

PhyCAGE: Physically Plausible Compositional 3D Asset Generation from a Single Image

Han Yan^{1*} Mingrui Zhang² Yang Li² Chao Ma¹ Pan Ji²

¹ MoE Key Lab of Artificial, AI Institute, Shanghai Jiao Tong University

² Tencent XR Vision Labs

<https://wolfball.github.io/phycage>

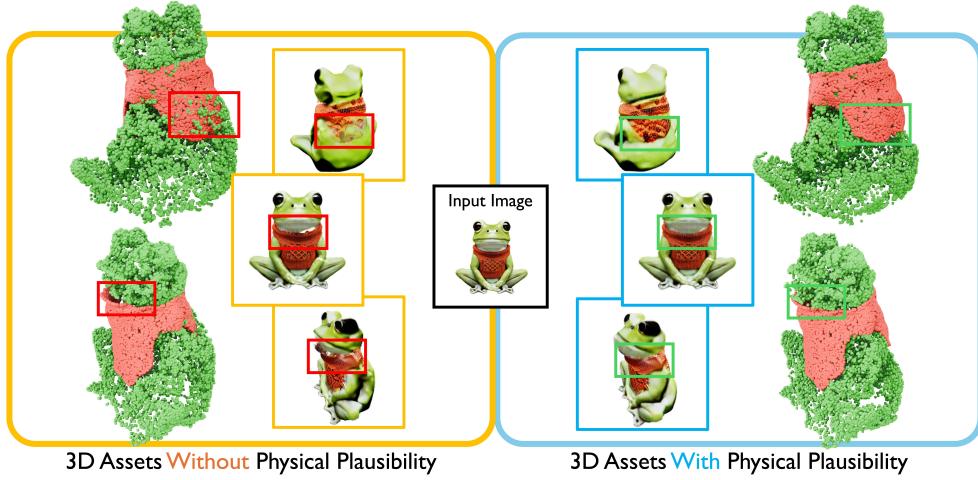


Figure 1. PhyCAGE can generate compositional 3D assets with interactive objects in a physically plausible manner. The generated 3D Gaussian Splatting shows better visual performance and physical plausibility under Material Point Method (MPM) simulation.

Abstract

We present **PhyCAGE**, the first approach for **Physically plausible Compositional 3D Asset Generation** from a single Image. Given an input image, we first generate consistent multi-view images for components of the assets. These images are then fitted with 3D Gaussian Splatting representations. To ensure that the Gaussians representing objects are physically compatible with each other, we introduce a **Physical Simulation-Enhanced Score Distillation Sampling (PSE-SDS)** technique to further optimize the positions of the Gaussians. It is achieved by setting the gradient of the SDS loss as the initial velocity of the physical simulation, allowing the simulator to act as a physics-guided optimizer that progressively corrects the Gaussians' positions to a physically compatible state. Experimental results demonstrate that the proposed method can generate physically plausible compositional 3D assets given a single image.

1. Introduction

Generating 3D shapes conditioned on 2D image input lies at the core of many applications, such as virtual reality (VR), augmented reality (AR), video gaming, and robotics. Recently, this field has seen remarkable progress, thanks to advancements in AI techniques, including transformers [40] and diffusion models [12].

While existing methods [13, 24, 36, 39, 43] mainly focus on the image-to-3D generation of a single object, this paper explores the more intricate challenge of generating compositional 3D assets: when presented with an image of an asset containing two compositional objects, our goal is to generate separate 3D representations of each component while ensuring that their relationships are semantically coherent and geometrically and physically plausible.

A simple strategy is to generate the entire assets as a holistic 3D mesh and subsequently use surface segmentation to separate the individual objects, as implemented in Part123 [21] and SAMPart3D [50]. However, mesh segmentation usually leads to incomplete surfaces and disre-

* Work done during internship at Tencent XR Vision Labs.

gards the relationships among objects. Alternative methods involve generating each component as an individual object and then combining them into a single model using estimated spatial placement, such as the similarity transformation that includes translation, rotation, and scaling. Examples of this approach can be found in [2, 6]. However, they struggle to manage complex spatial relationships that extend beyond simple similarity transformations. They fail in situations where non-rigid object deformation is required and often result in shape penetrations.

We observe that physical information, such as supporting relationships, stability, and affordance, can offer valuable clues for generating the shapes of interactive objects. For instance, objects in static scenes should exhibit stability. In a scene depicting “a frog wearing a sweater”, the frog should possess adequate body structure to support the sweater; otherwise, gravity will cause the sweater to fall off. To this end, we integrate differentiable physical simulations into the process of compositional 3D asset generation.

Specifically, given an input image, we generate consistent multi-view images for both the entire assets, a foreground component, and an inpainted occluded background. The multi-view images are subsequently fitted with 3D Gaussian Splatting [18] representations. Then, to ensure the physical plausibility of the assets, we build upon the Score Distillation Sampling (SDS) [29] method and introduce a physical simulation-enhanced SDS to further optimize the geometry (i.e., positions of Gaussians) for the objects. To ensure visual consistency with the input image, we incorporate image loss, i.e., the difference between the input image and rendered image from the generated object as a complement.

We observe that directly applying the SDS and image loss gradient to update Gaussians’ positions results in penetrations and non-physical artifacts. Our proposed physical simulation-enhanced SDS delegates updates of Gaussians’ positions to the physical simulation instead of the optimizer in the training process. By setting the loss gradient as the initial velocity of the physical simulation, the simulator serves as a physics-guided optimizer, which progressively corrects the particle positions by solving the physical system.

Experiments demonstrate the proposed method can generate physically plausible compositional 3D assets given a single image. Our contributions are as follows:

- We design a novel pipeline for image-based compositional 3D asset generation, particularly focusing on interactive objects with strong spatial coupling.
- We propose a physical simulation-enhanced Score Distillation Sampling to optimize 3D Gaussians in a physically plausible manner.
- We are the first to generate 3D compositional assets without penetration from a single image, facilitating down-

stream applications.

