ComboVerse: Compositional 3D Assets Creation Using Spatially-Aware Diffusion Guidance

Yongwei Chen[1][*][†], Tengfei Wang[2][*], Tong Wu[3],
Xingang Pan[1], Kui Jia[4], and Ziwei Liu[1]

[1] S-Lab, Nanyang Technological University
[2] Shanghai Artificial Intelligence Laboratory
[3] The Chinese University of Hong Kong
[4] The Chinese University of Hong Kong, Shenzhen
https://cyw-3d.github.io/ComboVerse/

Abstract. Generating high-quality 3D assets from a given image is highly desirable in various applications such as AR/VR. Recent advances in single-image 3D generation explore feed-forward models that learn to infer the 3D model of an object without optimization. Though promising results have been achieved in single object generation, these methods often struggle to model complex 3D assets that inherently contain multiple objects. In this work, we present **ComboVerse**, a 3D generation framework that produces high-quality 3D assets with complex compositions by learning to combine multiple models. **1)** We first perform an in-depth analysis of this "multi-object gap" from both model and data perspectives. **2)** Next, with reconstructed 3D models of different objects, we seek to adjust their sizes, rotation angles, and locations to create a 3D asset that matches the given image. **3)** To automate this process, we apply spatially-aware score distillation sampling (SSDS) from pretrained diffusion models to guide the positioning of objects. Our proposed framework emphasizes spatial alignment of objects, compared with standard score distillation sampling, and thus achieves more accurate results. Extensive experiments validate ComboVerse achieves clear improvements over existing methods in generating compositional 3D assets.

1   Introduction

Learning to create high-quality 3D assets from a single image is a long-standing goal in computer graphics and vision research, given its potential applications in AR/VR, movies, games, and industrial design. Over the years, a plethora of attempts have been made to leverage diffusion models [15] for 3D content creation.

Previously, limited 3D data availability led researchers to rely on pretrained 2D diffusion models for 3D guidance, with a score distillation sampling (SDS)

---

[*]Equal contribution.

[†]Work done when interning at Shanghai Artificial Intelligence Laboratory.

Fig. 1: *ComboVerse* can generate high-quality 3D models from a single image that contains multiple objects, *e.g.,* a squirrel sitting on a paper box. We show textured meshes of created 3D content, showcasing stunning reconstruction quality.

loss [38] transferring 3D-aware knowledge. More recently, alternative approaches focus on training feed-forward 3D diffusion models for fast generation, facilitated by large-scale 3D object datasets like Objaverse [7]. Once trained, these models can produce signed distance field [5], points [34], radiance fields [18,53], mesh [29], or multi-view images [27] through a single forward inference within one minute.

Despite compelling results on simple object generation, these feed-forward methods usually encounter difficulties when applied to more complex data, such as scenes with multiple objects and complex occlusion. Fig. 2 illustrates the drawbacks of existing models when dealing with such combining objects. However, upon generating each object separately, we observed that these models performed well. We perform an in-depth analysis of this "multi-object gap" and conjecture that this gap comes from the bias of their training data, *i.e.,* Objaverse. The scarcity of 3D assets containing multiple objects makes it challenging for trained models to manage composites beyond the training data distribution.

Given the observations made above, is it possible to design a generative system that can produce 3D content containing multiple objects? Typically, skilled human artists create each object separately before integrating them into a whole. This has motivated us to present a compositional generation paradigm termed *ComboVerse*, which generates each object individually and then focuses on automatically combining them to create a composite. A key advantage of our proposed paradigm is its ability to effectively manage complex assets containing multiple objects and occlusion.

Our approach comprises two stages: single-object reconstruction and multi-object combination. We first decompose and reconstruct each object within an image independently, using an occlusion removal module and an image-to-3D

model. In the second stage, we aim to automatically combine the generated 3D objects into a single model, accounting for various factors such as object scale, placement, and occlusion. However, this process poses a challenge due to depth-size ambiguity in the input image, leading to inaccurate composition.

To address this issue, we opt for pre-trained diffusion models as spatial guidance for object positioning. Unlike previous SDS-based methods [4, 25, 38, 48] that require optimizing both the shape and texture from scratch, we fix the 3D model of individual objects and focus only on achieving a reasonable spatial layout so that the optimization process would be much faster. However, we have found that the standard SDS is insufficient for accurately placing objects, as it tends to prioritize content over position to match the given text prompt (see Fig. 5). To address this issue, we introduce a spatially-aware SDS loss that places greater emphasis on the spatial relationships between objects. Specifically, we reweight [14] the attention map of the position tokens that indicate the spatial relation for score distillation. By prioritizing the awareness of position, the proposed loss can effectively distill the spatial knowledge from well-trained diffusion models for object placement.

To evaluate our method, we collect a benchmark consisting of 100 images that comprise a diverse range of complex scenes. We evaluate *ComboVerse* on this benchmark, and extensive experiments show clear improvements over previous methods in terms of handling multiple objects, occlusion, and camera settings. Our main contributions can be summarized as:

- We propose *ComboVerse*, an automatic pipeline that extends object-level 3D generative models to generate compositional 3D assets from an image.
- We perform an in-depth analysis of the "multi-object gap" of existing feed-forward models from both model and data perspectives.
- We propose spatially-aware diffusion guidance, enabling pre-trained image diffusion models to provide guidance on spatial layout for object placement.

