_i \in \mathbb{Z}^3$ and local shape-and-appearance features $\mathbf{f}_i \in \mathbb{R}^8$. The image features $\mathcal{V} \in \mathbb{R}^{16\times 16\times 1024}$ follow this grid order.

To combine these representations, the image branch needs the image location associated with each 3D voxel. We fit a projection $\pi_I : \mathbb{R}^3 \to \mathbb{R}^2$ by aligning the decoded mesh's silhouette with the

source-image foreground (Sec. 3.3). The projected voxel centers then locate the image features sampled from $\mathcal{V}$ (Sec. 3.2).

The two-branch predictor F_θ, with trainable parameters θ, maps these aligned features to the physical prediction at voxel i:

$$F_\theta(\mathcal{Z}, \mathcal{V}; \pi_I)_i = (\mathbf{p}_i, \hat{\rho}_i, \hat{E}_i, \hat{\nu}_i). \tag{2}$$

Here $\mathbf{p}_i \in [0,1]^{16}$ gives material probabilities with $\sum_{k=1}^{16} p_{ik} = 1$, where k indexes material classes. The scalar outputs are density $\hat{\rho}_i$ in $\mathrm{kg/m^3}$, Young's modulus $\hat{E}_i$ in Pa, and dimensionless Poisson ratio $\hat{\nu}_i$. Together, these predictions form an $N \times 19$ *spatial physical field*. The field is attached to the mesh decoded from the same Pred-SLat $\mathcal{Z}$ (Figure 2; Sec. 3.4).

3.2 Gated Latent–Image Physical Prediction

The predictor first constructs an SLat base field, then refines it with a gated residual derived from local SLat and image features, so that every voxel receives a base prediction before image evidence is applied.

SLat branch: base-field prediction. The SLat branch combines each voxel's location and native feature with context pooled over the object. For grid resolution R, we normalize the voxel center as $\tilde{\mathbf{x}}_i = 2(\mathbf{x}_i + \frac{1}{2})/R - 1$ and compute $\mathbf{h}_i^{\mathrm{loc}} \in \mathbb{R}^{192}$:

$$\mathbf{h}_i^{\mathrm{loc}} = \mathrm{Enc}_{\mathrm{SLat}}(\tilde{\mathbf{x}}_i, \mathbf{f}_i). \tag{3}$$

Mean–max pooling summarizes the object in $\mathbf{g} \in \mathbb{R}^{192}$. Fusing this summary with each local feature produces $\mathbf{h}_i^{\mathrm{ctx}} \in \mathbb{R}^{192}$ for base prediction:

$$\begin{aligned} \mathbf{g} &= \mathrm{Enc}_{\mathrm{obj}}\left(\left[\frac{1}{N}\sum_{j=1}^{N}\mathbf{h}_j^{\mathrm{loc}}; \max_{j=1,\dots,N}\mathbf{h}_j^{\mathrm{loc}}\right]\right), \\ \mathbf{h}_i^{\mathrm{ctx}} &= \mathrm{Fuse}_{\mathrm{ctx}}([\mathbf{h}_i^{\mathrm{loc}}; \mathbf{g}]), \qquad \mathbf{q}_i^{\mathrm{SLat}} = \mathrm{Head}_{\mathrm{SLat}}(\mathbf{h}_i^{\mathrm{ctx}}). \end{aligned} \tag{4}$$

Here $[\cdot;\cdot]$ denotes concatenation and j indexes active voxels. The raw output $\mathbf{q}_i^{\mathrm{SLat}} \in \mathbb{R}^{21}$ contains 16 material logits, two channels each for density and modulus, and one for Poisson ratio.

