Interact3D: Compositional 3D Generation of Interactive Objects

Hui Shan[1,2,3], Keyang Luo[†], Ming Li[1,2,3], Sizhe Zheng[2,3], Yanwei Fu[2,5], Zhen Chen[4], and Xiangru Huang[3⋆]

[1] Zhejiang University, China
[2] Shanghai Innovation Institute, China
[3] Westlake University, China
[4] Adobe, United States
[5] Fudan University, China
huangxiangru@westlake.edu.cn

Fig. 1: Given a text prompt and a user-given mesh, Interact3D synthesizes a high-quality, geometrically-compatible complementary mesh. It then seamlessly composes these two assets into an interactive 3D scene. Unlike existing approaches, our framework automatically generates collision-aware and physically-sound 3D environments.

Abstract. Recent breakthroughs in 3D generation have enabled the synthesis of high-fidelity individual assets. However, generating 3D compositional objects from single images–particularly under occlusions–remains challenging. Existing methods often degrade geometric details in hidden regions and fail to preserve the underlying object-object spatial relationships (OOR). We present a novel framework **Interact3D** designed to generate physically plausible interacting 3D compositional objects.

⋆ Corresponding author.
† Independent researcher from China.

Our approach first leverages advanced generative priors to curate high-quality individual assets with a unified 3D guidance scene. To physically compose these assets, we then introduce a robust two-stage composition pipeline. Based on the 3D guidance scene, the primary object is anchored through precise global-to-local geometric alignment (registration), while subsequent geometries are integrated using a differentiable Signed Distance Field (SDF)-based optimization that explicitly penalizes geometry intersections. To reduce challenging collisions, we further deploy a closed-loop, agentic refinement strategy. A Vision-Language Model (VLM) autonomously analyzes multi-view renderings of the composed scene, formulates targeted corrective prompts, and guides an image editing module to iteratively self-correct the generation pipeline. Extensive experiments demonstrate that **Interact3D** successfully produces promising collision-aware compositions with improved geometric fidelity and consistent spatial relationships. The code and dataset will be available at `https://github.com/SII-Hui/Interact3D`.

Keywords: 3D Generation · Interaction · Agentic Refinement

1 Introduction

Robotic manipulation shows immense potential in domestic assistance, industrial automation, and disaster response. In recent years, learning-based manipulation paradigms, particularly Reinforcement Learning (RL) [20, 33, 43], have achieved remarkable progress. However, deploying these algorithms directly in the real world is expensive and unsafe. Therefore, training agents in physically simulated virtual interactive environments (Sim2Real) has become a dominant paradigm [21, 22, 38]. To generalize to diverse real-world scenarios, these simulation platforms require a massive and diverse collection of interactive 3D assets with realistic geometry, plausible physical properties, and correct object–object relationships (OOR).

However, the availability of high-quality interactive 3D data remains severely limited. Curating these 3D datasets heavily relies on tedious manual modeling and annotation, making it difficult to scale. To bypass the scarcity of interactive 3D assets, recent works attempt to leverage large-scale human videos as an alternative supervision signal for robot learning [10, 13, 45]. While video-driven approaches provide rich semantic priors and action trajectories, 2D videos inherently lack precise 3D geometric grounding and physical collision constraints. This fundamental limitation hinders the robot's ability to comprehend complex spatial relationships and perform fine-grained interactions in 3D space.

Concurrently, 3D generation models have achieved impressive results in synthesizing high-fidelity 3D assets from text or image prompts [35, 36, 48]. However, these models typically output monolithic and fused geometries which lack OOR. The generated meshes are "baked" into a single geometry, severely lacking independent geometry and physically meaningful spatial relationships. Naively segmenting such geometry, for instance, directly using PartField [16], often in-

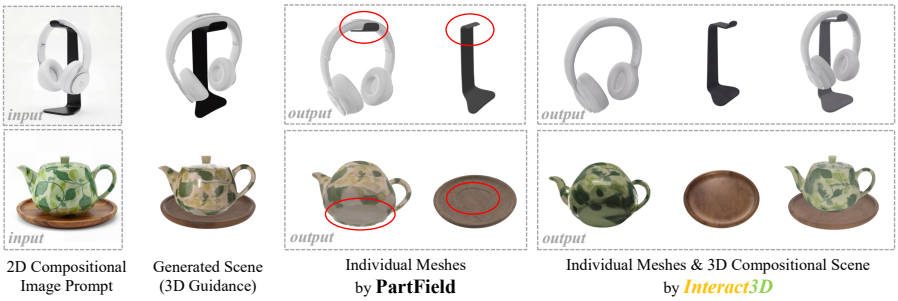

Fig. 2: Given a generated scene, **PartField** directly segments it to extract individual 3D assets, whereas **Interact3D** leverages it as 3D spatial guidance to compose high-quality independent geometries. (Though generating scenes and meshes via separate TRELLIS2 inferences yields minor texture variations, their spatial guidance remains robust to these discrepancies.)

troduces difficult-to-repair holes and inaccurate 3D assets (see Figure 2). Hunyuan3D [30] and Rodin [4] commercial websites can segment scenes and inpaint geometries, but suffer from the loss of the original texture in their remeshed geometry. Attempting to manually segment and optimize these monolithic assets is extremely labor-intensive and fails to generalize across diverse topologies. Therefore, the key challenge lies in how to reliably generate geometrically-compatible 3D compositional assets guided by user-given text or image prompts. An intuitive solution is to train a feed-forward network using large amounts of 3D compositional data [12, 41]. However, feed-forward methods are inherently bottlenecked by the scarcity of 3D data, resulting in poor generalization and high training costs.

