Inclusive Interactive Collisions for Multi-View Consistent Compositional 3D Generation

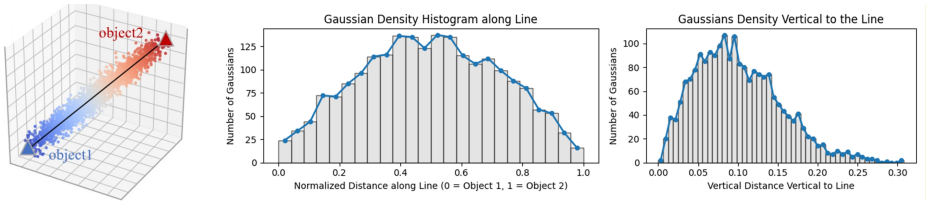
Chang Liu[1,3], Mingwen Shao[2,3*], Xiang Lv[3], Xinyuan Chen[3], Lingzhuang Meng[3], Qiao Zhang[3], Zhengyi Gong[3], and Jinghao Hu[4]

[1] College of Computer Science, Fudan University, Shanghai, 200438, China
[2] Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology, Shenzhen, 518107, China
[3] School of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, China
[4] College of Computer Science, Northwest University, Xi'an, 710127, China
upc_liuchang@163.com, smw278@126.com, lvxiang1997@126.com, xinyuan_sdnu@163.com, lzhmeng1688@163.com, bz24070008@s.upc.edu.cn, 17865420733@163.com, hujinghao@stumail.nwu.edu.cn

Abstract. Recent breakthroughs in 3D generation have advanced notably with the development of text-to-image diffusion model. However, existing methods remain two practical challenges: (1) They primarily generate single 3D object, but struggle to generate multi-object compositional 3D assets due to the lack of the modeling for Gaussian primitives in reasonable interactions. (2) They often suffer from cross-view inconsistency during 3D optimization, as Score Distillation Sampling inherently performs on each single view, inevitably resulting in cross-view hallucinations. To solve above issues, we propose I^{2}C-3D, a novel optimization-based method to generate multi-view consistent compositional 3D assets with reasonable interactions. Specifically, we propose an Inclusive Interactive Collisions strategy to guide Gaussian primitives appearing in reasonable interaction regions naturally, thereby ensuring objects in the compositional scene interact in a physically plausible and visually coherent way. Additionally, to enhance multi-view consistency, Multi-View Adaptive Score Distillation Sampling is devised to distill multi-view consistency prior and layout prior from pre-trained diffusion model by modulating attention map of instance token and spatial token across viewpoints. Benefiting from above elaborate designs, I^{2}C-3D not only generates high-fidelity multi-view consistent compositional 3D assets but also supports 3D editing flexibly, facilitating complex scene generation. Extensive experiments demonstrate our I^{2}C-3D outperforms existing methods in generation quality and multi-view consistency.

Keywords: Compositional 3D generation · Multi-object 3D generation · Interaction content generation · Multi-view consistency

* Corresponding author



(a) Visualization. **(b)** Analysis of the distribution of Gaussian primitives.

Fig. 1: Analysis and visualization of the statistical distribution of Gaussian primitives in the interaction region. (a) The Gaussian primitives are sparse at the line's ends and densely distributed with more collision near the midpoint of the line. (b) Through quantifying the Gaussian distribution along line and vertical line, further demonstrate most Gaussian primitives are distributed around the midpoint of line.

1 Introduction

Recently, generating high-quality 3D assets according to given text prompt or a single image is a growing demand in AR/VR [2,22], movies and industrial design [14]. Thanks to large scale text-to-image generative diffusion models [27, 29, 42], significant progress has also been made in the field of 3D generation. Previous works have been broadly divided into two categories: (1) Training-based generation [12,19,31], which is a feed-forward generation process by training an efficient reconstruction model on 3D data. (2) Optimization-based generation [21, 23, 32], which distillates prior knowledge from pre-trained text-to-image diffusion model via Score Distillation Sampling (SDS) [39] for 3D generation [3, 6, 33].

Although the above methods can achieve remarkable results in single-object 3D generation, they struggle to generate multi-object scene with complex interactions. Specifically, for training-based methods, they lack complex multi-object data in training dataset [7], which inevitably leads to failure on multi-object 3D generation. For optimization-based methods, they struggle to distill compositional prior knowledge from pre-trained text-to-image diffusion models, which is mainly due to the fact that existing pre-trained 2D diffusion models lack fine-grained controllability, further fail to precisely follow the text prompt involving multiple objects, attributes or spatial relationships [13, 20, 37].

