

Sim Anything: Automated 3D Physical Simulation of Open-world Scene with Gaussian Splatting

Haoyu Zhao^{* 1} Hao Wang^{* 2} Xingyue Zhao^{* 4} Hongqiu Wang⁵ Zhiyu Wu⁶
 Chengjiang Long^{†3} Hua Zou^{†1}

¹School of Computer Science, Wuhan University

²Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

³Meta Reality Lab ⁴School of Software Engineering, Xi'an Jiao Tong University

⁵The Department of Systems Hub, Hong Kong University of Science and Technology (Guangzhou)

⁶School of Computer Science, Fudan University



Figure 1. We develop an efficient method for simulating the dynamic movements of 3D objects with customizable behaviors, and synthesizing interactive 3D dynamics under arbitrary forces (red arrows). Compared to recent methods [14, 22, 45], our approach produces more realistic 3D dynamics with much faster inference times.

Abstract

Recent advancements in 3D generation models have opened new possibilities for simulating dynamic 3D object movements and customizing behaviors, yet creating this content remains challenging. Current methods often require manual assignment of precise physical properties for simulations or rely on video generation models to predict them, which is computationally intensive. In this paper, we rethink the usage of multi-modal large language model (MLLM) in physics-based simulation, and present **Sim Anything**, a physics-based approach that endows static 3D objects with interactive dynamics. We begin with detailed scene reconstruction and object-level 3D open-vocabulary segmentation, progressing to multi-view image in-painting. Inspired by human visual reasoning, we propose MLLM-based Physical Property Perception (MLLM-P3) to predict mean physical properties of objects in a zero-shot manner. Based on the mean values and the object's geometry, the Material Property Distribution Prediction model (MPDP) model then estimates the full distribution, reformulating

the problem as probability distribution estimation to reduce computational costs. Finally, we simulate objects in an open-world scene with particles sampled via the Physical-Geometric Adaptive Sampling (PGAS) strategy, efficiently capturing complex deformations and significantly reducing computational costs. Extensive experiments and user studies demonstrate our Sim Anything achieves more realistic motion than state-of-the-art methods within 2 minutes on a single GPU. Our project page is at <https://sim-gs.github.io/>.

1. Introduction

With the development in 3D representation, Neural Radiance Fields (NeRF) [25] and 3D Gaussian Splatting (3DGS) [17] offer new perspectives for 3D reconstruction and 3D representation [36, 37]. However, these approaches are unable to simulate interactions with 3D assets in simulation environments [33, 39], which is critical for generating realistic object responses to novel interactions, such as external forces or agent manipulations in many applications, e.g., virtual reality [16], embodied intelligence [24].

Some recent approaches aim to bridge the gap between

^{*} Equal contributions.

[†]Corresponding Author.

rendering and simulation integrating physics-based priors into 3D object representations using physical simulators [4, 7, 28]. For instance, PAC-NeRF [21] estimates the geometry and physical parameters of objects from multi-view videos and then integrates physical models with NeRF-based representations. Similarly, PhysGaussian [40] first injects physical parameters into 3DGS objects, and then predicts motion using a physics-based simulator. However, their ability to handle real objects is limited, as they require a predefined material model with manually assigned parameters or rely on multi-view videos to predict the physical properties of each objects.

To automatically set parameters, some approaches [14, 22, 45] leverage video generation models [2] which are trained on real-world video data to estimate physical material parameters. For example, PhysDreamer [45] employs stable video diffusion model to learn Young’s modulus of objects. However, learning material physical properties from video diffusion priors is computationally expensive and time-consuming in practice. Moreover, video diffusion models have limited controllability and often fail to obey physical laws [31, 50]. Additionally, these models are also generally restricted to non-rigid objects, making them unsuitable for deriving the physical properties of large rigid objects (such as cup, bowl, and chairs). However, humans are remarkably adept at predicting physical properties of objects based on visual information [8, 9]. We therefore ask this question: how can we develop models for perceiving physics from just visual data?

To this end, we rethink physics-based simulation and the usage of multi-modal large language model (MLLM), such as GPT-4V [41]. In this paper, we propose **Sim Anything**, a novel physics-based method that transforms static 3D objects into interactive ones capable of responding to new interactions, as shown in Fig. 1. We first segment objects in an open-word scene with priors from foundation models [19, 23, 46]. Inspired by how humans predict physical properties of objects through visual data, our Sim Anything leverages MLLM-based Physical Property Perception (MLLM-P3) to predict the mean values of physical properties. Unlike previous methods [14, 22, 45] iteratively refining each physical properties through video analysis, we reformulate this problem from a regression task to a probability distribution estimation task by predicting the full range of these properties based on the mean value and the object’s geometry, reducing computational demands. Finally, our approach simulates object interactions in an open-world scene with driving particles sampled by the Physical-Geometric Adaptive Sampling (PGAS) strategy, enabling a seamless integration of realistic physics with adaptable sampling precision. Extensive experiments and user studies demonstrate that *Sim Anything achieves more accurate physical property prediction and synthesizes more realistic motion with much faster inference time*.

	automatic parameter computation	fast inference time	physics-based deformation	static 3D object input	scene-wide physical simulation
PhysGaussian [40]	✗	✓	✓	✓	✗
DreamGaussian4D [31]	✗	✗	✗	✓	✗
Animate124 [50]	✗	✗	✗	✓	✗
PAC-NeRF [21]	✗	✓	✓	✗	✗
PIE-NeRF [7]	✗	✓	✓	✗	✗
Spring-Gaus [51]	✗	✓	✓	✗	✗
PhysDreamer [45]	✓	✗	✓	✓	✗
DreamPhysics [14]	✓	✗	✓	✓	✗
PhysDreamer [45]	✓	✗	✓	✓	✗
Physics3D [22]	✓	✗	✓	✓	✗
Ours	✓	✓	✓	✓	✓

Table 1. **Comparison to Concurrent Works.** Note that our method is the only one capable of simulating the entire scene at a much faster speed.

tic motion with much faster inference time. We provide an overview of the comparison to major prior works in Tab. 1. In summary, our work makes the following contributions:

- Sim Anything is the first to use MLLM for zero-shot physical property estimation of objects in 3D scenes.
- We reformulate physical property estimation as a probability distribution task, enabling adaptable physical simulations with PGAS in open-world scenes.
- Experiments show Sim Anything effectively predict physical properties and creates realistic 3D dynamics.