To address these challenges, we propose **Interact3D**, a novel pipeline for scalable and automated synthesis of interactive 3D scenes, see Figure 3. Our key insight is to leverage the spatial priors implicitly encoded in image-guided 3D generation as guidance for geometric composition. By fully exploiting the priors of advanced 2D and 3D generative models, we reformulate compositional generation—traditionally requiring complex geometric reasoning—as a structured 3D registration problem.

However, unlike traditional 3D registration, compositional generation requires not only accurate alignment but also physically plausible interactions between objects. To this end, we first anchor the primary object using a global-to-local registration strategy, and then integrate additional components through a Signed Distance Field (SDF)-based optimization that explicitly penalizes interpenetrations, resulting in collision-aware compositions.

Moreover, inherent 2D occlusions often cause generated assets with complex joints to suffer from severe geometric mismatches, which cannot be fixed by simple rigid transformations. To address this, we propose a VLM-based agentic refinement strategy that formulates corrective text prompts to guide 2D generative image editing, thereby iteratively fixing the geometric flaws. Extensive

experiments demonstrate the effectiveness of our methods in multi-object composition (see Figure 1). Our contributions are summarized in the following:

1. We propose **Interact3D**, a generate-then-compose framework for training-free compositional 3D generation, featuring a two-stage, extensible composition pipeline, that enables the creation of collision-aware interactive 3D scenes.
2. We introduce an interactive 3D dataset, which contains more than 8,000 interactive pairs and will be publicly released to facilitate future research.
3. We develop a VLM-based agentic refinement strategy that automatically resolves severe geometric mismatches and improves compositional coherence.

2 Related Work

3D Generation. Recent advances in 3D generation have increasingly shifted away from slow, optimization-based paradigms [1, 15, 17, 26, 47] to large-scale foundational models [14, 29, 35, 36, 46, 48], drastically improving both synthesis speed and geometric fidelity. CLAY [46] has advanced native 3D geometry generation, using 3D datasets from Objaverse [5] to produce highly controllable, realistic, and watertight assets. Hunyuan3D [48] achieves industrial-grade asset quality by effectively unifying text-to-3D and image-to-3D pipelines within robust large-scale flow-based diffusion transformers. TRELLIS [36] introduces a unified structured latent representation (SLAT) and rectified flow transformers that significantly mitigate multi-view inconsistencies and artifacts in complex geometries. Building upon this foundation, its recent successor, TRELLIS2 [35] further scales the transformer architecture within the 3D latent space, achieving state-of-the-art in high-resolution 3D generation.

3D Part Segmentation and Generation. The segmentation and generation of 3D parts, which are essential tasks in 3D generation, require a compositional understanding of the structures of 3D objects. SAMPart3D [39], PartField, and P^3-SAM [19] capture 3D features through feed-forward training and perform segmentation. Although they can segment parts generated by advanced 3D generation models, the resulting components often contain difficult-to-repair holes. In contrast, OmniPart [40] and SAM3D [2] are capable of directly generating complete 3D parts from masked images through expensive data curation and feed-forward training. Although the meshes they generate are watertight, their geometric accuracy, PBR materials, and alignment with image prompts still fall short when compared to 3D foundation models like TRELLIS2.

3D Shape Composition. 3D shape composition [37, 44] typically restores complete objects or scenes from multiple segmented components. Jigsaw [18], NSM [3] and RGL-NET [23] focus on fragment composition without relying on predefined semantic information, as fragments often contain complex and precise geometric details. Based on these, Puzzlefusion [11], Diffassemble [28], and Puzzlefusion++ [32] use diffusion models to refine the poses of fragments.

2BY2 [24] is the first to propose the daily pairwise object composition task and contributes the corresponding dataset for the task. Compared to fragment composition, daily pairwise components always lack clear geometric relationships. Therefore, COPY-TRANSFORM-PASTE [8] utilizes text-guided optimization with vision-language models to supervise composition and achieves promising results. Although 2BY2 introduces a daily composition dataset, it consists of only 517 pairs, most of which are selected from existing 3D datasets [5, 34]. In contrast, we propose a fully automated pipeline for generating large-scale brand-new 3D composition datasets.

3 Preliminary

Our method is designed for physically-sound 3D compositional generation. In this section, we first formally define the task of 3D compositional generation (section 3.1), then briefly formulate 3D point cloud registration task and introduce one classical solution, Iterative Closest Point (ICP) (section 3.2).

3.1 Problem Setup

Given a 3D mesh $\mathbf{M} = (\mathbf{V}, \mathbf{F})$, where $\mathbf{V} \in \mathbb{R}^{N\times 3}$ denotes the vertex positions and $\mathbf{F} \in \mathbb{N}^{N_f\times 3}$ denotes the face indices, together with a text prompt that specifies compositional semantics, our goal is to generate a complementary 3D component $\mathbf{M}_{\text{comp}}$ that forms a coherent compositional scene with $\mathbf{M}$.

In addition to synthesizing $\mathbf{M}_{\text{comp}}$, we optimize the transformation parameters $\boldsymbol{\theta} = (\boldsymbol{\tau}, \mathbf{R}, \mathbf{s})$ between both meshes to achieve geometrically compatible composition, where $\boldsymbol{\tau} \in \mathbb{R}^3$ denotes translation, $\mathbf{R} \in \mathcal{SO}(3)$ denotes rotation, and $\mathbf{s} \in \mathbb{R}^+$ denotes uniform scaling.