Source-image branch: voxel-aligned appearance features. The image branch maps each voxel center $\mathbf{u}_i = (\mathbf{x}_i + \frac{1}{2})/R - \frac{1}{2}$ to the source image through π_I. It bilinearly samples the frozen CLIP grid and adapts the resulting feature $\mathbf{v}_i \in \mathbb{R}^{1024}$ to $\mathbf{h}_i^{\mathrm{img}} \in \mathbb{R}^{192}$:

$$\begin{aligned} \mathbf{v}_i &= \mathrm{Bilinear}(\mathcal{V}, \pi_I(\mathbf{u}_i)), \\ \chi_i^{\mathrm{proj}} &= \mathbf{1}[d_i > 0]\,\mathbf{1}[\pi_I(\mathbf{u}_i) \in \Omega_I], \qquad \mathbf{h}_i^{\mathrm{img}} = \chi_i^{\mathrm{proj}}\,\mathrm{Adapt}_{\mathrm{img}}(\mathbf{v}_i). \end{aligned} \tag{5}$$

Here d_i is camera-space depth, Ω_I is the image domain, and $\chi_i^{\mathrm{proj}} \in \{0,1\}$ marks projections in front of the camera and inside the image. Invalid projections contribute a zero image feature. Depth-aware visibility is used separately when constructing supervision (Sec. 3.3).

Gated residual fusion. The fusion module concatenates the local SLat feature, adapted image feature, and projection-valid bit into $\mathbf{h}_i^{\mathrm{joint}} \in \mathbb{R}^{385}$. It predicts a residual feature $\mathbf{h}_i^{\mathrm{res}} \in \mathbb{R}^{192}$ and a scalar gate $\alpha_i \in [0,1]$ from this joint representation:

$$\begin{aligned} \mathbf{h}_i^{\mathrm{joint}} &= [\mathbf{h}_i^{\mathrm{loc}}; \mathbf{h}_i^{\mathrm{img}}; \chi_i^{\mathrm{proj}}], \\ \mathbf{h}_i^{\mathrm{res}} &= \mathrm{Fuse}_{\mathrm{res}}(\mathbf{h}_i^{\mathrm{joint}}), \qquad \alpha_i = \chi_i^{\mathrm{proj}}\,\mathrm{sigmoid}(\mathrm{Gate}(\mathbf{h}_i^{\mathrm{joint}})). \end{aligned} \tag{6}$$

The validity mask forces $\alpha_i = 0$ outside valid projections. A residual head then updates the complete SLat prediction:

$$\Delta\mathbf{q}_i^{\mathrm{vis}} = \mathrm{Head}_{\mathrm{vis}}(\mathbf{h}_i^{\mathrm{res}}), \qquad \mathbf{q}_i = \mathbf{q}_i^{\mathrm{SLat}} + \alpha_i \Delta\mathbf{q}_i^{\mathrm{vis}}. \tag{7}$$

Both heads output 21 channels, so $\Delta\mathbf{q}_i^{\mathrm{vis}}, \mathbf{q}_i \in \mathbb{R}^{21}$. A zero gate preserves the base field; a nonzero gate scales the correction before the raw outputs are decoded into physical properties.

Decoding physical properties. Material logits give $\mathbf{p}_i = \mathrm{softmax}(\mathbf{q}_i^{\mathrm{mat}})$. For $a \in \{\rho, E\}$, a voxel-specific offset is combined with material-conditioned priors in standardized log space and then converted to physical units:

$$\hat{\xi}_i^a = q_i^{a,\mathrm{mean}} + \sum_{k=1}^{16} p_{ik}\eta_k^a, \qquad \hat{a}_i = \exp(\sigma_a \hat{\xi}_i^a + \mu_a). \tag{8}$$

Here $q_i^{a,\mathrm{mean}}$ is a scalar offset, η_k^a is the standardized class-k prior, and μ_a and σ_a are the training-set mean and standard deviation of the natural-log property. The scalar $\hat{\xi}_i^a$ is its normalized prediction. The remaining channels give a bounded Poisson ratio and auxiliary log variances for density and modulus:

$$\hat{\nu}_i = 0.499\,\mathrm{sigmoid}(q_i^{\nu}), \qquad \hat{s}_i^a = \mathrm{clip}(q_i^{a,\mathrm{var}}, -6, 6), \quad a \in \{\rho, E\}. \tag{9}$$

The resulting 16 material probabilities and three physical scalars form the $N \times 19$ field introduced in Sec. 3.1; the auxiliary variances are used only by the training loss. Appendix B gives the normalization statistics and output-head layout.

