GaussianProperty: Integrating Physical Properties to 3D Gaussians with LMMs

Xinli Xu[1*] Wenhang Ge[1*] Dicong Qiu[1*] ZhiFei Chen[1] Dongyu Yan[1] Zhuoyun LIU[1]
Haoyu Zhao[1] Hanfeng Zhao[3] Shunsi Zhang[3] Junwei Liang[1,2] Ying-Cong Chen[1,2†]

HKUST(GZ)[1] HKUST[2] Quwan[3]

Abstract

Estimating physical properties for visual data is a crucial task in computer vision, graphics, and robotics, underpinning applications such as augmented reality, physical simulation, and robotic grasping. However, this area remains under-explored due to the inherent ambiguities in physical property estimation. To address these challenges, we introduce ***GaussianProperty****, a training-free framework that assigns physical properties to 3D Gaussians. Specifically, we integrate the segmentation capability of SAM with the recognition capability of GPT-4V(ision) to formulate a global-local physical property reasoning module for 2D images. Then we project the physical properties from multi-view 2D images to 3D Gaussians using a voting strategy. We demonstrate that 3D Gaussians with physical property annotations enable applications in physics-based dynamic simulation and robotic grasping. For physics-based dynamic simulation, we leverage the Material Point Method (MPM) for realistic dynamic simulation. For robot grasping, we develop a grasping force prediction strategy that estimates a safe force range required for object grasping based on the estimated physical properties. Extensive experiments on material segmentation, physics-based dynamic simulation, and robotic grasping validate the effectiveness of our proposed method, highlighting its crucial role in understanding physical properties from visual data. Online demo, code, and annotated datasets are available on the project page: https://Gaussian-Property.github.io*

1. Introduction

Estimating physical properties from visual data is a critical task in both computer vision and graphics, serving as the foundation for various fields, including augmented reality (AR) [2, 3, 15], robotic grasping [4, 7, 37], and physics-based dynamic simulation [8, 13, 14]. Recently, the integration of physical properties into 3D model has generated significant interest across these domains, underscoring the need for precise physical property estimation. However, this area remains under-explored due to the inherent ambiguities in physical property estimation. Key challenges include the difficulty of acquiring labeled ground-truth data, as intrinsic physical properties are not directly observable through visual means, and the ambiguity of the prediction task, which is further compounded by the limited number of observable surfaces.

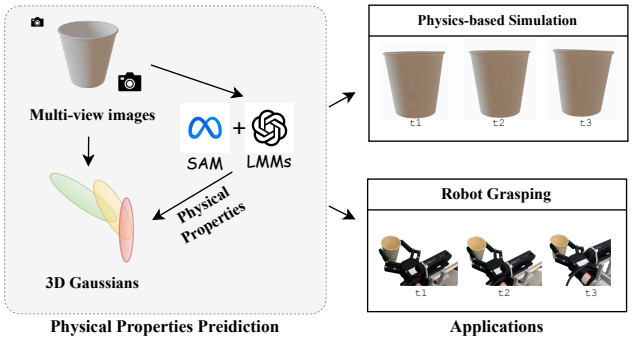

Figure 1. **GaussianProperty** is a training-free framework, aiming at adding physical properties to 3D Gaussians with the assistance of LMMs. By assigning physical properties to 3D Gaussians, it promotes several downstream tasks such as physical-based generative dynamics and robot grasping.

Prior studies in cognitive science [12, 26] reveal that humans infer physical properties by associating visual appearances with material experiences. Inspired by this mechanism, we propose a novel framework that mimics human reasoning through Large Multimodal Models (LMMs). This process allows us to intuitively gauge physical property such as weight, texture, and density from visual observation. Recently, Large Language Models (LLMs) have achieved impressive progress in nature language understanding. Based on this, Large Multimodal Models (LMMs) extend LLMs by further incorporating image modality into the model training. With a massive repository of prior knowledge, which covers the task of physical property estimation, showcasing a robust understanding and recognition

*Equal contribution.
†Corresponding author.

capabilities of visual data that mirrors human perception. We show an example in Figure 3.

In this study, we introduce a novel method called ***GaussianProperty***, designed to assign physical properties to 3D Gaussians using Segment Anything (SAM) and GPT-4V. Specifically, we leverage the recognition capabilities of GPT-4V to estimate physical properties from 2D images. However, predicting properties for complex scenes containing multiple components with distinct physical characteristics from a single global image presents significant challenges. To address this, we first use the robust segmentation capabilities of SAM [18] to segment each component within the global image. We then employ GPT-4V, incorporating both global and detailed local information from each segmented part and its spatial context, to achieve more precise physical property estimations.

After acquiring physical properties from 2D images, we project this information onto 3D Gaussians using a multi-view reconstruction approach and a voting strategy. The 3D Gaussians, representing an explicit 3D point cloud format, support effective reconstruction from multi-view images. To be specific, we first reconstruct the 3D Gaussian representation using multi-view images. We then project the spatial positions of the 3D Gaussians onto the visible 2D images to gather corresponding estimations. A voting strategy is subsequently employed to determine the final physical properties of the 3D Gaussians, effectively avoiding occasional errors that may occur in a single view.

We demonstrate that incorporating physical properties into 3D model enhances two downstream tasks: robotic grasping and physics-based dynamic simulation. For robotic grasping, we develop a grasping force prediction module. Based on the estimated physical properties, this module predicts the upper bound force to avoid object deformation and the lower bound force required to lift the object without slipping, ensuring proper grasping force estimation. We select some common objects from daily life to validate the effectiveness of the adaptively adjusted grasping force predicted by physical properties. We compare the grasping success ratio and deformation ratio with those obtained using a fixed force. For physical-based dynamic simulation, a key limitation of previous methods is the need for manual assignment of physical properties or estimation from videos, while our method can predict the physical properties of the object from 2D images, we enable more efficient and intuitive physics simulation.

