MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation

Zehuan Huang[1] Yuan-Chen Guo[2†] Xingqiao An[3] Yunhan Yang[4] Yangguang Li[2] Zi-Xin Zou[2]
Ding Liang[2] Xihui Liu[4] Yan-Pei Cao[2 ✉] Lu Sheng[1 ✉]

[1]Beihang University [2]VAST [3]Tsinghua University [4]The University of Hong Kong

Project page: https://huanngzh.github.io/MIDI-Page/

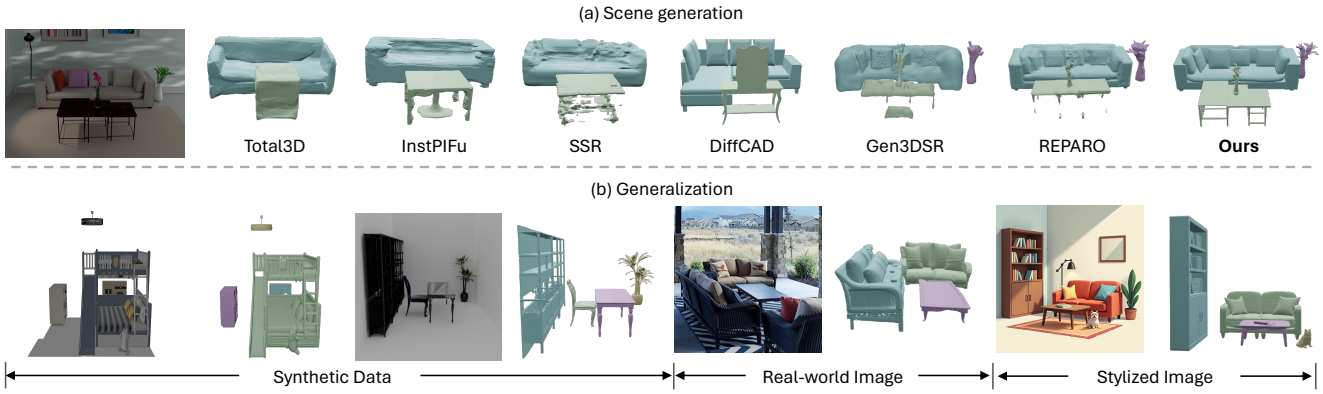

Figure 1. MIDI generates compositional 3D scenes from a single image by extending pre-trained image-to-3D object generation models to multi-instance diffusion models, incorporating a novel multi-instance attention mechanism that captures inter-object interactions. (a) shows our generated scenes compared with those reconstructed by existing methods. (b) presents our generated results on synthetic data, real-world images, and stylized images.

Abstract

This paper introduces MIDI, a novel paradigm for compositional 3D scene generation from a single image. Unlike existing methods that rely on reconstruction or retrieval techniques or recent approaches that employ multi-stage object-by-object generation, MIDI extends pre-trained image-to-3D object generation models to multi-instance diffusion models, enabling the simultaneous generation of multiple 3D instances with accurate spatial relationships and high generalizability. At its core, MIDI incorporates a novel multi-instance attention mechanism, that effectively captures inter-object interactions and spatial coherence directly within the generation process, without the need for complex multi-step processes. The method utilizes partial object images and global scene context as inputs, directly modeling object completion during 3D generation. During training, we effectively supervise the interactions between 3D instances using a limited amount of scene-level data, while incorporating single-object data for regularization, thereby maintaining the pre-trained generalization ability. MIDI demonstrates state-of-the-art performance in image-to-scene generation, validated through evaluations on synthetic data, real-world scene data, and stylized scene images generated by text-to-image diffusion models.

1. Introduction

Generating compositional 3D scenes from a single image is challenging due to the limited spatial clues captured from a partial point of view. In fact, accurately inferring the 3D geometry of each instance and the spatial relationships of multiple instances within a scene, requires extensive prior knowledge of the 3D visual world.

Existing methods can be broadly categorized into two classes, according to how the prior knowledge is processed. The former class [4, 6, 7, 18, 38, 48, 50, 77, 79] encodes 3D geometry by neural networks that are trained from scene-level 3D datasets, and then inferences the geometry in a new image with a feed-forward pass. Due to the scarcity of supervised data, these methods often suffer from poor reconstruction quality in unseen scenarios. The other class [17, 19, 28, 32–34] stores 3D models in a database, then retrieve and assemble similar 3D models to match the

†Project lead; ✉ corresponding author

input image. However, the limited geometric clues from a single image make it difficult to precisely identify and arrange the correct models. Moreover, since it is impractical for a 3D database to contain every possible model that exactly corresponds to the input image, the retrieved models can only approximately align with the objects, leading to inconsistencies. Therefore, methods in both classes lack accuracy and sufficient out-of-domain generalizability, in terms of novel object shapes and unseen scene layouts.

Recent image-to-3D object generation models [20, 24, 27, 29, 35, 39, 40, 43, 44, 64–66, 68–71, 73, 75, 78, 80], with powerful 3d prior and generalization capabilities, can generate high-quality geometry from a single object image. Building upon these pre-trained models, a common approach for scene generation involves using them as tools within a multi-step compositional generation process, which includes segmenting the scene image, completing individual object images, generating each object, and optimizing their spatial relationships [5, 21, 81], as shown in Fig. 2. While these methods leverage the priors of 3D object generation models, the generation process is inherently lengthy and prone to error accumulation – errors in intermediate steps can significantly distort the final result. Moreover, the optimization of spatial relationships cannot directly optimize 3D objects generated one by one by the previous stage that lacks global scene context, leading to misalignments between the generated instances and the overall scene. Therefore, if inter-object spatial relationships can be modeled directly within the 3D generation model, it is possible to construct an end-to-end pipeline that addresses these issues by generating all instances simultaneously with coherent spatial arrangements.

We propose MIDI, which extends pre-trained 3D object generation models to multi-instance diffusion models, establishing a new paradigm for compositional 3D scene generation. Our approach enables the simultaneous creation of multiple 3D instances with accurate spatial relationships from a single scene image, moving beyond independent object generation to a holistic understanding of the scene. Building upon large-scale pre-trained image-to-3D object generation models [35, 71, 78, 80], MIDI employs a novel multi-instance attention mechanism that effectively captures complex inter-object interactions and spatial coherence directly within the generation process, eliminating the need for complex multi-step procedures. This advanced design allows for the direct generation of cohesive 3D scenes, significantly enhancing both efficiency and accuracy. Due to the universal nature of spatial relationships between objects, we effectively supervise the interactions between 3D instances using a limited amount of scene-level datasets [15, 16] during training. Additionally, we incorporate single-object data for regularization, thereby maintaining the generalization ability of the pre-trained model.