3.3 Supervision and Training

Training connects reference-part annotations to generated voxels and uses the accepted labels to optimize the predictor. We first construct this correspondence, then define the masked property losses and joint optimization.

Supervision construction. Reference and generated surfaces can differ in shape, so we establish correspondence in their shared source view. Silhouette fitting aligns the mesh decoded from $\mathcal{Z}$ with the foreground of I and defines π_I. For accepted fits, we render reference-part labels in the same view and erode their boundaries to remove pixels near part transitions. Depth-aware rasterization assigns pixels to the frontmost active voxel of Pred-SLat. Each visible voxel receives the labels of the part with the largest pixel overlap, provided it has sufficient support and label purity. Purity is the fraction of supporting pixels assigned to that part. The scalar weight w_i combines purity, pixel support, and silhouette overlap so that better supported correspondences contribute more to training.

Training objective. The objective combines four property losses over voxels with accepted targets, weighted by w_i. We apply a separate validity mask to each property so that available annotations still contribute to training. Material-type prediction uses class-balanced cross entropy. Density and Young's modulus are supervised in the standardized log space defined in Sec. 3.2, using Smooth-L1 together with a heteroscedastic Gaussian term (Kendall & Gal, 2017) and squared log-variance regularization. Poisson ratio uses Smooth-L1. Each loss is normalized by its valid weight mass, and the total objective is

$$\mathcal{L} = \lambda_m \mathcal{L}_{\mathrm{mat}} + \lambda_\rho \mathcal{L}_\rho + \lambda_E \mathcal{L}_E + \lambda_\nu \mathcal{L}_\nu. \tag{10}$$

Optimization. The SLat branch, source-image branch, gated residual fusion module, and output heads are trained jointly, with the TRELLIS generator and CLIP encoder frozen. The residual heads are zero-initialized, so the initial combined raw output equals the SLat branch's base raw output. We apply object-level dropout to the source-image features during training.

3.4 Simulation Interface

The decoded mesh and predicted field share the coordinates of Pred-SLat $\mathcal{Z}$. We assign each mesh face the properties of its nearest active voxel and convert the density, Young's modulus, and Poisson ratio to per-particle parameters for the material point method (MPM). A fixed Neo-Hookean elastic model evolves the particles under the specified load. The resulting trajectories deform the textured mesh for rendering.

4 Experiments

We evaluate physical-property prediction, the contribution of source-image appearance, and the effect of transferring supervision through the source view. MPM simulations then test how field errors affect deformation under the same load.

Table 1: Physical-property prediction on CM-25. Overall MAE averages the two source-specific MAEs with equal weights. Fine F1 evaluates 16 material classes; Coarse F1* uses an approximate three-group mapping across material taxonomies. "–" denotes no native 16-class material output. The appendix gives the complete breakdown.

Method	Material F1 ↑		Overall MAE ↓			PhysX-Mobility MAE ↓			PixieVerse MAE ↓		
	Fine (16)	Coarse*	$\log \rho$	$\log E$	ν	$\log \rho$	$\log E$	ν	$\log \rho$	$\log E$	ν
NeRF2Physics-style	–	0.299	1.142	4.879	0.061	0.990	4.211	0.044	1.293	5.547	0.078
PIXIE	–	0.437	0.759	3.836	0.043	0.904	6.480	0.041	0.614	**1.192**	0.045
SLAT-Phys-style	0.189	0.382	0.660	1.985	0.035	0.832	2.354	0.029	**0.488**	1.616	0.041
LP^{2}SLAT	**0.667**	**0.738**	**0.419**	**1.151**	**0.032**	**0.286**	**0.713**	**0.027**	0.552	1.590	**0.037**

4.1 Experimental Setup

Data and splits. The reported experiments use PhysX–Pixie, a corpus assembled from 1,703 PhysX-Mobility objects (Cao et al., 2026a) and 922 processed PixieVerse objects (Le et al., 2025). It is split by object into 2,101/262/262 training/validation/test objects.