3.2 3D Point Clouds Registration

Given source and target point clouds $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^N$ and $\mathcal{Q} = \{\mathbf{q}_j\}_{j=1}^M$, we adopt *scale-aware ICP* [42] for similarity registration. The objective jointly estimates a uniform scale $s \in \mathbb{R}^+$, rotation $\mathbf{R} \in \mathcal{SO}(3)$, and translation $\boldsymbol{\tau} \in \mathbb{R}^3$:

$$\min_{\mathbf{s},\mathbf{R},\boldsymbol{\tau}} \sum_{(\mathbf{p}_i,\mathbf{q}_j)\in C} \|\mathbf{s}\cdot\mathbf{R}\cdot\mathbf{p}_i + \boldsymbol{\tau} - \mathbf{q}_j\|_2^2, \tag{1}$$

where C denotes point correspondences. At each iteration, correspondences are updated using the scheme mentioned in [42], and the optimal similarity transformation $(\mathbf{s}, \mathbf{R}, \boldsymbol{\tau})$ is computed in closed form using Umeyama's method [31].

4 Method

We present a comprehensive framework for generating physically-sound 3D shape components. After curating high-quality individual assets and a 3D guidance

scene, we process them through a novel two-stage composition pipeline (Section 4.1). A global-to-local geometric alignment (Section 4.2) is used to accurately anchor the reference object in stage one. A SDF-based physics-aware optimization (Section 4.3) is introduced to align subsequent objects while avoiding spatial intersections in stage two. For cases of severe and unavoidable collisions, we integrate an agentic, VLM-driven refinement loop (Section 4.4) to iteratively edit the scene until a low-collision configuration is achieved. The main notations are included in Table 1.

Table 1: Summary of main notations.

Symbol	Description
$\mathbf{M}, I_{\text{rendered}}$	Input 3D mesh, and corresponding rendered image
$I_{\text{scene}}, I_{\text{comp}}$	Generated compositional and complementary image
$\mathbf{M}_{\text{scene}}, \mathbf{M}_{\text{comp}}$	Scene and complementary mesh reconstructed from I_{scene} and I_{comp}
$\mathbf{M}_{\text{anchor}}, \mathbf{M}_{\text{remain}}$	Anchor mesh and remaining mesh selected from $\{\mathbf{M}, \mathbf{M}_{\text{comp}}\}$
$\mathbf{M}', \mathbf{M}'_{\text{comp}}$	Segmented meshes extracted from $\mathbf{M}_{\text{scene}}$
$\mathbf{M}'_{\text{anchor}}, \mathbf{M}'_{\text{remain}}$	Segmented counterparts from $\mathbf{M}_{\text{scene}}$
$\boldsymbol{\theta}(\mathbf{p}) = \mathbf{s} \cdot \mathbf{R}\mathbf{p} + \boldsymbol{\tau}$	Transformation function in terms of translation $\boldsymbol{\tau}$, rotation $\mathbf{R}$, scale $\mathbf{s}$
$\Phi_{\mathbf{M}}(\mathbf{p})$	Signed distance from point $\mathbf{p}$ to mesh $\mathbf{M}$

4.1 Data Curation and Workflow

We begin by rendering the input mesh $\mathbf{M}$ from a canonical frontal viewpoint, obtaining an image I_{rendered}. Conditioning on this rendered image and a user-provided textual prompt, we use *Nano Banana Pro* [9] to generate a compositional scene image I_{scene}. To obtain the complementary component, we further prompt the model to remove the original object from I_{scene}, producing an image I_{comp} that contains only the newly introduced geometry. Intuitively, I_{comp} captures the complementary part needed to complete the composition.

Subsequently, we use *TRELLIS2* to generate $\mathbf{M}_{\text{scene}}$ and $\mathbf{M}_{\text{comp}}$ meshes separately based on images I_{scene} and I_{comp}. To recover the spatial relationship between components, we apply *PartField* to segment $\mathbf{M}_{\text{scene}}$ into two parts: $\mathbf{M}'$ and $\mathbf{M}'_{\text{comp}}$. Although these segmented meshes often contain holes and geometric artifacts that make them unsuitable as final assets, they still provide reliable spatial cues. In particular, the transformation $\boldsymbol{\theta}$ extracted from $\mathbf{M}'$ and $\mathbf{M}'_{\text{comp}}$ serves as geometric guidance for composing the original mesh $\mathbf{M}$ and the high-quality complementary mesh $\mathbf{M}_{\text{comp}}$ through 3D point cloud registration (see Figure 3).

Two-stage Composition Pipeline. We decompose composition into two sequential stages with distinct roles.