We thoroughly investigate the statistical distribution of Gaussian primitives in the interaction region and observe that Gaussian primitives tend to concentrate around the midpoint of the connecting line between interacting objects, forming a roughly cylindrical interaction space. Specifically, Fig. 1a visualizes the distribution of Gaussian primitives around the connecting line of the two interacting objects: blue and red dots denote Gaussian primitives belonging to objects 1 and object 2, respectively. Lighter-colored regions indicate that the two kinds of Gaussian primitives interact and collide. The results in Fig. 1a show the Gaussian primitives at the ends of the connecting line are relatively sparse, and those located near the midpoint of the line are densely distributed with more collision. Moreover, we plot the Gaussian Density Histogram along Line

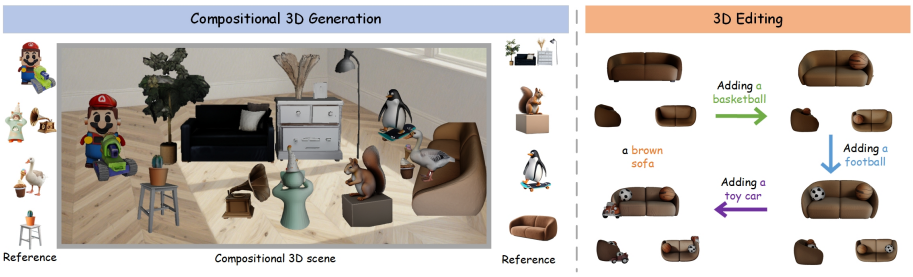


Fig. 2: Illustration of our I^{2}C-3D generated compositional 3D scene and 3D editing results. Our I^{2}C-3D not only generates multi-view consistent 3D scene with reasonable interaction region, but also achieves flexible 3D editing.

and Gaussians Density Vertical to the Line (displayed in Fig. 1b), which quantitatively indicate the Gaussian density near the midpoint of the line is larger. Inspired by these observations, we design a collision constraint that guides Gaussian primitives near the interaction boundary to a reasonable collision region, enabling realistic interaction-aware 3D generation.

In this paper, we propose the I^{2}C-3D, a novel framework to generate multi-view consistent compositional 3D assets with reasonable interactions, which is a coarse-to-fine optimization process. In coarse 3D generation stage, we elaborate a local-global 3D layout generator based on LLMs, then utilize pre-trained single-object 3D generation model to reconstruct and compose a coarse 3D scene. In refinement stage, we propose Inclusive Interactive Collisions strategy, which not only constrains most Gaussian primitives are in inside the bounding box to ensure generated compositional 3D scene aligns with the planned 3D layout, but also guide Gaussian primitives to collide naturally in interaction region for reasonable interaction generation. Subsequently, Multi-View Adaptive Score Distillation Sampling is designed to distill multi-view consistency prior and fine-grained layout prior from pre-trained diffusion model through adaptive attention modulation for enhancing multi-view consistency. Attribute to above elaborate designs, our I^{2}C-3D not only generates multi-view consistent compositional 3D scene with reasonable interaction, but also supports flexible 3D editing (see Fig. 2). Extensive quantitative and qualitative experiments suggest our I^{2}C-3D achieves state-of-the-art (SOTA) results, outperforming in high quality and multi-view consistency. Our main contributions are summarized as follows:

- We propose I^{2}C-3D, a novel optimization-based compositional 3D generation framework, which can generate multi-view consistent compositional 3D assets with reasonable interactions.
- We devise an Inclusive Interactive Collisions (I^{2}C) strategy to constrain gaussian primitives appearing in reasonable interaction regions, realizing physically plausible interaction.

- Multi-View Adaptive Score Distillation Sampling (MV-ASDS) is designed to distill cross-view prior and fine-grained layout prior for achieving multi-view consistency.
- Extensive experiments show our I^2C-3D qualitatively and quantitatively outperforms SOTA methods.

2 Related Work

2.1 Text-to-3D Generation

Significant advances have been made in text-to-3D generation with the development of generative diffusion models. A representative work, DreamFusion [25] creatively proposes Score Distillation Sampling to distill the prior from powerful 2D diffusion models for 3D generation. Inspired by DreamFusion, various excellent approaches have been proposed. On the one hand, researchers attempt to explore more efficient 3D representations, from NeRF [24] to 3D Gaussians (3DGS) [17], DMTet [28] and hybrid representations [30, 43]. On the other hand, they exploit more efficient supervision to improve over-saturation and over-smoothing in SDS, such as VSD [33] and CSD [41]. Specifically, Magic3D [21] devises a coarse-to-fine optimization strategy that utilizes multiple diffusion priors at different resolutions to optimize the 3D representation for high-resolution 3D Generation. Subsequently, GaussianDreamer [40] bridges 2D and 3D diffusion models for fast text-to-3D Generation. Meanwhile, gsgen [6] proposes a two-stage optimization strategy that first joints 2D and 3D diffusion guidance to shape a rough geometric structure, then refines appearance details through compactness-based densification. However, these methods can only generate single-object 3D asset, but our I^2C-3D supports multi-object 3D generation.

2.2 Compositional 3D generation

Compositional 3D generation aims to generate attribute-specific and spatially reasonable multi-object compositional 3D assets from text prompt or an image. Specifically, ComboVerse [4] proposes spatially-aware score distillation sampling (SSDS) to guide the position of objects for multi-object combination. Meanwhile, REPARO [11] designs differentiable rendering techniques to optimize the layout of meshes, ensuring coherent scene composition. GALA3D [47] utilizes layout priors to bridge text and compositional scene for end-to-end high-quality scene-level 3D content generation. Subsequently, CompGS [9] employs 3DGS initialization with 2D compositionality and Dynamic SDS Optimization for high-quality 3D generation. Layout-your-3D [45] exploits 2D layout as blueprints, enabling precise and plausible control over 3D generation. However, above methods struggle to generate multi-view consistent compositional 3D assets with plausible interaction. In contrast, our I^2C-3D constrains the distribution of Gaussian primitives in interaction region based on geometry theory for natural interaction generation.