2. Related Work

2.1. Dynamic 3D Animation

The demand for dynamic 3D animation creation has grown significantly across various applications, including video games, virtual reality, and robotic simulation [11, 47–49]. With the success of video generative models, some methods [50] have attempted to leverage video diffusion models to guide the prediction of 3D deformations. For instance, DreamGaussian4D [31] uses pre-generated videos to supervise the deformation of static scenes. However, the deformations produced by these methods may not always be accurate or physically plausible.

Recent works [26, 51] introduce physics simulation to the 3D deformation and enable synthesizing motions under any physical interactions. Virtual Elastic Objects [4] jointly reconstructs the geometry, appearances, and physical

parameters of elastic objects with multi-view data. Spring-Gaus [51] integrate a 3D Spring-Mass model into 3D Gaussian kernels, and then simulate elastic objects from videos of the object from multiple viewpoints. PAC-NeRF [21] and PhysGaussian [40] integrate physics-based simulations with NeRF [25] and 3DGS [17], respectively, to model the deformation of elastic objects. However, these methods either require manual setup of physical properties for 3D objects before simulation or depend on multi-view videos to predict physical properties.

To avoid manually setting parameters, some works estimate physical material parameters with video generation model [2] to estimate physical material parameters [14]. PhysDreamer [45] and DreamPhysics [14] leverage video generation models to estimate physical material parameter (e.g., Young’s modulus), while Physics3D [22] further optimizes a wider range of physical parameters for 3D objects. However, these methods are computationally expensive, as learning material properties through video diffusion priors is time-consuming. Moreover, the controllability of generated videos is limited, often deviating from physical laws [31, 50], which we further demonstrate in the experimental Section 5.4. Additionally, these models are typically restricted to non-rigid objects, making them unsuitable for determining the physical properties of large rigid objects, such as tables, chairs, and sofas. Inspired by how humans perceive physical properties of the objects [8, 9], we propose leveraging multi-modal large language models (MLLMs) to zero-shot predict the mean values of physical properties for objects in a 3D scene, enabling faster inference times. We then use the proposed MPDP model to predict the full distribution of these properties.

2.2. Visual Physics Perception

Physics perception is a long-standing challenging problem [38]. Previous studies demonstrate that deep learning models can potentially exhibit physical perception abilities similar to humans [1, 12]. Most prior research focuses on dynamically addressing object properties, either by observing the target’s behavior [21] or by interacting with it in a 3D physical engine [27, 42]. Other works also explore the estimation of material properties directly from static images [1, 34]. However, these works mostly focus on specific material properties, such as mass or tenderness, often relying on task-specific data. In contrast, we propose leveraging MLLM, such as GPT-4V [41], to generate a wide range of physical properties such as mass, Young’s modulus, and Poisson’s ratio in a zero-shot manner.

3. Preliminaries

3.1. Material Point Method

The Material Point Method (MPM) [13] is a popular simulation framework for multi-physics phenomena due to its capability to handle topology changes and frictional interactions. Unlike mesh-based methods, MPM represents the continuum using particles in a grid-based space, making it well-suited for point-based 3D Gaussian representation. Following PhysGaussian [40], we define each Gaussian kernel’s time-dependent state as:

$$x_i(t) = \Delta(x_i, t), \quad \Sigma_i(t) = F_i(t)\Sigma_iF_i(t)^T, \quad (1)$$

where $\Delta(\cdot, t)$ and $F_i(t)$ denote coordinate deformation and deformation gradient at timestep t . The viewpoint must also adjust with the continuum rotation $\Omega_i(t)$ to match the view direction of the spherical harmonic coefficient C_i .

3.2. 3D Gaussian Splatting (3DGS)

3D Gaussian Splatting (3DGS) represents scenes as point clouds, with each point modeled as a 3D Gaussian defined by a center point \mathcal{X} (mean) and a covariance matrix Σ . Each Gaussian at \mathcal{X} is given by $G(\mathcal{X}) = e^{-\frac{1}{2}\mathcal{X}^T\Sigma^{-1}\mathcal{X}}$. Σ is decomposed into a scaling matrix \mathcal{S} and rotation matrix \mathcal{R} , such that $\Sigma = \mathcal{R}\mathcal{S}\mathcal{R}^T$, with \mathcal{S} and \mathcal{R} stored as vectors $s \in \mathbb{R}^{N \times 3}$ and $r \in \mathbb{R}^{N \times 4}$, respectively. Differential splatting [44] applies a viewing transform W and Jacobian J to compute the transformed covariance $\Sigma' = JW\Sigma W^T J^T$, enabling novel view rendering. Each pixel color \mathcal{C} is obtained by blending N overlapping points:

$$\mathcal{C} = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (2)$$

where c_i and α_i denote color and opacity, derived from the Gaussian with covariance Σ and optimized parameters.

