Scale3D: Autoregressive Modeling for Large Outdoor Scene Generation

Di Qi, Zheng Sun, Xuanyang Zhang[†], and Gang Yu[‡]

StepFun, Shanghai, China
{qidi, sunzheng, zhangxuanyang, yugang}@stepfun.com

Abstract. We present Scale3D, an autoregressive model for generating high-quality outdoor 3D scenes at scale. Scaling 3D generation to vast environments like city blocks usually degrades local geometric details as the layout expands. Scale3D decouples this problem into a two-stage process. A causal autoregressive model first plans the global layout by predicting scene chunks over a 2D spatial grid. A masked autoregressive model then synthesizes the actual 3D geometry within each chunk. We reconstruct these local structures using a 3D chunk-based VAE designed to preserve sharp edges and high-fidelity surfaces. We also introduce a pipeline that extracts spatial relationships directly from geometry, allowing users to control complex scene layouts via simple text prompts. Unlike patch-based diffusion models that struggle with boundary artifacts, Scale3D maintains long-range spatial coherence across expansive outdoor datasets while rendering detailed 3D structures.

Keywords: Unbounded 3D generation · Hierarchical autoregressive models · High-quality 3D VAE

1 Introduction

Generating large-scale outdoor 3D scenes requires solving two problems: ensuring accurate local geometry and maintaining long-range structural coherence. High-quality 3D assets demand precise underlying geometry, rather than merely plausible multi-view appearances. When scaling these assets to vast environments like city blocks, local geometric details degrade as the spatial layout expands.

Existing paradigms generally fail on at least one of these fronts. Methods that lift 2D diffusion priors [1, 26] into 3D via multi-view reconstruction [4, 16, 34] often synthesize coherent appearances but distort the underlying geometry. On the other hand, direct 3D generation methods denoise explicit spatial representations, typically by using native 3D diffusion models to sequentially synthesize volumetric chunks [11, 22, 24, 32]. This paradigm encounters severe limitations when applied to unbounded scenes. Because the iterative denoising process inherently relies on a fixed, local context, these models cannot maintain long-range

[†]Project leader, [‡]Corresponding author.

structural coherence. Attempts to extend them via resampling-based inpainting [21] or chunk-based outpainting [11] inevitably produce visible boundary seams and geometric discontinuities. Beyond these structural failures, current implementations are further restricted by their underlying 3D VAEs, which often produce low-fidelity reconstructions, and by the absence of robust mechanisms for controllable synthesis from high-level commands.

Autoregressive(AR) models scale remarkably well in language [5, 10] and visual generation [28, 30], offering a natural way to model long-range spatial dependencies. But treating 3D geometry as a simple causal sequence fails in practice. A single causal model forced to predict both the global structure and the high-frequency local details usually collapses into producing overly smooth, low-quality meshes, as demonstrated by our ablations in Sec. 4.4.

To this end, we introduce Scale3D, a hierarchical framework designed to harness the global modeling strengths of autoregression while ensuring local geometric fidelity. The core of Scale3D is a hierarchical AR generator that decouples global planning from local synthesis: a causal AR model first establishes a coherent global layout, upon which a masked bidirectional AR model then synthesizes high-fidelity local details. This design directly resolves the fundamental tension between large-scale structure and fine-grained geometry. To support this generation, we build an enhanced 3D chunk-based VAE. It employs direct mesh partitioning for occupancy supervision at higher resolution, augmented by salient geometry guidance from sharp-edge sampled points [2] and surface normals. Additionally, to enable precise interactive control, we develop an automated annotation pipeline that produces a structured text dataset for spatial conditioning. Validated on large-scale unbounded 3D scenes, Scale3D demonstrates superior performance in both unconditional synthesis (measured by FPD/KPD) and text-conditioned generation.

2 Related Work

2.1 3D Scene Generation Methods

The field of 3D scene generation has evolved along two primary paradigms based on their geometric representations. The first employs 2D-to-3D lifting pipelines that leverage strong image generation priors. Early works pioneered perpetual view generation from single images [15, 18], which later extended to text-conditioned radiance field generation [3, 7]. More recent approaches use 2D diffusion models to generate multi-view content before lifting it to 3D [4,13,34,35,39]. While these methods create diverse scenes, they suffer from a fundamental limitation: geometric quality is compromised by accumulated depth prediction errors and view synthesis inconsistencies. These geometric shortcomings motivate the second paradigm: direct 3D generation. Within this paradigm, diffusion-based methods like BlockFusion [32], NuiScene [11], and LT3SD [22] synthesize scenes by learning to denoise structured latent representations such as triplanes or vector sets. While these methods achieve better geometric quality than 2D-to-3D approaches, they remain predominantly diffusion-based and struggle with

unbounded generation due to their reliance on local context modeling. These limitations are addressed by our hybrid AR framework, which decouples global layout planning from local detail synthesis to handle large-scale environments.

Current methods for controllable 3D generation often rely on structured priors that pose significant accessibility challenges for average users. One group relies on complex structural priors, such as the scene graphs required by GraphDreamer [6], CommonScenes [36], InstructLayout [17] and C3DO-SG [20], or the geographic data (like 2.5D OSM data for UrbanWorld [27] and HD maps for InfiniCube [18]) which are difficult to obtain. Another group, including Holodeck [33] and SceneCraft [9], utilizes large language models for spatial arrangement but is limited to manipulating predefined asset libraries and cannot generate new geometry. SSEditor [37] offers a mask-conditional approach for semantic scene generation and editing using a diffusion model, providing a different form of control. Even text-driven methods like Text2LiDAR [31] remain constrained by their output modality, generating sparse point clouds where text primarily controls semantic content rather than detailed spatial layout. This widespread reliance on hard-to-acquire priors or limited control modalities highlights a critical gap in achieving fine-grained control over both the generation and spatial arrangement of novel 3D geometry. Our work addresses this by introducing a purely text-driven system that generates complete 3D meshes from natural language.