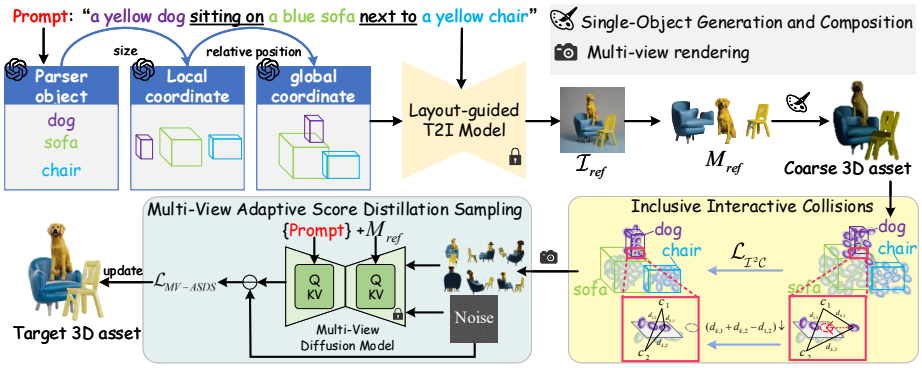


Fig. 3: Overview of our I²C-3D. We first utilize pre-trained single-object 3D generation model to reconstruct and compose a coarse 3D scene. Subsequently, leverage Inclusive Interactive Collisions strategy (I²C) to refine interaction region generation. Then Multi-View Adaptive Score Distillation Sampling (MV-ASDS) is applied to distill cross-view priors from multi-view diffusion model for achieving multi-view consistency. Through coarse-to-fine optimization, high-quality 3D asset has been generated.

3 Methodology

3.1 Preliminary

Score Distillation Sampling. DreamFusion first proposes Score Distillation Sampling (SDS) for distilling prior from pre-trained 2D diffusion models to optimize 3D representation. Specifically, given a text prompt y and an initialized 3D representation parameterized by θ, the rendered image x from viewpoint c can be obtained through a differentiable rendering process $g(\theta, c)$. Subsequently, add Gaussian noise ϵ to x and leverage a pre-trained 2D diffusion model ϕ to predict noise, then denoise it. By computing the difference between predicted noise $\epsilon_\phi(x_t, y, t)$ and initial noise ϵ, SDS constructs gradient and propagates backward it to optimize 3D representation, which can be expressed as:

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{\epsilon,t}\left[w(t)(\epsilon_\phi(x_t, y, t) - \epsilon)\frac{\partial g(\theta, c)}{\partial \theta})\right], \tag{1}$$

where t is timestep, $w(t)$ is a weighting function. DreamFusion utilizes NeRF [24] as 3D representation and renders the image via voxel rendering. While considering efficiency and editability, we choose 3DGS as the 3D representation for subsequent optimization.

3.2 Coarse 3D Generation Stage

3D layout generation. Unlike 2D layout generation, LLMs struggle to directly derive 3D coordinates from complex descriptions because their training data lack sufficient 3D-text pairs. Consequently, the generated 3D bounding boxes

often overlap or are overly spaced, preventing realistic object interactions. Based on this, we design a local-global 3D layout generation strategy. First, given a complex text description containing multiple objects, we utilize LLMs to parser the main objects' words and relative positional relationship. Then we explore the layout planning capability of LLMs to estimate approximate size of each object and return local coordinate. For example, *object 1, [length, width, height], in the left of object 2 [length, width, height] and smaller than object 2*. Subsequently, we place the estimated bounding boxes in global coordinate system according to their relative positional relationships and take into account the perspective prompt in the text to compute their global coordinate. Specific details for each stage will be detailed in the supplementary material.

3D Object Reconstruction. Given a text prompt y and generated 3D layout $B = \{(y_1, b_1), (y_2, b_2), \ldots, (y_n, b_n)\}$, where y_i denotes i-th object token and b_i is corresponding 3D bounding box. First, we project the 3D bounding box to 2D layout B^{2D} from the front viewpoint, then utilize a layout-guided text-to-image model $G(y, B^{2D})$ to generate the 2D reference image I_{ref}. Subsequently, the Segment-Anything Model (SAM) [18] is utilized to get each instance with broken interaction regions. To avoid inpainting leads to some unreasonable interaction region generation during 3D composition, we directly utilize Large Multi-View Gaussian Model for instance reconstruction without inpainting:

$$I_{ref} = G(y, B^{2D}); I_i = SAM(I_{ref}, y_i), \tag{2}$$

$$S_i = LGM(I_i). \tag{3}$$

3.3 Refinement Stage

As mentioned in Introduction, existing approaches struggle to generate a compositional 3D scene due to the challenges in synthesizing reasonable interaction region and maintaining multi-view consistency. Based on this, we propose an Inclusive Interactive Collisions strategy (I^2C) to ensure that 3D instances are inside the bounding box while allowing Gaussian primitives to collide in the interaction region naturally to achieve reasonable interaction generation. Subsequently, we devise a Multi-View Adaptive Score Distillation Sampling (MV-ASDS) to distill cross-view prior and layout prior from pre-trained diffusion model by modulating attention map of instance token and spatial token across viewpoints, further achieving multi-view consistency.