4. Method

Predicting various physical properties of 3D objects from static scene is an extremely challenging task due to limited supervisions. Instead of capturing physical data from generation models or multi-view videos [7, 14, 21, 22, 45, 51], we reformulate this task from a new perspective, decomposing it into a set of sub-tasks. Specifically, as shown in Fig. 2, we first segment the images with a set of foundation models [19, 23, 46] and lift these 2D segmented masks to segment 3D object in the scene via radiance fields rendering (Section 4.1). We propose MLLM-based Physical Property Perception (MLLM-P3) to predict the mean values of these properties (Section 4.2). We then use the Material Property Distribution Prediction (MPDP) model to estimate

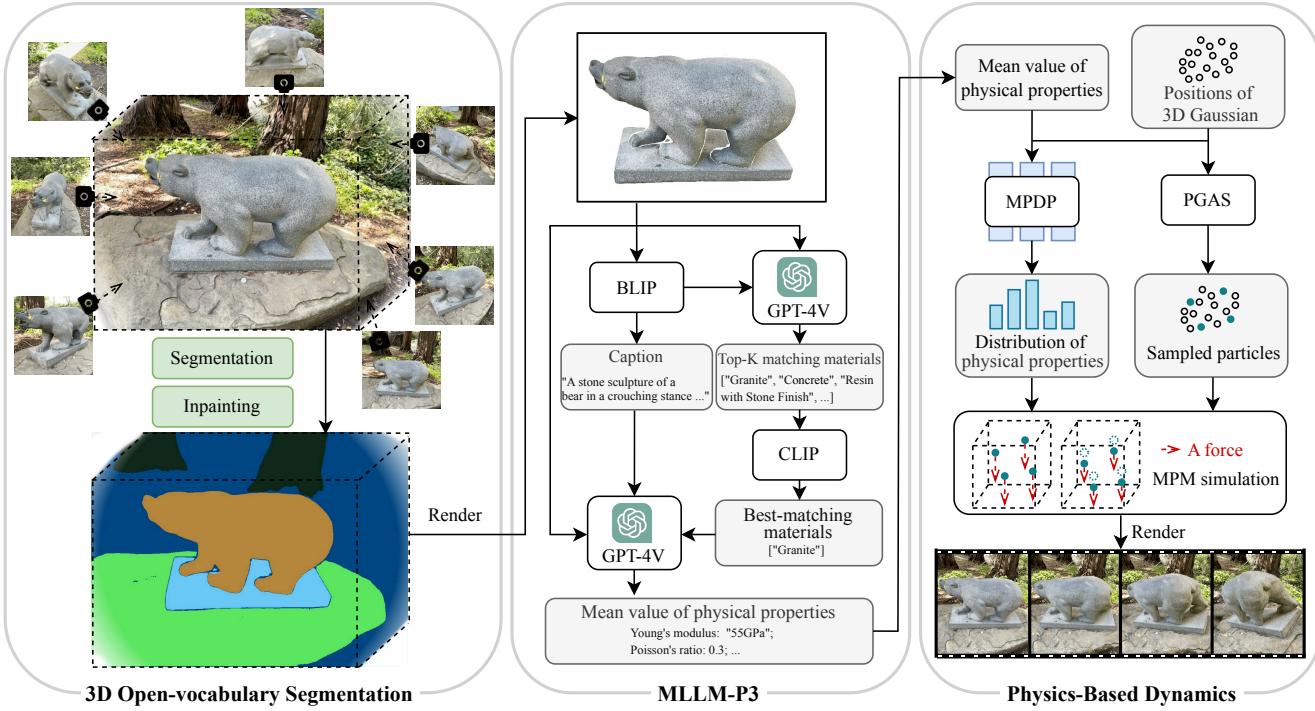


Figure 2. Overview of Sim Anything. Given a pre-trained 3D scene and its corresponding 2D images, we first perform object-level segmentation of the 3D scene with the prior from a set of foundation models [19, 23, 46]. We obtain the mean physical properties of the object from the proposed MLLM-P3, and based on this and the object’s geometry, we then derive the full distribution using the MPDP model. Finally, we animate the 3D objects using a physics-based simulator with driving particles sampled via the Physical-Geometric Adaptive Sampling (PGAS) strategy.

the full distribution, simulating object dynamics with driving particles sampled using the Physical-Geometric Adaptive Sampling (PGAS) strategy (Section 4.3).

4.1. 3D Open-vocabulary Segmentation

For each scene, we first train a 3DGS model on given images and camera poses. Inspired by prior work [32], we integrate 2D open-vocabulary models like Grounding DINO [23] for detection, SAM [19] for segmentation, and RAM [46] for tagging. These models automatically segment objects in images without textual input. Specifically, we use RAM to tag the image, Grounding DINO to create bounding boxes based on tags, and SAM to refine these boxes into precise masks. This approach enables full automatic image labeling using expert models.

After 2D open-vocabulary segmentation, each segmented image contains semantic features for each object. We project these 2D masks into 3D space using radiance field rendering. Inspired by recent work [43, 48], each Gaussian retains its original attributes, with a learnable semantic attribute added for encoding object semantics. Using a zero-shot tracker [5], we assign unique IDs to masks across views, helping distinguish categories within the 3D scene

through differentiable rendering (see Fig.2). Extracting objects from 3DGS introduces holes, which we inpaint using LaMa [35] to guide 3D Gaussian inpainting, keeping Gaussians outside holes fixed.