2.2 Large-Scale Unbounded Scene Generation

Generating unbounded 3D scenes while maintaining spatial coherence presents a challenge. Most diffusion-based methods divide scenes into equal chunks and adopt chunk-based generation with local context modeling [12, 19, 32]. These fixed-context approaches use overlapping windows and resampling-based inpainting techniques like RePaint [21] for scene extension. However, this strategy introduces artifacts at chunk boundaries due to insufficient global context propagation. While explicit outpainting models [11] can improve boundary coherence and inference speed, achieving seamless spatial transitions across large-scale environments remains an open problem. In contrast, our autoregressive framework models dependencies between chunks sequentially, ensuring global coherence by design rather than a post-hoc repair strategy like inpainting.

3 Methodology

3.1 3D Chunk VAE for Local Geometry

Scale3D VAE is built on the chunk VAE from NuiScene [11] (see A.2 of Appendix for preliminary). We propose an enhanced mesh chunk VAE featuring two techniques, Flexible Resolution Control (FRC) and Salient Geometry Guidance (SGG), to improve the reconstructed geometry details as shown in Fig. 1.

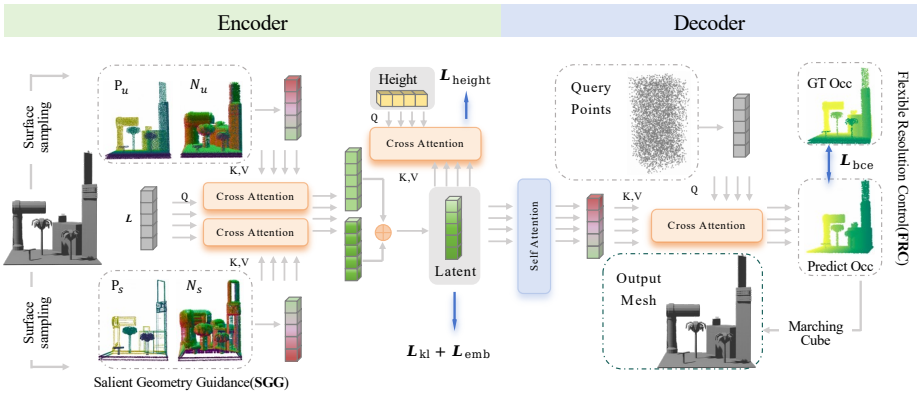

Fig. 1: Illustration of Scale3D chunk VAE. Scale3D VAE introduces high resolution supervision and salient geometry guidance to improve the reconstruction fidelity.

Flexible Resolution Control. The most straightforward approach to enhance chunk geometric details is to increase the resolution of the occupancy supervision. In NuiScene [11], the data pipeline for generating chunk occupancy proceeds as follows: the scene mesh is first converted into a Signed Distance Field (SDF), which is then thresholded to obtain occupancy values. Chunk occupancy is subsequently sampled using coordinates along the length (L) and width (W) dimensions. However, due to the technical limitation of the mesh2sdf tool [8], which only supports up to 32-bit representation, the maximum achievable resolution is constrained to approximately 1k voxels along each axis. Since the current scene resolution already approaches this upper limit, directly increasing the chunk resolution in the above data process pipeline is non-trivial.

To address the limitation, we propose a novel pipeline that directly samples the chunk mesh from the scene mesh. This chunk mesh is then converted into a chunk SDF, from which the high-resolution chunk occupancy is derived and more details can be found in A.1 of Appendix. This approach elevates the upper resolution limit of chunk occupancy from 50 to 1000 voxels per axis, thereby enabling more flexible and precise control over the chunk occupancy resolution to improve the reconstruction fidelity.

Salient Geometry Guidance. In addition to increasing the resolution, we also enhance geometric reconstruction details from the perspective of input points. We define the sharp-edge sampled points proposed in Dora [2] and point normal as the salient geometry guidance. Instead of only using uniformly sampled points, we further incorporate salient geometry guidance for VAE input feature to enhance geometry details. After that, we obtain the latent vector set Z_u and Z_s from uniform points and sharp-edge points by updating the VAE encoder as

follows:

$$Z_u = \text{CrossAttn}(L, \text{CAT}(\text{PosEmb}(P_u), \text{PosEmb}(N_u))), \tag{1}$$

$$Z_s = \text{CrossAttn}(L, \text{CAT}(\text{PosEmb}(P_s), \text{PosEmb}(N_s))), \tag{2}$$

where L is the learnable query set, P_u and P_s are the uniform points and sharp-edge points sampled from the scene mesh, N_u and N_s are the corresponding normal maps. The final latent set Z is formed by aggregating features from both sampling strategies via element-wise summation $Z = Z_u + Z_s$.