Inclusive Interactive Collisions. I^2C consists of two constraints: In-Box constraint and Interaction Collision constraint respectively. The former ensures generated compositional 3D assets align with planned 3D layout by judging the position relationship between most Gaussian primitive and its bounding box in a local coordinate system. The latter allows the few Gaussian primitives appearing in reasonable interaction regions, realizing physically plausible and visually coherent object interaction.

Firstly, we utilize the K-means algorithm to cluster the global Gaussian primitives $G = \{g_1, g_2, \ldots, g_n\}$, and the initial clustering center $C = \{c_1, c_2, \ldots, c_k\}$ is the center of the planned instance bounding box $B = \{B_1, B_2, \ldots, B_k\}$. Then, the clustering center C is assigned to each Gaussian primitive g_i, and the mean of the centers of all the Gaussian primitives in each cluster is computed as the new clustering center. Repeat the above steps until the clustering centers no longer change. The process can be expressed as follows:

$$c_i^{t+1} = \frac{1}{|G_i^t|} \sum_{g_j \in G_i^t} g_j, \tag{4}$$

where G_i^t is the set of Gaussian primitives belonging to the clustering center c_i at iteration t, and $|G_i^t|$ is the number of Gaussian primitives in the set G_i^t.

Subsequently, we constrain most Gaussian primitives to be inside the bounding box of the group which they belong. In addition, we hope the Gaussian primitives to fill the bounding box, avoiding generated 3D assets are too thin and narrow. Therefore, for the set of Gaussian primitives of i-th instance $g_i = \{g_i^1, g_i^2, \ldots, g_i^n\}$, B_i^{max} and B_i^{min} are its longest and shortest sides in local coordinate system, respectively, and $M_i = \{m_i^1, m_i^2, \ldots, m_i^n\}$ is the local coordinates of all Gaussian primitives of i-th instance. Then the In-Box loss can be expressed as follows:

$$\mathcal{L}_{IB} = \sum_{i=1}^{k} \sum_{j=1}^{N} \Big[\mathcal{C}\big(\max(m_i^j - B_i^{\max}, 0)\big) + \mathcal{C}\big(\max(B_i^{\min} - m_i^j, 0)\big) \Big], \tag{5}$$

where $\mathcal{C}$ is the clamp function to ensure that the difference value is non-negative. Minimizing the above In-Box loss until the proportion of the number of Gaussian primitives in the group is greater than threshold and stop the calculation.

The remaining Gaussian primitives which are on the outside of the bounding box and in the space of neighboring bounding boxes, are used to generate interaction regions. As shown in Fig. 1b, interaction collision tend to occur in the region near the midpoint of the line connecting the centers of the neighboring bounding boxes. Therefore, we design Interaction Collision constraints to guide the remaining Gaussian primitives toward the interaction region for collision and interaction. Based on triangle trilateral relationship theory in geometry, i.e., the sum of two sides of a triangle is greater than or equal to the third side, for the i-th set of Gaussian primitives beyond the bounding box, we compute the interaction collision loss with the neighboring j-th bounding box:

$$\mathcal{L}_{\mathcal{IC}} = \sum_{k=1}^{N} max(0, (d_{k,i} + d_{k,j}) - d_{i,j} - \tau), \tag{6}$$

where $d_{i,j}$ is the European distance between the i-th and j-th bounding boxes, $d_{k,i}$ and $d_{k,j}$ denote the distance from the k-th Gaussian primitive to the i-th

bounding box and the j-th bounding box. τ is a hyperparameter to avoid too many Gaussian primitives densely clustered in a small region, which is set to 5% of the maximum bounding-box diagonal.

The total loss in I^2C can be expressed as follows:

$$\mathcal{L}_{\mathcal{I}^2\mathcal{C}} = \lambda_{\mathcal{IB}}\mathcal{L}_{\mathcal{IB}} + \lambda_{\mathcal{IC}}\mathcal{L}_{\mathcal{IC}}, \tag{7}$$

where $\lambda_{\mathcal{IB}}$ and $\lambda_{\mathcal{IC}}$ are two hyperparameters.

Multi-View Adaptive Score Distillation Sampling. Traditional SDS cannot accurately achieve positional adjustment and spatial optimization. Inspired by [5, 46], we enhance the attention scores corresponding to the tokens describing spatial relations to improve spatial expressions in diffusion process. However, only enhancing the attention of spatial tokens while ignoring the instance tokens leads to unclear textures and geometric inconsistency. To solve this issue, we propose Adaptive Score Distillation Sampling (ASDS) to not only enhance the attention of spatial token but also modulate the cross attention map of instance token according to the segmentation region, which can be represented as:

$$A_i = softmax(\frac{Q_i K_i^T}{\sqrt{d}}), \tag{8}$$

$$A_i' := \begin{cases} k \cdot A_i, & \text{if} \quad i = i_{spa} \\ A_i \cdot M_i + \frac{A_i}{k}, \cdot (1 - M_i), & \text{if} \quad i = i_{ins} \\ A_i, & \text{otherwise} . \end{cases}, \tag{9}$$

where A_i and $A_i^{'}$ individually represent the original and modulated cross-attention maps corresponding to the i-th token. i_{spa} and i_{ins} are spatial token and instance token respectively. While M_i indicates the mask region of i-th token, and k is a constant, which is set to 1.25 in experiment.