Training setup. We jointly train the SLat branch, source-image branch, gated residual fusion module, and prediction heads of LP^{2}SLAT. The TRELLIS generator and CLIP image encoder remain frozen, using Microsoft's released TRELLIS-image-large and OpenAI's CLIP ViT-L/14 checkpoints, respectively.[1] Pred-SLat generation and predictor training run on a university GPU cluster. We optimize the predictor on a single NVIDIA V100S GPU (32 GB) using AdamW with mixed precision, a learning rate of 2×10^{-4}, a batch size of four objects, and a maximum of 50 epochs. During training, we randomly sample up to 8,192 voxels from each object's Pred-SLat and compute losses at sampled locations with valid labels. At test time, we process all Pred-SLat voxels and evaluate predictions at labeled locations. Of the 262 test objects, 225 carry accepted physical labels; per-metric support is reported in Appendix C.

Evaluation metrics. We report material macro-F1 with equal object weights to prevent objects with more labeled voxels from dominating the score. Continuous-property prediction is evaluated using mean absolute error (MAE) for natural-log density, natural-log Young's modulus, and Poisson ratio. To balance the contributions of the two datasets despite their different sizes, we compute MAE for each object, average it within each dataset, and then average the two dataset means.

4.2 Physical-Property Prediction Results

We compare material classification and continuous-property prediction across LP^{2}SLAT, PIXIE (Le et al., 2025), NeRF2Physics-style (Zhai et al., 2024), and SLAT-Phys-style (Das & Manocha, 2026) in Table 1. PIXIE uses its released checkpoints and 200 calibrated RGB views rendered from each generated asset, whereas LP^{2}SLAT uses only one source image and its generated Pred-SLat. NeRF2Physics-style is our single-view adaptation: source-view CLIP features and material proposals support candidate regression in place of multi-view NeRF/F3RM fitting. SLAT-Phys-style follows the published decoder and is trained on the same PhysX–Pixie split using the same single-image Pred-SLat input as LP^{2}SLAT; its architecture is described in Sec. 4.3.

Table 1 compares the four pipelines on CM-25, which contains 13 PhysX-Mobility and 12 PixieVerse objects. LP^{2}SLAT has the highest material F1 and the lowest source-averaged MAE. For material classification, Fine F1 reaches 0.667 compared with 0.189 for SLAT-Phys-style, while Coarse F1* increases from PIXIE's 0.437 to 0.738 (paired difference 0.301; 95% CI $[0.083, 0.504]$). For continuous-property prediction, log-density and log-modulus MAE decrease by 36.5% and 42.0% relative to SLAT-Phys-style, and by 44.8% and 70.0% relative to PIXIE, respectively. Overall Poisson-ratio MAE is also lowest at 0.032. On PhysX-Mobility, LP^{2}SLAT attains the lowest MAE for all three properties.

[1] TRELLIS: `https://huggingface.co/microsoft/TRELLIS-image-large`; CLIP: `https://huggingface.co/openai/clip-vit-large-patch14`.

Table 2: Representation and decoder ablation on PhysX–Pixie. Acc., Rec., and F1 denote material accuracy, macro recall, and macro-F1; Params is the number of trainable parameters (M).

Variant	Material ↑			MAE ↓			Params
	Acc.	Rec.	F1	$\log \rho$	$\log E$	ν	
Surface geometry	0.464	0.167	0.160	0.684	2.096	0.027	1.340
SLAT-Phys-style	0.502	0.206	0.216	0.501	1.355	0.025	6.393
Latent only	0.531	0.445	0.429	0.357	0.983	**0.021**	1.419
LP^{2}SLAT	**0.686**	**0.571**	**0.585**	**0.318**	**0.914**	0.022	2.364

Table 3: Gate-structure ablation under the complete Table 2 protocol. The shared scalar gate is the main LP^{2}SLAT design.

Gate design	Mat. F1 ↑	$\log \rho$ MAE ↓	$\log E$ MAE ↓	ν MAE ↓	Params
Material-only residual	0.579	0.369	0.996	0.021	2.364M
Property-specific gates	0.537	0.335	**0.863**	**0.020**	2.364M
Shared scalar gate (main)	**0.585**	**0.318**	0.914	0.022	2.364M

Table 4: Behavior of the main shared gate on the complete held-out set. Continuous "improved" is the fraction of target voxels whose absolute error decreases relative to the latent branch inside the same checkpoint; $\Delta|e|$ is latent-branch error minus full-model error.