Objective optimization. As for VAE decoder and training loss, we largely follow the formulation from NuiScene as described in the preliminary section except that, enabled by our FRC pipeline, the query points and occupancy supervision can be sampled at a much higher resolution. Chunk VAE optimizes the binary cross entropy loss $\mathcal{L}_{bce}$=BCE$(O, \tilde{O})$ between occupancy prediction and ground truth, a KL divergence regularization $\mathcal{L}_{kl}$ of Z, a embedding constraint $\mathcal{L}_{emb}$=MSE$(Z, \hat{Z})$ and height prediction loss $\mathcal{L}_{height}$=MSE$(h, \tilde{h})$. The total optimization loss is formulated as:

$$\mathcal{L}_{\text{VAE}} = \lambda_{bce}\mathcal{L}_{bce} + \lambda_{kl}\mathcal{L}_{kl} + \lambda_{emb}\mathcal{L}_{emb} + \lambda_{height}\mathcal{L}_{height}. \tag{3}$$

3.2 Hierarchical Autoregressive Generation

Our hierarchical autoregressive framework addresses large-scale 3D scene generation through a unified hierarchical architecture that seamlessly integrates global layout planning and local detail synthesis, enabling both unconditional large-scale generation and optional text-conditioned controllable synthesis.

Hierarchical Architecture Design We model a scene as a grid of $H \times W$ latent chunks, where each chunk $\mathbf{z}_i \in \mathbb{R}^{l \times d}$ is a set of l tokens encoded by VAE. Our framework generates this grid autoregressively, using a hierarchical approach (Fig. 2) that decomposes the task into two stages to balance global coherence and local detail.

Global Structure Generation. In the global stage, we treat each chunk as a unit and generate the scene structure following a Z-order spatial traversal. This naturally captures 2D spatial dependencies while maintaining a strict causal ordering. A causal transformer predicts the global embedding of each chunk conditioned on all l tokens of every previously generated chunk. For chunk i at spatial coordinates (x, y), the global autoregressive model predicts:

$$p(\mathbf{z}_i^{\text{global}} | \mathbf{z}_{<i}^{\text{global}}, \texttt{[BEGIN]}, \mathbf{c}) = f_{\text{global}}(\mathbf{z}_{<i}^{\text{global}}, \texttt{[BEGIN]}, \mathbf{c}, \text{RoPE}(x, y)), \tag{4}$$

where $\mathbf{z}_{<i}^{\text{global}}$ represents all previously generated chunk embeddings, `[BEGIN]` serves as a special start token for sequence initialization, $\mathbf{c}$ is the conditioning vector (text embedding or $\emptyset$ for unconditional generation), and RoPE(x, y) encodes 2D spatial coordinates using Rotary Positional Embeddings, which naturally capture geometric relationships between chunks across scene scales.

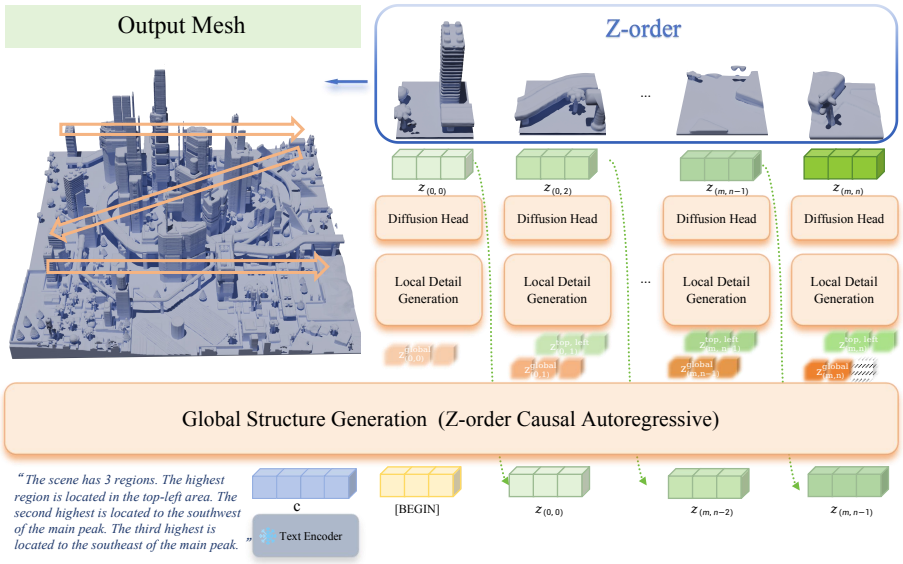

Fig. 2: Illustration of hierarchical autoregressive framework and the inference process.

Local Detail Generation. Given $\mathbf{z}_i^{\text{global}}$, the local stage generates the internal structure in chunk i. Since point cloud data lacks sequential ordering, we use masked autoregressive generation with bidirectional attention [14]. The model predicts masked tokens within the chunk, conditioned on the unmasked tokens, the global embedding, and tokens from neighboring chunks. This explicit neighbor conditioning ensures geometric continuity across chunk boundaries.

For chunk i, we randomly select a masking subset $\mathcal{M}_i \subset \{1, 2, \ldots, l\}$ and predict masked tokens autoregressively within the set:

$$p(\mathbf{z}_i^{\mathcal{M}_i} | \mathbf{z}_i^{\mathcal{U}_i}, \mathbf{c}_i^{\text{hierarchical}}) = \prod_{j \in \mathcal{M}_i} p(\mathbf{z}_i^j | \mathbf{z}_i^{\mathcal{U}_i}, \mathbf{z}_i^{k<j}, \mathbf{c}_i^{\text{hierarchical}}), \tag{5}$$

where $\mathcal{U}_i = \{1, 2, \ldots, l\} \setminus \mathcal{M}_i$ are unmasked indices, $\mathbf{z}_i^{k<j}$ denotes previously predicted masked tokens with $k \in \mathcal{M}_i, k < j$, and:

$$\mathbf{c}_i^{\text{hierarchical}} = \text{Concat}([\mathbf{z}_{\text{top}}, \mathbf{z}_{\text{left}}, \mathbf{z}_i^{\text{global}}]), \tag{6}$$

where $\mathbf{z}_{\text{top}}$ and $\mathbf{z}_{\text{left}}$ are tokens from the top and left neighboring chunks of chunk i, respectively. This hierarchical context provides: (1) structural guidance through $\mathbf{z}_i^{\text{global}}$, and (2) spatial continuity through neighboring chunks, ensuring geometric coherence across chunk boundaries.

