SLAT-Phys: Fast Material Property Field Prediction from Structured 3D Latents

Rocktim Jyoti Das[1], Dinesh Manocha[1]

***Abstract*— Estimating material property field of 3D assets is critical for physics-based simulation, robotics, and digital twin generation. Existing vision based approaches are either too expensive and slow or rely on 3D information. We present SLAT-Phys, an end-to-end method that predicts spatially varying material property fields of 3D assets directly from a single RGB image without explicit 3D reconstruction. Our approach leverages spatially organized latent features from a pretrained 3D asset generation model, that encode rich geometry and semantic prior, and trains a lightweight neural decoder to estimate Young's modulus (E), density(ρ), and Poisson's ratio(ν). The coarse volumetric layout and semantic cues of the latent representation about object geometry and appearance enables accurate material estimation. Our experiments demonstrates that our method provides competitive accuracy in predicting continuous material parameters when compared against prior approaches, while significantly reducing computation time. In particular, SLAT-Phys requires only ~ 9.9 seconds per object on an NVIDIA RTXA5000 GPU and avoids reconstruction and voxelization preprocessing. This results in $\sim 120\times$ speedup compared to prior methods and enables faster material property estimation from a single image.**

I. Introduction

Understanding the physical properties of objects from visual observations is a long-standing challenge in computer vision, physically-based simulation, and robotics [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Accurate estimation of spatially varying mechanical properties, such as Young's modulus(E), density(ρ), and Poisson's ratio(ν) is important for realistic deformation modeling, stable robotic manipulation, contact-rich interaction, and high-fidelity digital twin construction. In some robotics applications, incorrect material assumptions can lead to unstable grasps or inaccurate force prediction. In physics-based simulation and virtual environments, physically implausible material parameters result in unrealistic animations and non-physical behavior. Consequently, material property estimation [12], [7] has become a critical component in the construction of 3D representations that are simulation-ready of real-world objects.

Current 3D capture methods and widely-used 3D datasets [13], [14], [15] cover a wide range of 3D geometric shapes and mesh representations, but rarely contain any annotation of physical properties. This forces designers and simulation engineers to manually annotate or estimate the material properties, which can be inaccurate, expensive, and also leads to problems like subjectivity. Recently, there has been a lot of interest in developing algorithms [7], [16], [5], [8], [9] for physical property annotation of objects. Methods such as [7], [17], [16], [18], [10], [12], [19] employ differentiable physics solvers to iteratively optimize material parameters by matching simulated dynamics to observed signals or realism scores from generative models [20], [21]. However, the supervision signal is sparse for predicting physical parameters for hundreds of thousands of particles of MPM Simulation [22], [7] and is an extremely slow and difficult optimization process, often taking hours on a single scene.

Fig. 1: Different material field for different parts of the flower vase. The flower leaves are deformable whereas the vase is rigid.

In parallel, there have been efforts to develop vision based pipelines [5], [8], [9], [23] for physical property estimation, using semantic representation learned by 2D vision foundation models [24], [25]. These methods reconstruct the scene from multi-view images using 3D reconstruction techniques such as NeRF [26] or Gaussian Splatting [27]. Subsequently, a 3D semantic feature field is constructed by projecting dense 2D image features onto the resulting voxels from the reconstruction. While effective, such pipelines are computationally expensive require approximately 20 minutes of processing time on a NVIDIA RTXA5000 GPU for a single object as shown by our evaluation. The 3D reconstruction and feature extraction stages significantly increase inference time and limit scalability, especially when real-time

[1]University of Maryland, College Park
email: rocktimj@umd.edu

material inference is required for interactive robotics tasks or large-scale dataset processing. Many robotic applications require real-time understanding of physical properties to grasp objects without slipping, determine appropriate interaction forces [28], and estimate terrain traversability for navigation in outdoor environments [29], [30], [31]. Furthermore, fast single-view estimation of physical properties, combined with single-view 3D asset generation [32], [33], can accelerate both Real2Sim [34], [35] and Sim2Real [36] paradigms in robotics.

In this paper, we explore the use of large-scale 3D asset generation models to estimate the material properties. Large-scale 3D asset generation models [32], [33] have been useful for single-image 3D understanding. Trained on massive datasets [14], [15], these models learn structured latent representations that encode rich geometric and semantic priors. In particular, Structured LATent (SLAT) [32] representations which are structured latent features organized spatially in 3D, enabling high-quality 3D reconstruction from a single RGB image. These structured latents implicitly capture multi-view consistency, volumetric occupancy cues, and object-level semantics without requiring explicit test-time multi-view aggregation. Our goal is to leverage these pretrained generative priors for physically grounded inference, and thereby reduce the computational cost.