Subset / target	Voxels	Gate median [p5,p95]	Improved ↑	Mean $\Delta\|e\|$ ↑ / Spearman
All projection-valid sites	3,304,777	0.630 [0.032,0.970]	–	–
Material: corrected	56,111	0.840 [0.421,0.994]	–	–
Material: degraded	10,237	0.773 [0.520,0.955]	–	–
Density	460,321	0.583 [0.021,0.971]	52.7%	0.100 / 0.255
Young's modulus	458,111	0.581 [0.021,0.971]	56.1%	0.230 / 0.268
Poisson ratio	470,011	0.573 [0.021,0.970]	53.8%	0.0019 / 0.096

4.3 Ablation Studies

Three baselines test what the input features and decoder contribute (Table 2). *Surface geometry* maps face positions, normals, areas, and curvature to voxels and predicts properties with a local/global MLP, testing geometry-based prediction without latent or image features. *SLAT-Phys-style* applies a sparse-Swin decoder following Das & Manocha (2026) to Pred-SLat and uses our physical heads and objective, testing an alternative latent readout. *Latent only* retains our local SLat encoder and pooled object context but removes the image branch and gated residual fusion, isolating the contribution of source-image appearance.

The latent-only model raises F1 from 0.216 to 0.429 over SLAT-Phys-style and lowers density and modulus errors. Adding the image branch raises F1 to 0.585, a paired gain of 0.156 (95% CI $[0.081, 0.236]$). The clearest benefit of source appearance is material recognition.

Gate design. The main model uses one projection-valid scalar gate per voxel and separate zero-initialized residual heads. We test two alternatives under the same data and optimization budget: a property-specific design predicts four gates, one per property, while a material-only design keeps image residuals out of every continuous-property output.

The shared gate gives the highest material F1 and the lowest log-density MAE, while the property-specific variant slightly lowers the modulus and Poisson errors.

Learned gate behavior. Table 4 compares the full prediction with the base field inside the same trained model on all 225 held-out objects. The image residual revises 66,348 material decisions, of which 56,111 are corrections. On corrected decisions the median gate is 0.840, versus 0.773 on degraded ones. For the continuous properties, 52.7–56.1% of labeled voxels move closer to the reference, and error reduction grows with gate magnitude (Spearman correlations 0.255, 0.268, and 0.096 for density, modulus, and Poisson ratio).

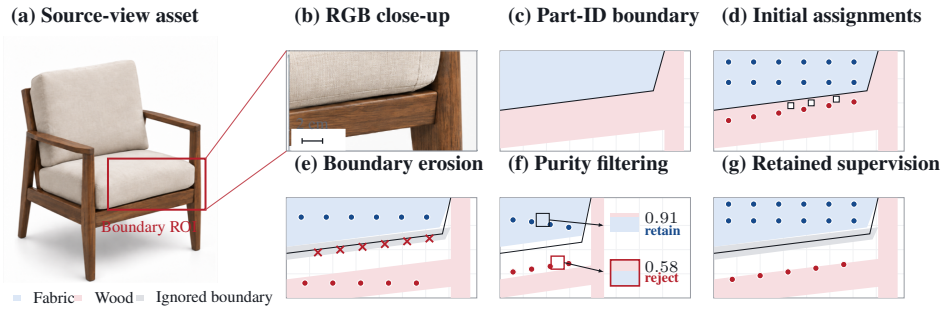

Figure 3: **Source-view supervision at an upholstery–wood boundary.** The RGB crop locates the part interface. Boundary erosion removes near-interface pixels, and label-purity filtering rejects voxel assignments with mixed part support.

Table 5: Source-view transfer ablation. Predictors are scored on identical test targets; voxel coverage reports each transfer's accepted supervision over the full test population.

Transfer	Voxel cov. ↑	Mat. F1 ↑	$\mathrm{MAE}_{\log E}$ ↓
ICP / 3D registration	50.7%	0.348	1.094
Silhouette-only	84.4%	0.533	**0.871**
w/o boundary/purity	18.2%	0.531	0.927
Full transfer	15.2%	**0.585**	0.914

Table 6: Runtime on CM-25, reported in seconds per asset. Stage-wise timings separate physics prediction from asset processing and export.