Training Objectives and Optimization We train Scale3D with three complementary objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diffusion}} + \lambda_{\text{coh}} \mathcal{L}_{\text{coherency}} + \lambda_{\text{dpo}} \mathcal{L}_{\text{DPO}}. \tag{7}$$

Diffusion-Enhanced Token Generation. Following MAR-Diffusion [14], our local generator is not a categorical token predictor. For each masked position, the bidirectional transformer outputs a conditioning vector $\mathbf{h}$, and a 12-layer diffusion head (Eq. 8) samples the continuous latent token via DDPM. We train this masked autoregressive framework using a diffusion objective for continuous embeddings. For the j-th latent token in chunk i, the diffusion objective is:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{i,j,t\sim[1,T],\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{z}_{i,t}^j, t, \mathbf{c}_i^{\text{hierarchical}})\right\|_2^2\right], \tag{8}$$

where $\mathbf{z}_{i,t}^j$ represents the noisy j-th token embedding in chunk i at diffusion timestep t, and ϵ_θ is the learned denoising network. During inference, tokens are initialized with Gaussian noise and iteratively denoised over T steps.

Coherency-Aware Regularization. To prevent mode collapse into repetitive patterns, a common failure in large-scale synthesis, we introduce neighborhood similarity regularization that encourages spatial diversity while preserving local coherence:

$$\mathcal{L}_{\text{coherency}} = \mathbb{E}_{i\sim\text{chunks}}\left[1 - \cos(\mathbf{z}_i^{\text{global}}, \mathbf{z}_{\text{neighbor}(i)}^{\text{global}})\right], \tag{9}$$

where neighbor(i) denotes spatially adjacent chunks (top and left). This cosine-based loss directly maximizes embedding diversity between neighboring chunks, promoting varied yet coherent spatial arrangements.

Direct Preference Optimization. For unconditional generation, which is prone to producing empty or repetitive scenes, we employ DPO to learn from human preferences over scene quality:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(\mathbf{c},\mathbf{Z}_w,\mathbf{Z}_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(\mathbf{Z}_w|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{Z}_w|\mathbf{c})} - \beta\log\frac{\pi_\theta(\mathbf{Z}_l|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{Z}_l|\mathbf{c})}\right)\right], \tag{10}$$

where $\mathbf{c}$ is the conditioning vector (text embedding or $\emptyset$ for unconditional generation), $(\mathbf{Z}_w, \mathbf{Z}_l)$ are preferred and dispreferred scene generation outputs, π_{ref} is the reference model, and $\mathcal{D}$ is a human-curated preference dataset with $\approx$400 pairs collected from ranked unconditional 8×8 generations. DPO effectively addresses mode collapse and repetitive patterns common in large-scale 3D generation, promoting spatially coherent and semantically meaningful scenes.

Training Details. We set $\lambda_{\text{coh}} = 0.1$, $\lambda_{\text{dpo}} = 0.05$, $\beta = 0.1$ for DPO. Detailed implementation details are provided in A.4 of Appendix.

Text-Conditioned Controllable Generation To extend beyond unconditional generation, we incorporate optional text conditioning to allow natural language control over scene layouts. This mechanism enables precise scene control through natural language while preserving the framework's capacity for purely unconditional synthesis.

Automated Scene Description Pipeline. Training this conditional model requires paired text-geometry data. To build these pairs, we develop an automated annotation pipeline that converts raw 3D geometry into natural language descriptions. The process begins by extracting 2D height maps from the 3D meshes. We then apply watershed segmentation to isolate individual objects, computing their explicit geometric properties and spatial relationships. Finally, an LLM synthesizes these structured spatial features into text captions.

Conditioning Integration and Guidance Input prompts are encoded via a frozen CLIP text encoder to produce a conditioning vector $\mathbf{c}$. This vector directly guides the global layout generation (Eq. 4) and subsequently influences the local geometry synthesis through the hierarchical context (Eq. 6). To enable classifier-free guidance, we randomly drop these text conditions during training with probability $p_{\text{drop}} = 0.1$.

4 Experiments

4.1 Dataset Collection

We construct our dataset using 13 large-scale scenes with dense urban structures for fair comparison with NuiScene [11]. Raw meshes undergo preprocessing to remove artifacts, normalize to $[-1, 1]$ range, scaled according to architectural anchors for consistency, and regularize ground planes through occupancy-based detection and mesh completion.

To extract chunks, we convert the meshes to signed distance fields (SDF) and apply spatial filtering to exclude regions with massive ground holes or excessive flatness. Qualified regions are segmented into 0.1×0.1 grid meshes using Blender's bisect plane operation. Each chunk yields SDF, occupancy grids, point clouds, and normals at higher resolution than NuiScene. We construct datasets with grid sizes of 2×2, 4×4, 8×8, 16×16, and 8×20 for both single-scene and 13-scene configurations as detailed in Tab. 1. More data collection details can be found in A.1 of Appendix.

Table 1: Statistics of the collected chunk data of both one scene and 13 scene across multi scale resolutions.Chunk diversity boosts generalization, and multi-resolution support ensures consistent multi-scale spatial generation.

Dataset	2×2	4×4	8×8	16×16	8×20
1 scene	100k	100k	30k	4k	2k
13 scenes	280k	200k	12k	5k	4k

4.2 Generation Results

Experiment setup

Table 2: Performance comparison between NuiScene variants (different receptive fields) and Scale3D on unconditional generation. Note that KPD*=KPD$\times 10^3$.