To summarize, our contributions are listed as follows.

- We present a framework exploring the use of Large Multimodal Models (LMMs), e.g. GPT-4V for physical property estimation for 3D model, showing robust results in physical properties estimation.
- We demonstrate two crucial downstream tasks that benefits from estimated physical properties, i.e.,robotic grasping and physical-based dynamic simulation.
- Extensive experiments including materials segmentation, real-world grasping and realistic dynamic simulation validate the effectiveness of our proposed method, showing superior performance and benefiting downstream tasks.

2. Related Work

2.1. Physical property estimation

In the burgeoning field of 3D modeling, the accurate estimation of physical properties such as density, elasticity, and thermal conductivity is a long-standing problem [1, 44], serving critical roles in downstream tasks like AR, robotics, and physical-based simulation. Although promising, existing work [30, 40] mostly tackles specific types of material properties, e.g. mass, by collecting corresponding task-dependent data with little generalization. Several works have explored LLMs for physical property estimation. For example, NeRF2Physics [50] leverages large language models to propose candidate materials for objects, constructing a language-embedded point cloud to estimate physical properties such as mass, friction, and hardness through a zero-shot kernel regression approach. Make-it-real [11] reasons the PBR materials including albedo, metallic, and roughness for 3D assets texture generation. Another category of methods, such as GIC [6] and PAC-NeRF [21], relies on multi-view videos for physical property estimation, which cannot achieve single-frame inference of physical properties, thus limiting their practical applications. Our method can generate diverse physical properties like mass density, friction, and hardness in a zero-shot manner with the recognition capability of LLMs.

2.2. Multimodal Large Language Models

Large Language Models (LLMs) have made significant strides in natural language understanding but traditionally lack visual reasoning due to missing image priors. To address this, recent efforts focus on developing Large Multimodal Models (LMMs) that integrate image modalities. State-of-the-art models [5, 22, 28, 41, 42] have shown strong performance in tasks like image captioning [25], PBR material estimation [11], and 3D grounding [43]. Among them, GPT-4V [28] stands out with its powerful 2D understanding and broad open-world knowledge. Though not designed for 3D input, it has been effectively applied to 3D evaluation via GPTEval3D [45], yielding results aligned with human judgment. The continued progress in LMMs opens new opportunities for vision-language integration across domains.

2.3. Material-sensitive Robot Grasping

Soft robotic grippers [20, 39, 49] leverage the deformation and compliance properties of soft materials enabling grip-

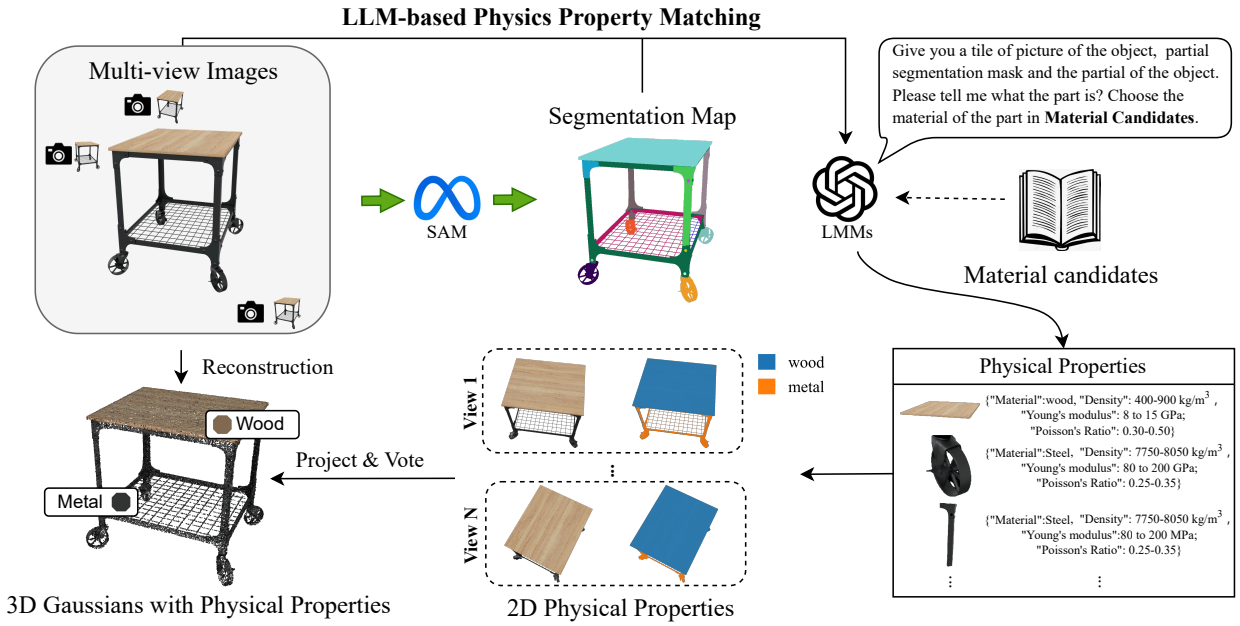

Figure 2. **Overall pipeline**. Our Gausssian-Property initially leverages SAM to get the segmentation map of the object. Then the original images and the masks are sent to the foundation models like GPT-4V(ision) to get the corresponding physical properties by inquiring the material candidates. After acquiring physical properties from 2D images, we using a multi-view approach and a voting strategy to add physical properties to the reconstruction 3D Gaussians.