4.2. MLLM-based Physical Property Perception

The variety of materials in the world is vast and hard to define, with many appearing identical and indistinguishable by local appearance alone. However, humans can infer material composition by combining high-level reasoning about object semantics with low-level visual cues. Recent research [6] has shown that multi-modal large language models (MLLM) excels in logical reasoning and decision-making for complex tasks. Inspired by how humans perceive and reason about physical properties of the objects they encounter, we propose MLLM-based Physical Property Perception (MLLM-P3) leveraging MLLM for open-vocabulary semantic reasoning about materials and their physical properties.

The segmented 3D scene in Section 4.1 is usually tightly related to the physical properties of the 3D objects in it. We first select a canonical view and render an object in 3D scene based on the 3D Gaussian’s semantic attribute

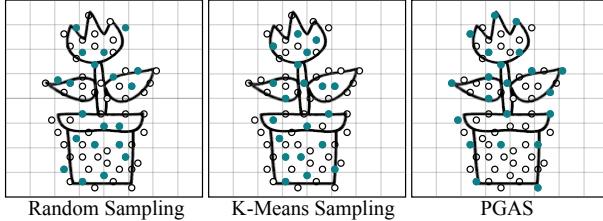


Figure 3. **Sampling.** We design a novel Physical-Geometric Adaptive Sampling (PGAS) strategy that captures the boundary of the object well. We employ PGAS to sample some “driving particles” (in green) and simulate only these particles. For rendering, each particle’s position and rotation are derived by fitting a local rigid body transformation based on neighboring driving particles.

introduce Section 4.1. Then we use a VQA model, such as BLIP [20] to produce a text description of the image. This description, along with the image, are then passed to a Multi-modal Large Language Model (MLLM) such as GPT-4V [41], prompting it to return a dictionary containing K candidate materials and information on whether the object is rigid (related to the sampling method in Section 4.3). we compute the CLIP [30] similarity score between the image and the materials in the dictionary to select the most matching material name. Finally, we prompt the MLLM with the selected material name, image, and text description to return a list of physical properties for the object, $M = \rho, E, \nu$, where ρ is the density, E is Young’s modulus, and ν is Poisson’s ratio.

Although it is theoretically possible for MLLM to propose the materials directly from the image, we find decomposing the task into two parts produces more reliable results in our experiments. We will further demonstrate this in the experimental Section 5.5.

4.3. Physics-Based Dynamics

Material Property Distribution Prediction. Even for object composed of a single material, local physical properties exhibit inherent variations across different regions of the object [3]. Additionally, the physical properties estimated by multi-modal large language model (MLLM) may not capture the 3D structure of the object. To address these challenges, we propose material property distribution prediction (MPDP), and reformulate the problem from a regression task to a probability distribution estimation task.

Specifically, we train a network \mathcal{D}_θ on part of synthesized dataset, using the object’s point cloud and predicted mean values (Section 4.2) as input, and supervised by the physical properties of all particles predicted by Physics3D [22]. The remaining synthesized data is reserved for comparison in later experiments. The network is designed to predict the geometry-aware probability distribution \mathcal{P} of physical properties across particles:

$$\mathcal{P} = \mathcal{D}_\theta(\mathcal{X}), \quad (3)$$

where \mathcal{X} is the position of 3D Gaussians of the object. We then scale the distribution \mathcal{P} by a global mean value predicted by the MLLM in Section 4.2 through element-wise multiplication, yielding the final physical property values for each point in the material field. This approach efficiently estimates per-point physical attributes, such as Young’s modulus and Poisson’s ratio, across the entire point cloud while avoiding the computational overhead of per-particle calculations.

Simulation with Physical-Geometric Adaptive Sampling. Rendering high-fidelity 3D scene often needs millions of 3D Gaussians, which is significant computational demands for simulation. To reduce this burden, we implement a sub-sampling approach. Specifically, we design a Physical-Geometric Adaptive Sampling (PGAS) strategy. The original Poisson disk sampling requires that the distance between any two particles be larger than a threshold r . Starting from an initial point, PDS then tries to fill a banded ring between r and $2r$ with new samples.

Our observation is that softer objects and those with complex shapes require more driving particles to accurately simulate their dynamics. To this end, we adaptively adjust the sample radius r based on the object’s Young’s modulus E predicted in Section 4.2 and curvature K . The curvature K is defined as:

$$C = \frac{1}{n} \sum_{j=1}^n (\mathcal{X}_j - \bar{\mathcal{X}})(\mathcal{X}_j - \bar{\mathcal{X}})^T, \quad (4)$$

$$K = \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3}, \quad (5)$$

where \mathcal{X}_j is the position of the j -th 3D Gaussian of the object, $\bar{\mathcal{X}}$ is the mean position of all 3D Gaussians, and $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of the covariance matrix C . Then, the sample radius r is adjusted as:

$$\hat{K} = \min(V_{max}, \max(V_{min}, K)), \quad (6)$$

$$\hat{r} = \min(r, k \sqrt{\frac{E}{\hat{K}}} r), \quad (7)$$

where we set $V_{max} = 10$, $V_{min} = 1$, and $k = \sqrt{10}$ in our paper. Our sampling ensures that the distance between a particle and its nearest neighbor is at least \hat{r} . By using smaller radii for softer materials and high-curvature areas, PGAS captures fine details more accurately, enhancing model resolution in deformation simulations and complex surface reconstruction, as shown in Fig. 3.