2. Related Work

2.1. Image conditioned 3D Generation

With the remarkable success of diffusion models [38] in the 2D domain [12, 35], numerous studies have started investigating how to build 3D generation models. One approach involves generating 3D assets by distilling knowledge from pre-trained 2D generators [29, 30]. DreamFusion [29] proposed Score distillation Sampling (SDS) to optimize a NeRF [26] model with images generated by a 2D generator. Meanwhile, Magic3D [30] employed a coarse-to-fine, two-stage strategy to enhance both the speed and quality of the generated models. The other technical solution involves directly training 3D generators using ground truth 3D data, and training denoising models to produce 3D shapes from image conditions. Notable works include Rodin [43], LAS [55], 3DShape2VecSet [51], and CLAY [52]. LRM [13] reformulated 3D generation as a deterministic 2D-to-3D reconstruction problem. Synthesizing multi-view consistent images enhances the capabilities of 3D generation or reconstruction, as shown in Zero123++ [36] and Syncdreamer [24].

The aforementioned approach generates 3D data in the form of a single, entangled representation, which is not ideal for numerous downstream applications that require semantically compositional shapes.

2.2. Compositional 3D Reconstruction and Generation.

ObjectSDF [45] and ObjectSDF++ [46] introduced an object-composition neural implicit representation, which allows separate reconstruction of each piece of furniture within a room, solely based on image inputs. DELTA [8] presented hybrid explicit-implicit 3D representations, designed for the joint reconstruction of compositional avatars. This includes the integration of components such as the face and body, or hair and clothing, respectively. Similar compositional avatars generation with a the SMPL [25] body prior works can be found in [5, 14, 41, 44]. AssetField [47] proposed to learn a set of object-aware ground feature planes to represent the scene and various manipulations could be performed to rearrange the objects. [4, 28] jointly optimized multiple NeRFs, each for a distinct object, over semantic parts defined by text prompts and bounding boxes. [6] and SceneWiz3D [53] eliminated the requirements for user-defined bounding boxes by simultaneously learning the layouts. Since the text could be problematically complicated when describing complex scenes, GraphDreamer [10] used scene graphs as input instead. Frankenstein [49] extended 3D diffusion approach for building a compositional scene generation tool.

In this paper, we adhere to the SDS-based methodology but incorporate physics simulation to address the inherent ambiguity of the 2D-to-3D problem and enhance the physical plausibility of the 3D assets.

2.3. Physics based 3D generation

Several attempts have been made to generate physically compatible objects. Aiming to generate physically compatible objects, [3] proposed an SDS-based method with rigid-body simulation, which can generate self-supporting objects from text. [11] presented a method of generating objects constrained by static equilibrium from a single image. In addition to object geometry generation from texts or images, there are existing works focusing on learning the objects' internal material parameters. In [54], an approach was proposed to distill dynamic priors from pre-trained video diffusion models by minimizing the discrepancy between physical simulation and diffusion-generated videos. [22] further utilized a more complex viscoelastic material model to simulate the objects and optimize the physical parameters via SDS. The above methods mainly focus on a single object, approaches are proposed for physically plausible scene reconstruction [27], language-grounded physics-based scene editing[31]. The existing methods above mainly focus on either single-object generation or rigid-body scene generation. In our work, we propose a novel approach for non-rigid compositional asset generation.

3. Preliminaries

3.1. Gaussian Splatting

3D Gaussian Splatting [18] (GS) has been proven efficient in 3D reconstruction tasks, due to its high inference speed and rendering quality.

Specifically, 3DGS represents 3D scenes as N Gaussians with attributes $G = \{\mu_i, \Sigma_i, q_i, \alpha_i, c_i\}_{i=1}^N$, where $\mu \in \mathbb{R}^3$ is the center, $\Sigma \in \mathbb{R}^3$ is the scaling factor, $q \in \mathbb{R}^4$ is the rotation quaternion, $\alpha \in \mathbb{R}$ is the opacity value, and $c \in \mathbb{R}^3$ is the color feature.

To render an image, all Gaussians are first projected onto an image plane. Then, volumetric rendering is performed for each pixel in front-to-back depth order to produce the alpha map A_{rd} and color map I_{rd} .

We use the following loss function to optimize the Gaussians:

$$\begin{aligned}\mathcal{L} &= (1 - \lambda_1)\mathcal{L}_1(I_{gt}, I_{rd}) + \lambda_1\mathcal{L}_{SSIM}(I_{gt}, I_{rd}) \quad (1) \\ &\quad + \lambda_2 A_{rd}(1 - A_{gt}), \quad (2)\end{aligned}$$

where I_{gt} and A_{gt} are ground-truth image and mask map, \mathcal{L}_1 is the L1 loss function, \mathcal{L}_{SSIM} is the structure similarity loss function, and $\lambda_{1,2}$ are the weighting factors.

Given a set of images $\{I_{gt,i}\}_{i=1}^M$, we can train 3DGS:

$$G = GaussianSplatting(\{I_{gt,i}\}_{i=1}^M), \quad (3)$$

where we eliminate the need for ground-truth mask maps since they can be extracted from images using background removal model ^{*}.

3.2. Physical Simulation

Continuum Mechanics. The motion of material is described by a mapping $\mathbf{x} = \phi(\mathbf{X}, t)$ from rest material space \mathbf{X} to a deformed space \mathbf{x} at time t . The Jacobian of the mapping $\mathbf{F} = \frac{\partial \phi}{\partial \mathbf{X}}(\mathbf{X}, t)$, i.e., deformation gradient measures the local rotation and strain [1]. Given the conservation of momentum and conservation of mass, the governing equations for describing the dynamics of an object are as follows:

$$\rho \frac{D\mathbf{v}}{Dt} = \nabla \cdot \boldsymbol{\sigma} + \mathbf{f}, \quad \frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{v} = 0, \quad (4)$$

where \mathbf{f} denotes an external force, $\boldsymbol{\sigma}$ is the internal stress, the \mathbf{v} and ρ denote the velocity and density respectively.