Main Results: In this work, we introduce **SLAT-Phys**, an end-to-end feedforward network that directly predicts continuous, spatially varying material property fields from structured 3D latent representations. Given a single RGB image, we extract spatially organized SLAT features from a pretrained 3D generation backbone. Unlike prior work which require constructing explicit geometry [5], [8], we utilize the semantic and geometric information encoded in the structured latent space and employ a lightweight neural decoder to regress Young's modulus (E), density(ρ), and Poisson's ratio(ν). Operating purely in latent space yields two major advantages. First, it removes expensive reconstruction stages such as multi-view rendering, voxelization, and 3D semantic feature construction, leading to much faster inference compared to reconstruction-based pipelines. Second, it allows us to test a broader hypothesis about representation learning: that structured generative latents encode sufficient physically grounded information to support downstream physical reasoning tasks.

Extensive experiments conducted under the Pixie [8] evaluation protocol demonstrate that SLAT-Phys achieves competitive accuracy in material prediction while reducing the inference time by $\sim$ 120 times. Our results indicate that structured 3D latents are not merely geometric priors for reconstruction but also encode physically informative signals that can be harnessed for simulation-ready digital twin generation from a single image. By bridging large-scale 3D generative modeling and material-aware physical inference, SLAT-Phys opens a new direction toward real-time, physically grounded 3D understanding. Rather than treating reconstruction and physics estimation as separate stages, our work suggests that structured latent representations can serve as a unified foundation for both geometry and physical parameter reasoning.

Our key contributions are summarized as follows:

- We introduce **SLAT-Phys**, the first approach to directly regress continuous material property fields from a single image without explicit 3D reconstruction or multi-view aggregation.
- We demonstrate that spatially organized SLAT features learned by large-scale 3D generation models encode physically meaningful information beyond geometry and appearance, enabling accurate estimation of Young's modulus (E), density (ρ), and Poisson's ratio (ν).
- Through extensive evaluation, we show that SLAT-Phys achieves competitive material prediction accuracy while enabling faster, simulation-ready digital twin generation from a single RGB image, achieving **120$\times$ faster inference**.

II. Related Work

A. Vision-based Physics Estimation

Inferring physical properties of objects from visual observations is a long-standing problem [1], [6], [3], [4]. Early work such as *image2mass* [2] demonstrated that object mass can be estimated from a single image by learning correlations between appearance, scale, and category priors. However, these approaches typically predict only global object-level properties and cannot recover spatially varying material fields. NeRF2Physics [5] leverages vision–language models and neural feature fields to infer material properties by reasoning over candidate materials suggested by language priors. Along similar lines, PhysGS [23] is combining Gaussian Splatting [27] with Bayesian inference [37] to give get prediction with an uncertainity estimate. Other approaches such as PhysDreamer [7], PGND [10] and PhysTwin [12] estimate physics parameters by matching simulated dynamics to observations using differentiable simulation or generative video models. While effective, these methods often require expensive per-object optimization and may produce simulator-specific parameters that do not generalize across different physics engines.

A recent line of work [9], [8] instead learns feed-forward mappings from visual or 3D features to spatially varying mechanical properties. The volumetric fields of Young's modulus, density, and Poisson's ratio is predicted from 3D representations using learned volumetric architectures. However, these methods typically rely on explicit 3D reconstructions or multi-view feature aggregation to obtain volumetric representations. In contrast, SLAT-Phys operates directly in the structured latent space of a pretrained 3D generative model [32], [33] and predicts spatially varying mechanical properties from a single image without explicit 3D reconstruction, enabling significantly faster inference while maintaining competitive accuracy.

B. Learned 3D Representations

Traditional 3D representations such as meshes, voxel grids, and signed distance fields (SDFs) explicitly encode object geometry but do not capture higher-level semantic or material information, limiting their usefulness for reasoning about physical properties. Early work [4] proposed learning material-aware local descriptors on 3D surfaces, where a projective CNN extracts view-based features around surface points to classify material labels, demonstrating that local geometric context can provide cues about object materials. More recent work such as F3RM [38] has focused on incorporating semantic information into spatial representations by aggregating features from pretrained vision models [25], [24]. This multi-view feature aggregation paradigm is also adopted by recent physics estimation frameworks such as NeRF2Physics [5] and Pixie [8], which rely on explicit 3D reconstruction and multi-view feature lifting to estimate material properties. However, these approaches require expensive reconstruction pipelines and multi-view aggregation. They take approximately 20 mins of inference time per object on an NVIDIA RTXA5000 GPU. More recently, large-scale 3D generation models such as TRELLIS [32], [33] proposes Structured LATents (SLAT), a spatially organized latent representation that encode geometric and semantic priors directly from data. In this work, we build on this representation and developed **SLAT-Phys** to directly predict spatially varying material property fields from a single image without explicit 3D reconstruction.