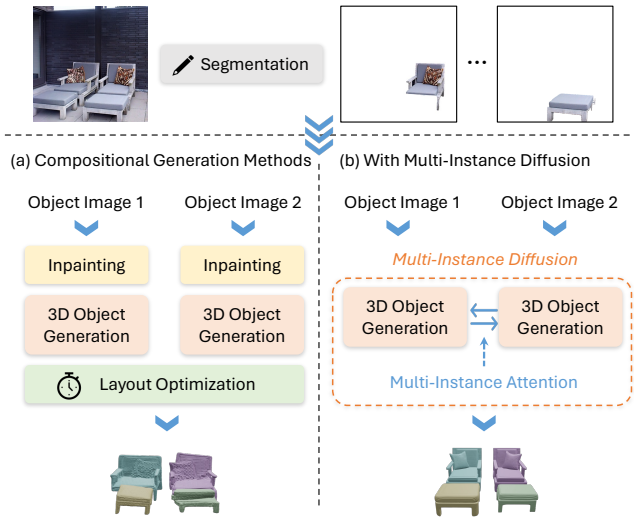

Figure 2. Comparison between our scene generation pipeline with multi-instance diffusion and existing compositional generation methods.

To validate the effectiveness of our proposed paradigm, we conduct experiments on synthetic datasets [15, 16], real-world scenes [8, 62], and various stylized scene images generated by text-to-image diffusion models [52, 57]. Results demonstrate that MIDI significantly advances the field of 3D scene generation by effectively modeling inter-object interactions through our multi-instance attention mechanism in the pre-trained 3D generation model. MIDI produces high-quality 3D scenes with accurate geometries and spatial layouts, while exhibiting strong generalization capabilities. In summary, our main contributions are as follows:

- We establish a new paradigm for compositional 3D scene generation by proposing a multi-instance diffusion model, which extends pre-trained image-to-3D object generation models to generate spatially correlated 3D instances.
- We introduce a novel multi-instance attention mechanism that effectively models cross-instance interactions, ensuring the coherence and accurate spatial relationships.
- Experiments demonstrate MIDI achieves state-of-the-art performance, significantly improving the generation of 3D scenes by accurately capturing inter-object relationships and providing better alignment with the input.

2. Related Work

2.1. Scene Reconstruction from a Single Image

Recovering the 3D structure of a scene from a single image is a fundamental challenge in computer vision. Existing methods can be categorized into feed-forward reconstruction methods [4, 6, 7, 18, 38, 48, 50, 77, 79], retrieval-based methods [17, 19, 28, 32–34], and recent compositional generation approaches [5, 11, 21, 63, 81].

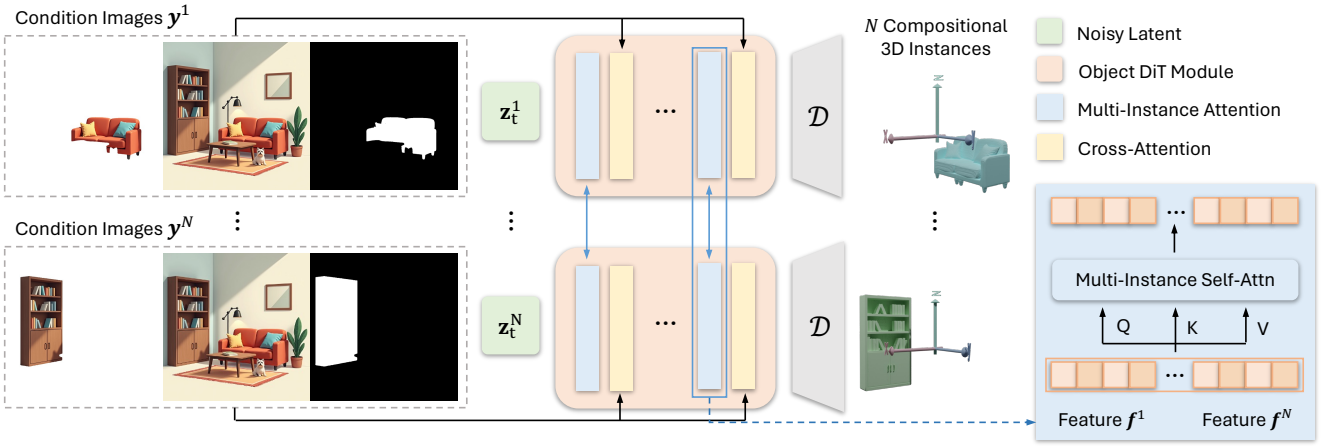

Figure 3. Method overview. Based on 3D object generation models, MIDI denoises the latent representations of multiple 3D instances simultaneously using a weight-shared DiT module. The multi-instance attention layers are introduced to learn cross-instance interaction and enable global awareness, while cross-attention layers integrate the information of object images and global scene context.

Feed-forward reconstruction methods. Feed-forward reconstruction methods [4, 6, 7, 18, 38, 48, 50, 77, 79] leverage 3D spatial knowledge and use 3D supervision to train end-to-end regression systems. They typically employ encoder-decoder architectures to predict scene properties such as geometry and instance labels from a single image. While jointly predicting scene layout and object poses ensures intrinsic correctness, these methods often suffer from limited reconstruction quality due to the scarcity of supervised 3D scene data and struggle to generalize to out-of-distribution images.

Retrieval-based methods. Retrieval-based methods [17, 19, 28, 32–34] reconstruct scenes by retrieving and aligning 3D models from a database based on the input image. For example, DiffCAD [17] trains diffusion models [23, 59–61] under synthetic data supervision, to model distributions of CAD object shapes, poses, and scales, which facilitates CAD model retrieval and alignment to the image input. Although these methods can produce detailed objects by leveraging existing 3D assets, they heavily depend on database diversity and often face retrieval errors due to insufficient information from single images, leading to misalignments.

Compositional generation methods. Recent compositional generation methods [5, 11, 21, 81] utilize large-scale perceptual and generative models in both image [31, 41, 47, 52, 54, 55, 57, 58] and 3D object [13, 29, 78] domains to improve scene reconstruction. These methods typically involve a multi-stage pipeline, including image segmentation [55], object completion [57], per-object generation [29, 78], and layout optimization. While they enhance generalization capabilities by leveraging pre-trained models, their complex pipelines are susceptible to error accumulation, and the lack of global scene context during per-object processing can lead to misaligned results. Our work addresses these issues by leveraging a pre-trained image-to-3D object generation model to simultaneously generate multiple 3D instances with interrelated relationships, improving robustness and maintaining strong generalization.