pers to automatically adapt to the geometries and various weights of the objects being grasped. This adaptability necessitates the careful selection of materials and mechanical designs tailored to specific applications, limiting the generality of such solutions across all scenarios. Optical tactile sensing approaches [16, 19, 24] requires a camera positioned within each fingertip of a gripper, situated behind a soft and transparent artificial skin, to convert optical observations of markers printed on the skin to force estimations; while electronic skins [36, 38] detects exerted forces from electric signals. However, these two approaches often face challenges related to durability, and some of those methods require significant additional installation space, limiting their practicality in certain applications. In this work, we propose integrating *GaussianProperty* to enable material-sensitive robot grasping, which takes merely the visual inputs from a camera to predict the composing materials and estimate corresponding physical properties of the object to grasp. Our approach can be easily adapted to a wide spectrum of robotic and industrial applications.

2.4. Physics-based Dynamic Rendering

Neural Radiance Fields (NeRF) [27] have garnered significant interest in recent years due to their remarkable capabilities in multi-view 3D reconstruction. An evolutionary advancement within the NeRF framework is the incorporation of a temporal dimension, enhancing the representation of dynamic scenes, such as D-NeRF [31] and NeRFies [29]. To achieve physical law-compliant deformation in NeRF, methods [23, 32, 48] have integrated physics-based deformation into the NeRF framework. However, the effectiveness of these methods relies on using the exported mesh from NeRF. To overcome this constraint, Gaussian Splatting [17] enable direct kinematic modeling of forward displacements without intermediary geometry conversions. PhysGaussian [46] and methods [35, 51, 52] have achieved dynamic simulation by integrating Newtonian physics directly into 3D Gaussian representations, using the Material Point Method (MPM) to enable realistic physical interactions. However, a key limitation in these approaches is the need for manual assignment of physical properties to each Gaussian point like material type. This manual assignment is time-consuming and unrealistic. Our method can predict the physical properties of objects, by coupling with PhysGaussian, we enable more efficient physics simulation.

3. Method

3.1. Problem Formulation

Given a well-reconstructed 3D Gaussian representation, our objective is to attribute physical properties to each Gaussian. The specific physical property can vary according to the downstream task. In this work, we demonstrate a potential application in physics-based dynamic simulation

via Material Point Method (MPM) and robotic grasping. The former application requires material density ρ, Young's modulus E, Poisson's ratio P, and material type T. And robotic grasping requires the material density ρ, volume V, friction coefficient μ, thickness d, maximal tolerable curvature κ, Young's modulus E. An overview of our framework is illustrated in Figure 2.

3.2. Part-Level Segmentation

Understanding an object's physical properties requires delving into the characteristics of its individual parts, as each part may present unique attributes. Considering this, we utilize SAM for image segmentation, adeptly predicts masks with precise boundaries that capture whole, part, and subpart levels, thereby reflecting the object's hierarchical semantic structure. In this work, we emphasize the significance of part-level information, which enables us to dissect an object into its constituent parts. This approach facilitates a more accurate and exhaustive comprehension of the physical properties of visual data. Our method not only harnesses the semantic stratification provided by SAM but also actively integrates it to remedy the ambiguity arising from objects possessing multiple physical attributes.

Concretely, for each image I within the observed set $\mathcal{I}^N$, we input a grid of 32×32 point prompts. SAM responds by segmenting precise masks at varying levels based on the prompts at these points. We operate using the part-level semantic mask $\mathbf{M}$, subsequently refining the segmentation by eliminating superfluous masks within each of the three mask sets. This culling is informed by predicted intersection-over-union (IoU) scores, stability scores, and the overlap rates between masks. The resulting segmentation maps meticulously trace the boundaries of objects at their respective hierarchical levels, effectively segmenting the scene into semantically coherent regions. Please refer to **Supplementary** for the detailed system prompt.

3.3. Physics Property Matching

After achieving precise part-level semantic segmentation, the next step is to match the segmented parts with their corresponding physical properties, a process we term Physics Property Matching. We discuss the establishment of material candidates in Section 3.3.1 and utilizing a combination of global and local knowledge in Section 3.3.2 to assist GPT-4V in recognizing the material properties of the object. Additionally, we discuss the Gradual Prompt Guidance in Section 3.3.3 to help the model progressively build an understanding of the entire object and discern the association between its parts and the whole.

3.3.1. Material Candidates

Our approach leverages a curated collection of candidate materials, consisting of fifteen ubiquitous material families and more than 600 materials, integral to everyday objects and structures. This library encompasses a wide range of materials, ensuring comprehensive coverage of various densities and material properties. The common object material library includes density ranges for a variety of materials. For instance, metals such as aluminum (2700 kg/m³), steel (7750-8050 kg/m³), and copper (8920-8960 kg/m³) are covered, as well as non-metals like glass (2200-2500 kg/m³), concrete (2300-2500 kg/m³), and plastics such as polyethylene (930-970 kg/m³). This diversity highlights the extensive range of physical properties found in commonly used substances.

This robust material database is the cornerstone of our physical property matching process. By offering a comprehensive material library, the material candidates simplify material retrieval for the LLM model. Additionally, it minimizes ambiguity in property predictions from different perspectives, ensuring accuracy. Reliable material identification thus provides a dependable reference.