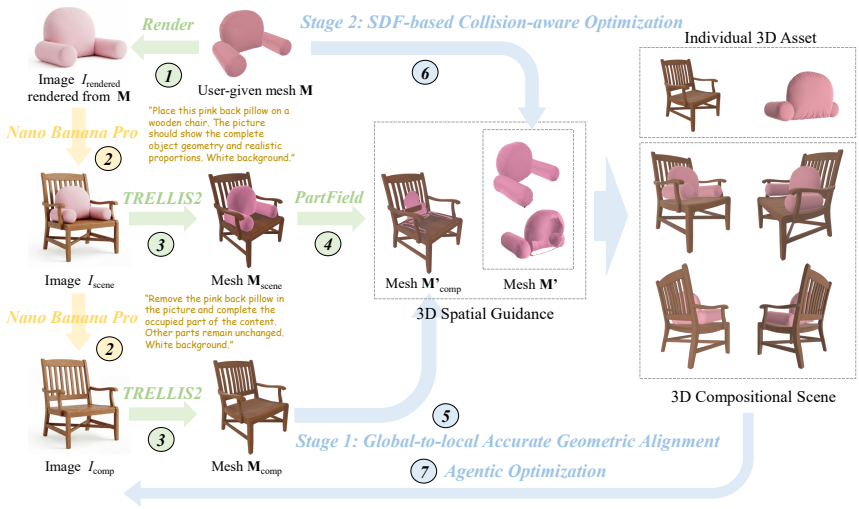

Fig. 3: Overview of Interact3D. Given a user-provided mesh $\mathbf{M}$, we render it and use Nano Banana Pro to synthesize a guided scene image I_{scene} and a complementary image I_{comp}. TRELLIS2 then reconstructs these into 3D meshes ($\mathbf{M}_{\text{scene}}$ and $\mathbf{M}_{\text{comp}}$) to provide spatial guidance. Relying on coarse parts segmented by PartField, we execute a two-stage composition. Stage 1 performs a global-to-local registration on the "largest" mesh to establish the anchor pose, while Stage 2 applies an SDF-based collision-aware optimization on the remaining mesh to resolve spatial intersections. (In this case, $\mathbf{M}_{\text{comp}}$ is considered as $\mathbf{M}_{\text{anchor}}$, $\mathbf{M}$ is considered as $\mathbf{M}_{\text{remain}}$.) Finally, a VLM-based agentic refinement handles unavoidable collisions, yielding a physically-sound interactive 3D scene.

Stage 1: Anchor Alignment. We first select an anchor object $\mathbf{M}_{\text{anchor}}$ from the candidate meshes ($\mathbf{M}$ or $\mathbf{M}_{\text{comp}}$). Empirically, the object with the larger projected area (measured by 2D bounding box size) in I_{scene} provides a more stable reference. We apply the global-to-local registration strategy (Section 4.2) exclusively to this anchor object to recover an accurate initial pose with respect to the guidance mesh $\mathbf{M}'_{\text{anchor}}$.

Stage 2: Collision-aware Composition. Once the anchor object is fixed, we optimize the pose of the remaining component(s) $\mathbf{M}_{\text{remain}}$ using the SDF-based collision-aware formulation (Section 4.3) with the guidance $\mathbf{M}'_{\text{remain}}$. This stage explicitly balances geometric alignment with collision avoidance, ensuring physically plausible composition.

4.2 Global-to-local Accurate Geometric Alignment

This stage is applied only to the selected anchor object mentioned in Section 4.1. Its objective is to recover a reliable initial pose under partial overlap, occlusion, and reconstruction noise.

In practice, the generated compositional image I_{scene} inevitably introduces occlusions, and the individual meshes are not perfectly consistent with $\mathbf{M}_{\text{scene}}$. Moreover, the subsequent segmentation process in PartField often produces structural holes and incomplete surfaces (see Figure 2). As a result, the resulting point cloud pairs ($\mathbf{M}_{\text{anchor}}$ and $\mathbf{M}'_{\text{anchor}}$) typically exhibit low-overlap ratios and contain a significant number of outlier points. Under such conditions, traditional local registration methods, such as ICP, are highly sensitive to initialization and may converge to undesirable local minima.

To effectively address the aforementioned challenges, we adopt a robust and accurate global-to-local registration paradigm. We begin by resolving the scale discrepancies using an Oriented Bounding Box (OBB) approach to estimate the geometric extents and compute the initial scale factor $\mathbf{s}$. Following this scale alignment, we employ GeoTransformer [25], a transformer-based 3D registration network, which is exceptionally robust in handling geometries with low-ratio overlaps. This step provides reliable global estimates of the initial global translation $\boldsymbol{\tau}$ and rotation $\mathbf{R}$. Finally, with these global estimates of $(\mathbf{s}, \mathbf{R}, \boldsymbol{\tau})$ as initialization, a scale-aware ICP algorithm is deployed to achieve precise alignment. In summary, this global initialization provides a warm start that prevents the algorithm from falling into local minima, allowing the subsequent local optimization to focus on minimizing residual geometric discrepancies for an accurate final alignment.

4.3 SDF-Based Collision-Aware Composition

After global-to-local registration, the primary object is fixed as the spatial anchor. The remaining components $\mathbf{M}_{\text{remain}}$ are then placed relative to this anchor. While we can use a similar global-to-local alignment method for the rest of the components to provide a reasonable pose estimate, it does not guarantee physically valid placement. Small geometric discrepancies may lead to interpenetration or unnatural spacing between objects.

To refine the placement, we introduce a SDF-guided optimization stage. Given the anchor mesh $\mathbf{M}_{\text{anchor}}$, we precompute its Signed Distance Field (SDF) $\Phi_{\text{anchor}}(\mathbf{p})$, which encodes the signed distance from a query point $\mathbf{p}$ to the surface of $\mathbf{M}_{\text{anchor}}$ and negative values indicate penetration into the anchor object.