Sole ASDS distills appearance feature and spatial layout information from the pre-trained text-to-image diffusion model, which cannot guarantee the consistency of multiple viewpoints, leading to the Janus problem. To alleviate this issue, we introduce Multi-View Adaptive Score Distillation Sampling (MV-ASDS) to distill cross-view prior from the multi-view diffusion model, which ensures the multi-view consistency of the generated compositional 3D scene. Unlike previous methods, we apply MV-ASDS not only to the compositional 3D scene but also to individual entity to further enhance the multi-view consistency. Similar to SDS, we render an image from the 3D representation at viewpoint v, then add random Gaussian noise to it. Subsequently, utilize Stable Zero123 ϵ_{ϕ^*} to predict noise and denoise. Then minimize the predicted noise $\epsilon_{\phi^*}(x_t; y, t, v)$ with the added noise to align the rendered image and the reference image at viewpoint v. The process can be expressed as follows:

$$\nabla_\theta \mathcal{L}_{\text{MV-ASDS}}(\epsilon_{\phi^*}, \theta, v, y) \;=\; \mathbb{E}_{t,\epsilon,v}\Big[w(t)\big(\epsilon_{\phi^*}(x_t; y, t, v) \;-\; \epsilon\big)\frac{\partial g(\theta, v)}{\partial \theta}\Big], \tag{10}$$

where $\epsilon_{\phi^*}(x_t; y, t, v)$ is the predicted noise calculated with the modulated attention maps which focus on the spatial words and instance tokens.

When applying MV-ASDS to every entity, we scale the 3D bounding box of an entity Gaussian θ_i with a global volumetric space B_{global}. Then we compute the shift parameter β and scale parameters γ to realize the transformation for the center position:

$$\beta = Mean(B_i); \gamma = B_{global}/B_i. \tag{11}$$

4 Experiments

4.1 Experiment Setting

Implementation details. Our I^{2}C-3D is implemented by ThreeStudio [10]. We leverage GLIGEN [20] to generate reference image and pre-trained LGM [31] is used to generate individual 3D object. Zero-123 is used to realize MV-ASDS and the guidance timestep is set to 50. We optimize Gaussian primitives through $\mathcal{L}_{\mathcal{I}^2\mathcal{C}}$ for 1000 iteration and the hyperparameter $\lambda_{\mathcal{IB}}$ and $\lambda_{\mathcal{IC}}$ are set to 0.15 and 1.0 respectively. Moreover, the threshold in In-Box Loss is set to 0.95.

Dataset. For a fair comparison, we utilize the dataset provided by ComboVerse [4] for evaluation, which consists of 50 textual descriptions, 50 corresponding real images, and 50 images generated by Stable Diffusion. We utilize the textual data to evaluate the performance in text-to-3D and use real images to evaluate the generation capability in single image-to-3D generation.

Baselines. To enable a more comprehensive evaluation, we conduct experiments on three tasks: text-to-3D, single image-to-3D and compositional 3D generation. For text-to-3D, we compare our method with representative approaches based on NeRF and 3DGS. Specifically, DreamFusion [25], Latent-NeRF [23], sjc [32] and Magic3D [21] adopt NeRF as the 3D representation, while GaussianDreamer [40], gsgen [6], and GraphDreamer [8] are built upon 3DGS. For single-image-to-3D, we select four competitive baselines for comparison, including InstantMesh [38], Hunyuan3D [16], ReconViaGen [1], and TRELLIS [36]. In addition, we compare with two compositional 3D generation methods: Layout-your-3D [45] and ComboVerse [4]. More comparisons with the 3D scene generation method (MIDI [15]) are provided in the **Supplementary Materials**.

Evaluation metrics. Following common practice [11], we choose CLIP-Score [26] to evaluate the semantic alignment between generated 3D assets and text. However, CLIP-Score struggles to assess the geometry consistency and spatial relationships. Therefore, following by [35, 39], we utilize GPT-4V to provide a more comprehensive assessment of 3D assets from four aspects: Prompt Alignment, Spatial Arrangement, Geometric Fidelity and Scene Quality. Additionally, to better evaluate Janus problem, we choose Janus Rate (JR) and consistency accuracy (ACC). JR measures the percentage of inconsistent objects among all generated 3D objects, while ACC denotes the proportion of fully consistent scenes. Moreover, for single image-to-3D, we leverage CLIP-Sim, PSNR, SSIM [34] and LPIPS [44] to evaluate the similarity of the rendered image with

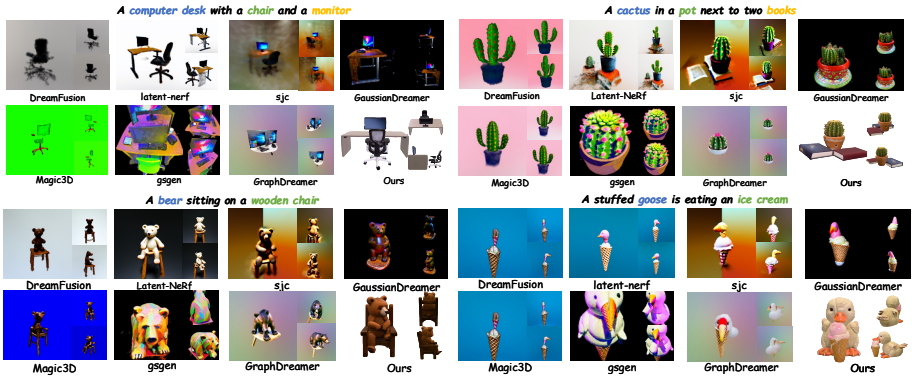


Fig. 4: Qualitative comparison between our approach and previous methods.