Method	Build	Feat.	Head	Bind	Total
NeRF2Physics-style	4.707	0.085	0.024	1.121	5.801
PIXIE	704.798	14.599	0.190	0.388	732.276
SLAT-Phys-style	4.707	—	0.025	**0.327**	**5.011**
LP^{2}SLAT	4.707	**0.073**	**0.008**	0.328	5.078

Efficiency. We measure inference efficiency on CM-25 using an NVIDIA A40 GPU with batch size 1. Table 6 separates asset construction, feature preparation, physics prediction, and binding/export. The three SLat-based pipelines share the same TRELLIS-generated assets and the same generation cost of 4.707 seconds, so generation dominates their runtime. The LP^{2}SLAT physics head takes 0.008 seconds versus 0.025 seconds for SLAT-Phys-style, and its readout-and-export subtotal is 0.401 seconds, keeping physical-field prediction and export well below one second per asset. PIXIE instead renders 200 calibrated views and fits F3RM, so its construction stage takes 704.798 seconds.

4.4 Source-View Supervision

We compare four ways to transfer reference-part annotations to generated Pred-SLat voxels. *ICP / 3D registration* aligns the reference and generated geometries and transfers the label of the nearest reference point. *Silhouette-only* projects the reference part map into the source view and accepts voxels inside the overlapping foreground, without testing visibility or part boundaries. *Source-view transfer w/o boundary/purity* adds fitted source-view alignment and depth-aware visibility, but keeps pixels near part boundaries and voxels supported by mixed part labels. *Full transfer* erodes projected part boundaries and accepts a voxel only when its supporting pixels agree on one part label. Figure 3 illustrates how the two filters act near an upholstery–wood boundary.

With evaluation targets fixed, full source-view supervision raises material F1 by 0.237 and lowers log-modulus MAE by 0.180 relative to ICP transfer. Boundary/purity filtering changes F1 from 0.531 to 0.585 while reducing voxel coverage from 18.2% to 15.2%.

4.5 Downstream MPM Evaluation

We compare 25 PixieVerse objects using a shared Warp-MPM solver (Xie et al., 2023); only the predicted (E, ν, ρ) fields differ. Relative to SLAT-Phys-style, LP^{2}SLAT reduces trajectory RMSE by at least 25% on 20 objects, with smaller reductions on four. In the eight-object four-method comparison, it attains the lowest RMSE on five assets. Figure 4 shows representative cases; Appendix D gives the complete comparison and time sequences.