III. Our Method: SLAT-Phys

A. Method Overview

Our goal is to estimate spatially varying material properties of a 3D object from a single RGB image. Specifically, we predict per-voxel mechanical parameters including Young's modulus E, density ρ, and Poisson's ratio ν, along with a discrete material label. Given an accurate and valid triplet (E, ν, ρ) along with a reasonable material model, a consistent numerical simulation such as MPM [39] can produce accurate predictions of an object's behavior under external force.

Our approach consists of three stages. First, a pretrained image-to-3D generation model extracts a sparse structured latent representation of the object from a single image. These structured latents, referred to as *Structured LATents (SLAT)*, encode geometric and semantic information at a set of occupied voxel locations. Second, a lightweight neural decoder predicts per-voxel physical properties directly from the SLAT features using a sparse transformer architecture. Finally, the predicted material fields are passed to a downstream physics simulator to generate physically plausible object behavior.

B. Problem Formulation

We consider the problem of estimating spatially varying physical material properties of a 3D object from a single RGB observation. Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input RGB image containing an object. Our goal is to estimate a volumetric field of physical parameters defined over a discretized voxel grid. We represent the object using a voxel grid

$$\mathcal{V} = \{v_i\}_{i=1}^{N}, \quad v_i \in \mathbb{R}^3, \tag{1}$$

where each voxel corresponds to a spatial location in a canonical 64^3 grid. For each voxel we aim to predict a vector of physical parameters

$$\mathbf{p}_i = [E_i, \rho_i, \nu_i, c_i], \tag{2}$$

where E_i denotes Young's modulus, ρ_i the density, ν_i the Poisson's ratio, and c_i a discrete material class label. Instead of predicting these parameters from dense 3D geometry, we leverage a structured latent representation produced by a pretrained 3D generative model. Given an input image I, the encoder of the 3D generation model produces a set of sparse latent features

$$\mathcal{Z} = \{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^{N}, \tag{3}$$

where $\mathbf{x}_i$ denotes the voxel coordinate and $\mathbf{z}_i \in \mathbb{R}^d$ denotes the latent feature vector. Our model learns a mapping

$$f_\theta : (\mathbf{x}_i, \mathbf{z}_i) \to (E_i, \rho_i, \nu_i, c_i), \tag{4}$$

where f_θ is implemented as a sparse transformer decoder operating on the structured latent representation.

C. SLAT Feature Extraction

Given an RGB image I, we use a pretrained image-to-3D model to extract a structured latent representation of the object. Our encoder predicts a sparse set of occupied voxel coordinates and associated latent features. For each object the model produces a set

$$\{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^{N}, \tag{5}$$

where $\mathbf{x}_i \in \{0, \dots, 63\}^3$ denotes the voxel coordinate and $\mathbf{z}_i \in \mathbb{R}^8$ represents an 8-dimensional latent feature vector. These voxels correspond to the predicted surface of the reconstructed object and form a sparse representation of the geometry. The TRELLIS encoder is kept frozen during training, allowing the model to benefit from geometric priors learned from large-scale 3D generative training.

D. Physics Decoder

The structured latent features from the pretrained image-to-3D model are processed by a neural decoder that predicts physical properties for each voxel. The decoder is implemented as a sparse Swin-style transformer [40] operating directly on the occupied voxel set. The network consists of an input projection layer, positional encoding, four sparse transformer blocks, and regression and classification heads. Each transformer block performs windowed self-attention over local voxel neighborhoods to capture spatial dependencies between nearby latent features. The final predictions are