Table 1: Quantitative comparisons of I²C-3D and other competing text-to-3d methods on ComboVerse benchmark. The highest and second scores in each section of the table are highlighted in **bold** and underlined respectively.

Method	CLIP-Score ↑	3D	GPT-4V				JR ↓	ACC ↑	Time ↓
			Prompt Alignment ↑	Spatial Arrangement ↑	Geometric Fidelity ↑	Scene Quality ↑			
DreamFusion	0.273	NeRF	50.6	51.3	50.9	28.5	55.6	42	6hours
Latent-NeRF	0.299	NeRF	67.9	70.6	68.4	67.7	58.6	44	**15min**
sjc	0.303	NeRF	65.2	69.7	67.3	71.2	61.1	48	2hours
Magic3D	0.276	NeRF	55.1	53.4	54.5	52.4	49.4	52	5.3hours
GaussianDreamer	0.295	3DGS	78.4	75.3	74.3	67.5	62.5	48	**15min**
gsgen	0.304	3DGS	80.9	72.9	73.7	70.2	64.2	50	3hours
GraphDreamer	0.286	3DGS	78.5	80.3	81.1	79.6	69.7	48	3hours
I²C-3D (Ours)	**0.314**	3DGS	**92.7**	**84.5**	**82.6**	**85.3**	**17.3**	**88**	28min

the reference image and generation quality of 3D asset. Additionally, we also calculate training time as an efficiency metric

.

4.2 Comparisons with Existing Methods

Quantitative results. Table 1 shows quantitative comparison results between our I²C-3D and various SOTA models on the benchmark, which demonstrates our method significantly outperforms previous approaches in Semantic Alignment, Spatial Arrangement, Geometric Consistency and Scene Quality. As for training time, compositional 3D generation needs longer time to optimize all objects, which is related with the number of objects. Despite a slightly longer training time than Latent-NeRF and GaussianDreamer, both qualitative and quantitative results indicate that our I²C-3D improve generation quality by a large margin. In addition, we provide more quantitative experiments compared with single image-to-3D methods in **Supplementary Materials**, which indicates that our proposed I²C-3D performs well in CLIP-Sim and PSNR, and achieves competitive results in SSIM and LPIPS.

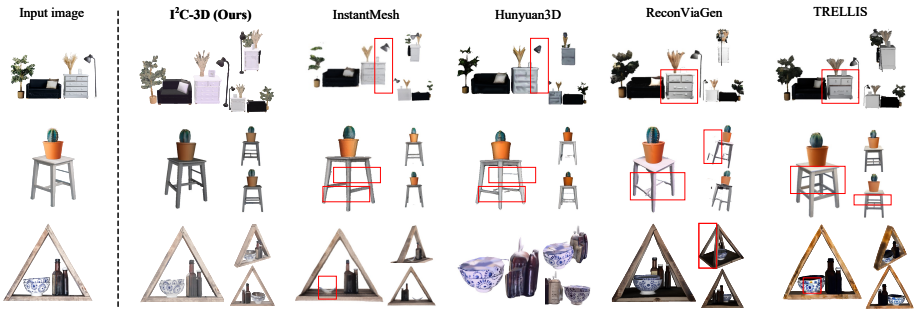


Fig. 5: Qualitative comparisons on single image-to-3D generation task.

Qualitative results. As shown in Fig. 4, our I^{2}C-3D can generate compositional 3D assets that not only enable reasonable interaction regions, but also maintain multi-view consistency. However, other methods struggle to generate interaction regions and suffer from the Janus Problem. For example, DreamFusion [25] fails to generate multiple objects based on given complex text, mainly due to the difficulty of SDS in optimizing multiple objects. While GaussianDreamer [40], Magic3D [21] and sjc [32] can improve the appearance of the 3D assets but only generate partial objects of given text. In addition, gsgen [6] and GraphDreamer [40] suffer from multi-view inconsistency and unreasonable interaction regions. Moreover, Latent-NeRF [23] and sjc [32] can generate good geometric structures by introducing the mesh prior, but produce artifacts and messy spatial relationships. In contrast, I^{2}C-3D can generate high-quality geometry and texture for each object while preserving good spatial relationships.

In single image-to-3D task, other methods suffer from object incompleteness and cross-view inconsistency (shown in Fig. 5). For instance, in the first example, InstantMesh and Hunyuan3D generate partially missing light structures, while cabinets generated by ReconViaGen and TRELLIS are both unfaithful. In the last example, ReconViaGen emerges severe multi-view inconsistency, InstantMesh and Hunyuan3D fail to reconstruct the entire scene, and the appearance of the bowl generated by TRELLIS is inaccurate. On the contrary, our method reconstructs geometrically consistent 3D assets with reasonable interaction, faithfully preserving object appearance as well as the spatial relationships depicted in the input image.

Moreover, we compared our method with open-source compositional 3D generation methods. As displayed in Fig. 6, the interaction area in our generated 3D assets is more reasonable. For instance, the wearing relationship between the tie and the beggle is expressed more naturally. Meanwhile, we also present some challenging interactive cases in Fig. 7 (e.g., more humanized interactions), which demonstrates that our method achieves competitive results in complex interaction tasks. What's more, we provide more visual comparisons in the **supplementary material** and support video material to display our generation results, you can click to watch the results more intuitively.