3.3.2. Combined Global-Local Reasoning Module

Our observation revealed that utilizing a global-to-local knowledge framework significantly improves the accuracy in assigning physical properties to each part. A straightforward method involves having the model understand the entire object first and then evaluate a part of the object. However, we found it challenging for the model to establish a connection between the whole and its parts, as shown in Figure 3 **(Left)**. Motivated by this insight, we built a bridge between global and local information, enabling the model to understand their connection. As shown in Figure 3 **(Right)**, the left image displays the original object, the middle image shows a partial segmentation with the mask highlighted in red, and the right image depicts a specific part of the object. Starting from this global perspective, GPT-4V then focuses on the details of each part, incorporating local cues such as texture, color, and contextual information from adjacent parts. This approach aids in accurately identifying each part and inferring its material composition.

3.3.3. Gradual Prompt Guidance

We design gradual prompt guidance to help the LMMs gradually build an understanding of the entire object and then discern the association between its parts and the whole through the segment map. The prompt instructs the LLM to first briefly describe the part based on the provided image and then identify the material of the part, specifying its mass density, Young's modulus, and Poisson's Ratio. The material types are selected from a predefined material candidates of common object. This structured approach ensures that the LMMs can effectively comprehend the context and specifics of each part, thereby enhancing its accuracy in identifying physical properties. The "Gradual Prompt Guidance" design thus provides a systematic method to improve the model's understanding and performance by leveraging

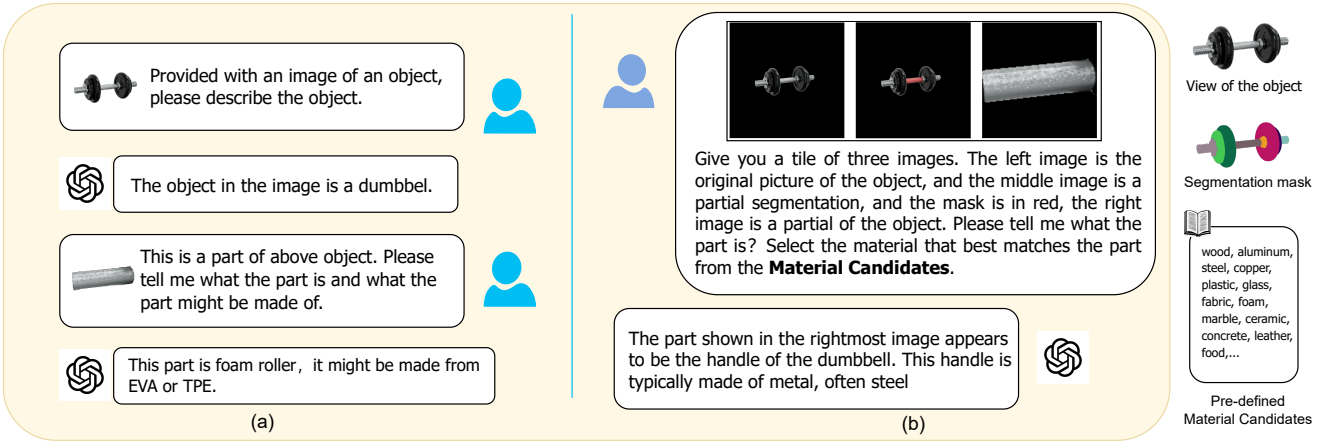

Figure 3. **Left**: GPT-4V(ision) struggles to recognize the material when directly provided with both global and partial image inputs. **Right**: Enhanced with combined global-local information and association, the agent accurately characterizes the component's properties.

both global and local information.

3.4. Lift 2D to 3D via Voting

3.4.1. 3D Reconstruction from Multi-view Images

3D Gaussian Splatting method has the advantage of providing an explicit 3D representation, making it easy to add any other properties. This method reparameterizes NeRF with a set of unstructured 3D Gaussian kernels $\{x_p, \sigma_p, A_p, C_p\}_{p\in P}$, where x_p, σ_p, A_p, and C_p denote the centers, opacities, covariance matrices, and spherical harmonic coefficients of the Gaussians, respectively. A differentiable rasterization rendering method is employed to project 3D Gaussians to 2D images to compare the rendered image with ground-truth image by

$$C = \sum_{k\in P} \alpha_k \mathrm{SH}(d_k; C_k) \prod_{j=1}^{k-1} (1-\alpha_j) \tag{1}$$

where α_k are the z-depth ordered opacity, and d_k is the view direction from the camera to x_k.

3.4.2. Frequency-based Voting Strategy

Through reconstruction, we obtain 3D Gaussians denoted as GS. Previous works [33, 53] incorporate CLIP features into 3D Gaussians through training, but the process is time-consuming and scene-specific, limiting downstream applications. Alternatively, we lift the 2D information to 3D models with a projection-based method. Each 3D Gaussian $\mathbf{s} \in GS$ is projected to each 2D image $I \in \mathcal{I}^N$, we determine the pixel coordinates (u, v) on 2D plane using the camera parameters. The projection is performed as

$$u, v = \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{s} \tag{2}$$

where $\mathbf{K}$ is the camera intrinsic matrix, $[\mathbf{R}|\mathbf{t}]$ represents the rotation and translation matrices (extrinsic parameters), and $\mathbf{s}$ is the coordinates of the point.

However, the projected pixel coordinates are meaningless if the point is invisible in the source image. Thus, We estimate the visibility using the Gaussian-estimated depth to determine if the point is visible of the image. The voting strategy involves projecting each Gaussian to all the visible views and retrieving the corresponding properties. To ensure consistency across multi-view images, we adopt a frequency-based voting strategy. The attribute with the highest frequency is chosen as the final predicted attribute. The voting process can be described as:

$$\hat{a} = \arg\max_{a} \sum_{i=1}^{N} \mathbb{I}(a_i = a) \tag{3}$$

where $\hat{a}$ is the predicted attribute, N is the number of views, a_i is the attribute observed in the i-th view, and $\mathbb{I}$ is the indicator function that equals 1 if the attribute matches and 0 otherwise.