2   Related works

3D Generation with 2D Diffusion Prior. Many methods opt to pretrained 2D diffusion models [15, 43] as a source of 3D guidance. Early works [38] proposed a score distillation sampling method to leverage the imaginative power of 2D diffusion for text-conditioned 3D content creation. Later works have improved the quality by using two-stage optimization [4, 25, 32], better score distillation [55], and stronger foundation diffusion models [24, 45]. Other works [31, 39, 48, 52, 56, 60] extend the approach to generate 3D models from a single image. Some works [47, 61] replace implicit representation with 3D gaussian splatting. Although the results are promising, creating a 3D model in this way can take several minutes to hours of optimization.

Feed-forward 3D Generative Models. Another line of approaches trained feed-forward models for fast generation, eliminating the need for per-case optimization. 3D-aware generative adversarial networks [1,2,11,35,36,59] have gained considerable research interest in early research. Later, many attempts have been

made to leverage diffusion models for image-conditioned and text-conditioned 3D generation. Once trained, they can produce signed distance field [5,6], points [34], radiance fields [12,16,18,21,53], gaussian splattings [22,62], mesh [29], or multi-view images [26–28,30,49] without optimization. Besides diffusion models, recent works also explore feed-forward 3D reconstruction with transformer architecture [17,23] or UNet architecture [46]. Despite fast generation, these methods are limited by the training data, restricting their ability to reconstruct complex 3D assets. We aim to build on these object-level generative models and extend them to handle more complex objects or scenes.

Compositional 3D Generation. Previous studies [35] have investigated the use of compositional neural radiance fields in an adversarial learning framework for the purpose of 3D-aware image generation. Additional studies have explored the concept of part-based shape generation, which involves assembling 3D parts into a 3D model. The seminal work [9] retrieves from a mesh database to find parts of interest and composite the cut parts to produce novel objects. Later, the following works involve probabilistic models for part suggestion [19], semantic attribute [3], fabrication [44], and CAD assembly [57]. With pretrained diffusion models as guidance, recent work [37] generates compositional 3D scenes with user-annotated 3D bounding boxes and text prompts. Concurrent works generate 3D scenes from text prompts by using large language models (LLMs) to propose 3D layouts as an alternative for human annotations [10,50,54], or jointly learning layout during optimization process [8]. These approaches can produce 3D scenes that match the text prompts, but texts can be unclear and imprecise when describing how objects are arranged in space. In contrast, our approach focuses on reconstructing complex 3D assets from a reference image. Unlike text descriptions, images represent the spatial relations among objects more accurately, requiring higher standards of composition quality.

3   ComboVerse

In this section, we will first analyze the "multi-object gap" of state-of-the-art image-to-3D generation methods trained on Objaverse, followed by a discussion of our compositional generative scheme. We then present the details of the two stages involved in our approach: single-object reconstruction and multi-object combination. The overview architecture is shown in Fig. 3.

3.1   Analysis of "Multi-Object Gap"

Most existing feed-forward models are trained on Objaverse [7]. As shown in Fig. 2 (also see *Supp.*), these methods suffer three typical failure cases for multiple objects generation due to data and model biases.

Camera Setting Bias. When setting up cameras, most image-to-3D methods assume that the object has a normalized size and is centered in the image. However, in scenarios with multiple objects, an object could appear in a corner or be
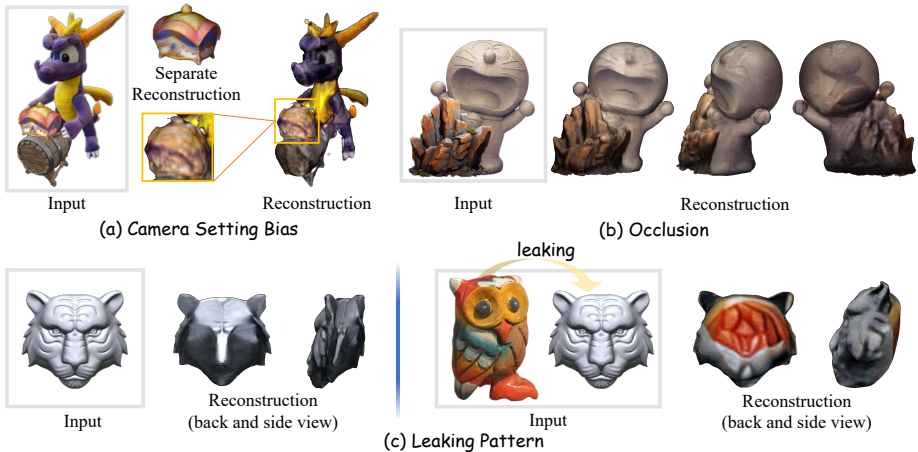
Fig. 2: "Multi-object gap" of models trained on Objaverse. (a) Camera Setting Bias. The reconstruction quality for small and non-centered objects will significantly downgrade compared to separate reconstruction. (b) Occlusion. The reconstruction results tend to blend when an object is occluded by another. (c) Leaking Pattern. The shape and texture of an object will be influenced by other objects in the input image. For example, in (c), the tiger's back face adopts the owl's color, and its back surface becomes convex instead of concave due to the owl's shape influence.

very small in the image, which does not conform to object-centric assumptions. Such a case can result in a significant decline in modeling quality.