Material Point Method. The Material Point Method (MPM) is a framework for multi-physics simulation. It utilizes the strengths of both Eulerian grids and Lagrangian particles which enables it to simulate phenomena with large deformation, topology changes, and frictional contacts. It is widely adopted for the simulation of a broad range of materials such as elastic objects, snow, sand, and cloth [7, 15–17, 33]. Gaussian splatting provides a particle-based explicit 3D representation, which is naturally suitable for serving as the spatial discretization of objects in physical simulation. Following [48], we run MPM on these particles directly. The MPM pipeline consists of three stages in general: particle-to-grid(P2G), grid-operation and grid-to-particle(G2P). In the P2G stage, the MPM transfers mass and momentum from particles to grids:

$$m_i^n = \sum_p w_{ip}^n m_p \quad (5)$$

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

where p and i denote the Lagrangian particles and Eulerian grid respectively. The term w_{ip}^n denotes the B-spline basis function defined on the i -th grid, evaluated at the point \mathbf{x}_p^n . The particles carry properties including position \mathbf{x}_p^n , velocity \mathbf{v}_p^n , deformation gradient \mathbf{F}_p^n , local velocity gradient C_p^n and mass m_p at timestep t_n . The grids are updated after the P2G stage:

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^n - \frac{\Delta t}{m_i} \sum_p \tau_p^n \nabla w_{ip}^n V_p^0 + \Delta t g, \quad (7)$$

^{*}<https://github.com/OPHoperHPO/image-background-remove-tool>

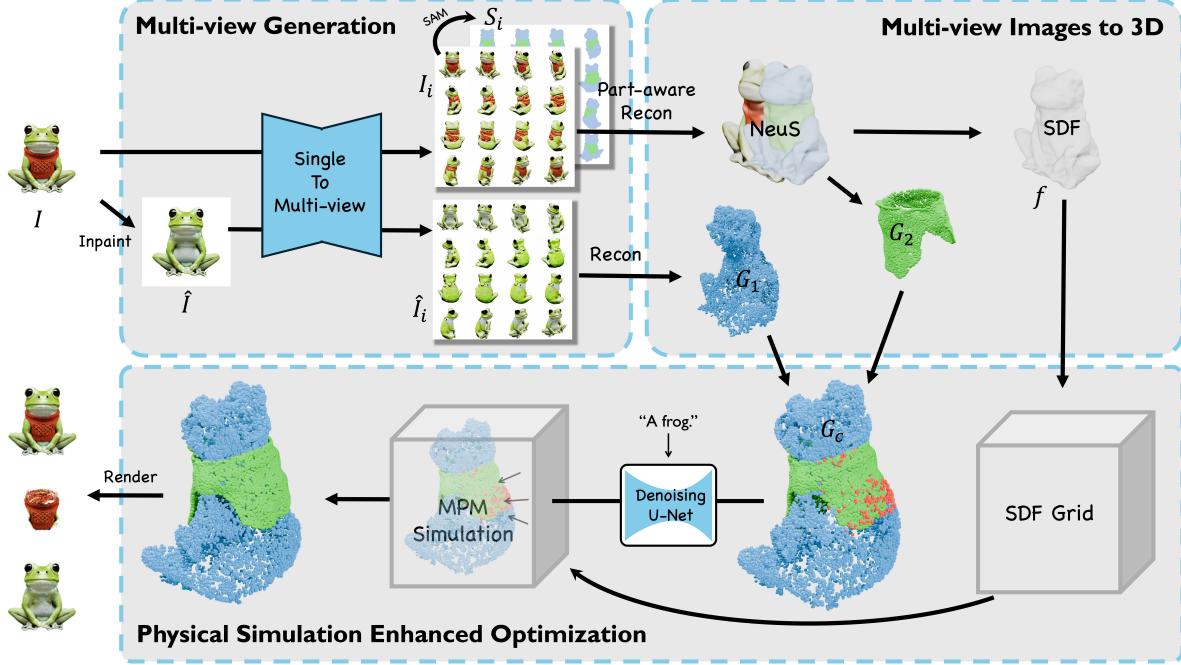


Figure 2. **The overview of PhyCAGE.** Given an input image, we first generate consistent multi-view images for the components of the assets (see Sec. 4.1). Then, we fit multi-view images with 3D Gaussian Splatting representations (see Sec. 4.2). Finally, we introduce a Physical Simulation-Enhanced SDS to further optimize the positions of the Gaussians (see Sec. 4.3).

where g denotes the gravity acceleration. The updated velocities are transferred back to the particles as well as updating the positions:

$$\mathbf{v}_p^{t+1} = \sum_i N(\mathbf{x}_i - \mathbf{x}_p^t) \mathbf{v}_i^t \quad (8)$$

$$\mathbf{x}_p^{t+1} = \mathbf{x}_p^t + \Delta t \mathbf{v}_p^{t+1}. \quad (9)$$

We utilize the MPM to simulate the interactions of compositional objects in the assets.

4. Method

Given an image $I \in \mathbb{R}^{H \times W}$ of an asset with two compositional objects $\{O_1, O_2\}$ described by text prompts τ_1 and τ_2 , we would like to reconstruct a 3D representation of the two objects individually. Here we denote O_1 and O_2 as background and foreground objects respectively. We segment the foreground object in image space using Grounded-SAM [19] to obtain a semantic map. The image after segmentation is inpainted to complete the background object. For reconstruction, the multi-view images and the inpainted background images are generated using SyncDreamer [24]. We then reconstruct two Gaussian Splatting representations for background and foreground objects, denoted as G_1 and G_2 . A physical simulation-enhanced Score Distillation Sampling (SDS) is then applied to optimize the Gaussians for obtaining a physically plausible representation.