2.2. 3D Object Generation from a Single Image

Advancements in diffusion models [23, 60] and large-scale datasets [9, 10] have propelled progress in 3D generation [12, 24, 27, 35, 37, 39, 40, 43, 44, 46, 56, 64, 66, 68–73, 78, 80]. Several image-to-3D object generation methods [26, 43, 44, 64, 66, 68, 69, 73] adopt a two-stage pipeline that involves generating multi-view images and then reconstructing 3D objects. They fine-tune pre-trained image [52, 57] or video [2] diffusion models to produce multi-view images and employ large reconstruction models [24, 64, 74, 82] or optimization-based methods [67] to recover geometries. Another group of work [35, 36, 71, 78, 80] focuses on generating 3D native geometry by training large-scale generative models, which typically comprise a variational autoencoder [30] and a latent diffusion transformer (DiT) [51]. These models produce high-quality geometries with strong generalization due to training on diverse datasets. Building upon these advancements, we fine-tunes such an object geometry generator to create compositional instances while retaining generalization ability.

3. Preliminary: 3D Object Generation Models

Large-scale 3D object generation models [35, 36, 71, 78, 80] are the foundation of our approach. These models often comprise three main components: 1) a transformer-based variational autoencoder (VAE) [30] with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, which compress 3D geometric representations

into a low-dimensional latent space, and 2) a denoising transformer network ϵ_θ, trained on the compressed latent space to transform noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ into the original 3D data distribution $\mathbf{z}_0$, and 3) a group of condition encoders, such as CLIP [53] and DINO encoders [49] for encoding text or image conditions, which are then passed to the denoising network by cross-attention mechanism.

At inference time, the denoising process generates samples in the latent space, and the decoder $\mathcal{D}$ produces geometric representations like SDF values or tri-plane features, which can be converted into a 3D mesh by applying marching cubes [45] or using an additional mapping network.

4. MIDI: Multi-Instance 3D Generation

MIDI lifts 3D object generation to compositional 3D instance generation, enabling the creation of 3D scenes with accurate spatial relationships from a single image. Specifically, given a scene image, our objective is to generate spatially correlated 3D latent tokens $\{\mathbf{z}_0^i\}_{i=1}^N$ corresponding to the N instances present in the image. These latent tokens can be decoded and directly combined to obtain high-quality 3D scenes.

In this section, Sec. 4.1 first introduces the overall framework of multi-instance diffusion models, detailing how it generalizes single-object diffusion models to handle multiple interacting instances. Sec. 4.2 then elaborates on the multi-instance attention mechanism that models cross-instance relationships in 3D space. Finally, Sec. 4.3 presents the training procedure of MIDI.

4.1. Multi-Instance Diffusion Models

As demonstrated by Fig. 3, our multi-instance diffusion models extend the original DiT modules of 3D object generation models in three aspects: 1) the latent representations of multiple 3D instances are denoised simultaneously (*i.e.* in parallel) using a shared denoising network, 2) a novel multi-instance attention mechanism is introduced into the DiT modules to learn cross-instance interaction and enable global awareness, and 3) a simple yet effective method for encoding image inputs, including local object images and global scene context.

Overview of framework. Our multi-instance diffusion model builds upon existing 3D object diffusion models by extending them to denoise the 3D representations of multiple instances simultaneously. Specifically, we retain the VAE of the base model to compress the 3D geometric representations of multiple instances into low-dimensional latent features $\{\mathbf{z}_0^i\}_{i=1}^N$. We extend the denoising network ϵ_θ to condition on the global scene image $\boldsymbol{c}_g$, the RGB images of the N local objects $\{\boldsymbol{c}_l^i\}_{i=1}^N$, and their corresponding masks $\{\boldsymbol{m}_l^i\}_{i=1}^N$. The denoising network learns to transform noise $\{\boldsymbol{\epsilon}^i \sim \mathcal{N}(0, I)\}_{i=1}^N$ into the 3D data distribution, effectively capturing the spatial configurations of the instances.

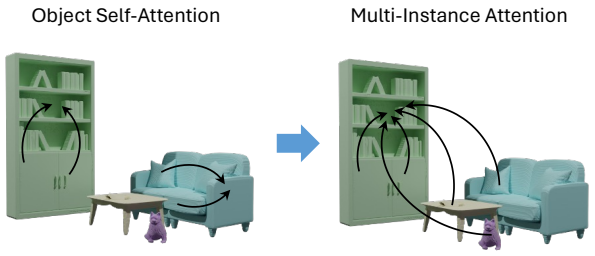

Figure 4. Multi-instance attention. We extend the original object self-attention, where tokens of each object query only themselves, to multi-instance attention, where tokens of each instance query all tokens from all instances in the scene.

Cross-instance interaction. Compositional 3D instance generation requires that the generated multiple instances exhibit interactive relationships in 3D space. To achieve this, we introduce a multi-instance attention mechanism within the denoising process, which models cross-instance interactions in the latent feature space during denoising. The integration of this mechanism transforms the generation of multiple objects from independent processes into a synchronous interactive process, enhancing global scene coherence and ensuring that the spatial relationships among objects are accurately represented.

Image conditioning. To encode all the image conditions, we propose a simple yet effective method, involving 1) the encoding of both global scene information and local instance details and locations with a ViT-based image encoder τ_θ [49], and 2) integrating the image embeddings using cross-attention layers. Specifically, for each instance z^i, we concatenate its RGB image $\boldsymbol{c}_l^i$, mask $\boldsymbol{m}_l^i$, and the global scene image $\boldsymbol{c}_g$ along the channel dimension, resulting in a composite representation $\mathbf{y} \in \mathbb{R}^{h \times w \times c}$, where $c = 7$. The composite image is then passed into a ViT-based encoder with extended input channels to extract a sequence of image features. Finally, we use a cross-attention mechanism in the transformer-based denoising network to integrate the conditioning image features.

4.2. Multi-Instance Attention

We now introduce the multi-instance attention mechanism, which is the key of MIDI to enforce spatial relationship across multiple 3D instances. This mechanism extends original object self-attention layers by connecting different instances within the attention computation (See Fig. 3).

Specifically, we transform the K original object self-attention layers into multi-instance attention layers by integrating the features of all instances $\{\boldsymbol{f}^i\}_{i=1}^N$ into the attention process, formulated as:

$$\boldsymbol{f}_{\text{out}}^i = \text{Attention}\left(\boldsymbol{f}^i, \{\boldsymbol{f}^j\}_{j=1}^N\right), \tag{1}$$

where $\boldsymbol{f}^i$ is the feature of instance i, and Attention$(\cdot)$ de-

notes the attention function that allows each instance to attend to the features of all instances in the scene, including itself. Therefore, as illustrated in Fig. 4, each token within a particular instance now queries information from tokens of all instances in the scene. This enables the attention mechanism to effectively model cross-instance interactions by considering the collective set of tokens, thereby capturing inter-object relationships and spatial dependencies.