Dataset Bias. The Objaverse dataset predominantly features single-object assets, which poses a challenge for models trained on it to generalize to complex composites. Additionally, the near absence of occlusion in Objaverse results in these models struggling to handle occluded objects. As a result, generated objects often blend together due to occlusion ambiguity.

Leaking Pattern. Existing methods tend to exhibit leakage issues when generating multiple objects simultaneously, where the geometry and appearance of one object can affect another. This issue may stem from the model's biases, as it is trained to generate a single object where different parts are consistent. However, in scenes with multiple objects, different objects may have different geometry and texture. If they still affect each other, it can lead to bleeding patterns.

Motivation. As shown in Fig. 2, though the current methods have difficulties in generating compositional objects, we have observed that these methods are successful in reconstructing each component object. This observation suggests the possibility of generating each object separately (Sec. 3.2) and subsequently combining them to form the desired compositional object (Sec. 3.3).

3.2   Single-Object Reconstruction

Components Decomposition. Given an input image I, we first specify each object's 2D bounding box $\{b_i \in \mathbb{Z}^4\}$, indicating the coordinates of the upper left
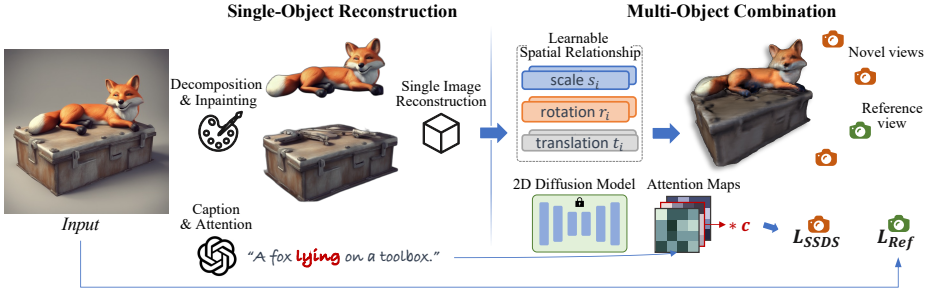
Fig. 3: Overview of our method. Given an input image that contains multiple objects, our method can generate high-quality 3D assets through a two-stage process. In the single-object reconstruction stage, we decompose every single object in the image with object inpainting, and perform single-image reconstruction to create individual 3D models. In the multi-object combination stage, we maintain the geometry and texture of each object while optimizing their scale, rotation, and translation parameters $\{s_i, r_i, t_i\}$. This optimization process is guided by our proposed spatially-aware SDS loss $\mathcal{L}_{\text{SSDS}}$, calculated on novel views, emphasizing the spatial token by enhancing its attention map weight. For example, considering the prompt *"A fox lying on a toolbox."* given to the 2D diffusion model, we emphasize the spatial token "lying" by multiplying its attention map with a constant c ($c > 1$). Also, we utilize the reference loss $\mathcal{L}_{\text{Ref}}$, calculated on a reference view for additional constraints.

and bottom right corners. Given bounding boxes $\{b_i \in \mathbb{Z}^4\}$ for different objects, we use SAM [20] to segment each object as follows:

$$O_i, M_i = \text{SAM}(I, b_i), \tag{1}$$

where O_i and M_i are the RGB channels and binary mask of i-th object.

Object Inpainting. To complete O_i that is possibly occluded by another object, we utilize Stable Diffusion (SD) [43] for objects inpainting. However, we face the challenge of not having a known mask to identify occluded regions. To address this issue, we design a strategy for completing the occluded parts of objects. First, to avoid generating a white or black border around the objects when inpainting, we replace the background of image O_i with random noise, and the noised image I_i is generated as follows:

$$I_i = O_i + noise * (\sim M_i), \tag{2}$$

where the $\sim M_i$ is the background region of O_i. The noised image I_i is illustrated in Fig. 4. Second, the background region and bounding box b_i are combined to generate an inpainting mask m_i for each object, which indicates the inpainting region for each one. Specifically, for each inpainting mask m_i, the pixels that lie in the bounding box but outside the foreground object are set to 1, and the others are set to 0. That is:

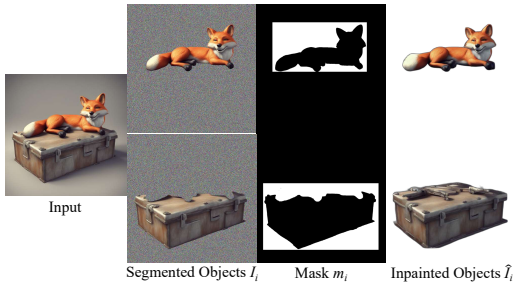$$m_i = (\sim M_i) \cap b_i, \tag{3}$$

Fig. 4: Objects decomposition and inpainting. In this stage, given an input image, we segment each separate object and get segmented objects with noise background image I_i and bounding-aware mask m_i, then I_i and m_i are input to Stable Diffusion to obtain the inpainted objects $\hat{I}_i$.
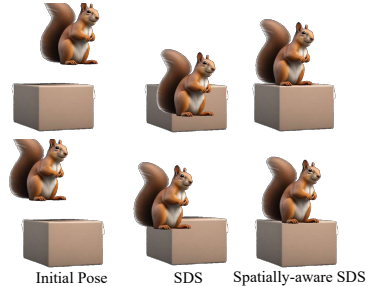

Fig. 5: 2D toy examples. We randomly initialize the squirrel with two different initial positions (left), and optimize the position parameters to match the prompt "a squirrel is sitting on a box". Compared to standard SDS, spatially-aware SDS produces better results.

where the $\sim M_i$ is the background region of i-th object O_i, and b_i indicates its bounding box. The mask m_i is illustrated in Fig. 4. Finally, we input I_i and m_i that contain bounding box information to SD, to complete the object I_i inside the bounding box b_i: $\hat{I}_i = SD(I_i, m_i)$, where $\hat{I}_i$ is the completed object, which is illustrated in Fig. 4. For better completion, we input a text prompt *"a complete 3D model"* to SD when inpainting. After that, each inpainted object $\hat{I}_i$ can be reconstructed by image-to-3D methods to produce single 3D models.

3.3   Multi-Object Combination

At this stage, we seek to combine separate 3D models by optimizing their scale, rotation, and translation parameters $\{s_i, r_i, t_i\}$, such that they align with the input image I and semantic spatial relationships. We begin by initializing each object's scale, rotation, and translation based on I, and then refine them using the proposed spatially-aware diffusion priors and guidance from the reference image. We will first introduce a spatially-aware diffusion distillation scheme, followed by a discussion on its application for automatic object combinations.

Spatially-Aware Diffusion Guidance. DreamFusion [38] presents a method that optimizes 3D representations from textual descriptions, by employing a pre-trained 2D diffusion model. The subject is represented as a differentiable parameterization [33], where a differentiable MLP renderer g renders 2D images $x = g(\theta)$ from a neural radiance field parameterized as θ. It leverages a diffusion model ϕ to provide a score function $\hat{\epsilon}_\phi(x_t; y, t)$, which predicts the sampled noise ϵ given the noisy image x_t, text-embedding y, and noise level t. This score function guides the direction of the gradient for updating the neural parameters

θ, and the gradient is calculated by Score Distillation Sampling (SDS):

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, x) = \mathbb{E}_{t,\epsilon} \left[w(t)(\hat{\epsilon}_\phi(x_t; y, t) - \epsilon) \frac{\partial x}{\partial \theta} \right], \tag{4}$$

while $w(t)$ is a weighting function.

However, we find that SDS is unstable for position adjustment in our case. We use a text prompt *"a squirrel is sitting on a box"* and an image of a squirrel and a box as a toy example, and aim to test the ability of SDS to adjust the position of the image elements according to the text prompt. As shown in Fig. 5, SDS does not produce the correct placement, as the image content (squirrel and box) already matches the prompt and SDS does not push the adjustment of position. We thus propose spatially-aware SDS to emphasize the position adjustment when calculating the gradient.

Recall that in SDS, for each text embedding y and time step t, we use a UNet to predict the noise $\hat{\epsilon}_\phi(x_t; y, t)$. The features of noisy image $\phi(x_t)$ are projected to a query matrix $Q = \mathcal{F}_Q(\phi(x_t))$, and the textual embedding is projected to a key matrix $K = \mathcal{F}_K(y)$ and a value matrix $V = \mathcal{F}_V(y)$, via the learned linear projections $\mathcal{F}_Q$, $\mathcal{F}_K$ and $\mathcal{F}_V$. The attention maps are then calculated by:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} \right), \tag{5}$$

where the A_j indicates the attention map of j-th token, and d is the latent projection dimension of the keys and queries.

To prioritize the refinement of spatial layout, we strengthen the key token that describes the spatial relationship. The key token can be the word describing spatial relationships, such as "front", "on," and "below," or the word describing object interaction, such as "riding" and "holding", which can be extracted by LLMs or indicated by the user. For example, consider the prompt *"a squirrel is sitting on a box"*, we want to strengthen the effect of the word *"sitting on"*, which describes the relationship between the squirrel and paper box. To achieve this spatially-aware optimization, we scale the attention maps of the assigned tokens $j^\star$ with a constant c $(c > 1)$, similar to [14], resulting in a stronger focus on the spatial relationship. The rest of the attention maps remain unchanged:

$$A := \begin{cases} c \cdot A_j & \text{if } j = j^\star \\ A_j & \text{otherwise.} \end{cases} \tag{6}$$

The spatially-aware SDS loss (SSDS) can be formulated as:

$$\nabla_\theta \mathcal{L}_{\text{SSDS}}(\phi^\star, x) = \mathbb{E}_{t,\epsilon} \left[w(t)(\hat{\epsilon}_{\phi^\star}(x_t; y, t) - \epsilon) \frac{\partial x}{\partial \theta} \right], \tag{7}$$

where the $\hat{\epsilon}_{\phi^\star}(x_t; y, t)$ is the predicted noise calculated with the strengthened attention maps which focus on the spatial words. For timesteps, we sample t from a range with high noise levels, as these steps have a bigger impact on the spatial layout of a generated image.