Dataset	Method	2×2		8×8		32×32	
		KPD* (↓)	FPD (↓)	KPD* (↓)	FPD (↓)	KPD* (↓)	FPD (↓)
1 scene	NuiScene(2x2) [11]	**0.21**	0.07	0.57	0.32	1.92	0.87
	NuiScene(16x16) [11]	0.22	**0.06**	0.54	0.29	1.38	0.76
	Scale3D (**Ours**)	0.24	0.07	**0.52**	**0.24**	**0.91**	**0.71**
13 scenes	NuiScene(2x2) [11]	0.34	**0.11**	0.67	0.43	2.65	0.97
	NuiScene(16x16) [11]	0.34	0.12	0.63	0.34	2.36	0.82
	Scale3D (**Ours**)	**0.33**	0.12	**0.59**	**0.27**	**1.97**	**0.75**

Dataset. We conduct experiments on the dataset of 13 large-scale scenes, partitioned into training and testing sets. For unconditional generation, we evaluate geometric fidelity on both a single scene and the full 13-scene set, using high-resolution chunks $(256, h_{vox}, 256)$. For conditional generation, we assess spatial reasoning and instruction following on the full dataset using lower-resolution chunks $(50, h_{vox}, 50)$ to facilitate large-scale evaluation.

Baselines. Our primary baseline is NuiScene [11]. We evaluate two NuiScene variants: (1) the original model with its native 2×2 receptive field, and (2) a version we retrained with a maximum receptive field of 16×16 with mixed 2×2, 4×4, 8×8, and 16×16 multi-scale training. The latter ensures a fair comparison by matching the contextual window of our method. Since NuiScene does not support text conditioning, we evaluate our conditional generation results qualitatively. We also train XCube [25] and LT3SD [22] on our 13-scene benchmark. Constrained by training-inference resolution binding and global voxelization, we restrict our qualitative comparison of these two baselines to 8×8 scale. More implementation details can be found in A.8 of Appendix.

Evaluation protocol. For unconditional generation, we employ Fréchet PointNet++ Distance (FPD) and Kernel PointNet++ Distance (KPD) [23] with 61440 sampled points per chunk. We evaluate at 2×2, 8×8, and 32×32 scales, generating 10k, 640, and 40 scenes respectively. For conditional generation, we report the CLIP score, Uni3D cosine similarity [38], and Spatial Relation Accuracy on prompts with explicit directional and height-rank constraints. We use 100 diverse prompts from the test set, generating 8 scenes per prompt for scales of 2×2, 4×4, 8×8, and 16×16.

Generation Results

Unconditional Generation. As shown in Tab. 2, Scale3D demonstrates competitive performance against NuiScene baselines. The advantage is most pronounced at larger scales (8×8 and 32×32), demonstrating our method's superior ability to maintain global consistency. At 8×8, Scale3D also outperforms the adapted

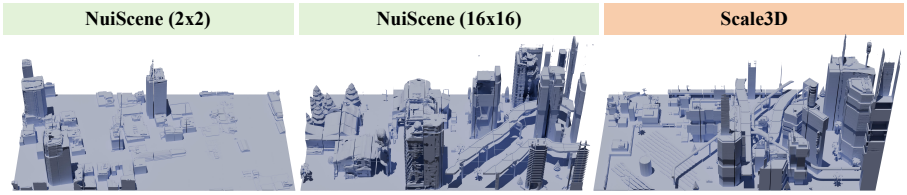

Fig. 3: Qualitative comparisons on generated geometry at 8x16 scale.

XCube (FPD/KPD* = 0.45/0.72) and LT3SD baselines quantitatively. A.6 reports a scaling curve showing that the performance gap grows monotonically from 4×4 for scales $\geq$ 8×8. Qualitatively, Fig. 3 shows that Scale3D generates scenes with more organized layouts and finer geometric details. Notably, retraining NuiScene with a larger receptive field substantially improves its performance, validating the importance of a large contextual window for scene coherence. This strong performance holds on the more diverse 13-scene dataset, where metrics show a slight decrease as expected, reflecting the increased complexity. Furthermore, as shown in Fig. 4, the adapted XCube and LT3SD fail to maintain structural integrity even at their maximum viable 8×8 scale, exhibiting discontinuous ground planes and severe voxel-blurring artifacts respectively. Additional visual comparisons are provided in A.8 of Appendix, and we also show generated 3D scenes with texture in A.9 of Appendix.

Conditional Generation. Our method demonstrates robust and scalable spatial control via text conditioning. Tab. 3 shows high CLIP scores (0.98 at 2 × 2 to 0.91 at 16 × 16) and strong Uni3D similarities. Spatial Relation Accuracy on 50 test prompts further reaches 0.78/0.71 for direction/height-rank at 8×8 (random baselines: 0.25/0.17). As shown in Fig. 6, Scale3D excels at rendering complex relative arrangements, correctly preserving directional and height relationships even as the scene scales. These results validate our framework's effectiveness in multi-modal spatial control. Additional visual comparisons are provided in A.7 of Appendix.

Table 3: Quantitative evaluation for spatial-controlled large-scene generation. The CLIP and Uni3D scores quantify text-geometry consistency and geometric plausibility, respectively.