Following our global-to-local approach, we first obtain the initial transformation $\boldsymbol{\theta}$ between $\mathbf{M}_{\text{remain}}$ and $\mathbf{M}'_{\text{remain}}$ using OBB and GeoTransformer as a starting point. For any point $\mathbf{p}$ in $\mathbf{M}_{\text{remain}}$, we define the collision loss $\mathcal{L}_{\text{col}}$ using the ReLU operator $[\cdot]_+$ after transformation $\boldsymbol{\theta}$:

$$\mathcal{L}_{\text{col}}(\boldsymbol{\theta}) = \sum_{\mathbf{p}\in\mathbf{M}_{\text{remain}}} \left(\underbrace{[-\Phi_{\text{anchor}}(\boldsymbol{\theta}(\mathbf{p}))]_+^2}_{\text{Hard Penalty}} + \lambda \cdot \underbrace{[\epsilon - \Phi_{\text{anchor}}(\boldsymbol{\theta}(\mathbf{p}))]_+}_{\text{Soft Repulsion}} \right). \quad (2)$$

Here, ϵ denotes the safety margin and λ is set to a minimal value to ensure a smooth transition of the penalty field near the boundary of the object. The final

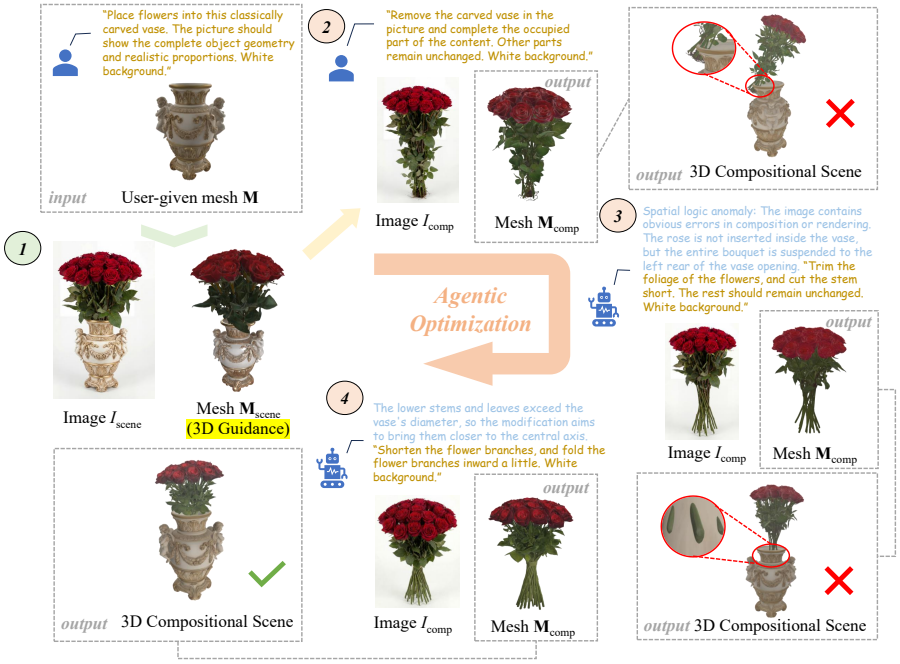

Fig. 4: Agentic Refinement. When generating flowers inside a vase, severe occlusion in image I_{scene} causes the loss of object-object spatial relationships during TRELLIS2 reconstruction. As a result, the flower stem in the complementary image I_{comp} (generated by Nano Banana Pro) may not align correctly with the vase mesh $\mathbf{M}$. This leads to geometric intersections after composition (see *top-right* image). To resolve this, we render multi-view images, including internal cross-sections. These renderings are analyzed by a VLM, which generates a corrective text prompt to update the complementary image via Nano Banana Pro. This process continues until no more geometric intersections are found or the maximum iteration limit is reached.

placement is obtained by minimizing:

$$\min_{\boldsymbol{\theta}} \sum_{p \in \mathbf{M}_{\text{remain}}} \|\boldsymbol{\theta}(\mathbf{p}) - \mathbf{p}'\|^2 + \beta^{(k)} \mathcal{L}_{\text{col}}(\boldsymbol{\theta}), \tag{3}$$

where the first term preserves the alignment between $\mathbf{M}_{\text{remain}}$ and $\mathbf{M}'_{\text{remain}}$, and $\beta^{(k)}$ balances geometric consistency and collision avoidance. Specifically, $\beta^{(k)}$ is initialized as 0, so that it is equivalent to the scale-aware ICP solver at the beginning and increases linearly to β_{max} throughout the optimization process. At the k-th optimization stage, $\mathbf{p}'$ is first chosen from $\mathbf{M}'_{\text{remain}}$ for each $\boldsymbol{\theta}(\mathbf{p})$ following [42], and then optimize Equation (3) with $\mathbf{p}'$ fixed. This schedule allows the optimization to prioritize global geometric alignment in the early stages and progressively enforce physical plausibility as the pose converges.

By treating the anchor as a fixed spatial reference and refining the remaining components through SDF-guided optimization, our method generates geometrically compatible and physically plausible composition.

4.4 Agentic Optimization

Although SDF-based optimization effectively resolves moderate geometric discrepancies, it cannot correct fundamentally incompatible geometries (See Figure 4). In practice, the complementary mesh generated from 2D guidance may deviate significantly from the intended spatial configuration, leading to persistent collisions or implausible object relationships that cannot be resolved through transformation alone.

To address this limitation, we introduce an agentic refinement mechanism that operates at the semantic level. Instead of further adjusting pose parameters, we revise the complementary geometry itself. Specifically, we render the current compositional scene from multiple viewpoints and feed the rendered images, along with the original compositional prompt, into a Vision-Language Model (VLM, e.g., Gemini 3 Pro [9]). The VLM analyzes spatial inconsistencies and generates corrective feedback in the form of targeted editing instructions. These instructions are then used to guide the 2D image editing model to refine the complementary component. The updated 2D image is subsequently reconstructed into a new 3D mesh and re-integrated into the composition pipeline.

This closed-loop process enables semantic-level correction beyond purely geometric optimization. By iteratively combining geometric alignment and language-guided refinement, our framework can resolve severe mismatches that are otherwise difficult to handle with traditional registration or SDF-based methods alone.