$$\hat{p}_i = (\hat{E}_i, \hat{\rho}_i, \hat{\nu}_i, \hat{c}_i), \tag{6}$$

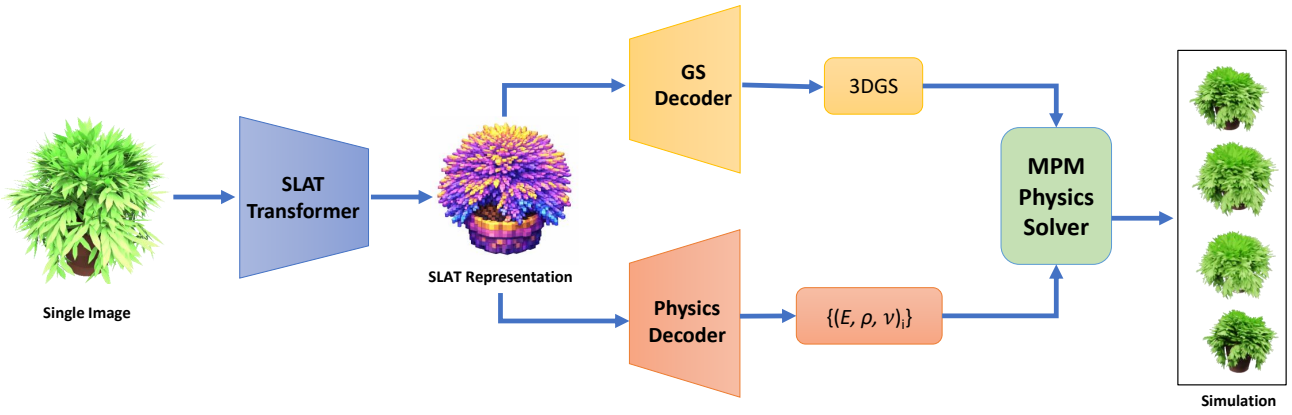

Fig. 2: **Overview of SLAT-Phys:** Given a single image of a 3D asset, our framework first encodes the image using the Trellis encoder to obtain a structured 3D latent (SLAT) representation. From this latent, two decoders predict the corresponding 3D Gaussian Splatting (3DGS) representation and spatially varying material properties (E, ρ, ν) at each voxel. The resulting Gaussian representation together with the predicted physical parameters are then passed to an MPM solver to perform physics-based simulation.

computed using two output heads. The regression head predicts continuous physical parameters, while the classification head predicts discrete material labels.

E. Training Data

We train our model using the PixieVerse dataset introduced in Pixie [8]. PixieVerse is a large-scale dataset of 3D objects paired with spatially varying physical material annotations. The dataset contains 1,624 high-quality single-object assets spanning 10 semantic categories, including organic objects (trees, shrubs, flowers), deformable toys (e.g., rubber ducks), sports equipment (balls), granular materials (sand, snow, mud), and hollow containers (e.g., soda cans and metal crates). The objects are sourced from the Objaverse repository [14], [15] and curated through a filtering pipeline to ensure geometric and visual quality. Each asset is annotated with both discrete material categories and continuous mechanical parameters, including Young's modulus (E), Poisson's ratio (ν), and mass density (ρ), enabling simulation-ready material fields. The annotations are generated through a semi-automatic labeling pipeline that leverages vision-language models [41] and CLIP feature fields [25], combined with manual verification to ensure physical plausibility. These annotations provide voxel-level supervision for learning spatially varying material property fields from visual observations.

F. Training Objective

Following Pixie [8], our model is trained to jointly predict continuous physical parameters and material classes. For each voxel with valid annotation we compute regression losses between the predicted and ground truth per voxel continuous physics:

$$\mathcal{L}_E = \|\hat{E}_i - E_i\|_2^2, \tag{7}$$

$$\mathcal{L}_\rho = \|\hat{\rho}_i - \rho_i\|_2^2, \tag{8}$$

$$\mathcal{L}_\nu = \|\hat{\nu}_i - \nu_i\|_2^2, \tag{9}$$

along with a material classification loss

$$\mathcal{L}_{mat} = \mathrm{CE}(\hat{c}_i, c_i), \tag{10}$$

where CE denotes cross-entropy loss. The final training objective is

$$\mathcal{L} = \lambda_E \mathcal{L}_E + \lambda_\rho \mathcal{L}_\rho + \lambda_\nu \mathcal{L}_\nu + \lambda_{mat} \mathcal{L}_{mat}. \tag{11}$$

IV. Experiments

Our experiments are designed to assess the performance of SLAT-Phys in terms of performance and runtime speed.

A. Implementation Details

a) SLAT Generation: For extracting structured latent features, we employ the TRELLIS image-to-3D generation framework [32]. Given a single RGB image of an object, the model produces a Structured LATent (SLAT) representation consisting of voxel coordinates and an 8-dimensional latent feature for each occupied voxel on a 64^3 grid. The coordinates indicate the spatial position of voxels, while the latent vectors capture geometric and semantic cues learned by the pretrained TRELLIS model. Throughout our pipeline, the TRELLIS encoder is kept frozen and used purely as a feature extractor.

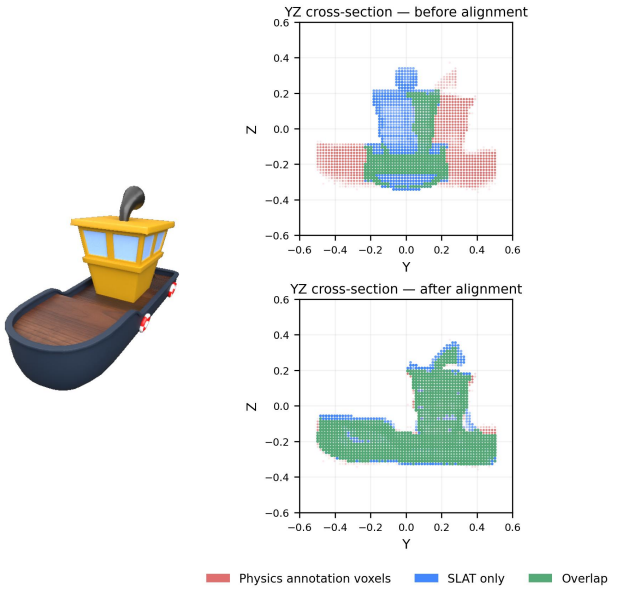

Fig. 3: **Physics-to-SLAT voxel alignment.** Physics annotations predicted by Pixie and the SLAT voxel grid produced by TRELLIS may exhibit a rigid rotational offset due to independent reconstruction pipelines. The figure illustrates the voxel grids before alignment and after applying the estimated rigid transformation using ICP.