Combine the objects. We begin with a coarse initialization of scale, rotation and translation $\{s_i, r_i, t_i\}$ from the bounding box b_i and estimated depth d_i. Specifically, the scale s_i is decided by the ratio of the bounding box size and image size:

$$s_i = max \left\{ \frac{W_{b_i}}{W_I}, \frac{H_{b_i}}{H_I} \right\}, \tag{8}$$

where the W_{b_i}, H_{b_i}, W_I, H_I are the width and height of bounding box b_i and input image I respectively. As for translation t_i, we use a monocular depth prediction model [41] to estimate the average depth d_i for i-th object, which is set as z-dimension in t_i, and the x and y dimensions are initialized by the center coordinates of bounding box b_i and image size. That is:

$$\begin{aligned} t_i &= (X_{b_i} - \tfrac{W_I}{2}, Y_{b_i} - \tfrac{H_I}{2}, d_i), \\ d_i &= \text{Average}(\text{Depth}(O_i)), \end{aligned} \tag{9}$$

where X_{b_i}, Y_{b_i} are the center coordinates of bounding box b_i, d_i is the average depth of each pixel that lie in i-th object O_i. The rotation angles in three dimensions r_i are initialized as $(0, 0, 0)$.

However, due to depth-size ambiguity in a single-view image, the predicted depth can be inaccurate, leading to an unreasonable initialization of depth and size. To alleviate the single-view ambiguity, we refine spatial parameters $\{s_i, r_i, t_i\}$ with the proposed spatially-aware SDS loss (SSDS) as novel-view supervision. To stabilize the optimization, we also constrain the reconstruction error between reference-view rendering $\hat{I}$ in and input image I:

$$\mathcal{L}_{\text{Ref}} = \lambda_{\text{RGB}} \left| \hat{I}_{\text{RGB}} - I_{\text{RGB}} \right| + \lambda_{\text{A}} \left| \hat{I}_{\text{A}} - I_{\text{A}} \right|, \tag{10}$$

where λ_{RGB} and λ_{A} are weights for RGB and alpha channels. The total loss is a weighted summation of $\mathcal{L}_{\text{Ref}}$ and $\mathcal{L}_{\text{SSDS}}$.

4   Experiments

4.1   Implementation Details

We set the guidance scale to 7.5 and the number of inference steps to 30 when inpainting the image with Stable Diffusion. We use Pytorch3D [42] as our differentiable rendering engine. We randomly sample timestep t between 800 and 900. We use Adam as our optimizer, and the learning rate of translation on the z dimension is 0.01, while the others are set to 0.001. The loss weight λ_{Ref}, λ_{SSDS}, λ_{RGB}, λ_{A}, are set to 1, 1, 1,000, 1,000 respectively. We set the multiplier c for attention map to 25. We downsample each separate mesh to 50,000 faces in the multi-object combination stage and render 10 views for each iteration. The SSDS optimization takes approximately 2 minutes.
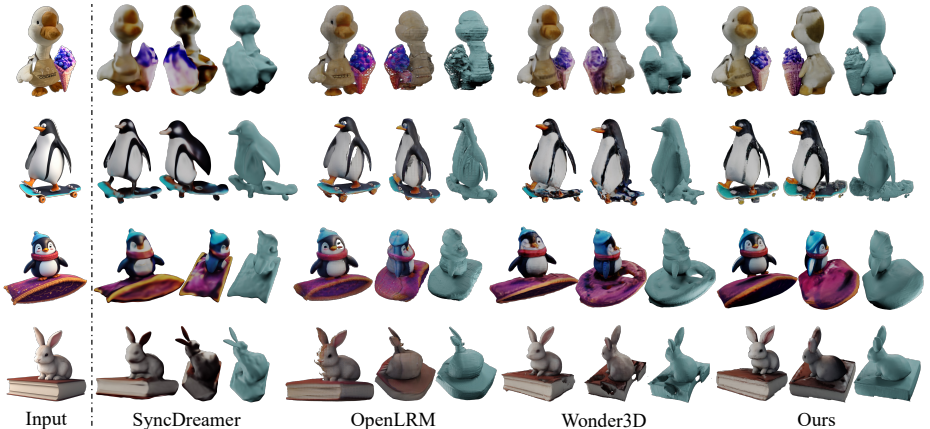
Fig. 6: Visual comparison for image-to-3D generation. Given an input image, previous methods reconstruct inaccurate geometry and blurry texture, especially in novel views. Our method produces higher-fidelity 3D models with the proposed compositional generation scheme.

Table 1: Quantitative comparison.