MPM-Driven Physics-Based Dynamics. To model physical properties, we employ MLS-MPM [13] as our simulator. In MPM, a continuum is represented by particles distributed in a grid-based space, offering a distinct advantage

over mesh-based methods. MPM can be seamlessly applied to 3D Gaussian Splatting (3DGS) in point-based representations. Building on PhysGaussian [40], we define a time-dependent state for each Gaussian kernel as follows:

$$x_i(t) = \Delta(x_i, t), \quad \Sigma_i(t) = F_i(t)\Sigma_i F_i(t)^T, \quad (8)$$

where $\Delta(\cdot, t)$ and $F_i(t)$ represent the coordinate deformation and deformation gradient at time t . Additionally, to account for continuum rotation $\Omega_i(t)$, the rendering viewpoint is adjusted to align with the view direction of the spherical harmonic coefficient C_i .

5. Experiments

5.1. Implementation Details

We initiate the process by reconstructing 3D Gaussians from multi-view images and execute internal particle filling operations to refine the representation further. Each Gaussian kernel is then associated with a set of physical properties targeted for optimization following [40, 45]. We then discretize the foreground region into a grid structure, typically sized at 64^3 . For the MPM simulation, we use 768 sub-steps per interval between video frames, resulting in a sub-step duration of 4.34×10^{-5} seconds to ensure precision and accuracy in simulation dynamics. All experiments are conducted on a single NVIDIA 3090 GPU. For more implementation details, please refer to the Supp.Mat.

5.2. Datasets

Open-world dataset. To evaluate the physical simulation accuracy in open-world 3D scenes, we chose some 3D scenes from LERF [18] and Instruct-NeRF2NeRF [10].

PhysDreamer [45]. We also conduct experiments on the physical simulation of single objects on four real-world static scenes from PhysDreamer [45] for fair comparison. Each scene includes an object and a background.

Synthesized dataset [22]. Following [22], we utilize BlenderNeRF [29] to synthesize several scenes. Five cases are used to train the proposed MPDP model (as introduced in Section 4.3), while the remaining four cases are reserved for subsequent comparisons.

5.3. Evaluation Metrics

We mainly focus on the motion realism and aesthetic quality of the synthesized 3D object motion in this task. To evaluate these aspects, we conduct a user study in which 42 participants rate each video based on motion realism. Additionally, we assess the aesthetic quality, particularly the naturalness of the videos, using the LAION aesthetic predictor, following [15]. For further details about the user study, please refer to the Supp.Mat.

Table 2. **Quantitative comparisons on PhysDreamer [45].** RS (Realism Score) represents the realism ratings assigned by participants in the user study, while AS (Aesthetic Score) is predicted using the LAION aesthetic predictor. Time represents the inference time required for physics-based 4D generation on an RTX 4090. The best and the second best results are denoted by pink and yellow.

Method:	RS	AS	Time
PhysGaussian [40]	4.50	7.56	-
PhysDreamer [45]	4.54	7.71	-
DreamGaussian4D [31]	4.57	7.28	0.1h
Sim Anything	4.66	7.89	2min

Table 3. **Quantitative comparisons on synthesized dataset [22].**

Method:	RS	AS	Time
PhysGaussian [40]	4.94	7.35	-
DreamPhysics [14]	5.05	7.92	1.5h
Physics3D [22]	5.10	8.01	1.5h
DreamGaussian4D [31]	4.98	6.81	0.1h
Sim Anything	5.10	8.20	2min

5.4. Comparison with SOTA Methods

We chose the performance from real-world static scenes from PhysGaussian [40] for fair comparison. We extensively compare our method with three the most recent methods: PhysGaussian [40], DreamGaussian4D [31], and PhysDreamer [45]. Tab. 2 presents the user study results (RS) and aesthetic score (AS) predicted by LAION aesthetic predictor following [15]. Since PhysDreamer [45] has not released its training code, we only compare the four evaluation scene and are unable to report its inference time. PhysDreamer [45] scores lower than DreamGaussian4D [31] in RS and PhysGaussian [40] in AS, which indicates that pre-generated videos may not be a proper ground truth for supervision. Our Sim Anything achieves better performance in both metrics, which demonstrates that Sim Anything generates videos that are both realistic and physically plausible, with a high degree of naturalness.

Following [45], we also compare the results with real captured videos in Fig. 4. We utilize space-time slices to present our comparisons, which depict time along the vertical axis and spatial slices of the object along the horizontal axis, as indicated by the red lines in the “object” column. Through these visualizations, we aim to elucidate the magnitude and frequencies of the oscillating motions under scrutiny. Sim Anything generates smooth and realistic motion patterns, accurately capturing the natural flow and details of real-world movements. Please see our project website videos for more video visualization.