Metrics	2×2	4×4	8×8	16×16
Clip score	0.98	0.98	0.94	0.91
Uni3D	0.36	0.32	0.31	0.29

4.3 Component Analysis

High-Fidelity VAE Reconstruction

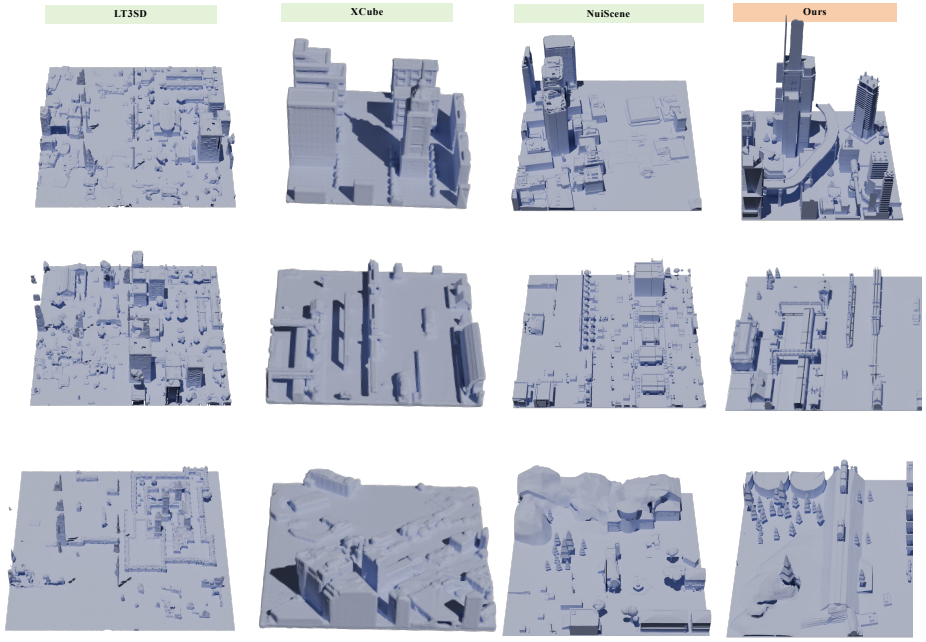

Fig. 4: Qualitative comparison with baselines (XCube and LT3SD) at 8×8 scale.

Dataset. We evaluate Scale3D-VAE through a pilot study on a single-scene dataset with 2×2 grid configuration, followed by an extended assessment of its generalization capability using a 13-scene dataset under the same grid setting. Each quad-chunk was subdivided into four sub-chunks for VAE training, yielding 400k and 1,120k training samples respectively. From each dataset, 1k samples were held out for quantitative evaluation.

Baselines. We adapt the NuiScene VAE with vector set as our baseline. To overcome the reconstruction limitations imposed by fixed occupancy resolution in the original pipeline, we introduce a revised approach: instead of splitting chunks directly from scene-level occupancy, we first extract mesh chunks from the scene mesh and then convert them to occupancy chunks at a flexible resolution. This allows increasing the occupancy resolution from (50, h_{vox}, 50) to (256, h_{vox}, 256), significantly improving the quality of reconstruction. We train the NuiScene VAE and our Scale3D VAE in the new collected datasets and more implementation details can be found in A.4 of Appendix.

Evaluation protocol. We evaluate the quality of the reconstruction using a comprehensive set of metrics: volumetric IoU [29], point cloud Chamfer distance (CD_p), normal Chamfer Distance (CD_n), and F-score. All metrics are computed between the reconstructed mesh and ground truth using 10k sampled points and

normals. Before evaluation, both meshes are normalized to the range $[-1, 1]$. The voxel size for IoU and the threshold for F-score are set to $0.05m$.

Table 4: Comparisons between Nuiscene VAE and Scale3D VAE across 1 scene and 13 scenes.

Dataset	Method	IoU (↑)	CD-p (↓)	CD-n (↓)	F-Score (↑)
1 scene	Nuiscene VAE [11]	0.4690	0.0062	0.0212	0.8314
	Scale3D VAE (**Ours**)	**0.6686**	**0.0009**	**0.0087**	**0.9696**
13 scenes	Nuiscene VAE [11]	0.4066	0.0071	0.0223	0.7923
	Scale3D VAE (**Ours**)	**0.6114**	**0.0013**	**0.0094**	**0.9567**

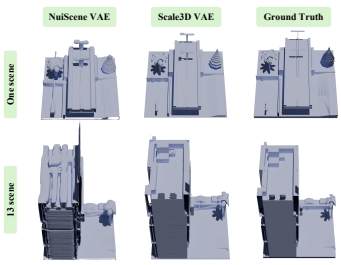

Fig. 5: Qualitative comparisons on reconstructed geometry. Zoom-in can see more details.

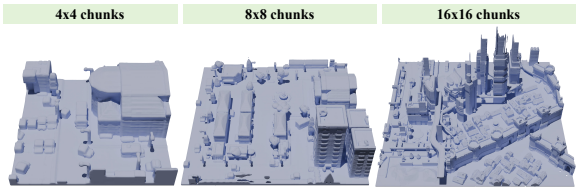

Fig. 6: Example of Spatial Control Generation. The prompt is: "The scene has 3 regions. The highest region located in the top-right area. The second highest region is located to the west of the main peak. The third highest region is located to the southeast of the main peak."

Reconstruction results. We evaluate the Scale3D VAE from both quantity and quality results across 1 scene and 13 scene dataset. As Tab. 4 shows, Scale3D VAE consistently achieves superior performance than NuiScene VAE from IoU, CD_p, CD_n and F-score metrics. These results demonstrate the effectiveness of our data collection pipeline and VAE architecture. We also render the reconstructed mesh to check the actual visualization improvement. As shown in Fig. 5, compared with the ground truth mesh, Scale3D VAE is capable of reconstructing more accurate geometric details than NuiScene VAE.