Method	CLIP-Score ↑	GPT-3DScore ↑
SyncDreamer [28]	81.47%	13.54%
OpenLRM [17]	83.65%	53.12%
Wonder3D [30]	85.57%	56.25%
Ours	**86.58**%	**65.63**%

4.2   Main Results

Benchmark. To evaluate our method, we built a test benchmark containing 100 images covering a variety of complex 3D assets. The benchmark includes 50 images generated with stable diffusion, and 50 images constructed from real images with PhotoShop. Each image has a foreground mask represented as the alpha channel, a set of bounding boxes for objects, and a text caption. We use GPT4 to propose text prompts and spatial tokens, followed by manual filtering. We will make this benchmark publicly available.

Comparison Methods. We compare our method with three state-of-the-art single-image reconstruction methods: 1) SyncDreamer [28], which we implement using the official code. The textured meshes are derived from the NeuS [51] representation. 2) LRM [17], which we implement using the publicly available code [13]. 3) Wonder3D [30], which we implement using the official code and also use as our base model for image-to-3D reconstruction.

Qualitative Comparison. As shown in Fig. 6, our method can accurately reconstruct each object and preserve good spatial relationships among them. Other methods often struggle to generate high-quality geometry and texture for

Fig. 7: Qualitative results. *ComboVerse* can generate high-quality 3D models from a single image that contains multiple objects.

Input	Full	w/o Random background	w/o Bounding-aware mask	w/o Prompt

Fig. 8: Analysis for objects inpainting. We produce compelling inpainting results with random background (EQ. 2), bounding-aware mask proposal (EQ. 3), and text prompting. More analysis can be found in supplemental materials.

small objects in the input image, and the junction parts of different objects often blend. More qualitative results are shown in Fig. 7.

Quantitative Comparison. We use CLIP-Score [40] to measure semantic similarities between novel-view images and the reference image. We also involve GPT-based evaluation following [58]. We conduct pair-wise comparisons for each method across all samples, and report the probability of success for each method. Table 1 shows that our method outperforms comparison methods in both semantic similarity and GPT evaluation.

4.3   Ablation Study

Effectiveness of Object Inpainting. We performed a study to examine designs in object completion. We first investigate the effect of noisy backgrounds on object completion. When using a black or white background, the inpainting model tends to generate a corresponding black or white border around the object to ensure compatibility with the background, and the inpainted resuts were not complete, as shown in the third column of Fig. 8. Then, we simply used the background mask for inpainting instead of a bounding-aware mask, and the inpainted results had extra parts, which caused inconsistency in the 3D reconstruction with the input image. We also removed the text prompt "a complete 3D model" for diffusion models during inpainting, and the results also degraded.

Effectiveness of Spatially-Aware Diffusion Guidance. As shown in Fig. 9, we use an example of "an astronaut is riding a red horse" to analyze different guidance settings in object combination. *Base* only enforces reconstruction loss in the reference view without additional guidance in novel views, and thus yields incorrect relative depth between the astronaut and the horse. With a standard *SDS loss* or *depth loss* from a depth prediction model as spatial guidance, the interaction between the astronaut and the horse improves, but it was still far from accurate. By strengthening the attention to the word "riding" with the proposed *SSDS loss (full)*, the full model achieves the best result. This confirms
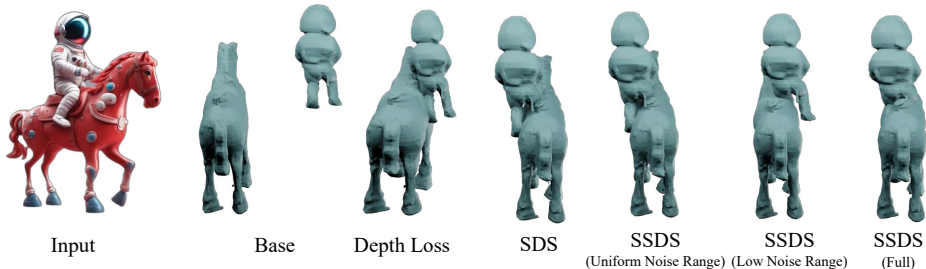
Fig. 9: Analysis for objects combination. Compared with standard SDS and depth constrain, SSDS provides stronger guidance on object positioning.

the improved spatial control capability of the proposed method over the standard SDS. As discussed in Sec. 3.3, we sampled from a high noise range ([800, 900]) for Stable Diffusion when performing SSDS, as these steps have a bigger impact on the spatial layout of a generated image. We also experiment with SSDS with different sample ranges of noise timesteps, *low noise range* ([100, 200]), and *uniform noise range* ([20, 980]) and observe a performance drop. We also give quantitative ablation results in *supplementary materials.*.

4.4 Application in Scene Reconstruction

Besides the generation of 3D assets with two objects, we also validate the generalization ability of the proposed method to multiple objects. As illustrated in Fig. 10, we use the proposed method to reconstruct 3D scenes consisting of multiple (> 2) objects. Previous methods that work on the object level have difficulties in generating scenes and produce obvious artifacts in both geometry and texture. Also, for small objects such as the dog in the last example, existing methods tend to ignore it due to the training bias mentioned before. In contrast, our method achieves realistic and high-quality reconstruction.