4.3. Training

To train MIDI, we extend the loss of our base model, which utilizes the rectified flow [42] architecture, from single-object to multi-instance. For each training step, we sample a shared noise level t from 0 to 1 for all the instances $\{\mathbf{z}^i\}_{i=1}^N$, perturbing them along a simple linear trajectory:

$$\{\mathbf{z}_t^i\}_{i=1}^N = t\{\mathbf{z}_0^i\}_{i=1}^N + (1-t)\{\boldsymbol{\epsilon}^i\}_{i=1}^N, \tag{2}$$

where $\boldsymbol{\epsilon}^i \sim \mathcal{N}(0, I)$. Then we employ the following loss function to fine-tune the denoising network ϵ_θ and the image encoder τ_θ:

$$\mathbb{E}_{\{\mathbf{z}^i\}_{i=1}^N, \mathbf{y}, \{\boldsymbol{\epsilon}^i\}_{i=1}^N, t}\Big[\sum_{i=1}^N \|\mathbf{z}_0^i - \boldsymbol{\epsilon}^i - \epsilon_\theta(\mathbf{z}_t^i, t, \tau_\theta(\mathbf{y}))\|_2^2\Big]. \tag{3}$$

Since our training dataset is much smaller than the pre-training dataset of single-object 3D generation models, we incorporate additional 3D object datasets for training to retain the original generalization capability. In practice, with a 30% chance, we train the multi-instance diffusion model as a simple image-to-3D object generation model on a subset of Objaverse dataset [9] by turning off the multi-instance attention.

5. Experiments

5.1. Setup

Implementation details. We implemented MIDI based on our own image-to-3D object generation model, which utilizes the rectified flow architecture [42] and employs 21 attention blocks to construct the denoising transformer network, developed from existing 3D object generation methods [78, 80]. We initialize the image encoder τ_θ in MIDI using DINO [49], and expand the channel dimension of the input projection layer to accommodate 7 channels, corresponding to the concatenated inputs $\mathbf{y}$ (*i.e.* scene images, object images and masks). We set the resolution of $\mathbf{y}$ to 512. During training, we adopt the Low-Rank Adaptation (LoRA) technique [25] to fine-tune the pre-trained model efficiently. For the multi-instance attention mechanism, we set the number of multi-instance attention layers K to 5. Note that we focus on generating the instances in the scene and their spatial relationships. Planar background structures like floors and walls are not part of our generation scope, and they can be easily generated using existing methods [11, 81].

Datasets. We trained MIDI on the 3D-Front dataset [15], which is a synthetic 3D dataset of indoor 3D scenes with rich annotations. We performed cleaning by filtering out scenes with unreasonable object placements, such as intersecting or floating objects, resulting in approximately 15,000 high-quality scenes. The dataset is split into training and testing sets, with 1,000 scene images randomly selected as the test set. We evaluate MIDI on four widely used 3D scene reconstruction benchmarks, which includes synthetic datasets (*i.e.* test set of 3D-Front [15], BlendSwap [1]) and real-world datasets (*i.e.* Matterport3D [3], ScanNet [8]). To further validate the generalization ability of MIDI, we also test on scene images with various styles generated by the text-to-image diffusion model [52].

Baselines. We mainly compare our method with the state-of-the-art methods in scene reconstruction from single images, which includes feed-forward reconstruction methods PanoRecon [7], Total3D [48], InstPIFu [38] and SSR [4], retrieval-based methods DiffCAD [17], and compositional generation methods Gen3DSR [11] and REPARO [21].

Metrics. Following existing scene reconstruction methods [48, 81], we use Chamfer Distance and F-Score with the default threshold of 0.1 to evaluate the whole scenes. To further evaluate the geometric quality of individual 3D objects, we compute the Chamfer Distance and F-Score at the *object level* for each object within the scene, assessing the fidelity of each object's geometry independently. Additionally, we calculate Volumetric Intersection over Union (Volume IoU) between the bounding boxes of objects in the reconstructed or generated scene and those in the ground truth scene to assess the accuracy of object layouts and spatial arrangements. We also report the average runtime for each method to generate one scene.

5.2. Scene Generation on Synthetic Data

Tab. 1 reports quantitative comparisons on synthetic datasets, including 3D-Front [15] and BlendSwap [1]. Our method, MIDI, achieves the best performance among the state-of-the-art methods across all evaluated metrics without incurring much time consumption. Specifically, at the **object level**, our method significantly outperforms existing methods [4, 7, 11, 17, 21, 38, 48] due to our novel design based on pre-trained 3D object prior. Our MIDI, utilizing pre-trained object generation models, achieves a substantial leap in quality compared to methods that rely solely on reconstruction from limited data. At the **scene level**, metrics assessing the overall scene reconstruction quality and the alignment of object locations with ground truth demonstrate that our multi-instance diffusion models exhibit better

Table 1. Quantitative comparisons on synthetic datasets [1, 15] in scene-level Chamfer Distance (CD-S) and F-Score (F-Score-S), object-level Chamfer Distance (CD-O) and F-Score (F-Score-O), and Volume IoU of object bounding boxes (IoU-B).

Method	3D-Front					BlendSwap					
	CD-S↓	F-Score-S↑	CD-O↓	F-Score-O↑	IoU-B↑	CD-S↓	F-Score-S↑	CD-O↓	F-Score-O↑	IoU-B↑	Runtime↓
PanoRecon [7]	0.150	40.65	0.211	35.05	0.240	0.427	19.11	0.713	13.06	0.119	32s
Total3D [48]	0.270	32.90	0.179	36.38	0.238	0.258	37.93	0.168	38.14	0.328	39s
InstPIFu [38]	0.138	39.99	0.165	38.11	0.299	0.129	50.28	0.167	38.42	0.340	32s
SSR [4]	0.140	39.76	0.170	37.79	0.311	0.132	48.72	0.173	38.11	0.336	32s
DiffCAD [17]	0.117	43.58	0.190	37.45	0.392	0.110	52.83	0.169	38.98	0.457	64s
Gen3DSR [11]	0.123	40.07	0.157	38.11	0.363	0.107	60.17	0.148	40.76	0.449	9min
REPARO [21]	0.129	41.68	0.160	40.85	0.339	0.115	62.39	0.151	42.84	0.410	4min
Ours	**0.080**	**50.19**	**0.103**	**53.58**	**0.518**	**0.077**	**78.21**	**0.090**	**62.94**	**0.663**	40s

