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

Supplementary Material

1. Background

Base model. Following scalable 3D object generation methods [5, 8, 10, 11], 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 [9], 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 [6], the original data $\mathbf{z}_0$ is perturbed along a simple linear trajectory:

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

for $t = 1, \cdots, T$, where T represents the number of steps in the diffusion process. In practice, we adopt logit-normal sampling [1] 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:

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

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

2. 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 [2], 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 [1] 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 1. 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 [7] 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 [3], 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.

3. Additional Discussions

MIDI vs. compositional generation methods. As show in Fig. 1, 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. 1). 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 1 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 [4] to generate texture for each instance with the partial image of instance image as input. The visualization results are shown in Fig. 2. It is recommended to interactively experience the generated 3D scenes in our project page.

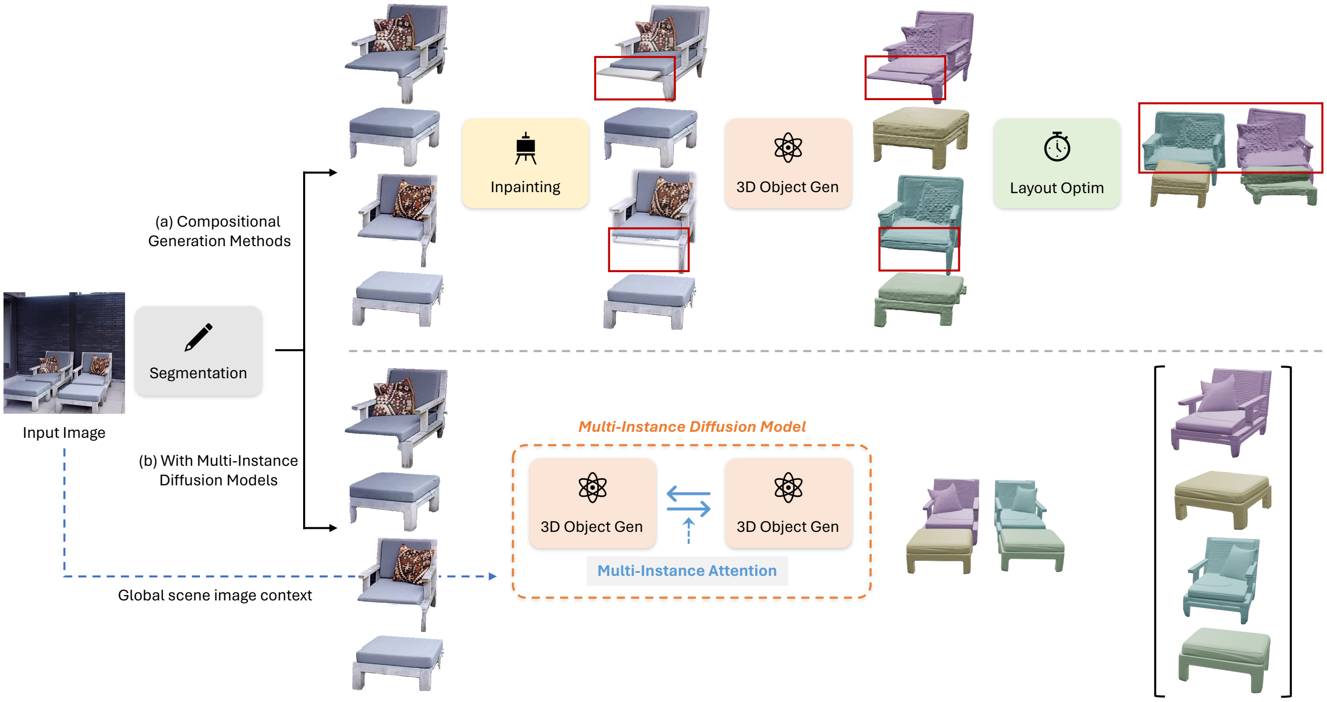
(a) Compositional Generation Methods
Inpainting
3D Object Gen
Layout Optim
Segmentation
Input Image
(b) With Multi-Instance Diffusion Models
Multi-Instance Diffusion Model
3D Object Gen
3D Object Gen
Multi-Instance Attention
Global scene image context

Figure 1. Detailed comparison between existing compositional generation methods and our multi-instance diffusion.

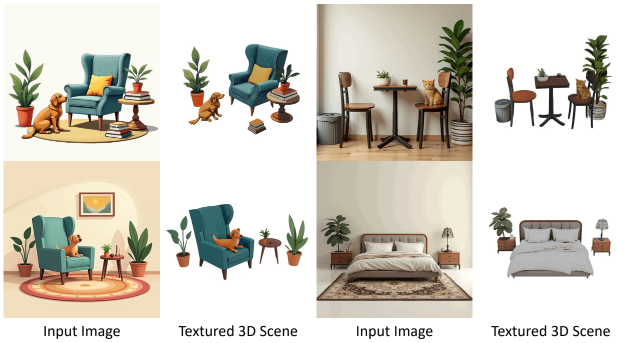

Figure 2. Visualization results of textured 3D scene generation with MV-Adapter [4].

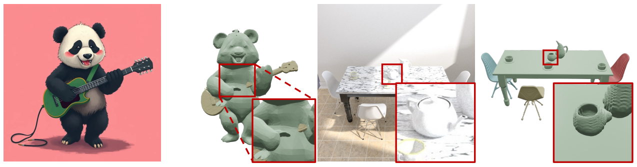

Figure 3. Failure cases.

4. Limitations

We present two typical failure examples of MIDI in Fig. 3. 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.

References

[1] 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

[2] 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. 1

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

[4] 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. 1, 2

[5] 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. 1

[6] 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. 1

[7] 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. 1

[8] 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. 1

[9] 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

[10] 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. 1

[11] 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. 1