3.5. Material-sensitive Robot Grasping

The diversity of objects in the real world, composed of various materials and physical properties, makes it impractical to use a single grasping force for all. An adaptive strategy is essential to calibrate the grasping force according to the specific materials of the object being manipulated. The grasping force applied by the robotic gripper must be sufficient to lift the target object without slipping while remaining below a threshold to prevent damage or deformation. These two criteria effectively define the lower bound $F_{\min}$ and the

upper bound $F_{\max}$ of the grasping force F.

$$
\begin{aligned}
&F_{\min} \leq F \leq F_{\max}, \\
&F_{\min} = \sum_{i=1}^{M} \frac{1}{2} \rho(i) V(i) g \left(\frac{\cos\theta}{\mu(s)} - \sin\theta \right), \\
&F_{\max} = \min \left[A\sigma_y(s), \frac{1}{2} AE(s) d(s) \kappa_{\max}(s) \right],
\end{aligned} \tag{4}
$$

where the object consists of M parts, and even physical property distribution of material is assumed within each part; $s \in \{1, \cdots, M\}$ refers to the object part containing the force bearing surface; $\rho(\cdot)$ and $V(\cdot)$ are respectively the density and the volume of a part; θ is the lifting angle of the gripper; $\mu(\cdot)$ is the friction coefficient between the gripper tips and a surface; A is the area of a force bearing surface; $d(\cdot)$ is the thickness of a surface, $\kappa_{\max}(\cdot)$ is maximal tolerable curvature of a surface; $E(\cdot)$ is Young's modulus of electricity of the material of a part; and $g \approx 9.8m/s^2$ is the gravity constant. Specific values of ρ, μ and E relate directly to the predicted material of each part, while those of V and d can be estimated from object reconstruction, and A is approximated with the area of the gripper finger tips.

To maximize the grasping reliability, confining the grasping force within the robotic gripper capability, and attempting to avoid the gripper executing commands close to its input bounds with $0 \leq \eta \leq 1$ margin, an optimal choice of grasping force

$$
F^* = \begin{cases} \left[\bar{F}\right]_{[F_{\min}]_G + \eta\Delta F}^{[F_{\max}]_G - \eta\Delta F} & F_{\min} < F_{\max}, \\ \left[\bar{F}\right]_G & F_{\min} \geq F_{\max}, \end{cases} \tag{5}
$$

with $[\cdot]_G$ and $[\cdot]_{\min}^{\max}$ clipping a force within the input range of robotic gripper G and between some lower and upper bounds. $\Delta F = \max\left[0, [F_{\max}]_G - [F_{\min}]_G\right]$. And F^* remains optimality in extreme situations where $F_{\min} > F_{\max}$. See the **Supplementary** for detailed derivation.

3.6. Physics-based Dynamic Simulation

A key limitation in current approaches like Physgaussian [46] is the need for manual assignment of physical properties to each Gaussian point, which is time-consuming and unrealistic. To address this inefficiency, our method can directly predicts the physical properties of each Gaussian point, thus eliminating the need for manual assignment. Specifically, we employ a combination of multi-view 2D-to-3D projection and frequency-based voting to derive these properties from observed images. For each Gaussian point in the 3D representation, our model predicts essential physical attributes, including density (ρ), Young's modulus (E), Poisson's ratio (P, among others. This prediction process begins with segmenting observed images at the part level to ensure each segment's unique physical characteristics are represented accurately. We then apply a voting strategy to integrate physical properties across multiple views, ensuring consistency and robustness in the 3D representation. By automating the assignment of these properties through *GaussianProperty*, we significantly reduce the time required for dynamic simulations, streamline the simulation workflow, and enable scalable applications in complex environments. We show some cases in Figure 6.

4. Experiments

4.1. Datasets and Evaluation Protocol

Datasets. We evaluated the quantitative and qualitative performance using both synthetic and real-captured data from the Amazon Berkeley Objects (ABO) dataset [9] and the MVImgNet dataset [47]. Following [50], we selected 100 validation objects from the ABO dataset. For MVImgNet, we also selected 210 objects. The criterion for selection was to ensure coverage of a diverse range of material categories, and we filtered out cases that could not be accurately classified. Finally, we manually annotated detailed material labels for each part of the objects. This process resulted in a final set of 78 labeled cases in the ABO dataset and 100 cases in MVImgNet. Moreover, we also captured 16 objects composed of various materials for robotic grasping. Further details can be found in the Supplementary.

Evaluation protocol. To evaluate the accuracy of material prediction after adding physical properties to 3D Gaussians. We select an angle from which the object can be better observed. The 3D Gaussians render the material information into 2D to form a material segmentation map of the angle. Similar to 2D evaluations, we use mean Intersection over Union (mIoU) metric [10] as an indicator to assess the accuracy of the material segmentation. For robotic grasping, the Picked-up Rate (PUR) and the No-damage Rate (NDR) evaluate respectively whether objects are picked up without slipping and whether no damages to objects are caused. A final success requires both criteria being met, yielding a final Success Rate (SR).

4.2. Implementation Details

For each object, we collected 30 views with camera centers randomly distributed over a hemisphere around the object. We implemented the officially released 3D Gaussian Splatting [17] for reconstruction, following the default parameter settings. To accelerate the part-level segmentation and property matching process, we selected only 10 views. The total training iteration is 30,000, which costs 5/6 minutes on NVIDIA A6000/A40. For multi-modal model processing, we used GPT-4V as the large multimodal model. Here we conducted all our experiments using GPT-4V API.