Hierarchical AR Modeling

Settings. To decouple the contribution of our hierarchical AR framework from VAE improvements, we conduct a controlled experiment on the single-scene dataset. Specifically, we train the Scale3D AR generator on the original frozen

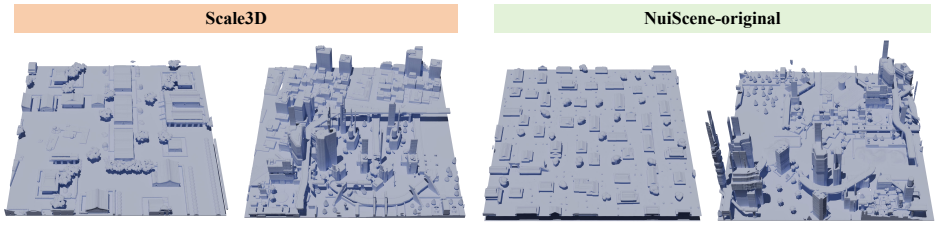

Fig. 7: Qualitative comparison using frozen NuiScene VAE on 13-scene data. Even with identical VAE quality, Scale3D generates more coherent layouts, fewer repetitive patterns, and better global consistency compared to NuiScene.

NuiScene VAE at its native resolution $(50, h_{vox}, 50)$ and compare it directly against the original NuiScene baseline. The evaluation protocol is identical to Sec. 4.2.

Generation Results. Table 5 presents the results on single-scene data. Despite identical reconstruction baselines, Scale3D outperforms the NuiScene baseline significantly at large scales (32×32). While performance is comparable at small scales (2×2), Scale3D exhibits superior scalability as scene complexity increases. Qualitative comparisons in Fig. 7 (13-scene data) further demonstrate that Scale3D generates coherent layouts with reduced repetitive patterns. These results confirm that our hierarchical design is the primary contributor to long-range coherence, independent of VAE fidelity.

Table 5: Architectural decomposition using frozen NuiScene VAE (single-scene). Scale3D AR framework outperforms NuiScene at large scales even with identical VAE.

Method	2×2		8×8		32×32	
	FPD (↓)	KPD* (↓)	FPD (↓)	KPD* (↓)	FPD (↓)	KPD* (↓)
NuiScene (Original)	**0.049**	**0.11**	0.37	0.48	1.26	1.33
Scale3D+NuiScene VAE	0.06	0.22	**0.33**	**0.41**	**0.93**	**0.98**

4.4 Ablation study

VAE key module ablations Scale3D VAE mainly introduces two main techniques, which consist of flexible resolution control (FRC) and input with salient geometry guidance (SGG). Through our data collection pipeline, we can increase the occupancy supervision resolution from $(50, h_{vox}, 50)$ to $(256, h_{vox}, 256)$. We ablate the effectiveness of RFC and SGG in Tab. 6. We observe that high resolution occupancy supervision and salient feature can both improve the reconstruct results but high-resolution supervision leads to a more substantial improvement.

Table 6: Ablations of RFC and SGG in Scale3D VAE. High supervision resolution and salient geometry guidance can both improve reconstructed results.

Settings		IoU (↑)	CD-p (↓)	CD-n (↓)	F-Score (↑)
FRC	SGG				
×	×	0.4690	0.0062	0.0212	0.8314
✓	×	0.6631	0.0009	0.0090	0.9657
×	✓	0.4692	0.0059	0.0225	0.8362
✓	✓	**0.6686**	**0.0009**	**0.0087**	**0.9696**

We also ablate the effectiveness of latent number in A.3 of Appendix and find the number of latent (tested with 8, 16, 32, and 64) exhibits minimal impact on the reconstruction results, which can be attributed to the relatively simple geometry of the scene data. Consequently, a default latent dimension of 16 was adopted in this paper.

Unconditional Generation To efficiently validate our AR design, we conduct ablation studies on a single scene, with each chunk at a $(50, h_{vox}, 50)$ resolution. More visual results on 13 scene can be found in A.7 of Appendix.

Hybrid Framework. We compare our hierarchical design against two monolithic alternatives:

(1) Block-wise Causal AR model, which generates an entire chunk embedding in one step, performs the worst across all scales. As shown in Fig. 8 (a), its outputs are consistently coarse, suggesting that simultaneously modeling high-level structure and fine-grained detail is too complex for a single generative head.

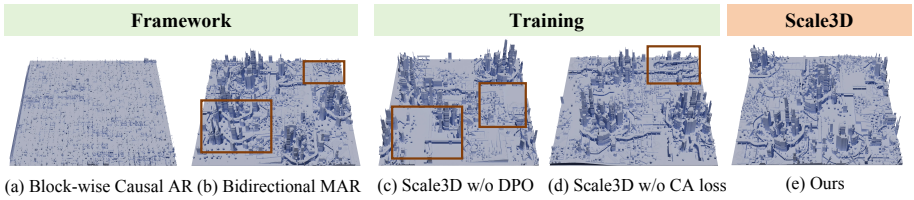

Fig. 8: Visual ablation of our hierarchical autoregressive method at 32x32 scale.

(2) Bidirectional MAR model, which treats generation as a token-level masking task, is competitive on small 2×2 scenes. However, its performance degrades substantially on larger scales, producing repetitive and incoherent structures, as shown in Fig. 8 (b). Specifically, it highlights discontinuous element repetition (red box, bottom-left) and structural collapse in large scenes (top-right). This indicates that while effective for local synthesis, MAR struggles to maintain global consistency without explicit high-level guidance.

Table 7: Ablation study of our hierarchical autoregressive method for unconditional generation.