4.1. Multi-view Generation

To reconstruct the object described in the image, we generate the multi-view images from I . **First**, we use Grounded-SAM [19, 23, 34] to segment out the masks of both objects:

$$\{M_1, M_2\} = \text{GroundedSAM}(I; \tau_1, \tau_2), \quad (10)$$

where $M_1, M_2 \in \mathbb{R}^{H \times W}$. **Second**, suppose that O_1 is occluded by O_2 , we use inpainting model [35] to complete the image of O_1 (see Fig. 3):

$$\hat{I} = \text{Inpainting}(I * (\sim M_2) + I_{\text{noise}} * M_2; \tau_1), \quad (11)$$

where $\hat{I} \in \mathbb{R}^{H \times W}$ is the inpainted image, $I_{\text{noise}} \in \mathbb{R}^{H \times W}$ is an image with random noise sampled from the normal distribution. **Third**, we use SyncDreamer [24] to generate

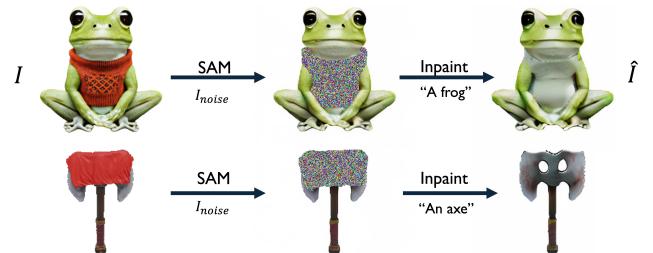


Figure 3. **Object inpainting with image diffusion models.**

images in 16 different views from I and \hat{I} :

$$\{I_i\}_{i=1}^{16} = \text{SyncDreamer}(I), \quad (12)$$

$$\{\hat{I}_i\}_{i=1}^{16} = \text{SyncDreamer}(\hat{I}), \quad (13)$$

where $I_i, \hat{I}_i \in \mathbb{R}^{H \times W}$. **Furthermore**, we obtain the semantic maps $S_i \in \{-1, 1, 2\}^{H \times W}$ of each I_i using Grounded-SAM, where -1 refers to the background, 1 refers to O_1 and 2 refers to O_2 .

4.2. Multi-view Images to 3D

We now have 1) the multi-view images and semantic maps $\{I_i, S_i\}_{i=1}^{16}$ of both O_1 and O_2 , and 2) multi-view images $\{\hat{I}_i\}_{i=1}^{16}$ of only O_1 . The target is to reconstruct 3D representations from these images and propagate the semantics from 2D images to 3D shapes. Since GroundedSAM does not guarantee multi-view consistent semantic segmentation, we leverage Part123 [21] to integrate the multi-view semantic maps into a 3D consistent one. Specifically, Part123 optimizes a semantic aware NeuS [42] from $\{I_i, S_i\}_{i=1}^{16}$:

$$\{f, g\} = \text{Part123}(\{I_i, S_i\}_{i=1}^{16}), \quad (14)$$

where $f : \mathbb{R}^3 \mapsto \mathbb{R}$ is the SDF field of both O_1 and O_2 , and $g : \mathbb{R}^3 \mapsto \mathbb{R}$ is the 3D semantic field. By marching cube algorithm, the mesh vertices can be extracted, denoted as $V = \{v_1, \dots, v_N\}$. Then V can be split into two groups $V = V_1 + V_2$ given the semantics from g . According to our assumption, V_2 denotes the mesh vertices of the foreground object O_2 . We fit GS for both O_1 and O_2 :

$$G_1 = \text{GaussianSplatting}(\{\hat{I}_i\}_{i=1}^{16}), \quad (15)$$

$$G_2 = \text{GaussianSplatting}(\{I_i\}_{i=1}^{16}; \mu \in V_2), \quad (16)$$

where we keep Gaussian centers of G_2 unchanged, i.e., based on the positions of V_2 , to keep its surface consistent with the extracted SDF. The SDF is utilized as a boundary constraint for the following MPM simulation [9].

4.3. Physical Simulation-Enhanced Optimization

Score Distillation Sampling Loss. To ensure the generated Gaussian Splatting G_1 is semantically consistent with the description of the inpainted image, we adopt Score Distillation Sampling (SDS) [29] to further optimize its representation. The SDS loss is defined as:

$$\nabla_{\theta} \mathcal{L}_{SDS} = \mathbb{E}_{t, \epsilon} \left[w(t)(\epsilon_{\phi}(I_t^p; y, t) - \epsilon) \frac{\partial I_t^p}{\partial \theta} \right], \quad (17)$$

where $w(t)$ denotes the time-dependent weighting function, ϵ_{ϕ} represents the pre-trained 2D diffusion model, I_t^p is the predicted image at timestep t . Here we reuse the text prompt y for inpainting as the condition for generation. θ denotes the parameters of the target Gaussian Splatting representation i.e., $\{\mu, \Sigma, q, \alpha, c\}$ as mentioned in section 3.1. Among

these parameters, μ represents the center position for each particle, which is the only key property to take care for ensuring physical plausibility. We freeze opacity α and color c during the optimization to prevent SDS from changing the appearance of the object. Therefore we divided the parameters into three groups $\theta = \{\theta_{\mu}, \theta_t, \theta_a\}$, where θ_t denotes the scaling factor and rotation quaternion for the Gaussian particles. θ_a represents the frozen appearance-related parameters.

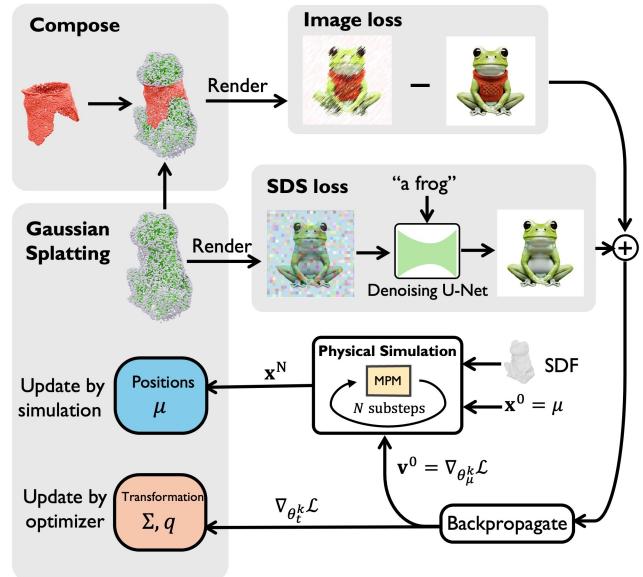


Figure 4. **The overview of our PSE-SDS.** The gradients come from the SDS and image loss are divided into two streams during the backpropagation. Specifically, $\nabla_{\theta_{\mu}} \mathcal{L}$ is utilized as the initial velocity of the physical simulation for updating the positions μ of Gaussians.