b) Dataset: Training targets are obtained from the PixieVerse dataset introduced in Pixie [8], which provides volumetric predictions of physical material properties on a 64^3 voxel grid. For each object, PixieVerse provides normalized predictions of density (ρ), Young's modulus (E), and Poisson's ratio (ν), along with an 8-class material label represented as one-hot channels. The physical parameters are stored in normalized form within the range $[-1, 1]$, where density and Young's modulus are encoded in log-space and Poisson's ratio is encoded linearly. These voxel-level annotations serve as the supervision signal for training the physics decoder (as described in Sections III-D and III-F) that maps SLAT features to material property fields.

c) Physics-to-SLAT Annotation Alignment: The physics annotations used for supervision originate from the Pixie [8], while the SLAT representation is generated independently by the TRELLIS image-to-3D model. Although both representations share the same 64^3 voxel resolution, they may differ by a rigid rotational offset due to the independent reconstruction pipelines. As illustrated in Fig. 3, we align the physics voxel grid to the SLAT coordinate frame before training.

We compute the alignment using the Iterative Closest Point (ICP) algorithm. The occupied SLAT voxels serve as the reference surface, while the boundary voxels of the Pixie physics occupancy mask are used as the source point cloud. To avoid poor local minima, we evaluate 64 candidate orientations generated from combinations of $\{0°, 90°, 180°, 270°\}$ rotations around the three axes and select the initialization with the highest ICP fitness score. A final ICP optimization then estimates a rigid transformation $T \in SE(3)$ that maps the physics annotations into the SLAT coordinate frame. The aligned material properties (E, ρ, ν) are subsequently queried at the corresponding SLAT voxel locations to provide supervision during training.

d) Decoder Architecture: Following the 3D decoder design of TRELLIS [32], we implement a lightweight sparse Swin-style transformer [40] operating on the SLAT voxels. Each occupied voxel is represented by its coordinate in a 64^3 grid and an 8-dimensional latent feature vector. The latent features are projected to a 256-dimensional embedding and processed by eight transformer blocks with sixteen attention heads. The resulting voxel features are passed to two prediction heads: a regression head that predicts the normalized physical parameters (E, ρ, ν) using a $\tanh$ activation, and an auxiliary classification head that predicts one of eight material classes.

e) Training Details: The model is trained using a combination of regression and classification losses. Mean squared error (MSE) losses are applied to the normalized physical parameters (E, ρ, ν), while a cross-entropy loss supervises the auxiliary material classification task. The total loss is defined as a weighted sum of these components, with weights 1.0 for each regression term and 0.5 for the classification loss. Training is performed using the AdamW optimizer [42] with a learning rate of 10^{-4} and cosine annealing scheduling. We employ mixed precision training and gradient accumulation to improve training efficiency.

f) Physics-Based Simulation: We evaluate the predicted material fields using the Material Point Method (MPM), a particle–grid hybrid approach for simulating deformable materials [22]. The MPM solver takes as input a set of particles representing the object geometry along with their associated material parameters and external force specifications. In our approach, both geometry and physics originate from the same SLAT representation produced by TRELLIS. A Gaussian Splatting model is decoded from the SLAT features to reconstruct the object geometry, where each Gaussian naturally corresponds to a simulation particle [22]. The predicted material properties, Young's modulus (E), density (ρ), and Poisson's ratio (ν), are transferred from the SLAT voxel grid to the Gaussian particles via nearest-neighbor interpolation. The resulting set of particles with spatially varying material parameters is then passed to the MPM solver to simulate object deformation under external forces. We highlight the results for different models in Fig. 4.

B. Ablation Study: Physics Decoder Architecture

Model	log E err ↓	log ρ err ↓	ν error ↓	Mat. Acc. ↑
Small	0.0285	0.0630	0.0674	0.9107
Medium	0.0195	0.0433	0.0573	0.9355
Large	**0.0173**	**0.0361**	**0.0540**	**0.9453**

TABLE I: Ablation over Physics Decoder architectures of increasing capacity. All models are trained on the same data with identical optimization settings.