We also evaluate our Sim Anything using the synthesized

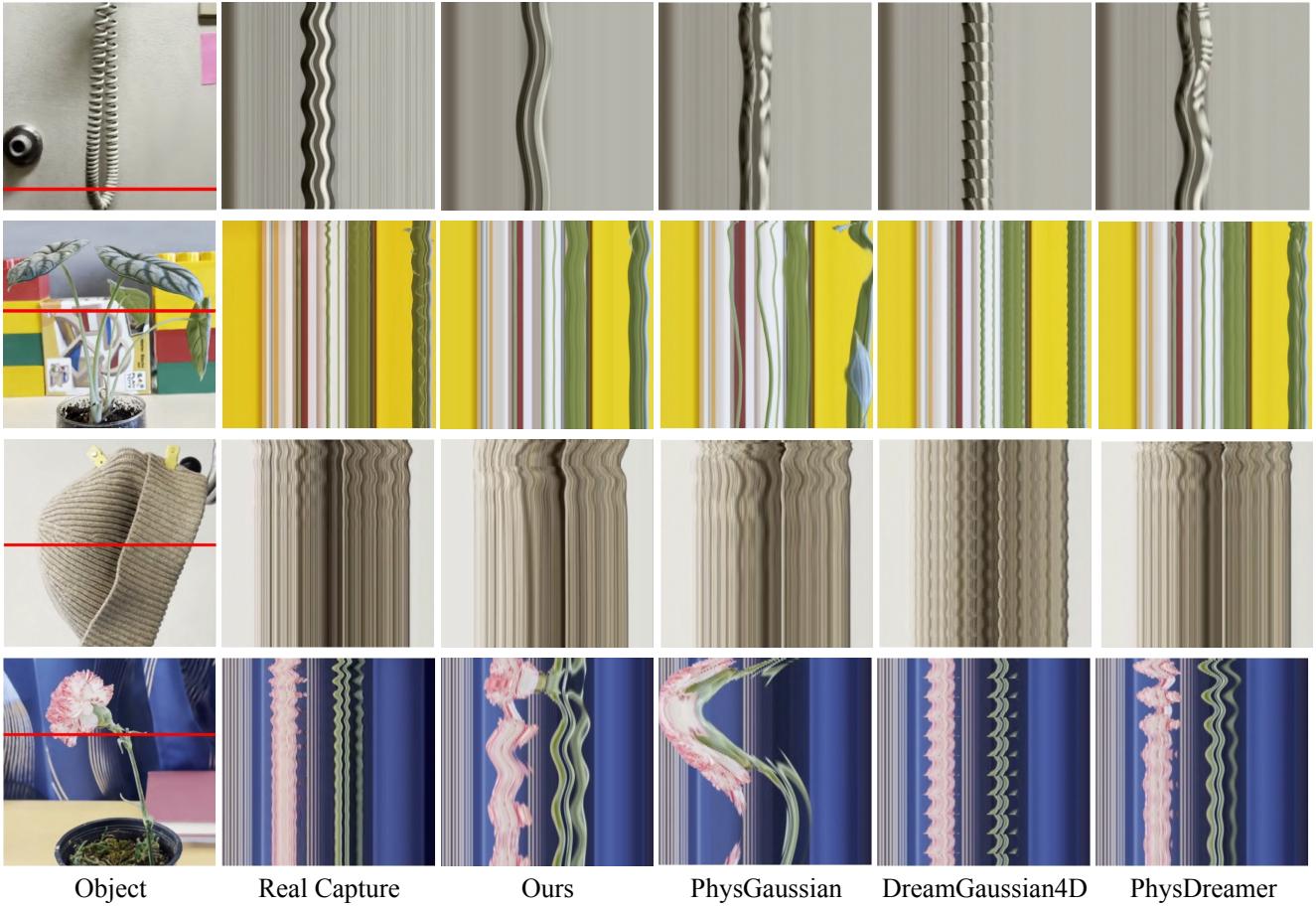


Figure 4. **Qualitative Comparison on PhysDreamer [45]**. We compare our results with real captured videos, and some recent SOTA methods [22, 31, 40, 45]. Our Sim Anything produces more realistic damping, closely matching real-world capture.

Table 4. **Ablation Study** on PhysDreamer [45] dataset. AS denotes the average aesthetic quality score predicted using the LAION aesthetic predictor.

GPT	BLIP	CLIP	w/o MPDP	PGAS	AS
✓			✓	✓	4.47
✓	✓		✓	✓	4.59
✓	✓	✓		✓	4.64
✓	✓	✓	✓		4.62
✓	✓	✓	✓	✓	4.66

dataset [22]. We report the quantitative results against recent methods [14, 22, 31, 40] in Tab. 3. Our method still generates the most consistent and natural motions. The visual results are shown in Fig. 5.

5.5. Ablation study

In this section, we conduct ablation experiments using PhysDreamer [45] dataset to evaluate the effectiveness of our proposed modules.

Model for physical property perception. In Fig. 6 we compare two methods with our proposed MLLM-P3: 1) GPT: Predicting physical properties using only the image; 2) GPT+BLIP: Predicting properties with both the image and a text description from BLIP; 3) GPT+BLIP+CLIP (MLLM-P3): Generating a dictionary of K candidate materials with GPT, selecting the best match via CLIP, and then predicting properties using the image, description, and chosen material. As shown in Tab. 4, MLLM-P3 performs best because there is inherent uncertainty in predicting materials based on just visual appearance or text description, as shown in Tab. 4. .

Material property distribution prediction. Material property distribution prediction is designed for complex physical properties distribution. As shown in Fig. 6 and Tab. 4, it is required to achieve optimal performance.

Sampling strategy selection. We also compare our proposed Physical-Geometric Adaptive Sampling (PGAS) strategy with K-Means sampling which is used in PhysDreamer [45]. The space-time slices of K-Means sampling