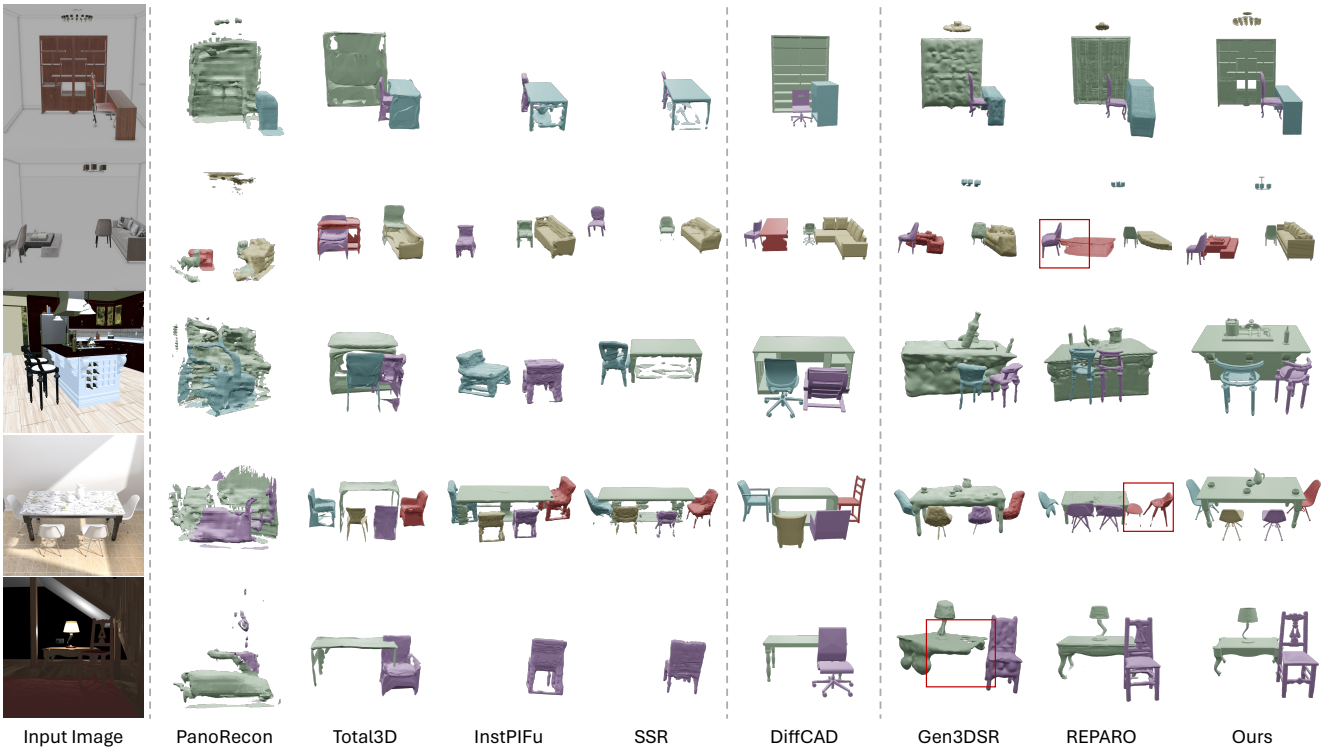

Figure 5. Qualitative comparisons on synthetic datasets, including 3D-Front [15] and BlendSwap [1].

robustness and accuracy compared to multi-stage object-by-object generation methods. MIDI effectively models global scene knowledge and the spatial relationships between objects, resulting in coherent and accurately arranged scenes.

The qualitiative comparison is shown in Fig. 5. Existing feed-forward reconstruction methods [7, 38, 48] often produce inaccurate geometry and misaligned scene layouts. Retrieval-based methods [17] produce results that do not accurately align with the input image. Multi-stage object-by-object generation methods [11, 21] generate instances that fail to align correctly with the overall scene due to the absence of scene context constraints during object image completion and 3D generation. In contrast, MIDI produces high-quality geometries and preserves accurate spatial configurations among multiple instances, due to our utilization of pre-trained object priors and effective multi-instance attention mechanism.

5.3. Scene Generation from Real Images

We further evaluate MIDI on Matterport3D [3] and ScanNet [8] using real images. For a qualitative comparison with other methods, we select 10 scenes from the test set of these two datasets, and sample one image from each scene as the input. We show the visual comparisons in Fig. 6, where we successfully generate scenes from real images and sig-

Figure 6. Qualitative comparisons on real-world data, including Matterport3D [3] and ScanNet [8].

nificantly outperform the previous works in the accuracy and completeness. This demonstrates the huge potentials and generalization capabilities of the multi-instance diffusion models in generating real-world 3D scenes.

5.4. Scene Generation from Stylized Images

To further assess the generalization capabilities of MIDI, we utilize the text-to-image diffusion model SDXL [52] to generate scene images with diverse styles, and test our method on them. Due to the limitations of existing methods in handling such diverse inputs, we compare MIDI exclusively with REPARO [21]. As shown in Fig. 8, MIDI generates accurate and coherent 3D scenes from the varied input images, demonstrating its strong generalization ability.

5.5. Ablation Study

We conduct ablation studies on 3D-Front [15] dataset to evaluate the impact of key components in MIDI. Specifically, we examine: 1) the number of multi-instance attention layers K, 2) the inclusion of the global scene image as conditioning input, and 3) the use of single-object dataset [9] for mixed training.

Base model without any design. We start with a baseline that directly fine-tunes the object generation model on the scene dataset without any design. However, the baseline model can not generate separable multi-instances, and shows weak modeling of spatial relationships (see Fig. 7) due to limited scene data for training.

Table 2. Ablation studies. We evaluate the number of multi-instance attention layers ($\#K$), the inclusion of global scene image (S.) input, and the use of Objaverse [9] (O.) for mixed training.