Fig. 4: Physics-based simulation results using the predicted physical parameters computed using SLAT-Phys. Given a single image, our method predicts spatially varying Young's modulus (E), Poisson's ratio (ν), and density (ρ) from SLAT features. The same SLAT representation is also decoded to obtain 3D Gaussian splats for geometry and appearance reconstruction. The resulting geometry and predicted physical parameters are then used in an MPM simulator to generate physically plausible dynamics. We highlight the simulation results for three objects with different material behaviors: a snow man and a rubber duck falling under gravity and a flower under the influence of wind.

We study the impact of model capacity by varying the number of channels, transformer blocks, and attention heads in the Physics Decoder. All models share the same Swin-attention-based sparse voxel architecture operating on a 64^3 grid with 8 latent input channels, and are trained for 2,000 steps under identical optimization settings (AdamW, lr=1e-4, cosine annealing). The Small model (64 channels, 4 blocks, 4 heads, 0.20M parameters) serves as the baseline. The Medium model (128 channels, 6 blocks, 8 heads, 1.19M parameters) doubles the channel width and adds two additional transformer blocks, yielding roughly a 6× parameter increase. The Large model (256 channels, 8 blocks, 16 heads, 6.32M parameters) further doubles the channel width and adds two more blocks, resulting in a 31× parameter increase over Small.

The Medium and Large models require fewer steps and we selected checkpoints before model starts overfitting and the results are reported in Table I. We can observe that increasing model capacity from Small to Large leads to consistent improvements across all regression metrics and material classification accuracy, indicating that higher representational capacity is beneficial for modeling physical properties. The Large model provides the strongest overall performance.

C. Quantitative Evaluation

a) Evaluation Metrics: Following PIXIE [8], we evaluate the physics decoder using mean squared error (MSE) in normalized space for each continuous material property, and classification accuracy for the discrete material class. Specifically, Young's modulus E and density ρ are first $\log_{10}$-transformed, then all three properties ($\log E$, $\log \rho$, ν) are linearly normalized to $[-1, 1]$ using the dataset min/max statistics. MSE is computed per voxel, averaged per object, then averaged globally. The aggregate continuous MSE is the mean of the three per-property MSEs. Material accuracy is the fraction of occupied voxels whose predicted material class matches the ground-truth label. These metrics measure the deviation between predicted and ground-truth physical properties at corresponding voxel locations.

b) Baselines: We compare our method with two representative prior approaches for estimating spatially varying mechanical properties from visual observations: NeRF2Physics [5], and Pixie [8]. NeRF2Physics assigns material properties by querying a large language model

Stage	NeRF2Physics	Pixie	SLAT-Phys
Preprocessing	Blender rendering: ∼65s	Blender rendering: ∼65s	Image preprocessing + DINO encoding: 0.138s
3D Reconstruction	NeRF training: 1020s	F3RM training: 575.3s	Sparse structure sampling: 4.69s
Feature Extraction	CLIP feature fusion: 96.8s	F3RM rendering + Voxelization: 606.6s	SLAT feature sampling: 5.03s
Physics Decoding	BLIP-2 captioning + LLM proposal + Property prediction: 14s	Neural inference: 14.1s	Physics decoder: 0.004s
Total per object	∼1196s (∼20 min)	∼1261s (∼21 min)	**∼9.9s**
Speedup vs Ours	121×	128×	1×

(a) Runtime performance of our approach with prior methods, and analyze the running time in different modules. We observe almost 120X speedup over prior methods.

Method	Mat. Acc. ↑	Avg. Cont. MSE ↓	log E err ↓	ν error ↓	log ρ err ↓
NeRF2Physics	0.274 ± 0.01	0.858 ± 0.109	1.115 ±0.165	0.462 ± 0.106	0.997 ± 0.162
Pixie	**0.985**± 0.011	0.056 ±0.005	0.022± 0.004	**0.034** ± 0.006	0.112 ± 0.009
SLAT-Phys (Ours)	0.9453 ± 0.112	**0.036**± 0.049	**0.017** ± 0.025	0.054± 0.074	**0.036** ± 0.049

(b) Material property prediction accuracy comparison.

TABLE II: Comparison with prior methods. (a) Runtime comparison across pipelines. (b) Material property prediction performance.

(LLM) for plausible materials and propagating these values to 3D points using CLIP feature similarities. In contrast, Pixie is feedforward models [43] that predict volumetric material fields including Young's modulus (E), density (ρ), and Poisson's ratio (ν), but both rely on an additional 3D reconstruction stage to obtain geometry-aware representations before physics prediction.

c) Mechanical Property Estimation: Table IIb compares SLAT-Phys with prior methods on the PixieVerse dataset. Compared to NeRF2Physics [5], our method achieves substantially better performance across all metrics.