Image Loss. As we constrain SDS loss for optimizing object geometry only, to ensure visual consistency, we utilize an image loss as a complement to penalize the L1-norm difference between the rendered from the generated composed object $G_c = \{G_1, G_2\}$ and the original input image:

$$\mathcal{L}_{Image} = (1 - \lambda_1) \mathcal{L}_1(I^c, I) + \lambda_1 \mathcal{L}_{SSIM}(I^c, I), \quad (18)$$

where I^c is the image rendered from the generated composed object, I denotes the original input image, and \mathcal{L}_{SSIM} refers to the structural similarity loss function.

Physical Simulation-Enhanced SDS. The final objective is to find parameters θ_{μ} and θ_t , by minimizing the total loss \mathcal{L} :

$$\mathcal{L} := \mathcal{L}_{Image}(\theta_{\mu}, \theta_t) + \lambda_3 \mathcal{L}_{SDS}(\theta_{\mu}, \theta_t), \quad (19)$$

where the \mathcal{L}_{SDS} and \mathcal{L}_{Image} are designed to penalize discrepancy in geometry and visual appearance respectively, between the generated objects and the input image. λ_3 is the weighting factor.

We observed that directly applying the loss gradient to update particle positions μ results in penetrations and artifacts as shown in Figure 8. To ensure the physical plausibility, we propose physical simulation-enhanced SDS (shown in Figure 4). We delegate the updates of μ to the physical simulation. Here we use the MLS-MPM [15] as the physical simulator. One sub-step of the simulation process can be formalized as follows:

$$\mathbf{x}^{n+1}, \mathbf{v}^{n+1} = \text{MPM}(\mathbf{x}^n, \mathbf{v}^n, \Delta t, \psi), \quad (20)$$

where \mathbf{x}^n and \mathbf{v}^n represent particle position and velocity at timestep n , ψ denotes all other properties such as the particle mass, particle volume and materials parameters. Note we omit the subscript p for clarity compared to the notations mentioned in section 3.2.

As described in algorithm 1, given K steps of optimization, we set the $\nabla_{\theta_\mu^k} \mathcal{L}$ i.e., loss gradient with respect to particle position as the initial velocity of particles for the MPM based physical simulation. The MPM outputs the updated μ^{k+1} after N sub-step simulations.

Algorithm 1 Physical Simulation-Enhanced SDS

Require: Given K steps of optimization, N sub-steps MPM simulation, learning rate γ

- 1: **for** $k = 1$ to K **do**
- 2: Compute $\nabla_{\theta^k} \mathcal{L}$ according to Eqn.19
- 3: $\nabla_{\theta^k} \mathcal{L} = \{\nabla_{\theta_\mu^k} \mathcal{L}, \nabla_{\theta_t^k} \mathcal{L}\}$
- 4: $\mathbf{x}^0 = \mu^k, \mathbf{v}^0 = \nabla_{\theta_\mu^k} \mathcal{L}$
- 5: $\Delta t = \gamma/N$
- 6: **for** $n = 0$ to N **do**
- 7: $\mathbf{x}^{n+1}, \mathbf{v}^{n+1} = \text{MPM}(\mathbf{x}^n, \mathbf{v}^n, \Delta t, \psi)$
- 8: **end for**
- 9: $\mu^{k+1} = \mathbf{x}^N$
- 10: $\theta_t^{k+1} = \theta_t^k - \gamma \nabla_{\theta_t^k} \mathcal{L}$
- 11: **end for**

Intuitively, at the first sub-step of the simulation, the MPM advances the particles' positions according to the initial velocity (i.e., loss gradient), which is equivalent to one step of vanilla optimization using gradient descent with a step size Δt . The following simulation sub-steps are then performed to progressively correct the particles positions by solving the physical system.

5. Experiment

5.1. Implementation Details

We use Stable-Diffusion-XL-1.0 as an inpainting model with a guidance scale in $\{7.5, 8.0, 9.0, 12.5\}$. During

SDS optimization, we decrease timestep t from 100 to 20. We train NeuS with 1k steps, fit G_2 with 30k steps and G_1 with 3k steps, and perform the physical simulation-enhanced optimization with 500 steps. We empirically set $\lambda_1 = 0.2, \lambda_2 = 1.0, \lambda_3 = 0.00001$.

5.2. Evaluation

We assess the results with the following metrics: 1) Peak Signal-to-Noise Ratio (PSNR), which quantifies the similarity between the rendered image and the input image at the reference view; 2) CLIP score [32] for various comparisons, including between novel-view images and the input image (CLIP_{mv}), between the reference view of O_1 and the inpainting prompt (CLIP_{text}), between the reference view of O_1 and the inpainted image (CLIP_{ip}), and between the novel-view images of O_1 and the inpainted image (CLIP_{ip}^{mv}).

Table 1. Quantitative comparison with previous work.

Method	PSNR(dB)↑	CLIP _{mv} (%)↑
Part123 [21]	17.52	79.60
ComboVerse [2]	16.22	85.23
Ours	30.70	87.29

5.3. Comparison with Baseline

We compare our approach with the following baselines: 1) Part123 [21], which generates a holistic mesh with semantics from a single image; 2) ComboVerse [2], which generates each component in the image separately, and assembles them with estimated similarity transformations. Fig. 5 and Tab. 1 shows the qualitative and quantitative results.

Overall, our method produces the most superior 3D compositional assets, taking into account both visual quality and physical plausibility. Part123 generates the entire assets as a single mesh, leading to incompletely segmented objects. ComboVerse can not address penetrations between objects. Our method achieves better consistency with the input image and effectively resolves the penetration problem.

5.4. Ablation Study

We conduct the following ablation study to validate the effectiveness of our Physical Simulation-Enhanced SDS (PSE-SDS): 1) **PPPS** uses physical simulation as a post-processing procedure after the asset generation. 2) **SDS** denotes the vanilla SDS optimization that relying solely on visual supervision. 3) **PPPS + SDS** represents performing simulation and the vanilla SDS alternately. A comparison of how these variants integrate information through physical simulation is shown in Figure 6.