$\#K$	S.	O.	CD-S↓	F-Score-S↑	CD-O↓	F-Score-O↑	IoU-B↑
0	✗	✗	0.152	41.16	–	–	–
0	✓	✓	0.145	40.94	0.096	54.16	0.327
5	✓	✓	**0.080**	**50.19**	**0.103**	**53.58**	**0.518**
21	✓	✓	0.127	44.88	0.141	48.55	0.423
5	✗	✓	0.134	41.49	0.102	52.91	0.459
5	✓	✗	0.137	42.00	0.126	51.62	0.502

Number of multi-instance attention layers K. We experiment with $K = 0$, $K = 5$, and $K = 21$. Quantitative results in Tab. 2 and qualitative examples in Fig. 7 indicate that $K = 5$ achieves the best performance. With $K = 0$, the model fails to capture correct spatial relationships, leading to incoherent scene layouts, demonstrating the importance of our proposed multi-instance attention. When $K = 21$, excessive attention layers cause overfitting and distorted object geometries due to disruption of the pre-trained 3D prior after the model is trained on a relatively small scene

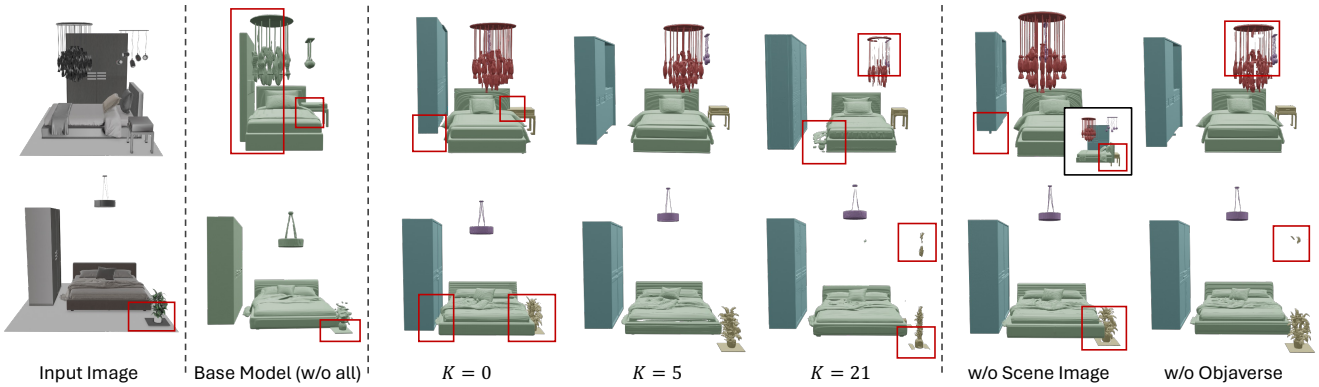

Figure 7. Qualitative ablation studies on the number of multi-instance attention layers K, and the use of global scene image conditioning, and mixed training with single-object dataset.

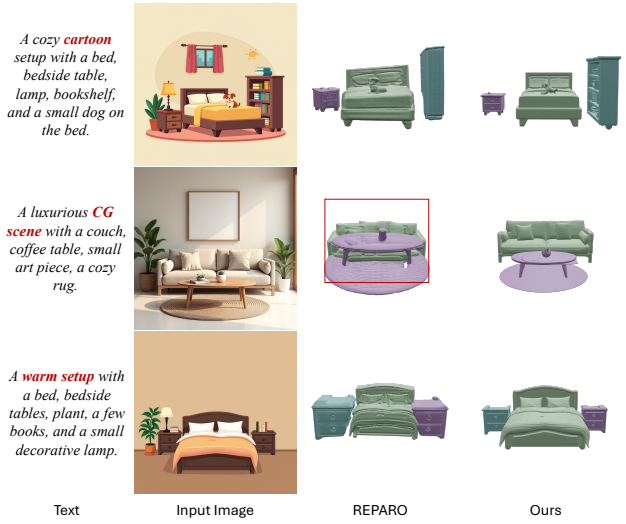

Figure 8. Qualitative comparisons on stylized images that are generated by text-to-image diffusion models.

dataset. We choose $K = 5$, where only a subset of the self-attention layers are converted to multi-instance attention, balancing between modeling interactions and preserving the pre-trained prior.

Global scene image conditioning. We remove the global scene image from the input and condition the model solely on local object images and masks. As shown in Tab. 2 and Fig. 7, excluding the global scene context significantly impairs the model's ability to generate coherent 3D scenes. The resulting scenes exhibit incorrect object placements and lack proper spatial relationships among instances.

Mixed training with single-object dataset. We explore the effect of mixed training by incorporating the Objaverse dataset [9] into the training process. Results in Tab. 2 and Fig. 7 show that, without this regularization, the model tends to produce objects with inferior geometry, as it overfits on the smaller scene dataset. Including single-object data helps preserve the object-level knowledge, enabling the model to generate high-quality geometries while effectively modeling inter-object interactions.

6. Conclusion

Limitations and future works. MIDI performs relatively poorly for tiny-resolution image input and complex inter-instance interaction, as shown in the supplementary materials. Building upon our proposed multi-instance diffusion for compositional 3D scene generation, future work can explore several directions: 1) extending the approach to model more complex interactions in compositional scenes, such as characters interacting with objects (*e.g.* "a panda playing a guitar"), which requires specialized datasets; 2) incorporating explicit 3D geometric knowledge to develop more efficient and expressive multi-instance attention mechanisms; 3) investigating the latent, implicit 3D perception capabilities of scene generation models; and 4) scaling the framework to handle a larger number of objects and operate in open-world environments.

Conclusion. This paper introduces MIDI, an innovative approach that significantly advances 3D scene generation from a single image. By extending pre-trained image-to-3D object generation models to multi-instance diffusion models and incorporating a novel multi-instance attention mechanism, MIDI effectively captures complex inter-object interactions and spatial coherence directly within the generation process. This enables the simultaneous generation of multiple 3D instances with accurate spatial relationships, leading to high-quality 3D scenes with precise geometries and spatial layouts. Extensive experiments demonstrate that MIDI achieves state-of-the-art performance while exhibiting strong generalization capabilities.

Acknowledgment

This work was supported by National Natural Science Foundation of China (62132001), and the Fundamental Research Funds for the Central Universities.