For dynamic simulation, we implemented Physgaussian [46] with assigning estimated materials for each 3D

Table 1. Comparison of material segmentation with NeRF2Physics [50] across different categories on ABO and MVImgNet dataset. Our method achieves a more comprehensive and accurate understanding of the object and achieve more precise material segmentation. Both methods utilize the same numbers of ten candidate materials. Due to the limited availability of objects composed of selected materials in the ABO dataset, we present results for the five most prevalent material categories.

Method	ABO dataset						MVImgNet										
	Wood	Metal	Plastic	Fabric	Ceramic	Average	Wood	Metal	Plastic	Glass	Fabric	Foam	Food	Ceramic	Paper	Leather	Average
NeRF2Phycics	27.87	13.01	8.38	40.26	38.44	25.59	6.39	3.63	6.70	1.15	1.11	0.38	2.40	6.54	6.73	5.20	4.02
Ours	**61.53**	**33.41**	**38.26**	**67.57**	**78.40**	**55.83**	**41.96**	**38.85**	**39.50**	**18.87**	**27.12**	**23.18**	**84.89**	**19.74**	**30.23**	**23.96**	**34.83**

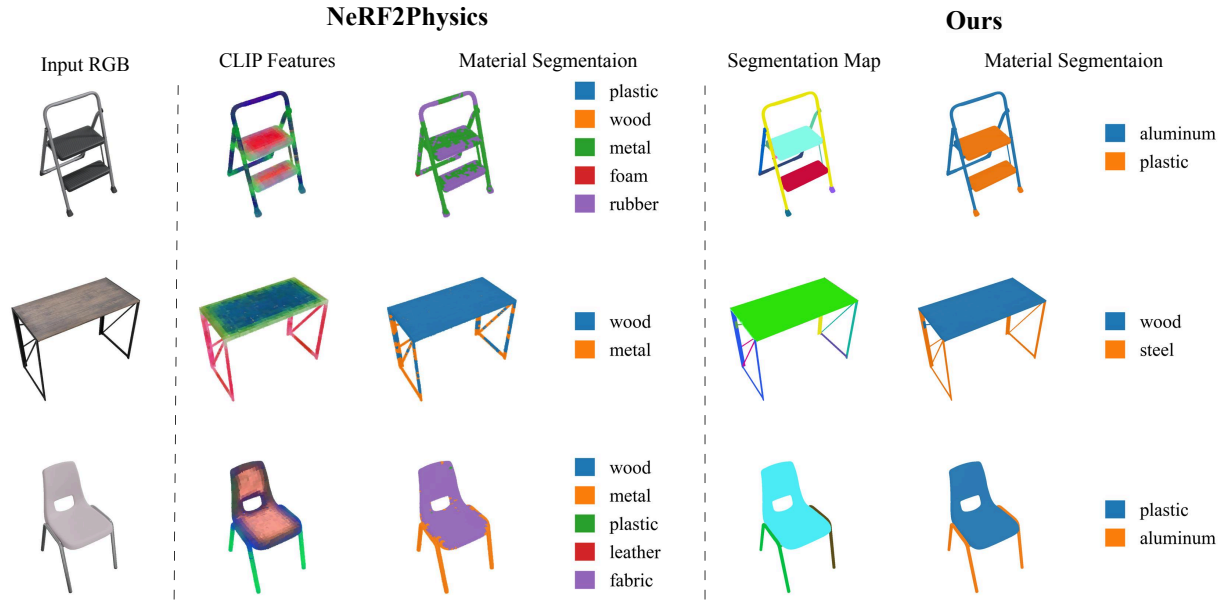

Figure 4. **Qualitative results of Material Segmentation**. Our model makes boundary-accurate physical material predictions.

Gaussian. In robot grasping experiments, we utilized a Jacobi.ai JSR-1 robot platform [34] equipped with a TEK CTAG2F90-C robotic gripper that has a maximum grasping force up to $40N$. The force-bearing surface at the tip of the gripper is measured to encompass an area of $A = 0.00011\text{m}^2$. And a maximum allowable bending curvature $\kappa_{\max} = 0.5$ is used. The robotic gripper's grasping force has been calibrated with its normalized input $15 \leq N_{GF} \leq 100$ before experiment.

4.3. Material Segmentation.

We compared material segmentation performance with the recent work Nerf2Physics [50], we present both qualitative and quantitative comparisons in Figure 4 and Table 1. Our method significantly outperforms Nerf2physcis on both synthetic and real-captured data. We also conducted mass and hardness estimation as Nerf2Physics. More results can be found in the **Supplementary**.

Table 2. Results of robot grasping experiments on 16 objects. MinGF, MidGF and MaxGF are baselines with minimum ($N_{GF} = 15$), medium ($N_{GF} = 60$) and maximum ($N_{GF} = 100$) grasping forces applied by the robotic gripper. **Bold**: best results.