5 Experiments

5.1 Experimental Setup

We use Nano Banana Pro to generate images and adopt TRELLIS2 (and partially Hunyuan3D) for image-conditioned 3D generation. Our proposed framework is entirely training-free, and the inference process is conducted on a single NVIDIA H200 GPU. Specifically, we set $\lambda = 0.003$ to smooth the penalty field in Equation (2), set $k_{max} = 100, \beta_{max} = 3.0$ in Equation (3) to progressively enforce physical plausibility. Furthermore, we set the maximum number of iterations in agentic optimization to 5. We empirically find our method to be robust to these hyperparameters within a reasonable range.

Baselines. We evaluate the compositional performance of Interact3D against three categories of baselines.

1. **Feed-forward baselines.** We compare our approach with Jigsaw [18] (for fractured object assembly), 2BY2 [24] (for daily pairwise object composition), and MIDI [12]. Specifically, Jigsaw and 2BY2 take the user-provided

Fig. 5: Results of two-part composition. Data-driven baselines (MIDI, Jigsaw, 2BY2) struggle to generalize to novel inputs. 3D segmentation of generated scenes (TRELLIS2 + PartField) inherently produces severe geometric holes (circled regions). Furthermore, "PartField + RANSAC" suffers from orientation inversion (1st and 2nd rows) and severe geometry intersections (circled region in the corresponding result column). In contrast, our framework consistently synthesizes physically plausible, and collision-aware components.

mesh $\mathbf{M}$ and the TRELLIS2-generated complementary mesh $\mathbf{M}_{\text{comp}}$ as input, without additional image or text conditioning. Notably, both are inherently constrained by the inability to optimize the object scale $\boldsymbol{s}$. We adopt the official pretrained weights for Jigsaw and train 2BY2 from scratch following its official protocol. In contrast, MIDI serves as an image-conditioned baseline, which predicts the constituent 3D assets and their spatial positions from a compositional scene image.

2. **Direct segmentation baseline.** To evaluate the impact of our generate-then-compose design, we construct a baseline that directly segments the TRELLIS2-generated scene mesh $\mathbf{M}_{\text{scene}}$ using PartField. This approach does not require an extra pose estimation and is based solely on object-level segmentation. However, the resulting meshes typically contain geomet-

Fig. 6: More than two parts composition results. Given a cabinet mesh, we render an image from it, add three pairs of shoes to the image using Nano Banana Pro and generate it with TRELLIS2 to obtain the object-object spatial relationships (OOR). Next, we use Nano Banana Pro again to extract each individual object in the shoe cabinet and generate them with TRELLIS2. Finally, using the OOR information, we sequentially add them to the cabinet mesh to form an interactive 3D scene.

ric holes that are difficult to repair, inevitably altering the user-provided assets. Comparison against this baseline highlights the advantage of our method in preserving high-fidelity individual geometries.

3. **Classical registration baseline.** To further demonstrate the effectiveness of our proposed registration approach (Section 4.2 and 4.3), we employ the RANSAC [6] algorithm to independently register the two inputs, $\mathbf{M}$ and $\mathbf{M}_{\text{comp}}$, based on $\mathbf{M}'$ and $\mathbf{M}'_{\text{comp}}$, yielding the final compositional result (PartField + RANSAC). This baseline isolates the contribution of our robust registration and collision-aware optimization design.

5.2 Qualitative Evaluation

Two-part Composition. Figure 5 presents qualitative comparisons of two-part composition results produced by different methods. MIDI, trained on 3D-FRONT [7], is inherently bottlenecked by the scarcity of 3D data, leading to poor generalization in generating individual 3D assets. Similarly, both Jigsaw and 2BY2 rely heavily on their specific training data distributions, which limits the generalization capability of their position predictions. Consequently, their

Table 2: Quantitative comparison of 3D compositional generation. We conduct a comprehensive evaluation using several unitless metrics, including semantic alignment measured by text- and image-based CLIP scores, human and VLM-based ratings, and geometric intersection rates computed over both surface and volume. The symbol "/" denotes metrics that are not applicable to a given baseline. The best results are highlighted in **bold**.

	Jigsaw	2BY2	PartField	PartField+RANSAC	MIDI	**Ours**
Text CLIP (Avg) ↑	0.3004	0.2798	/	0.3094	0.3079	**0.3314**
Image CLIP (Avg) ↑	0.7338	0.6982	/	0.7910	0.7840	**0.8236**
Human Rating (Avg) ↑	4.21	2.87	/	6.28	5.27	**8.33**
VLM Rating (Avg) ↑	5.89	4.08	/	6.59	5.84	**8.57**
$\mathbf{R}_{\text{surface}}$ ($\times 10^{-3}$)(Avg) ↓	2.7792	1.5827	/	2.2287	22.185	**0.6939**
$\mathbf{R}_{\text{volume}}$ ($\times 10^{-3}$)(Avg) ↓	16.837	9.1734	/	7.9280	116.70	**3.8762**

compositional outputs often exhibit severe misalignments or implausible spatial configurations. Directly applying PartField to segment scene meshes generated by TRELLIS2 introduces visible geometric artifacts and structural holes, as the generated geometries are typically fused. Moreover, the RANSAC-based alignment procedure struggles to achieve accurate geometric registration and may produce incorrect poses, such as the inverted pen inside the pen holder shown in rows 1 and 2. Finally, independently registering the two objects fails to explicitly account for collisions, leading to severe geometric interpenetrations in the final compositions, as observed in rows 3, 4, 5, 6, 9, and 10.