SLAT-Phys and Pixie [8] are comparable with each other. These results demonstrate that accurate material property estimation can be achieved directly from structured latent representations without requiring explicit multi-view reconstruction.

d) Run-Time Comparison: Table IIa compares the per-object runtime of our method with NeRF2Physics [5] and Pixie [8]. All experiments are conducted on a single NVIDIA RTX A5000 GPU. Both prior approaches are dominated by per asset optimization-based 3D reconstruction pipelines, including NeRF or F3RM training, followed by rendering and voxelization, which account for the majority of their runtime.

In contrast, our method operates directly on the structured SLAT representation produced by TRELLIS, eliminating the need for reconstruction, dense rendering, and voxelization. As a result, our approach reduces the total runtime to 9.9 seconds per object, compared to ∼1196 seconds for NeRF2Physics and ∼1261 seconds for Pixie. Overall, this yields a speedup of approximately 121× over NeRF2Physics and 128× over Pixie, while maintaining competitive accuracy in material property prediction.

D. Qualitative Evaluation

Figure 4 shows qualitative simulation results using the physical parameters predicted by SLAT-Phys. We consider three scenarios: a falling snowman, a flower subjected to wind, and a falling rubber duck. Each example demonstrates distinct material behaviors. The snowman fractures and collapses upon impact, reflecting brittle and weak structural properties. Importantly, while the snow body breaks apart, the attached stick arms remain intact and rigid throughout the motion, highlighting the model's ability to capture heterogeneous material properties within a single object. The flower responds to wind forces with smooth bending and oscillatory motion, capturing flexible dynamics. The rubber duck exhibits deformable motion during free fall, consistent with elastic, rubber-like materials.

In addition to physically plausible dynamics, the rendered assets maintain high visual fidelity. The Gaussian splats are directly derived from the SLAT representation, which is also used for material prediction, without any additional training or fine-tuning. This highlights that the structured latent representation captures sufficient geometric and semantic information to support both realistic rendering and physically consistent simulation from a single image.

V. Conclusion and Limitations

We presented **SLAT-Phys**, a feedforward framework for spatially varying material property field estimation from a single RGB image. Unlike prior approaches that construct volumetric 3D features through multi-view projection and voxel aggregation, our method could generate material property field directly from single image. Through experiments following the Pixie [8] evaluation protocol, we demonstrate that SLAT-Phys achieves competitive material estimation

accuracy while requiring only a single image as input and reducing inference time by more than two orders of magnitude (over 120× faster than prior reconstruction-based pipelines). Such fast estimation of physical properties is important for many robotic applications, including adjusting grasp force to prevent slipping, reasoning about interaction forces during manipulation [28], and estimating terrain traversability in outdoor navigation [29], [30], [31]. Moreover, combining fast single-view physical property inference with single-view 3D asset generation [32], [33] can further enable scalable Real-to-Sim [34], [35] and Sim-to-Real [36] pipelines in robotics.

Despite these promising results, several limitations remain. First, the material annotations from Pixie [8] used for training are generated through a vision-language model (VLM) based pipeline, which require extensive prompt engineering and prompts are designed for specific objects to get reliable annotation. In contrast, recent work such as VoMP [9] adopts a more reliable annotation strategy by grounding material parameters in physically measured ranges curated from sources such as Wikipedia [44] and engineering databases [45]. At the time of our experiments, the codebase and dataset of VoMP [9] were not publicly available. We plan to analyze how our framework performs on that data in future work. Second, our current framework relies solely on visual input. However, vision alone may not fully capture certain physical cues related to material properties. Prior work such as ObjectFolder [28] demonstrates that multi-sensory learning, incorporating tactile and other sensing modalities, can significantly improve material understanding. Extending SLAT-Phys to incorporate additional sensory modalities such as touch or audio represents an important direction for improving robustness and physical reasoning in real-world robotic applications.

REFERENCES

[1] E. H. Adelson, "On seeing stuff: the perception of materials by humans and machines," in *IS&T/SPIE Electronic Imaging*, 2001. 1, 2

[2] T. S. Standley *et al.*, "image2mass: Estimating the mass of an object from its image," in *CoRL*, 2017. 1, 2

[3] S. Bell *et al.*, "Material recognition in the wild with the materials in context database," *CVPR*, 2014. 1, 2

[4] H. Lin *et al.*, "Learning material-aware local descriptors for 3d shapes," *3DV*, 2018. 1, 2, 3

[5] A. J. Zhai *et al.*, "Physical property understanding from language-embedded feature fields," *CVPR*, 2024. 1, 2, 3, 6, 7

[6] J. Wu *et al.*, "Physics 101: Learning physical object properties from unlabeled videos," in *BMVC*, 2016. 1, 2

[7] T. Zhang *et al.*, "Physdreamer: Physics-based interaction with 3d objects via video generation," *arXiv*, 2024. 1, 2