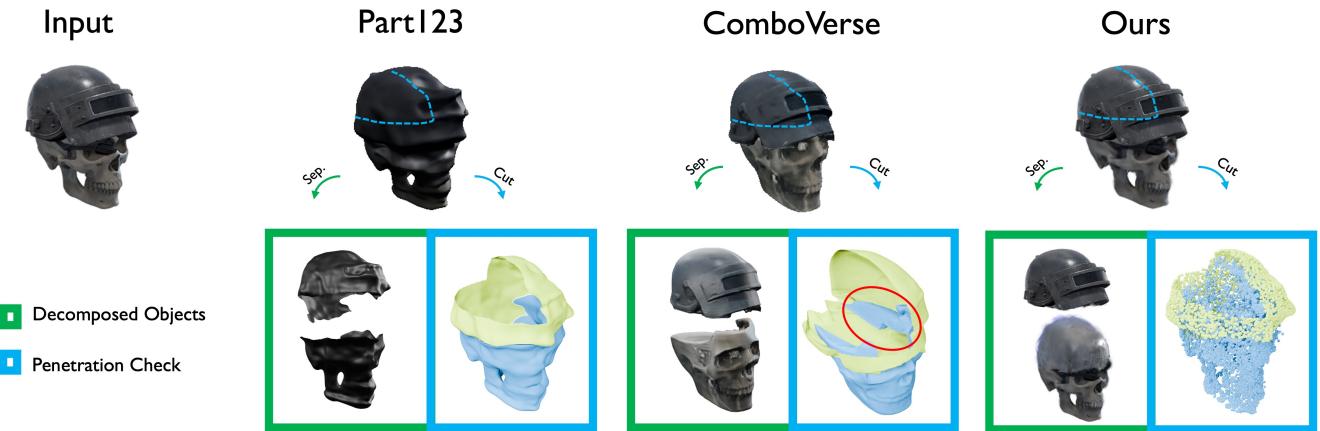


Figure 5. **Qualitative comparison with previous work.** The green box illustrates the decomposed objects, while the blue box highlights the physical relationships, such as whether the components are in penetration (in red circle). Since we use 3DGS representation, we convert Gaussian centers to point clouds for geometry visualization.

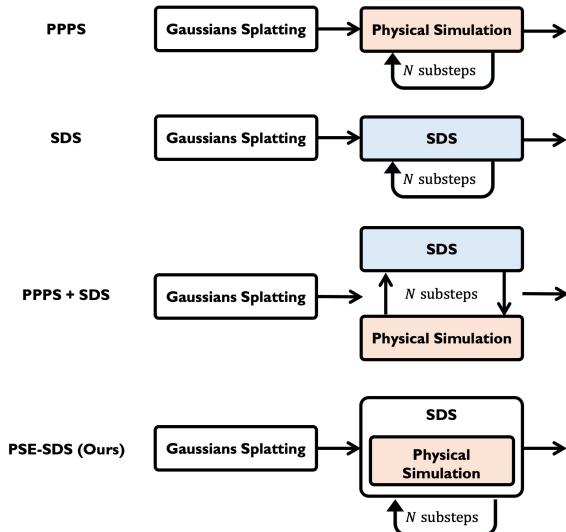


Figure 6. **Various methods of integrating interactive information through physical simulation**

Table 2. **Quantitative results of ablation studies on PSE-SDS.** The unit for PSNR is decibels (dB), and that for CLIP scores is percentage (%).

Method	PSNR↑	CLIP _{text} ↑	CLIP _{ip} ↑	CLIP _{mv} ↑	CLIP _{ip} ^{mv} ↑
PPPS	25.02	29.17	93.66	87.36	86.86
SDS	29.13	28.35	89.76	87.90	84.83
PPPS + SDS	20.74	28.17	89.46	88.06	84.69
PSE-SDS (Ours)	29.79	<u>28.66</u>	<u>92.68</u>	88.30	86.93

Effectiveness of Physical Simulation-Enhanced SDS. The output generated in Stage 2 (Sec. 4.2) encounters penetration issues (indicated by red boxes in the second column of Fig. 8), due to the omission of interactive information in

the process. Tab. 2 and Fig. 8 provide both quantitative and qualitative insights into the ablation studies examining various methods of integrating interactive information through physical simulation. 1) **PPPS** overlooks visual plausibility, since physical simulation treats every point as material without considering semantics. 2) **SDS** disregards physical plausibility; even though the overall asset aligns well with the input image, individual objects may collapse. 3) **PPPS+SDS** can still result in object collapse without adequate physical constraints. 4) Our **PSE-SDS** yields superior outcomes in terms of both visual and physical plausibility. We present more examples in Figure 9 to demonstrate that our method can generate assets with diverse compositional layouts.

Are SDS and Image Loss both necessary? We further assess the individual contributions of \mathcal{L}_{SDS} and \mathcal{L}_{Image} , respectively (See Fig. 7). Excluding \mathcal{L}_{Image} (w.o. IL) leads to the appearance of extraneous object, attributable to the variability inherent in SDS. Omitting \mathcal{L}_{SDS} (w.o. SDS) results in poor visual plausibility within occluded areas.