References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 5, 6

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 5, 6, 7

[4] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. In *2024 International Conference on 3D Vision (3DV)*, pages 1456–1467. IEEE, 2024. 1, 2, 3, 5, 6

[5] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. *arXiv preprint arXiv:2403.12409*, 2024. 2, 3

[6] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2023. 1, 2, 3

[7] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34: 8282–8293, 2021. 1, 2, 3, 5, 6

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 5, 6, 7

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3, 5, 7, 8

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[11] Andreea Dogaru, Mert Özer, and Bernhard Egger. Generalizable 3d scene reconstruction via divide and conquer from a single view. *arXiv preprint arXiv:2404.03421*, 2024. 2, 3, 5, 6

[12] Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. Tela: Text to layer-wise 3d clothed human generation. In *European Conference on Computer Vision*, pages 19–36. Springer, 2025. 3

[13] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 3

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[15] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2, 5, 6, 7, 1

[16] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 2

[17] Daoyi Gao, Dávid Rozenberszki, Stefan Leutenegger, and Angela Dai. Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 1, 2, 3, 5, 6

[18] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. Learning 3d object shape and layout without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1695–1704, 2022. 1, 2, 3

[19] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4022–4031, 2022. 1, 2, 3

[20] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, et al. threestudio: A unified framework for 3d content generation, 2023. 2

[21] Haonan Han, Rui Yang, Huan Liao, Jiankai Xing, Zunnan Xu, Xiaoming Yu, Junwei Zha, Xiu Li, and Wanhua Li. Reparo: Compositional 3d assets generation with differentiable 3d layout alignment. *arXiv preprint arXiv:2405.18525*, 2024. 2, 3, 5, 6, 7

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[24] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2, 3

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[26] Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024. 3, 1, 2

[27] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 2, 3

[28] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5134–5143, 2017. 1, 2, 3

[29] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2, 3

[30] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[32] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 260–277. Springer, 2020. 1, 2, 3

[33] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Patch2cad: Patchwise embedding learning for in-the-wild shape retrieval from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12589–12599, 2021.

[34] Florian Langer, Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Sparc: Sparse render-and-compare for cad model alignment in a single rgb image. *arXiv preprint arXiv:2210.01044*, 2022. 1, 2, 3

[35] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 2, 3, 1

[36] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025. 3

[37] Anran Liu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Zhiyang Dou, Hao-Xiang Guo, Ping Luo, and Wenping Wang. Part123: part-aware 3d reconstruction from a single-view image. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[38] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *European Conference on Computer Vision*, pages 429–446. Springer, 2022. 1, 2, 3, 5, 6

[39] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 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*, pages 10072–10083, 2024. 2, 3

[40] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[42] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5, 1

[43] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3

[44] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 2, 3

[45] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 4

[46] Quan Meng, Lei Li, Matthias Nießner, and Angela Dai. Lt3sd: Latent trees for 3d scene diffusion. *arXiv preprint arXiv:2409.08215*, 2024. 3

[47] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 16784–16804, 2022. 3

[48] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 2, 3, 5, 6

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 5

[50] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems*, 34:12013–12026, 2021. 1, 2, 3

[51] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[52] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 5, 7

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[55] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3, 1

[56] Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Angela Dai, and Matthias Nießner. L3dg: Latent 3d gaussian diffusion. *arXiv preprint arXiv:2410.13530*, 2024. 3

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[62] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[63] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20507–20518, 2024. 2

[64] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 2, 3

[65] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

[66] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 2, 3

[67] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 3

[68] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 2, 3

[69] Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, Yu Qiao, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion. *arXiv preprint arXiv:2406.03184*, 2024. 3

[70] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024.

[71] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 2, 3, 1

[72] Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024.

[73] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3

[74] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 3

[75] Wangbo Yu, Li Yuan, Yan-Pei Cao, Xiangjun Gao, Xiaoyu Li, Wenbo Hu, Long Quan, Ying Shan, and Yonghong Tian. Hifi-123: Towards high-fidelity one image to 3d content generation. In *European Conference on Computer Vision*, pages 258–274. Springer, 2024. 2

[76] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023. 1

[77] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 1, 2, 3

[78] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 2, 3, 5, 1

[79] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3d: A universal model for panoptic 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2023. 1, 2, 3

[80] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 5, 1

[81] Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Zero-shot scene reconstruction from single images with deep prior assembly. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3, 5

[82] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024. 3

MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation

Supplementary Material

7. Background

Base model. Following scalable 3D object generation methods [35, 71, 78, 80], we firstly trains a VAE to compress 3D geometric representations into a low-dimensional latent space. Specifically, $\mathbf{x} \in \mathbb{R}^{L\times 6}$, which represents positions and normals of L points, are mapped to latent space by $\mathbf{z} = \mathcal{E}(\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^{l\times c}$, and l denotes the length of the tokens after compression. The latents are converted back to the 3D space by regressing signed distance function (SDF) values using $\mathbf{s} = \mathcal{D}(\mathbf{z})$. Following 3DShape2Vecset [76], the VAE comprises of several transformer blocks.

Next, the denoising network ϵ_θ is trained in the compressed latent space to transform noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ into the original 3D data distribution. During training, following the rectified flow architecture [42], the original data $\mathbf{z}_0$ is perturbed along a simple linear trajectory:

$$\mathbf{z}_t = t\mathbf{z}_0 + (1-t)\boldsymbol{\epsilon} \tag{4}$$

for $t = 1, \cdots, T$, where T represents the number of steps in the diffusion process. In practice, we adopt logit-normal sampling [14] to increase the weight for intermediate steps. The denoising network ϵ_θ, featuring 21 attention blocks with residual connections, is trained to approximate the slope of the distribution transformation trajectory by minimizing the following loss:

$$\mathbb{E}_{\mathbf{z},\mathbf{y},\boldsymbol{\epsilon}\sim\mathcal{N}(0,I),t}[\|\mathbf{z}_0 - \boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{y}))\|_2^2] \tag{5}$$

where τ_θ is the image encoder, and $\mathbf{y}$ is the conditioning image, incorporated into the denoising transformer via cross-attention mechanism.

8. Implementation Details

Training. we trained MIDI to simultaneously generate up to $N = 7$ instances. We selected this value based on an analysis of the 3D-FRONT dataset [15], where we observed that scenes containing five or fewer objects constitute the majority, while scenes with more than five objects are relatively rare. Instead of excluding scenes with more than 5 objects, we employed a clustering method to select five representative objects from such scenes for training. During training, we randomly dropped the image conditioning with a probability of 0.1. We adopted the same strategy as in the training of the base model, utilizing logit-normal sampling [14] to increase the weight of intermediate diffusion steps, which helps the model focus on the more challenging stages of the generation process. For the training configuration, we used a learning rate of 5×10^{-5} and trained MIDI for 5 epochs on 8 NVIDIA A100 GPUs.

Table 3. Training costs. (Batch size is set to 1)

Number of Instances N	VRAM (GB)	Speed (iter/s)
$N = 1$	15	1.50
$N = 3$	17	0.83
$N = 5$	19	0.55
$N = 7$	21	0.40

Inference. In our experimental setup, we first used Grounded-SAM [55] to segment the scene images, obtaining masks for individual objects. We then applied our multi-instance diffusion model to generate compositional 3D instances using classifier-free guidance [22], which enhances the fidelity and coherence of the generated scenes. We set the number of inference steps to 50 and the guidance scale to 7.0. The entire process of generating a 3D scene from a single image takes approximately 40 seconds on an NVIDIA A100 GPU.