Method	Picked Up (%↑)	No Damage (%↑)	Success (%↑)
MinGF	50.00	**100.00**	50.00
MidGF	81.25	81.25	62.50
MaxGF	**93.75**	75.00	68.75
Ours*	**93.75**	**100.00**	**93.75**

4.4. Applications

4.4.1. Robot Grasping

To evaluate the effectiveness and performance of our proposed method, we collect 16 objects composed of diverse materials, and implemented three robot grasping baselines with fixed grasping forces, which are widely adopted force-sensitive grasping strategies in robotics. Table 2 shows our

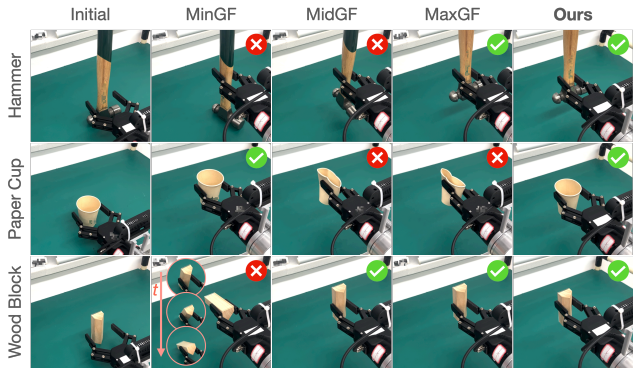

Figure 5. **Robot Grasping** is a downstream application of *GaussianProperty*. Several sample cases from robot grasping experiments are presented, where we compare our proposed method (right) against three baselines (middle columns), starting from initial configurations (left).

method on material-sensitive grasping with *GaussianProperty* outperforms all the baselines. The failure case observed using our method is when the robot tried to lift a *dumbbell* that is too heavy requiring a grasping force larger than the mechanical limit of our robot. Several sample cases are shown in Figure 5. Full object list and experiment results can be found in the **Supplementary**.

4.4.2. Generative Dynamics

Physical simulation is a crucial application of our method because it allows us to directly add all predicted physical properties to the Gaussian points without the need for manual querying and annotation. This integration speeds up dynamic rendering significantly. Figure 6 illustrates some examples showing that the physical properties predicted by our approach can be directly applied in simulation.

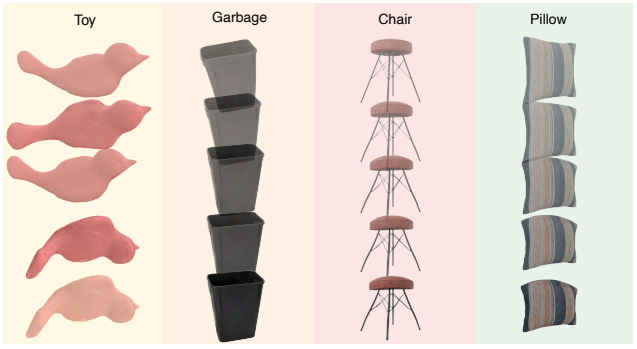

Figure 6. **Generative Dynamics**. We present a potential downstream task of 3D Gaussians with physical property, i.e., the generative dynamics. By imposing force, the 3D Gaussians generate corresponding motion. For example, in the first row, we applied a top-down force, the bird exhibited a movement corresponding to the applied force.

Table 3. Ablation study of Global-to-Local Knowledge Integration and Frequency-Based Voting.

Global-to-local	Voting	Average mIoU (%↑)
	✓	22.17
✓		51.28
✓	✓	**55.83**

4.5. Ablation Study

Global-to-Local Knowledge Utilization. Table 3 demonstrates the impact of incorporating global-to-local knowledge in material segmentation. Without this module, the method only utilizes images of each individual local part of the object for material querying. In contrast, with global-to-local knowledge, the method benefits from a broader context, enabling it to more accurately segment and classify materials. This approach enhances the understanding of the object's overall structure and finer details, leading to more precise predictions of materials.

Frequency-based Voting Strategy. Table 3 demonstrates that implementing a frequency-based voting strategy can improve the accuracy of property estimation. By projecting onto multi-view images, we can identify the most frequently occurring material for each part. This frequency-based approach ensures consistency and reliability in the predicted properties by effectively aggregating information from different viewpoints, minimizing errors, and enhancing overall prediction accuracy. We provide an example to demonstrate the design in the **Supplementary**.

5. Conclusion and Limitation

Limitation Despite the promising result of our method on 2D material segmentation, our method struggles to distinguish surface with ambiguous materials. We show an example in the Supplementary.

Conclusion In this paper, we explore the issue of estimating physical properties for 3D models. The inherent ambiguity and the challenge of acquiring labeled ground-truth data can significantly hinder the estimation. Our method, *GaussianProperty*, effectively addresses this challenge by leveraging the recognition capability of large multimodality models and segmentation capability of SAM to achieve a combined global-local reasoning module on 2D space. Then, a voting strategy is employed to project the 2D material property estimation results to 3D Gaussians, a effective and efficient 3D representation, supporting multi-view reconstruction and real-time rendering. We show two potential downstream applications, i.e., physics-based dynamic simulation and robotic grasping. Extensive experiments on manually annotated material segmentation dataset and real-world robot grasping experiments validate the effectiveness.

References

[1] Edward H Adelson. On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*, pages 1–12. SPIE, 2001. 2

[2] Syed Shah Alam, Samiha Susmit, Chieh-Yu Lin, Mohammad Masukujjaman, and Yi-Hui Ho. Factors affecting augmented reality adoption in the retail industry. *Journal of Open Innovation: Technology, Market, and Complexity*, 2021. 1

[3] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 1997. 1

[4] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings*, 2000. 1

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020. 2

[6] Junhao Cai, Yuji Yang, Weihao Yuan, Yisheng He, Zilong Dong, Liefeng Bo, Hui Cheng, and Qifeng Chen. Gic: Gaussian-informed continuum for physical property identification and simulation. *arXiv preprint arXiv:2406.14927*, 2024. 2

[7] Shehan Caldera, Alexander Rassau, and Douglas Chai. Review of deep learning methods in robotic grasp detection. *Multimodal Technologies and Interaction*, 2018. 1

[8] Eric C Carlson. Don't gamble with physical properties for simulations. *Chemical engineering progress*, 1996. 1