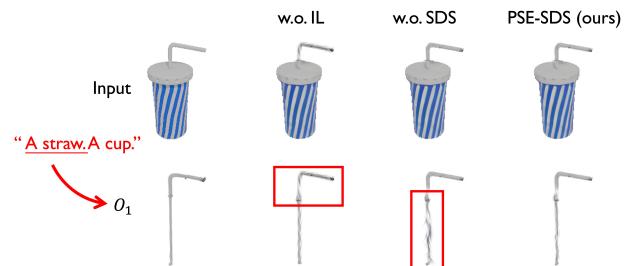


Figure 7. **Ablation studies on \mathcal{L}_{Image} and \mathcal{L}_{SDS} .**

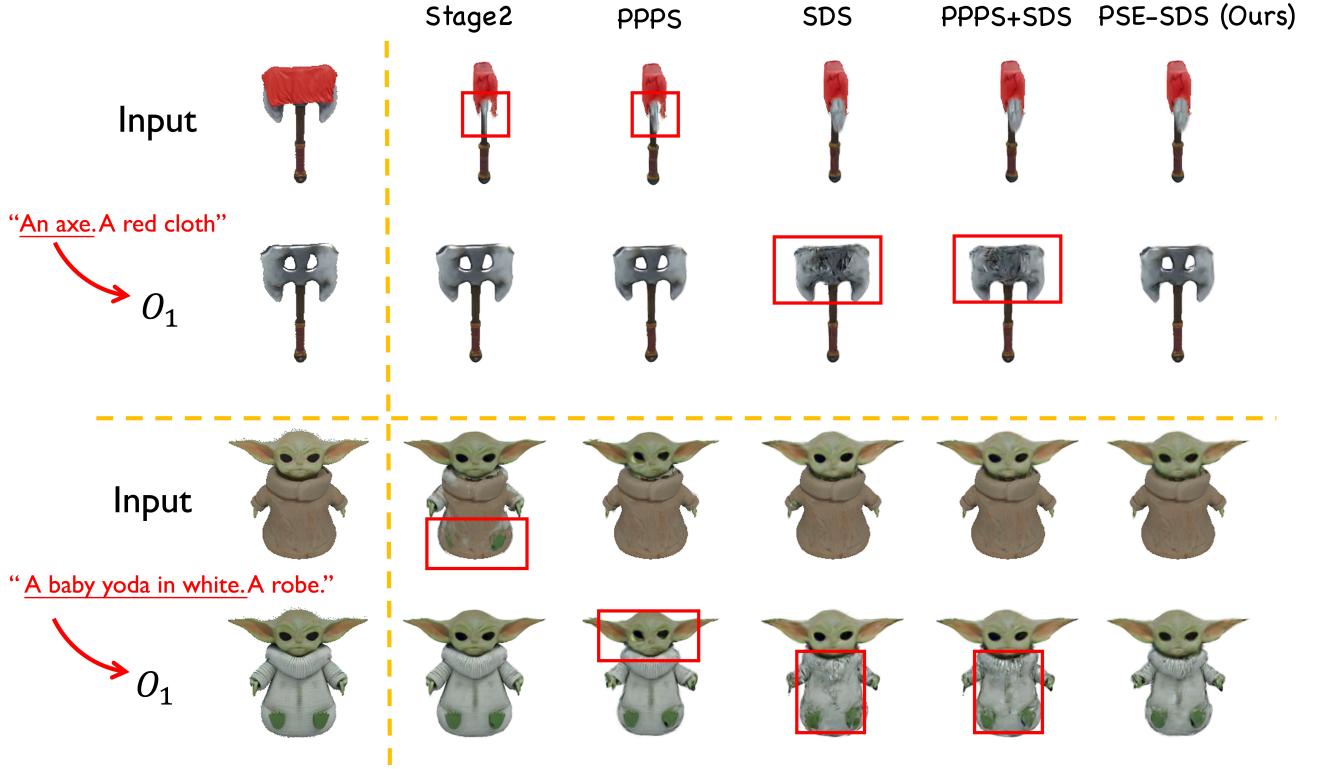


Figure 8. Qualitative results of ablation studies on PSE-SDS.

6. Conclusion

In this paper, we present PhyCAGE, the first approach to generate physically plausible compositional 3D assets from

a single Image. Our method incorporates a novel Physical Simulation-Enhanced Score Distillation Sampling (PSE-SDS) technique, which leverages a physical simulator as a physics-guided optimizer. This optimizer iteratively corrects the positions of the reconstructed Gaussians to achieve a physically compatible state. The experiments demonstrate that PhyCAGE is capable of generating various 3D assets in diverse compositional layouts. We believe our method represents a significant first step toward physics-aware 3D scene generation.

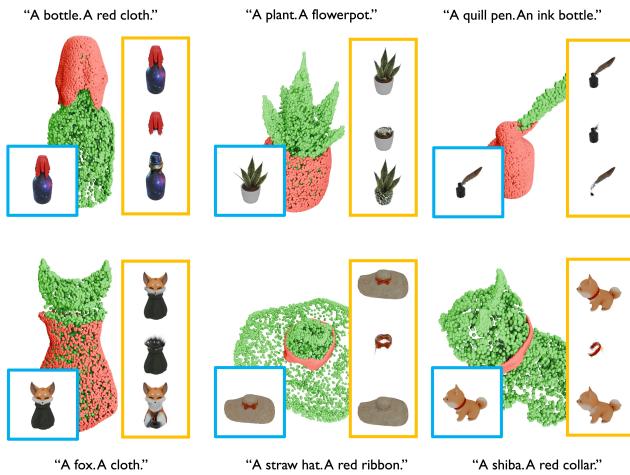


Figure 9. **More results.** The blue boxes depict the input image, whereas the orange boxes showcase the generated outcomes featuring decomposed objects. The colored point clouds provide visualizations of our generated 3DGs.

Limitations and future work. Our approach mainly focuses on assets consisting of two objects currently. While it has the potential to be extended to scenarios with more objects by iteratively treating one object as the foreground and the remaining ones as the background during each generation sub-routine. The quality of the final output depends on the performance of the multi-view generation method. We expect that our approach can be further improved by leveraging more robust reconstruction model in the future. We aim to further develop our method to generate mesh-based assets, thereby supporting more simulation techniques such as the Finite Element Method (FEM) [37] and sophisticated collision handling methods [20].