[8] L. Le *et al.*, "Pixie: Fast and generalizable supervised learning of 3d physics from pixels," *arXiv*, 2025. 1, 2, 3, 4, 5, 6, 7, 8

[9] R. Dagli *et al.*, "Vomp: Predicting volumetric mechanical property fields," *arXiv*, 2025. 1, 2, 8

[10] K. Zhang *et al.*, "Particle-grid neural dynamics for learning deformable object models from rgb-d videos," *RSS*, 2025. 1, 2

[11] A. Pumarola *et al.*, "D-nerf: Neural radiance fields for dynamic scenes," *CVPR*, 2021. 1

[12] H. Jiang, H.-Y. Hsu, K. Zhang, H.-N. Yu, S. Wang, and Y. Li, "Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos," *ICCV*, 2025. 1, 2

[13] Z. Dong *et al.*, "Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset," *CVPR*, 2025. 1

[14] M. Deitke *et al.*, "Objaverse: A universe of annotated 3d objects," *CVPR*, 2022. 1, 2, 4

[15] M. Deitke, R. Liu *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *arXiv*, 2023. 1, 2, 4

[16] T. Huang *et al.*, "Dreamphysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors," *arXiv*, 2024. 1

[17] K. M. Jatavallabhula *et al.*, "gradsim: Differentiable simulation for system identification and visuomotor control," *arXiv*, 2021. 1

[18] X. Li *et al.*, "Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification," *arXiv*, 2023. 1

[19] Y. Lin *et al.*, "Omniphysgs: 3d constitutive gaussians for general physics-based dynamics generation," *arXiv*, 2025. 1

[20] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, and D. Lorenz, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv*, 2023. 1

[21] J. Bruce *et al.*, "Genie: Generative interactive environments," *arXiv*, 2024. 1

[22] T. Xie, Z. Zong, Y. Qiu, X. Li, Y. Feng, Y. Yang, and C. Jiang, "Physgaussian: Physics-integrated 3d gaussians for generative dynamics," *CVPR*, 2023. 1, 5

[23] S. Chopra, J. Liang, G. Seneviratne, and D. Manocha, "Physgs: Bayesian-inferred gaussian splatting for physical property estimation," *CVPR*, 2026. 1, 2

[24] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv*, 2023. 1, 3

[25] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021. 1, 3, 4

[26] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *CACM*, 2021. 1

[27] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM TOG*, 2023. 1, 2

[28] R. Gao *et al.*, "The object folder benchmark : Multisensory learning with neural and real objects," *CVPR*, 2023. 2, 8

[29] G. Seneviratne *et al.*, "Cross-gait: Cross-attention-based multimodal representation fusion for parametric gait adaptation in complex terrains," *IROS*, 2024. 2, 8

[30] K. Weerakoon, A. J. Sathyamoorthy, J. Liang, T. Guan, U. Patel, and D. Manocha, "Graspe: Graph based multimodal fusion for robot navigation in outdoor environments," *RAL*, 2023. 2, 8

[31] M. B. Elnoor *et al.*, "Vlm-gronav: Robot navigation using physically grounded vision-language models in outdoor environments," *ICRA*, 2025. 2, 8

[32] J. Xiang *et al.*, "Structured 3d latents for scalable and versatile 3d generation," *CVPR*, 2024. 2, 3, 4, 5, 8

[33] Y. nuo Yang *et al.*, "Sam3d: Segment anything in 3d scenes," *arXiv*, 2023. 2, 3, 8

[34] Z. Xie *et al.*, "Vid2sim: Realistic and interactive simulation from video for urban navigation," *CVPR*, 2025. 2, 8

[35] A. Escontrela *et al.*, "Gaussgym: An open-source real-to-sim framework for learning locomotion from pixels," *arXiv*, 2025. 2, 8

[36] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," *arXiv*, 2021. 2, 8

[37] C. P. Robert, J.-M. Marin, and J. Rousseau, "Bayesian inference," 2010. 2

[38] B. W. Shen *et al.*, "Distilled feature fields enable few-shot language-guided manipulation," in *CoRL*, 2023. 3

[39] Y. Hu *et al.*, "A moving least squares material point method with displacement discontinuity and two-way rigid body coupling," *ACM TOG*, 2018. 3

[40] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *ICCV*, 2021. 3, 5

[41] G. Team, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv*, 2025. 4

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2017. 5

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv*, 2015. 7

[44] Wikipedia contributors, "Density," Wikipedia, 2024, accessed: 2025-06-25. [Online]. Available: https://en.wikipedia.org/wiki/Density 8

[45] Engineering ToolBox, "Engineering materials properties," 2024, accessed: 2025-06-25. [Online]. Available: https://www.engineeringtoolbox.com/engineering-materials-properties-d_1225.html 8