Fig. 6: Comparisons between our I^{2}C-3D and compositional 3D generation methods.

Fig. 7: Challenging cases that contain complex interactions.

User Study. We conducte a questionnaire to explore users' preference through scoring our method and other baselines across 20 prompts including more than 50 objects in terms of Prompt Alignment, Spatial Arrangement, Geometric Fidelity, Scene Quality and Multi-View Consistency. The visualization of human preference are shown in Fig. 8, which shows that our model significantly outperforms other methods in above five aspects.

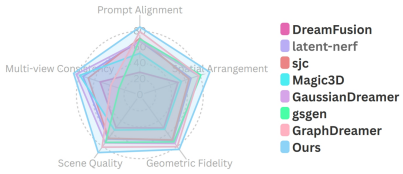


Fig. 8: Human preference results.

4.3 Extended Applications: 3D Editing

As illustrated in Fig. 9, our approach supports progressive 3D editing for compositional scenes by incrementally inserting new objects according to textual instructions. Starting from a single object prompt (a double bed), our I^{2}C-3D first reconstructs a stable 3D asset. Subsequently, additional objects (nightstand, flowerpot and apple) can be inserted step by step. Importantly, each newly added object is integrated while preserving the geometric structure, spatial relationships and interaction consistency of the existing scene. This progressive compositional editing capability demonstrates that our method is not limited to one-shot generation, but can flexibly support iterative scene generation.

Fig. 9: Progressive 3D Editing examples of our I^{2}C-3D.

4.4 Ablation Study

Effectiveness of Multi-View Adaptive Score Distillation Sampling. Results shown in Fig. 10 indicate the poor multi-view consistency and the disappeared texture appearance when discarding the MV-ASDS. For example, a pillow with a round front becomes square on the sides and the texture appearance of objects is missing.

Effectiveness of Inclusive Interactive Collisions. Without I^{2}C, the generated 3D asset is not full and will have many outlier Gaussian primitives and the interaction regions are expressed ambiguously. As shown in column 2 of Fig. 10, there are black artifacts in the interaction region between the pillow and stool.

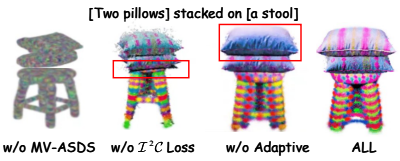


Fig. 10: Ablations on key components.

Effectiveness of Adaptive Score Distillation Sampling. It can also be observed in the column 3 of Fig. 10, without ASDS, the generated 3D asset lacks texture details. For instance, the floral detail on the pillows is missing and the texture of the stool is blurred.

Hyperparameter Ablation on τ. We conduct an ablation study on the interaction margin hyperparameter τ, which controls the minimum separation distance between object bounding boxes during optimization. In our implementation, τ is set to 5% of the maximum diagonal of bounding box, which provides a scale-adaptive margin across different scenes. The visual ablation result is shown in Fig. 11. When τ is too small ($\tau = 0.02$), the margin constraint becomes insufficient, leading to noticeable object penetration and unrealistic physical contact. When τ is too large ($\tau = 0.5$), the enforced separation becomes overly strong, causing visible gaps between interacting objects. In contrast, a moderate value ($\tau = 0.2$)

Fig. 11: Ablation on τ.

yields stable contact relationships and physically plausible interactions. More quantitative ablation results are available in the **supplementary materials**.

4.5 More experiments in challenging scenarios

We conduct more experiments to further evaluate the robustness of our I^{2}C-3D under more challenging scenarios. As illustrated in Fig. 12(a), we progressively increase scene complexity by inserting additional objects. Starting from a input with 3 object, our I^{2}C-3D successfully extends the scene to 4 and then 6 objects while preserving spatial relationships and interaction coherence. Despite the increased object count and potential occlusions, the generated 3D assets remain geometrically stable and semantically consistent with the input instructions. What's more, the generation results under real-world and generated scenarios are shown in Fig. 12(b), which further demonstrates our I^{2}C-3D maintains stable interaction quality and cross-view consistency in challenging scenarios.

Fig. 12: More experiments in challenging scenarios.

5 Conclusion

In this paper, we propose a novel framework for multi-view consistent compositional 3D generation, termed I^{2}C-3D. Unlike existing methods that suffer from unreasonable interaction generation, our I^{2}C-3D can generate physically plausible and visually coherent interaction region. Specifically, we propose an Inclusive Interactive Collisions strategy to constrain Gaussian primitives in the interaction region appearing at reasonable interaction locations naturally for interaction region generation. Additionally, Multi-View Adaptive Score Distillation Sampling is designed to distill multi-view consistency prior and layout prior from pretrained diffusion model by adaptive attention modulation across viewpoints, further achieving multi-view consistency. Extensive experiments indicate that our I^{2}C-3D outperforms existing methods in terms of multi-view consistency and rational interaction generation.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62376285 and 61673396), the Innovation Fund Project for Graduate Students at China University of Petroleum (East China), and the Fundamental Research Funds for the Central Universities (No. 26CX04038A).