References

- [1] Javier Bonet and Richard D Wood. *Nonlinear Continuum Mechanics for Finite Element Analysis*. Cambridge University Press, 1997. 3
- [2] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. *arXiv:2403.12409*, 2024. 2, 6
- [3] Yunuo Chen, Tianyi Xie, Zeshun Zong, Xuan Li, Feng Gao, Yin Yang, Ying Nian Wu, and Chenfanfu Jiang. Atlas3d: Physically constrained self-supporting text-to-3d for simulation and fabrication. *arXiv:2405.18515*, 2024. 3
- [4] Dana Cohen-Bar, Elad Richardson, Gal Metzger, Raja Giryes, and Daniel Cohen-Or. Set-the-scene: Global-local training for generating controllable nerf scenes. In *ICCV*, 2023. 2
- [5] Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. Tela: Text to layer-wise 3d clothed human generation. *arXiv:2404.16748*, 2024. 2
- [6] Dave Epstein, Ben Poole, Ben Mildenhall, Alexei A Efros, and Aleksander Holynski. Disentangled 3d scene generation with layout learning. *arXiv:2402.16936*, 2024. 2
- [7] Yu Fang, Minchen Li, Ming Gao, and Chenfanfu Jiang. Silly rubber: an implicit material point method for simulating non-equilibrated viscoelastic and elastoplastic solids. *ACM TOG*, 38(4):1–13, 2019. 3
- [8] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. Learning disentangled avatars with hybrid 3d representations. *arXiv:2309.06441*, 2023. 2
- [9] Arnulph Fuhrmann, Gerrit Sobotka, and Clemens Groß. Distance fields for rapid collision detection in physically based modeling. In *Proceedings of GraphiCon*, pages 58–65, 2003. 5
- [10] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *CVPR*, 2024. 2
- [11] Minghao Guo, Bohan Wang, Pingchuan Ma, Tianyuan Zhang, Crystel Elaine Owens, Chuang Gan, Joshua B Tenenbaum, Kaiming He, and Wojciech Matusik. Physically compatible 3d object modeling from a single image. *arXiv:2405.20510*, 2024. 3
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2
- [13] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv:2311.04400*, 2023. 1, 2
- [14] Shoukang Hu, Fangzhou Hong, Tao Hu, Liang Pan, Haiyi Mei, Weiye Xiao, Lei Yang, and Ziwei Liu. Humanliff: Layer-wise 3d human generation with diffusion model. *arXiv:2308.09712*, 2023. 2
- [15] 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 TOG*, 37(4):1–14, 2018. 3, 6
- [16] C. Jiang, C. Schroeder, A. Selle, J. Teran, and A. Stomakhin. The affine particle-in-cell method. *ACM TOG*, 34(4):1–10, 2015.
- [17] C. Jiang, T. Gast, and J. Teran. Anisotropic elastoplasticity for cloth, knit and hair frictional contact. *ACM TOG*, 36(4):1–14, 2017. 3
- [18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. 2, 3
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4
- [20] Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M. Kaufman. Incremental potential contact: Intersection- and inversion-free large deformation dynamics. *ACM TOG*, 39(4), 2020. 8
- [21] Anran Liu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Zhiyang Dou, Hao-Xiang Guo, Ping Luo, and Wenping Wang. Part123: Part-aware 3d reconstruction from a single-view image. In *ACM SIGGRAPH*, pages 1–12, 2024. 1, 5, 6
- [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:2406.04338*, 2024. 3
- [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:2303.05499*, 2023. 4
- [24] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv:2309.03453*, 2023. 1, 2, 4
- [25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 2
- [26] 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. In *ECCV*, 2020. 2
- [27] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. *arXiv:2404.16666*, 2024. 3
- [28] Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. *arXiv:2303.12218*, 2023. 2
- [29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv:2209.14988*, 2022. 2, 5
- [30] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv:2306.17843*, 2023. 2

- [31] Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Feature splatting: Language-driven physics-based scene synthesis and editing. [arXiv:2404.01223](#), 2024. 3
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. [arXiv:2103.00020](#), 2021. 6
- [33] D. Ram, T. Gast, C. Jiang, C. Schroeder, A. Stomakhin, J. Teran, and P. Kavehpour. A material point method for viscoelastic fluids, foams and sponges. In [SCA](#), 2015. 3
- [34] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kun-chang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. [arXiv:2401.14159](#), 2024. 4
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In [CVPR](#), 2022. 2, 4
- [36] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. [arXiv:2310.15110](#), 2023. 1, 2
- [37] Eftychios Sifakis and Jernej Barbic. Fem simulation of 3d deformable solids: a practitioner’s guide to theory, discretization and model reduction. In [ACM SIGGRAPH 2012 Courses](#), pages 1–50, 2012. 8
- [38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In [ICML](#), 2015. 2
- [39] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. [arXiv:2309.16653](#), 2023. 1
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In [NeurIPS](#), 2017. 1
- [41] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Xin Li, Wenping Wang, Rong Xie, and Li Song. Disentangled clothed avatar generation from text descriptions. [arXiv:2312.05295](#), 2023. 2
- [42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. [arXiv:2106.10689](#), 2021. 5
- [43] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In [CVPR](#), 2023. 1, 2
- [44] Yi Wang, Jian Ma, Ruizhi Shao, Qiao Feng, Yu-Kun Lai, Yebin Liu, and Kun Li. Humancoser: Layered 3d human generation via semantic-aware diffusion model. [arXiv:2312.05804](#), 2023. 2
- [45] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In [ECCV](#), 2022. 2
- [46] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In [ICCV](#), 2023. 2
- [47] Yuanbo Xiangli, Lining Xu, Xingang Pan, Nanxuan Zhao, Bo Dai, and Dahua Lin. Assetfield: Assets mining and reconfiguration in ground feature plane representation. In [ICCV](#), 2023. 2
- [48] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In [CVPR](#), 2024. 3
- [49] Han Yan, Yang Li, Zhennan Wu, Shenzhou Chen, Weixuan Sun, Tai Zhang Shang, Weizhe Liu, Tian Chen, Xiaqiang Dai, Chao Ma, et al. Frankenstein: Generating semantic-compositional 3d scenes in one tri-plane. [arXiv:2403.16210](#), 2024. 2
- [50] Yunhan Yang, Yukun Huang, Yuan-Chen Guo, Liangjun Lu, Xiaoyang Wu, Edmund Y. Lam, Yan-Pei Cao, and Xi-hui Liu. Sampart3d: Segment any part in 3d objects. [arXiv:2411.07184](#), 2024. 1
- [51] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. [ACM TOG](#), 42(4):1–16, 2023. 2
- [52] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. [arXiv:2406.13897](#), 2024. 2
- [53] Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. Scenewiz3d: Towards text-guided 3d scene composition. [arXiv:2312.08885](#), 2023. 2
- [54] 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. In [ECCV](#), 2025. 3
- [55] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. [arXiv:2305.04461](#), 2023. 2