9. Additional Discussions

MIDI vs. compositional generation methods. As show in Fig. 9, existing compositional generation methods involve a multi-step process, generating 3D objects one by one and then optimizing their spatial relationships. However, this type of methods lack the contextual information of the global scene when generating objects, thus generating inaccurate or mismatched 3D objects. In addition, it is very difficult to optimize the accurate scene layout based on a single image, and the position of similar objects will be reversed when there are similar objects in the scene (as shown in Fig. 9). In contrast, our method models object completion, 3D generation and spatial relationships in a multi-instance diffusion model, thus generating coherent and accurate 3D scenes.

Training costs. Table 3 presents the training costs for MIDI. As the number of instances N increases, both GPU memory requirements and training time increase. However, even when $N = 7$, resource utilization remains manageable, demonstrating the scalability of MIDI.

Texture generation. To generate textured 3D scene from single images, we firstly synthesize 3D geometry with our MIDI, and then leverage MV-Adapter [26] to generate texture for each instance with the partial image of instance im-

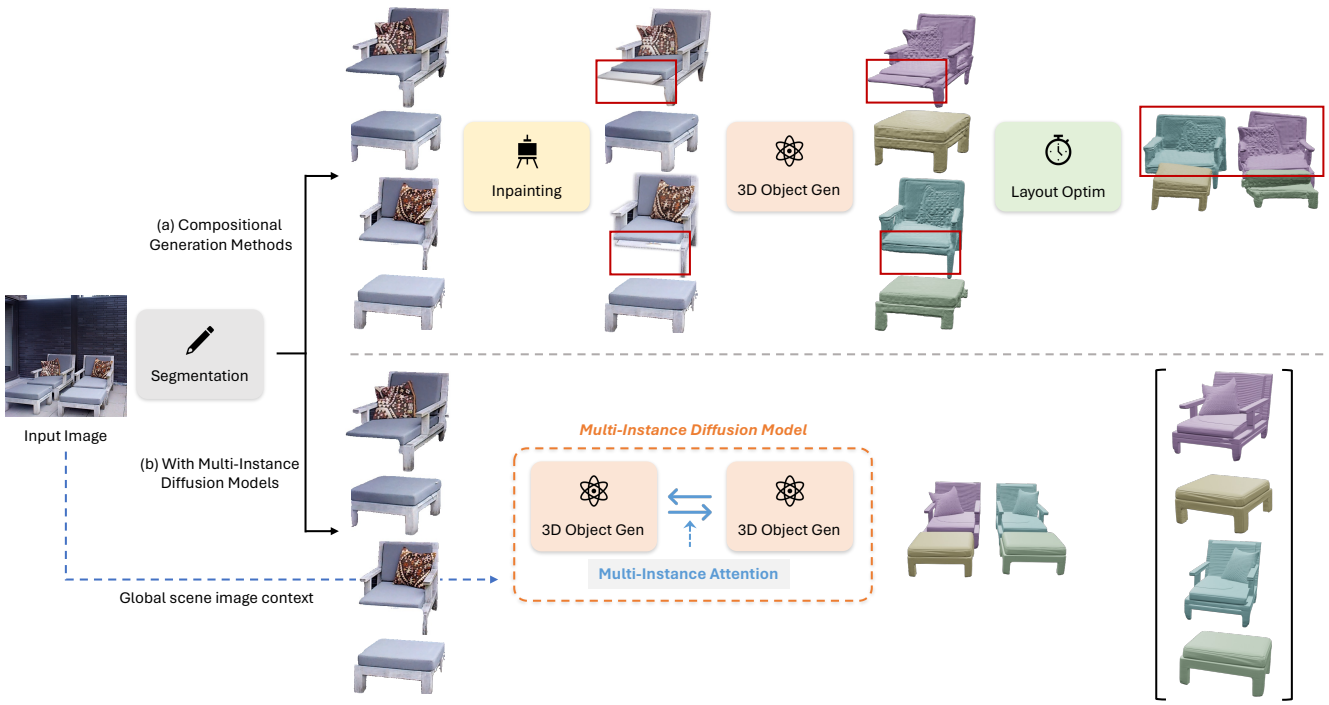

Figure 9. Detailed comparison between existing compositional generation methods and our multi-instance diffusion.

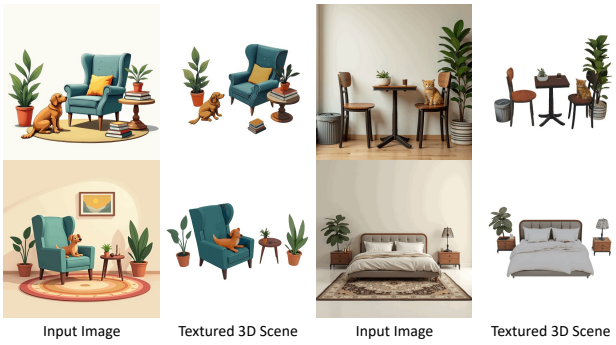

Figure 10. Visualization results of textured 3D scene generation with MV-Adapter [26].

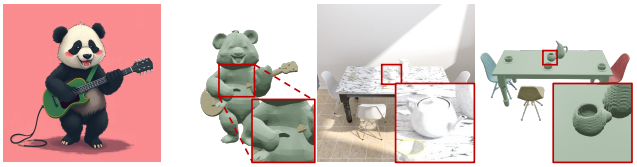

Figure 11. Failure cases.

age as input. The visualization results are shown in Fig. 10. It is recommended to interactively experience the generated 3D scenes in our project page.

10. Limitations

We present two typical failure examples of MIDI in Fig. 11. While MIDI generates 3D instances within the global scene coordinate system—specifically, a normalized space ranging from -1 to 1—this approach causes smaller objects to occupy a relatively minor portion of the overall space. Consequently, these small objects may have lower resolution compared to objects generated in their canonical spaces, where the entire capacity of the model can focus on a single object. We believe that enhancing the multi-instance diffusion model to generate objects in their canonical spaces, along with their spatial positions within the scene, could address this issue by allowing each object to be generated at optimal resolution.

Also, our model is constrained by the simplicity of interaction relationships present in existing scene datasets. As a result, MIDI may struggle to generate scenes featuring intricate interactions, such as objects with dynamic interplays. We anticipate that introducing more complex and diverse training data, encompassing a wider variety of object interactions and spatial relationships, would enhance the model's capacity to generalize at the level of object spatial interactions. This improvement would enable the generation of scenes with more sophisticated and realistic inter-object dynamics.