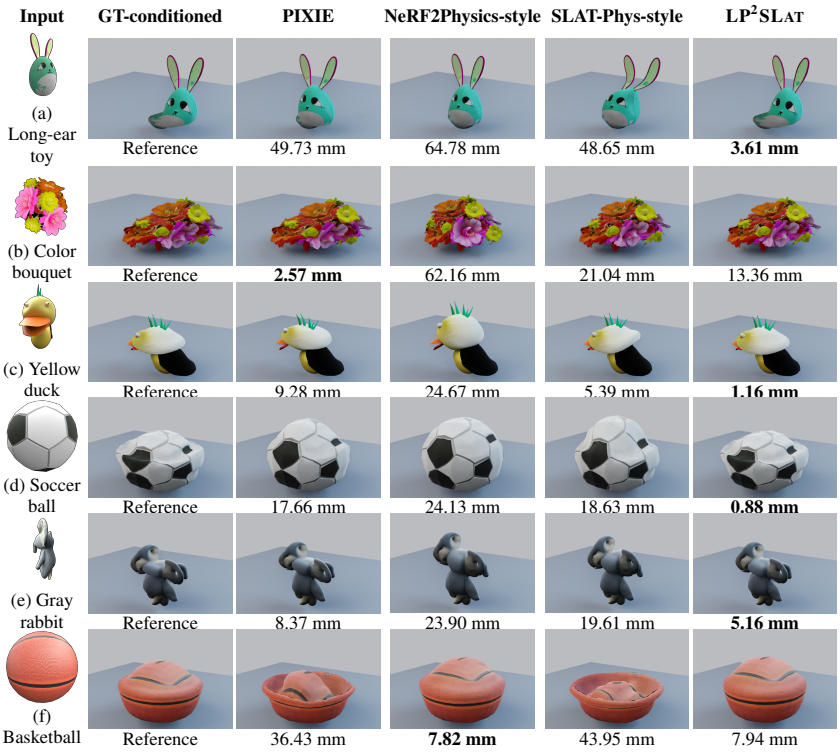

Figure 4: **MPM responses on four representative assets.** All columns share particles, load, solver, camera, and appearance; only (E, ν, ρ) changes. Full cases and time sequences are in Appendix D.

The 25-object pool covers toys, sport balls, cans and crates, plants, and granular-appearance assets, each with a source image and exported field. All runs use a 96^3 grid, 48 frames, 20 substeps, one fixed Neo-Hookean elastic family, and identical particles, boundary conditions, and load; we fix the load with the reference-conditioned fields alone, choosing among four candidate magnitudes the stable one closest to 0.14 m peak displacement. Faces without an accepted reference annotation receive the nearest valid value, so reference and prediction runs share the same particle set, and a shared Gaussian appearance, camera, and particle-to-vertex map turn the solver trajectories into the textured sequences of Figure 4.

4.6 Discussion

The experiments support three observations. Read from Pred-SLat alone, a structured latent already places a physical field on the generated surface: material F1 rises from 0.216 to 0.429 and both continuous errors fall. The source image adds the local material evidence the latent under-represents, lifting F1 to 0.585, and gate corrections are associated with lower error. Transferring annotations through the shared source view is what makes supervision work across assets: F1 rises by 0.237 over 3D registration, and boundary/purity filtering recovers material F1 at part interfaces. Driving an unmodified MPM solver with everything else held fixed, the predicted fields reduce trajectory error on 24 of the 25 tested assets.

5 Conclusion

LP^{2}SLat predicts spatial physical priors by combining the generated SLat with source-image appearance: the SLat branch defines a base field, gated image residuals refine it, and shared-view correspondence supplies training labels across reference and generated geometry. Image refinement primarily improves material recognition, while source-view supervision improves prediction over ICP-based transfer. PhysSLat-227 extends the available part-level physical-prior collections and will be released with the implementation.

Reproducibility statement

The appendix documents the data sources, the object-disjoint split, target-transfer rules, baseline specifications, hyperparameters, checkpoint selection, metric aggregation, confidence intervals, and the runtime protocol. The implementation, PhysSLat-227, and the raw evaluation outputs will be released with the trained checkpoints.

AI use statement

The authors used AI-based writing assistance for language editing and manuscript formatting. All research content—the method design, dataset construction, experiments, statistical analyses, and all reported numbers—was produced and verified by the authors, who take full responsibility for the content of this submission.

Ethics statement

This study uses publicly available 3D asset sources and physically simulated responses; it involves no personal data or human subjects.

References

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

Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-Omni: Unified simulation-ready physical 3d generation for rigid, deformable, and articulated objects. *arXiv preprint arXiv:2605.21572*, 2026b. URL https://arxiv.org/abs/2605.21572v1.

Rocktim Jyoti Das and Dinesh Manocha. Slat-phys: Fast material property field prediction from structured 3d latents. *arXiv preprint arXiv:2603.23973*, 2026.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, 2025.

Chunji Lv, Zequn Chen, Donglin Di, Weinan Zhang, Hao Li, Wei Chen, Yinjie Lei, and Changsheng Li. PhysGM: Large physical gaussian model for feed-forward 4d synthesis. *arXiv preprint arXiv:2508.13911*, 2025. URL https://arxiv.org/abs/2508.13911v4.

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

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023. URL https://arxiv.org/abs/2311.12198.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16262–16272, 2024.

Yunhan Yang, Chunshi Wang, Junliang Ye, Yang Li, Zanxin Chen, Zehuan Huang, Yao Mu, Zhuo Chen, Chunchao Guo, and Xihui Liu. Physforge: Generating physics-grounded 3d assets for interactive virtual world. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.