5 Conclusion

In this paper, we present *ComboVerse*, a novel method for creating high-quality compositional 3D assets from a single image. With an in-depth analysis of the "multi-object gap", we build on object-level 3D generative models and extend them to deal with more complex objects. With reconstructed 3D models of various objects, we seek to adjust their sizes, rotation angles, and locations to create a 3D asset that matches the given image. To this end, we proposed spatially-aware score distillation sampling from pretrained diffusion models to guide the placement of objects. Our approach can be a valuable step for complex 3D object reconstruction and pave the way for future 3D scene generation.

Limitations. The proposed method performs well for assets consisting of two or a few (usually < 5) objects. However, similar to existing text-based works, our
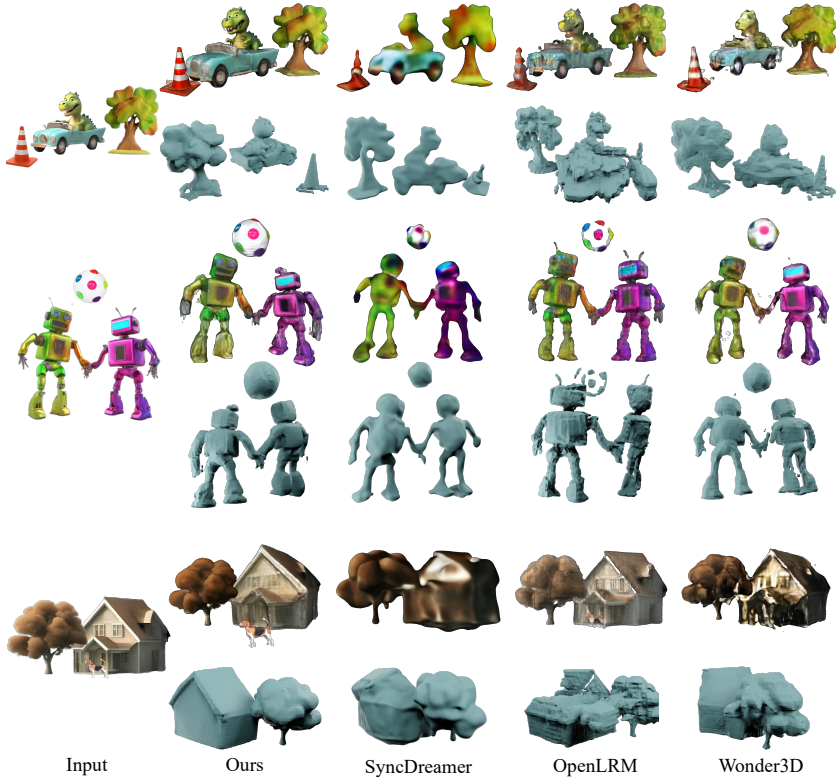
Fig. 10: Comparison of scene reconstruction. We show some challenging cases that contain more than two objects. The first example involves four objects: *a car, a dinosaur, a tree, and a cone*. The second example involves three examples: *two robots and a ball*. The third example involves three examples: *a house, a dog, and a tree*. Our method achieves compelling reconstruction quality with the compositional scheme.

method still faces challenges in creating very complex scenes with more objects. Another limitation of our approach is the lack of optimization for the geometry and texture in the combination process. Thus, the quality of the final results relies on the performance of the image-to-3D method that we use as backbone. We expect that our methods can be further enhanced with more robust backbone methods in the future.

Acknowledgement

References

1. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
2. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5799–5809 (2021)
3. Chaudhuri, S., Kalogerakis, E., Giguere, S., Funkhouser, T.: Attribit: content creation with semantic attributes. In: Proceedings of the 26th annual ACM symposium on User interface software and technology. pp. 193–202 (2013)
4. 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 (ICCV). pp. 22246–22256 (2023)
5. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A., Gui, L.: SDFusion: Multi-modal 3d shape completion, reconstruction, and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
6. Chou, G., Bahat, Y., Heide, F.: Diffusion-sdf: Conditional generative modeling of signed distance functions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2262–2272 (2023)
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 (CVPR). pp. 13142–13153 (2023)
8. Epstein, D., Poole, B., Mildenhall, B., Efros, A.A., Holynski, A.: Disentangled 3d scene generation with layout learning. In: International Conference on Machine Learning (ICML) (2024)
9. Funkhouser, T., Kazhdan, M., Shilane, P., Min, P., Kiefer, W., Tal, A., Rusinkiewicz, S., Dobkin, D.: Modeling by example. ACM Transactions on Graphics (TOG) **23**(3), 652–663 (2004)
10. 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 (CVPR). pp. 21295–21304 (2024)
11. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In: International Conference on Learning Representations (ICLR) (2022)
12. Gupta, A., Xiong, W., Nie, Y., Jones, I., Oğuz, B.: 3dgen: Triplane latent diffusion for textured mesh generation. arXiv preprint arXiv:2303.05371 (2023)
13. He, Z., Wang, T.: Openlrm: Open-source large reconstruction models. `https://github.com/3DTopia/OpenLRM` (2023)
14. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: International Conference on Learning Representations (ICLR) (2023)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) **33**, 6840–6851 (2020)