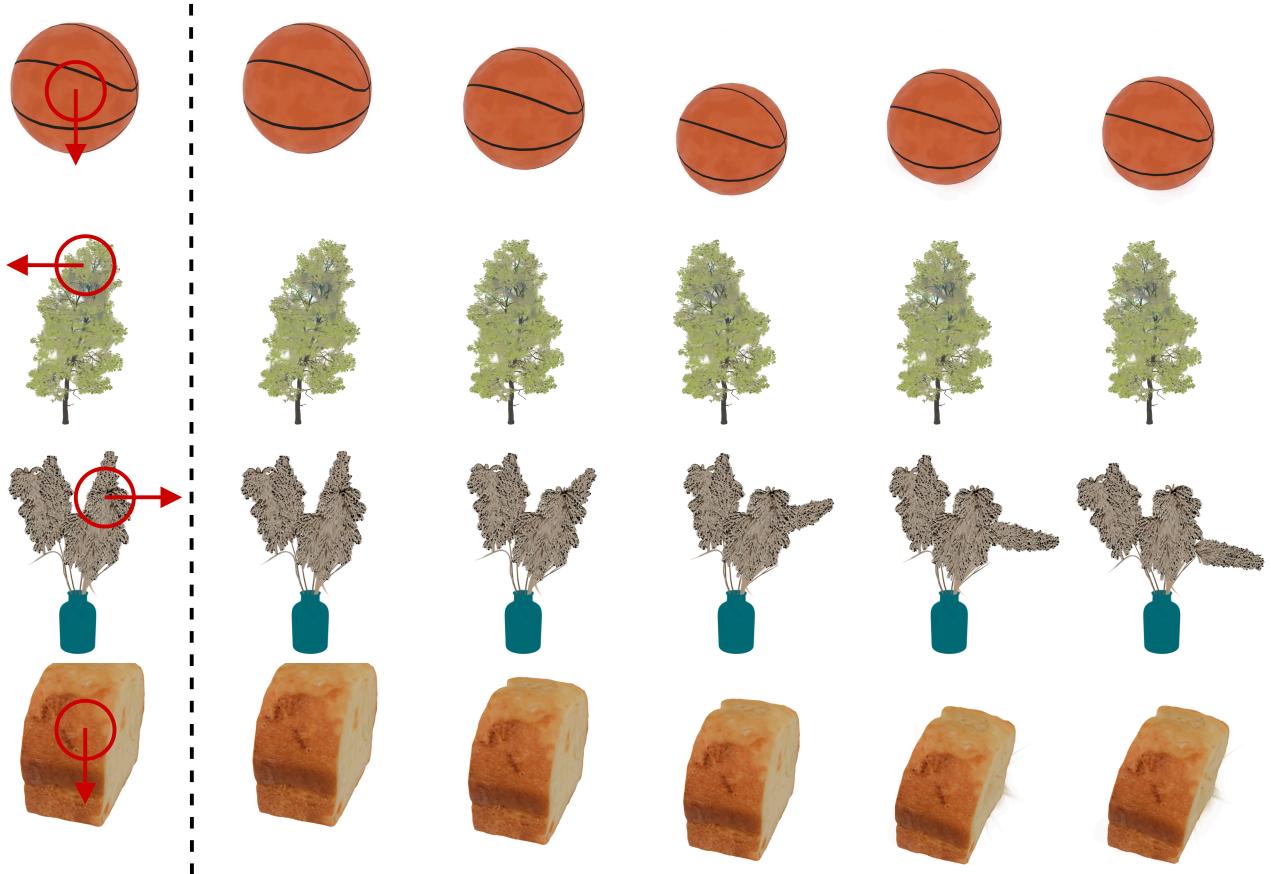


Figure 5. **Visual results on synthesized dataset** [22] with an external force (red arrows). Sim Anything is able to generate realistic scene movement while maintaining good motion consistency.

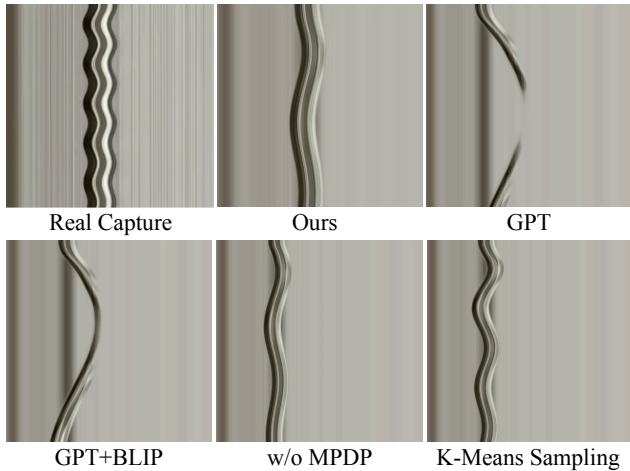


Figure 6. **Ablation study**. Visualization of space-time slices for ablation study on PhysDreamer [45]. Our method can generate closer content compared with the real capture.

is not quite consistent with the ground truth, while our final

method can produce 4D content that is competitive to real-captured videos, as shown in Fig. 6. The results in Tab.4 further confirm this, highlighting our method’s superiority in visual quality.

6. Conclusion

In this work, we introduce a framework, called Sim Anything, which generates physics-based dynamics and photo-realistic renderings. We begin with precise scene reconstruction and object-level 3D open-vocabulary segmentation, followed by multi-view image in-painting. Then, we propose MLLM-based Physical Property Perception (MLLM-P3) to predict mean physical properties of objects. Using these mean values and object geometry, the Material Property Distribution Prediction model (MPDP) then estimates the complete distribution, reframing the task as probability distribution estimation to reduce computational costs. Finally, we simulate objects in an open-world scene with particles sampled via the Physical-Geometric Adaptive Sampling (PGAS) strategy. Extensive experiments and user studies show that Sim Anything produces more realis-

tic motion than state-of-the-art methods within much faster inference time. We believe that Sim Anything represents a meaningful advance toward more engaging and immersive virtual environments, unlocking diverse applications from realistic simulations to interactive virtual experiences.

Limitation and future work. In complex environments with partially occluded objects, our Sim Anything is unable to segment the entire object, resulting in unnatural simulations, which is not efficient for more real applications. In the future, we aim to utilize generation model to reconstruct the occluded parts of these objects, which will takes a significant step to open up a wide range of applications from realistic simulations to interactive virtual experiences.