References

1. Chang, J., Ye, C., Wu, Y., Chen, Y., Zhang, Y., Luo, Z., Li, C., Zhi, Y., Han, X.: Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. arXiv preprint arXiv:2510.23306 (2025)
2. Chen, T., Ding, C., Zhang, S., Yu, C., Zang, Y., Li, Z., Peng, S., Sun, L.: Rapid 3d model generation with intuitive 3d input. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12554–12564 (2024)
3. Chen, Y., Pan, Y., Yang, H., Yao, T., Mei, T.: Vp3d: Unleashing 2d visual prompt for text-to-3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4896–4905 (2024)
4. Chen, Y., Wang, T., Wu, T., Pan, X., Jia, K., Liu, Z.: Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In: European Conference on Computer Vision. pp. 128–146. Springer (2024)
5. Chen, Y., Wang, T., Wu, T., Pan, X., Jia, K., Liu, Z.: Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In: European Conference on Computer Vision. pp. 128–146. Springer (2024)
6. Chen, Z., Wang, F., Wang, Y., Liu, H.: Text-to-3d using gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21401–21412 (2024)
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
8. Gao, G., Liu, W., Chen, A., Geiger, A., Schölkopf, B.: Graphdreamer: Compositional 3d scene synthesis from scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21295–21304 (2024)
9. Ge, C., Xu, C., Ji, Y., Peng, C., Tomizuka, M., Luo, P., Ding, M., Jampani, V., Zhan, W.: Compgs: Unleashing 2d compositionality for compositional text-to-3d via dynamically optimizing 3d gaussians. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18509–18520 (2025)
10. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., et al.: threestudio: A unified framework for 3d content generation (2023)
11. Han, H., Yang, R., Liao, H., Xing, J., Xu, Z., Yu, X., Zha, J., Li, X., Li, W.: Reparo: Compositional 3d assets generation with differentiable 3d layout alignment. arXiv preprint arXiv:2405.18525 (2024)
12. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)

13. Hu, T., Li, L., van de Weijer, J., Gao, H., Shahbaz Khan, F., Yang, J., Cheng, M.M., Wang, K., Wang, Y.: Token merging for training-free semantic binding in text-to-image synthesis. In: Advances in Neural Information Processing. vol. 37, pp. 137646–137672 (2024)
14. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023)
15. Huang, Z., Guo, Y.C., An, X., Yang, Y., Li, Y., Zou, Z.X., Liang, D., Liu, X., Cao, Y.P., Sheng, L.: Midi: Multi-instance diffusion for single image to 3d scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23646–23657 (2025)
16. Hunyuan3D, T., Yang, S., Yang, M., Feng, Y., Huang, X., Zhang, S., He, Z., Luo, D., Liu, H., Zhao, Y., et al.: Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. arXiv preprint arXiv:2506.15442 (2025)
17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
19. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214 (2023)
20. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
21. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023)
22. Liu, Y., Li, X., Li, X., Qi, L., Li, C., Yang, M.H.: Pyramid diffusion for fine 3d large scene generation. In: European Conference on Computer Vision. pp. 71–87. Springer (2024)
23. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12663–12673 (2023)
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
25. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Rombach, et al.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

28. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing. vol. 34, pp. 6087–6101 (2021)
29. Song, et al.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
30. Sun, Y., Zhang, D., Zhu, J., Cheng, H., Li, Z., Li, P., Fang, C., Dong, Y., Chen, L.: Tri-efficient transfer learning for point cloud videos (2026), `https://arxiv.org/abs/2606.24175`
31. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
32. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12619–12629 (2023)
33. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In: Advances in Neural Information Processing. vol. 36, pp. 8406–8441 (2023)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
35. Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., Wetzstein, G.: Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22227–22238 (2024)
36. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)
37. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)
38. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)
39. Yang, L., Zhang, Z., Han, J., Zeng, B., Li, R., Torr, P., Zhang, W.: Semantic score distillation sampling for compositional text-to-3d generation. arXiv preprint arXiv:2410.09009 (2024)
40. Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6796–6807 (2024)
41. Yu, X., Guo, Y.C., Li, Y., Liang, D., Zhang, S.H., Qi, X.: Text-to-3d with classifier score distillation. arXiv preprint arXiv:2310.19415 (2023)
42. Zhang, Q., Wang, C., Siarohin, A., Zhuang, P., Xu, Y., Yang, C., Lin, D., Zhou, B., Tulyakov, S., Lee, H.Y.: Towards text-guided 3d scene composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6829–6838 (2024)

43. Zhang, Q., Wang, C., Siarohin, A., Zhuang, P., Xu, Y., Yang, C., Lin, D., Zhou, B., Tulyakov, S., Lee, H.Y.: Towards text-guided 3d scene composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6829–6838 (2024)
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
45. Zhou, J., Li, X., Qi, L., Yang, M.H.: Layout-your-3d: Controllable and precise 3d generation with 2d blueprint. arXiv preprint arXiv:2410.15391 (2024)
46. Zhou, J., Li, X., Qi, L., Yang, M.H.: Layout-your-3d: Controllable and precise 3d generation with 2d blueprint. arXiv preprint arXiv:2410.15391 (2024)
47. Zhou, X., Ran, X., Xiong, Y., He, J., Lin, Z., Wang, Y., Sun, D., Yang, M.H.: Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. arXiv preprint arXiv:2402.07207 (2024)