16. Hong, F., Tang, J., Cao, Z., Shi, M., Wu, T., Chen, Z., Wang, T., Pan, L., Lin, D., Liu, Z.: 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. arXiv preprint arXiv:2403.02234 (2024)

17. 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. In: International Conference on Learning Representations (ICLR) (2024)

18. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. https://arxiv.org/abs/2305.02463 (2023)

19. Kalogerakis, E., Chaudhuri, S., Koller, D., Koltun, V.: A probabilistic model for component-based shape synthesis. ACM Transactions on Graphics (TOG) **31**(4), 1–11 (2012)

20. 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 (ICCV). pp. 4015–4026 (2023)

21. Lan, Y., Hong, F., Yang, S., Zhou, S., Meng, X., Dai, B., Pan, X., Loy, C.C.: Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)

22. Lan, Y., Tan, F., Qiu, D., Xu, Q., Genova, K., Huang, Z., Fanello, S., Pandey, R., Funkhouser, T., Loy, C.C., Zhang, Y.: Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)

23. 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. In: International Conference on Learning Representations (ICLR) (2024)

24. Li, W., Chen, R., Chen, X., Tan, P.: Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. In: International Conference on Learning Representations (ICLR) (2024)

25. 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 (CVPR) (2023)

26. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10072–10083 (2024)

27. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9298–9309 (2023)

28. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. In: International Conference on Learning Representations (ICLR) (2024)

29. Liu, Z., Feng, Y., Black, M.J., Nowrouzezahrai, D., Paull, L., Liu, W.: Meshdiffusion: Score-based generative 3d mesh modeling. In: International Conference on Learning Representations (ICLR) (2023)

30. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9970–9980 (2024)

31. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: RealFusion: 360 reconstruction of any object from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
32. 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 (CVPR). pp. 12663–12673 (2023)
33. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
34. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-E: A system for generating 3d point clouds from complex prompts. https://arxiv.org/abs/2212.08751 (2023)
35. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11453–11464 (2021)
36. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13503–13513 (2022)
37. Po, R., Wetzstein, G.: Compositional 3d scene generation using locally conditioned diffusion. In: International Conference on 3D Vision (3DV). pp. 651–663. IEEE (2024)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: DreamFusion: Text-to-3d using 2d diffusion. In: International Conference on Learning Representations (ICLR) (2023)
39. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In: International Conference on Learning Representations (ICLR) (2024)
40. 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 (ICML). pp. 8748–8763. PMLR (2021)
41. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
42. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv:2007.08501 (2020)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
44. Schulz, A., Shamir, A., Levin, D.I.W., Sitthi-Amorn, P., Matusik, W.: Design and fabrication by example. ACM Transactions on Graphics (TOG) **33**(4) (2014)
45. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. In: International Conference on Learning Representations (ICLR) (2024)
46. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. arXiv preprint arXiv:2402.05054 (2024)

47. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In: International Conference on Learning Representations (ICLR) (2024)
48. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
49. Tang, S., Chen, J., Wang, D., Tang, C., Zhang, F., Fan, Y., Chandra, V., Furukawa, Y., Ranjan, R.: Mvdiffusion++: A dense high-resolution multi-view diffusion model for single to sparse-view 3d objectreconstruction. arXiv preprint arXiv:2402.12712 (2024)
50. Vilesov, A., Chari, P., Kadambi, A.: Cg3d: Compositional generation for text-to-3d via gaussian splatting. arXiv preprint arXiv:2311.17907 (2023)
51. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Advances in Neural Information Processing Systems (NeurIPS) **34**, 27171–27183 (2021)
52. Wang, P., Shi, Y.: Imagedream: Image-prompt multi-view diffusion for 3d generation. arXiv preprint arXiv:2312.02201 (2023)
53. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4563–4573 (2023)
54. Wang, Z., Li, M., Chen, C.: Luciddreaming: Controllable object-centric 3d generation. arXiv preprint arXiv:2312.00588 (2023)
55. 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. Advances in Neural Information Processing Systems (NeurIPS) **36** (2024)
56. Wang, Z., Wang, T., Hancke, G., Liu, Z., Lau, R.W.: Themestation: Generating theme-aware 3d assets from few exemplars. ACM SIGGRAPH (2024)
57. Willis, K.D., Jayaraman, P.K., Chu, H., Tian, Y., Li, Y., Grandi, D., Sanghi, A., Tran, L., Lambourne, J.G., Solar-Lezama, A., et al.: Joinable: Learning bottom-up assembly of parametric cad joints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15849–15860 (2022)
58. 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 (CVPR). pp. 22227–22238 (2024)
59. Xiang, J., Yang, J., Deng, Y., Tong, X.: Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2195–2205 (2023)
60. Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z.: Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4479–4489 (2023)
61. Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: GaussianDreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. https://arxiv.org/abs/2310.08529 (2023)
62. Zhang, B., Cheng, Y., Yang, J., Wang, C., Zhao, F., Tang, Y., Chen, D., Guo, B.: Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. arXiv preprint arXiv:2403.19655 (2024)