[9] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022. 6

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 6

[11] Ye Fang, Zeyi Sun, Tong Wu, Jiaqi Wang, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. Make-it-real: Unleashing large multimodal model's ability for painting 3d objects with realistic materials. *arXiv preprint arXiv:2404.16829*, 2024. 2

[12] Roland W Fleming. Visual perception of materials and their properties. *Vision research*, 94:62–75, 2014. 1

[13] Pierre Goovaerts. Estimation or simulation of soil properties? an optimization problem with conflicting criteria. *Geoderma*, 2000. 1

[14] Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, and Frédo Durand. Difftaichi: Differentiable programming for physical simulation. *arXiv preprint arXiv:1910.00935*, 2019. 1

[15] Nicolas Imbert, Frederic Vignat, Charlee Kaewrat, and Poonpong Boonbrahm. Adding physical properties to 3d models in augmented reality for realistic interactions experiments. *Procedia Computer Science*, 2013. 1

[16] Chengpeng Jiang, Zhang Zhang, Jing Pan, Yancheng Wang, Lei Zhang, and Liming Tong. Finger-skin-inspired flexible optical sensor for force sensing and slip detection in robotic grasping. *Advanced materials technologies*, 6(10):2100285, 2021. 3

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 3, 6

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[19] Nathan F Lepora. Soft biomimetic optical tactile sensing with the tactip: A review. *IEEE Sensors Journal*, 21(19): 21131–21143, 2021. 3

[20] Shuguang Li, John J Stampfli, Helen J Xu, Elian Malkin, Evelin Villegas Diaz, Daniela Rus, and Robert J Wood. A vacuum-driven origami "magic-ball" soft gripper. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7401–7408. IEEE, 2019. 2

[21] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. *arXiv preprint arXiv:2303.05512*, 2023. 2

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[23] Ruiyang Liu, Jinxu Xiang, Bowen Zhao, Ran Zhang, Jingyi Yu, and Changxi Zheng. Neural impostor: Editing neural radiance fields with explicit shape manipulation. *arXiv preprint arXiv:2310.05391*, 2023. 3

[24] Sandra Q Liu and Edward H Adelson. Gelsight fin ray: Incorporating tactile sensing into a soft compliant robotic gripper. In *2022 IEEE 5th International Conference on Soft Robotics (RoboSoft)*, pages 925–931. IEEE, 2022. 3

[25] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[26] Stephen McAdams. Recognition of sound sources and events. *Thinking in sound: The cognitive psychology of human audition*, 1993. 1

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 3

[28] OpenAI. Gpt-4v(ision) system card, 2023. https://openai.com/research/gpt-4v-vision. 2

[29] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[30] Divya Patel, Amar Nath, and Rajdeep Niyogi. Adding material embedding to the image2mass problem. In *International*

Conference on Computational Science and Its Applications, pages 77–90. Springer, 2022. 2

[31] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 3

[32] Yi-Ling Qiao, Alexander Gao, and Ming Lin. Neuphysics: Editable neural geometry and physics from monocular videos. *Advances in Neural Information Processing Systems*, 35:12841–12854, 2022. 3

[33] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. *arXiv preprint arXiv:2312.16084*, 2023. 5

[34] Dicong Qiu, Wenzong Ma, Zhenfu Pan, Hui Xiong, and Junwei Liang. Open-vocabulary mobile manipulation in unseen dynamic environments with 3d semantic maps, 2024. 7

[35] Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Feature splatting: Language-driven physics-based scene synthesis and editing. *arXiv preprint arXiv:2404.01223*, 2024. 3

[36] Ye Qiu, Shenshen Sun, Xueer Wang, Kuanqiang Shi, Zhiqiang Wang, Xiaolong Ma, Wenan Zhang, Guanjun Bao, Ye Tian, Zheng Zhang, et al. Nondestructive identification of softness via bioinspired multisensory electronic skins integrated on a robotic hand. *npj Flexible Electronics*, 6(1):45, 2022. 3

[37] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 2008. 1

[38] Benjamin Shih, Dylan Shah, Jinxing Li, Thomas G Thuruthel, Yong-Lae Park, Fumiya Iida, Zhenan Bao, Rebecca Kramer-Bottiglio, and Michael T Tolley. Electronic skins and machine learning for intelligent soft robots. *Science Robotics*, 5(41):eaaz9239, 2020. 3

[39] Jun Shintake, Vito Cacucciolo, Dario Floreano, and Herbert Shea. Soft robotic grippers. *Advanced materials*, 30(29): 1707035, 2018. 2

[40] Trevor Standley, Ozan Sener, Dawn Chen, and Silvio Savarese. image2mass: Estimating the mass of an object from its image. In *Conference on Robot Learning*, pages 324–333. PMLR, 2017. 2

[41] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[43] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023. 2

[44] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in neural information processing systems*, 28, 2015. 2

[45] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation, 2024. 2

[46] Tianyi Xie, Zeshun Zong, Yuxin Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023. 3, 6

[47] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 6

[48] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022. 3

[49] Shadab Zaidi, Martina Maselli, Cecilia Laschi, and Matteo Cianchetti. Actuation technologies for soft robot grippers and manipulators: A review. *Current Robotics Reports*, 2 (3):355–369, 2021. 2

[50] 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. *arXiv preprint arXiv:2404.04242*, 2024. 2, 6, 7

[51] Mingtong Zhang, Kaifeng Zhang, and Yunzhu Li. Dynamic 3d gaussian tracking for graph-based neural dynamics modeling. *arXiv preprint arXiv:2410.18912*, 2024. 3

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

[53] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. *arXiv preprint arXiv:2312.03203*, 2023. 5