References

- [1] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3479–3487, 2015. 3
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3
- [3] P Ponte Castañeda and John R Willis. The effect of spatial distribution on the effective behavior of composite materials and cracked media. *Journal of the Mechanics and Physics of Solids*, 43(12):1919–1951, 1995. 5
- [4] Hsiao-yu Chen, Edith Tretschk, Tuur Stuyck, Petr Kadlecak, Ladislav Kavan, Etienne Vouga, and Christoph Lassner. Virtual elastic objects. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 15827–15837, 2022. 2
- [5] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with de-coupled video segmentation. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 1316–1326, 2023. 4
- [6] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palme: An embodied multimodal language model. In *Proc. of Intl. Conf. on Machine Learning*, 2023. 4
- [7] Yutao Feng, Yintong Shang, Xuan Li, Tianjia Shao, Chenfanfu Jiang, and Yin Yang. PIE-NeRF: Physics-based interactive elastodynamics with nerf. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4450–4461, 2024. 2, 3
- [8] Roland W Fleming. Visual perception of materials and their properties. *Vision research*, 94:62–75, 2014. 2, 3
- [9] Roland W Fleming, Christiane Wiebel, and Karl Gegenfurtner. Perceptual qualities and material classes. *Journal of vision*, 13(8):9–9, 2013. 2, 3
- [10] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 19740–19750, 2023. 6
- [11] Jennifer Healey, Wang, and et al. A mixed-reality system to promote child engagement in remote intergenerational storytelling. In *International Symposium on Mixed and Augmented Reality Adjunct*, pages 274–279, 2021. 2
- [12] Diane Hu, Liefeng Bo, and Xiaofeng Ren. Toward robust material recognition for everyday objects. In *Proc. of British Machine Vision Conference*, volume 2, page 6, 2011. 3
- [13] Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3, 5
- [14] Tianyu Huang, Yihan Zeng, Hui Li, Wangmeng Zuo, and Rynson WH Lau. DreamPhysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors. *arXiv preprint arXiv:2406.01476*, 2024. 1, 2, 3, 6, 7
- [15] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6
- [16] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, et al. VR-GS: a physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH*, pages 1–1, 2024. 1
- [17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 3
- [18] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LeRF: Language embedded radiance fields. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 19729–19739, 2023. 6
- [19] 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 *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 4015–4026, 2023. 2, 3, 4
- [20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. of Intl. Conf. on Machine Learning*, pages 12888–12900, 2022. 5
- [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. In *Proc. of International Conference on Learning Representations*, 2023. 2, 3
- [22] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3D: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024. 1, 2, 3, 5, 6, 7, 8
- [23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun

- Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 4
- [24] Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwén Lu, and Yansong Tang. ManiGaussian: Dynamic gaussian splatting for multi-task robotic manipulation. *arXiv preprint arXiv:2403.08321*, 2024. 1
- [25] 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*, 65(1):99–106, 2021. 1, 3
- [26] Vismay Modi, Nicholas Sharp, Or Perel, Shinjiro Sueda, and David IW Levin. Simplicits: Mesh-free, geometry-agnostic elastic simulation. *ACM Transactions on Graphics (TOG)*, 43(4):1–11, 2024. 2
- [27] Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. The curious robot: Learning visual representations via physical interactions. In *Proc. of European Conf. on Computer Vision*, pages 3–18, 2016. 3
- [28] 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. 2
- [29] Maxime Raafat. BlenderNeRF, May 2023. 6
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Intl. Conf. on Machine Learning*, pages 8748–8763, 2021. 5
- [31] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. DreamGaussian4D: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2, 3, 6, 7
- [32] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4
- [33] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 9339–9347, 2019. 1
- [34] Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H Adelson. Recognizing materials using perceptually inspired features. *Intl. Journal of Computer Vision*, 103:348–371, 2013. 3
- [35] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2149–2159, 2022. 4
- [36] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1
- [37] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Proc. of Advances in Neural Information Processing Systems*, 36, 2024. 1
- [38] 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. *Proc. of Advances in Neural Information Processing Systems*, 28, 2015. 3
- [39] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 1
- [40] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 2, 3, 6, 7
- [41] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of Imms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 2, 3, 5
- [42] Shaoxiong Yao and Kris Hauser. Estimating tactile models of heterogeneous deformable objects in real time. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 12583–12589, 2023. 3
- [43] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. *arXiv preprint arXiv:2312.00732*, 2023. 4
- [44] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Özireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38(6):1–14, 2019. 3
- [45] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. *arXiv preprint arXiv:2404.13026*, 2024. 1, 2, 3, 6, 7, 8
- [46] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 2, 3, 4
- [47] Haoyu Zhao, Hao Wang, Chen Yang, and Wei Shen. CHASE: 3d-consistent human avatars with sparse inputs via gaussian splatting and contrastive learning. *arXiv preprint arXiv:2408.09663*, 2024. 2
- [48] Haoyu Zhao, Chen Yang, Hao Wang, Xingyue Zhao, and Wei Shen. SG-GS: Photo-realistic animatable human avatars with semantically-guided gaussian splatting. *arXiv preprint arXiv:2408.09665*, 2024. 2, 4

- [49] Haoyu Zhao, Xingyue Zhao, Lingting Zhu, Weixi Zheng, and Yongchao Xu. HFGS: 4d gaussian splatting with emphasis on spatial and temporal high-frequency components for endoscopic scene reconstruction. *arXiv preprint arXiv:2405.17872*, 2024. [2](#)
- [50] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhen-guo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023. [2](#), [3](#)
- [51] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. *arXiv preprint arXiv:2403.09434*, 2024. [2](#), [3](#)