More-than-two-part Composition. Our framework naturally extends to compose an arbitrary number of objects, enabling the generation of rich interactive scenes. Similarly to our standard pipeline, after the user provides an initial 3D mesh, we first render it into an image and use Nano Banana Pro to edit the rendered image and introduce additional objects in the image space. We then use TRELLIS2 to generate a corresponding 3D scene. This provides spatial relationship priors that guide the placement of all components. Our multi-part composition is achieved through sequential two-object composition. As illustrated in Figure 6, the user-provided mesh $\mathbf{M}$ is first treated as the anchor object $\mathbf{M}_{\text{anchor}}$. A complementary object (e.g., leather shoes) is generated and treated as the remaining component $\mathbf{M}_{\text{remain}}$, which is aligned to the anchor using our composition pipeline. The resulting composed scene is then treated as a new anchor, and another object (e.g., high-heeled shoes) is introduced as the remaining component. This process is repeated until all objects are incorporated into the final scene. More cases are provided in Appendix.

5.3 Quantitative Evaluation

We quantitatively evaluate composition results across 150 test cases (including cases in Figures 5, 7, and 10) based on two key aspects: the semantic fidelity, and the physical validity (see Table 2).

Compositional Semantic Fidelity. Since PartField does not support compositional generation, our quantitative comparison focuses on the remaining four baselines. We use CLIP [27] to measure the semantic alignment between 2D renderings of the final compositional 3D scene, captured from predefined viewpoints, and both the user-provided text prompt and the 2D image generated by Nano Banana Pro. Higher CLIP scores denote stronger semantic consistency. To enable a more robust evaluation of OOR, we further include both human and VLM-based assessments using Qwen3, with ratings collected on a 1-to-10 scale.

Geometric Intersection Rate. To evaluate collision avoidance, we measure interpenetration from both surface and volumetric perspectives, where lower values indicate better physical compatibility.

Surface Intersection Ratio. We compute the total surface area involved in triangle–triangle intersections between meshes A and B. Let $\mathcal{A}_{\text{int}}$ denote the area of intersected surface regions, and $\mathcal{A}_A$, $\mathcal{A}_B$ the total surface areas of A and B. The surface intersection rate is defined as:

$$\mathbf{R}_{\text{surface}} = \frac{\mathcal{A}_{\text{int}}}{\mathcal{A}_A + \mathcal{A}_B}. \tag{4}$$

Volume Intersection Ratio. Since the compositional scene consists of solid objects represented by surface meshes, we estimate volumetric interpenetration via Monte Carlo sampling. Specifically, we uniformly sample $N(= 10^6)$ points within a bounding box enclosing both meshes. The volumetric intersection ratio is defined as the fraction of sampled points that lie inside both objects among those that lie inside at least one object:

$$\mathbf{R}_{\text{volume}} = \frac{\#\text{ points inside both } A \text{ and } B}{\#\text{ points inside } A \text{ or } B}. \tag{5}$$

6 Conclusion

We present **Interact3D**, a training-free generate-then-compose framework for interactive 3D scene synthesis. By leveraging spatial guidance from generative models, our method reformulates compositional generation as a collision-aware geometric alignment problem. Combining global-to-local registration, SDF refinement, and VLM-driven correction, our pipeline produces geometrically compatible and physically plausible scenes. We also release a curated compositional 3D dataset to facilitate future research in structured scene generation and robotic simulation.

Future Work. Although Interact3D generalizes well to daily objects, it struggles with fine-grained components (e.g., screws or tightly coupled joints). In these cases, severe occlusions in the 2D guidance images cause 3D geometric ambiguities, highlighting the inherent limitations of relying on 2D spatial priors for complex components. Future work will explore native 3D compositional generation to reason directly in 3D space, bypassing 2D dependencies and robustly handling intricate structures.

References

1. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22246–22256 (2023)
2. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025)
3. Chen, Y.C., Li, H., Turpin, D., Jacobson, A., Garg, A.: Neural shape mating: Self-supervised object assembly with adversarial shape priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12724–12733 (2022)
4. Deemos Technology: Hyper3d.ai. `https://hyper3d.ai/` (2026), accessed: 2026-02
5. 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)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
7. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021)
8. Gatenyo, R., Fried, O.: Copy-transform-paste: Zero-shot object-object alignment guided by vision-language and geometric constraints. arXiv preprint arXiv:2601.14207 (2026)
9. Google: Gemini. `https://gemini.google.com/` (2026), accessed: 2026-02
10. Hoque, R., Huang, P., Yoon, D.J., Sivapurapu, M., Zhang, J.: Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709 (2025)
11. Hossieni, S.S., Shabani, M.A., Irandoust, S., Furukawa, Y.: Puzzlefusion: Unleashing the power of diffusion models for spatial puzzle solving. Advances in Neural Information Processing Systems **36**, 9574–9597 (2023)
12. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23646–23657 (2025)
13. Jain, V., Attarian, M., Joshi, N.J., Wahid, A., Driess, D., Vuong, Q., Sanketi, P.R., Sermanet, P., Welker, S., Chan, C., et al.: Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. arXiv preprint arXiv:2403.12943 (2024)
14. Li, W., Zhang, X., Sun, Z., Qi, D., Li, H., Cheng, W., Cai, W., Wu, S., Liu, J., Wang, Z., et al.: Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. arXiv preprint arXiv:2505.07747 (2025)
15. 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)

16. Liu, M., Uy, M.A., Xiang, D., Su, H., Fidler, S., Sharp, N., Gao, J.: Partfield: Learning 3d feature fields for part segmentation and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9704–9715 (2025)
17. Liu, Z., Li, Y., Lin, Y., Yu, X., Peng, S., Cao, Y.P., Qi, X., Huang, X., Liang, D., Ouyang, W.: Unidream: Unifying diffusion priors for relightable text-to-3d generation. In: European Conference on Computer Vision. pp. 74–91. Springer (2024)
18. Lu, J., Sun, Y., Huang, Q.: Jigsaw: Learning to assemble multiple fractured objects. Advances in Neural Information Processing Systems **36**, 14969–14986 (2023)
19. Ma, C., Li, Y., Yan, X., Xu, J., Yang, Y., Wang, C., Zhao, Z., Guo, Y., Chen, Z., Guo, C.: P3-sam: Native 3d part segmentation. arXiv preprint arXiv:2509.06784 (2025)
20. Ma, Y.J., Liang, W., Wang, G., Huang, D.A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L., Anandkumar, A.: Eureka: Human-level reward design via coding large language models. arXiv preprint arXiv:2310.12931 (2023)
21. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021)
22. Narang, Y., Storey, K., Akinola, I., Macklin, M., Reist, P., Wawrzyniak, L., Guo, Y., Moravanszky, A., State, G., Lu, M., et al.: Factory: Fast contact for robotic assembly. arXiv preprint arXiv:2205.03532 (2022)
23. Narayan, A., Nagar, R., Raman, S.: Rgl-net: A recurrent graph learning framework for progressive part assembly. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 78–87 (2022)
24. Qi, Y., Ju, Y., Wei, T., Chu, C., Wong, L.L., Xu, H.: Two by two: Learning multi-task pairwise objects assembly for generalizable robot manipulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17383–17393 (2025)
25. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Ilic, S., Hu, D., Xu, K.: Geotransformer: Fast and robust point cloud registration with geometric transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9806–9821 (2023)
26. Qiu, L., Chen, G., Gu, X., Zuo, Q., Xu, M., Wu, Y., Yuan, W., Dong, Z., Bo, L., Han, X.: Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9914–9925 (2024)
27. 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)
28. Scarpellini, G., Fiorini, S., Giuliari, F., Moreiro, P., Del Bue, A.: Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28098–28108 (2024)
29. Shan, H., Li, M., Yang, H., Zheng, K., Zheng, S., Fu, Y., Huang, X.: Nitex: Non-isometric image-based garment texture generation. arXiv preprint arXiv:2511.18765 (2025)
30. Tencent: Tencent hunyuan3d. `https://3d.hunyuan.tencent.com/` (2026), accessed: 2026-02

31. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on pattern analysis and machine intelligence **13**(4), 376–380 (2002)
32. Wang, Z., Chen, J., Furukawa, Y.: Puzzlefusion++: Auto-agglomerative 3d fracture assembly by denoise and verify. arXiv preprint arXiv:2406.00259 (2024)
33. Wu, P., Escontrela, A., Hafner, D., Abbeel, P., Goldberg, K.: Daydreamer: World models for physical robot learning. In: Conference on robot learning. pp. 2226–2240. PMLR (2023)
34. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11097–11107 (2020)
35. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. arXiv preprint arXiv:2512.14692 (2025)
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. Xiong, Y., Ma, W.C., Wang, J., Urtasun, R.: Learning compact representations for lidar completion and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1074–1083 (2023)
38. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., et al.: Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4737–4746 (2023)
39. Yang, Y., Huang, Y., Guo, Y.C., Lu, L., Wu, X., Lam, E.Y., Cao, Y.P., Liu, X.: Sampart3d: Segment any part in 3d objects. arXiv preprint arXiv:2411.07184 (2024)
40. Yang, Y., Zhou, Y., Guo, Y.C., Zou, Z.X., Huang, Y., Liu, Y.T., Xu, H., Liang, D., Cao, Y.P., Liu, X.: Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
41. Yao, K., Zhang, L., Yan, X., Zeng, Y., Zhang, Q., Xu, L., Yang, W., Gu, J., Yu, J.: Cast: Component-aligned 3d scene reconstruction from an rgb image. ACM Transactions on Graphics (TOG) **44**(4), 1–19 (2025)
42. Ying, S., Peng, J., Du, S., Qiao, H.: A scale stretch method based on icp for 3d data registration. IEEE Transactions on automation science and engineering **6**(3), 559–565 (2009)
43. Yuan, Y., Cui, H., Huang, Y., Chen, Y., Ni, F., Dong, Z., Li, P., Zheng, Y., Hao, J.: Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. arXiv preprint arXiv:2508.13998 (2025)
44. Zakka, K., Zeng, A., Lee, J., Song, S.: Form2fit: Learning shape priors for generalizable assembly from disassembly. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9404–9410. IEEE (2020)
45. Zhang, C., Wang, J., Gao, Z., Su, Y., Dai, T., Zhou, C., Lu, J., Tang, Y.: Clap: Contrastive latent action pretraining for learning vision-language-action models from human videos. arXiv preprint arXiv:2601.04061 (2026)

46. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. ACM Transactions on Graphics (TOG) **43**(4), 1–20 (2024)
47. Zhang, Y., Liu, Y., Xie, Z., Yang, L., Liu, Z., Yang, M., Zhang, R., Kou, Q., Lin, C., Wang, W., et al.: Dreammat: High-quality pbr material generation with geometry- and light-aware diffusion models. ACM Transactions on Graphics (TOG) **43**(4), 1–18 (2024)
48. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025)