Method	2×2 Chunks		8×8 Chunks		32×32 Chunks	
	FPD (↓)	KPD (↓)	FPD (↓)	KPD (↓)	FPD (↓)	KPD (↓)
Block-wise Causal AR	0.13743	12.34	0.17264	14.57	0.21043	17.42
Pure Bidirectional MAR	**0.00022**	0.05	0.00038	0.13	0.00265	0.68
w/o DPO	0.00031	0.07	0.00072	0.16	0.00132	0.47
w/o CA Loss	0.00026	0.05	0.00043	0.13	0.00098	0.38
Scale3D (Full)	0.00025	**0.05**	**0.00039**	**0.13**	**0.00095**	**0.36**

These results demonstrate the necessity of our hierarchical design, which successfully decouples global planning from local synthesis.

Direct Preference Optimization (DPO). A common failure mode in chunk-wise generation is the model defaulting to simplistic patterns like repetitive flat terrain or empty regions, as shown in Fig. 8 (c). We introduce DPO to mitigate this issue by encouraging structurally diverse outputs. As Tab. 7 shows, the inclusion of DPO significantly improves performance across all scales. It effectively steers the model away from low-complexity outputs and aligns its distribution with a preference for more structurally diverse scenes, proving essential for generating high-quality, large-scale results.

Coherency-Aware Regularization. Removing our Coherency-Aware Regularization (`w/o CA Loss`) consistently degrades performance across all scales (Tab. 7). Qualitatively, its absence introduces interruptions at chunk boundaries, accompanied by slight incoherent element repetition. Fig. 8 (d) shows incoherent repetition of "bridge" elements. This confirms its critical role in ensuring seamless transitions and preserving structural integrity.

5 Conclusion

We present Scale3D, an autoregressive framework that resolves the trade-off between global consistency and local fidelity in large-scale 3D scene generation. The key is a hierarchical decomposition: a causal autoregressive model first generates a global layout scaffold, which then guides a masked autoregressive model in synthesizing high-fidelity geometric details. Our method sets a new state of the art in unconditional generation and also enables controllable synthesis that precisely adheres to text-specified spatial constraints. We believe Scale3D establishes a robust paradigm for virtual world generation, with promising future directions in accelerating inference and enabling richer semantic control.

Acknowledgements

This work was supported by StepFun.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
2. Chen, R., Zhang, J., Liang, Y., Luo, G., Li, W., Liu, J., Li, X., Long, X., Feng, J., Tan, P.: Dora: Sampling and benchmarking for 3d shape variational auto-encoders. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16251–16261 (2025)
3. Chen, Z., Wang, G., Liu, Z.: Scenedreamer: Unbounded 3d scene generation from 2d image collections. IEEE transactions on pattern analysis and machine intelligence **45**(12), 15562–15576 (2023)
4. Chung, J., Lee, S., Nam, H., Lee, J., Lee, K.M.: Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384 (2023)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
6. Gao, G., Liu, W., Chen, A., Geiger, A., Schölkopf, B.: Graphdreamer: Compositional 3d scene synthesis from scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21295–21304 (2024)
7. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7909–7920 (2023)
8. Hu, Y., Li, T.M., Anderson, L., Ragan-Kelley, J., Durand, F.: Taichi: a language for high-performance computation on spatially sparse data structures. ACM Transactions on Graphics (TOG) **38**(6), 201 (2019)
9. Hu, Z., Iscen, A., Jain, A., Kipf, T., Yue, Y., Ross, D.A., Schmid, C., Fathi, A.: Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In: Forty-first International Conference on Machine Learning (2024)
10. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
11. Lee, H.H., Han, Q., Chang, A.X.: Nuiscene: Exploring efficient generation of unbounded outdoor scenes. arXiv preprint arXiv:2503.16375 (2025)
12. Lee, J., Lee, S., Jo, C., Im, W., Seon, J., Yoon, S.E.: Semcity: Semantic scene generation with triplane diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28337–28347 (2024)
13. Li, H., Shi, H., Zhang, W., Wu, W., Liao, Y., Wang, L., Lee, L.h., Zhou, P.Y.: Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In: European Conference on Computer Vision. pp. 214–230. Springer (2024)
14. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
15. Li, Z., Wang, Q., Snavely, N., Kanazawa, A.: Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In: European conference on computer vision. pp. 515–534. Springer (2022)

16. Liang, H., Cao, J., Goel, V., Qian, G., Korolev, S., Terzopoulos, D., Plataniotis, K.N., Tulyakov, S., Ren, J.: Wonderland: Navigating 3d scenes from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 798–810 (2025)
17. Lin, C., Lin, Y., Pan, P., Zhang, X., Mu, Y.: Instructlayout: Instruction-driven 2d and 3d layout synthesis with semantic graph prior. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
18. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14458–14467 (2021)
19. Liu, Y., Li, X., Li, X., Qi, L., Li, C., Yang, M.H.: Pyramid diffusion for fine 3d large scene generation. In: European Conference on Computer Vision. pp. 71–87. Springer (2024)
20. Liu, Y., Li, X., Zhang, Y., Qi, L., Li, X., Wang, W., Li, C., Li, X., Yang, M.H.: Controllable 3d outdoor scene generation via scene graphs. arXiv preprint arXiv:2503.07152 (2025)
21. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
22. Meng, Q., Li, L., Nießner, M., Dai, A.: Lt3sd: Latent trees for 3d scene diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 650–660 (2025)
23. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
24. Ren, J., Xu, M., Wu, J.C., Liu, Z., Xiang, T., Toisoul, A.: Move anything with layered scene diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6380–6389 (2024)
25. Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4209–4219 (2024)
26. 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. pp. 10684–10695 (2022)
27. Shang, Y., Lin, Y., Zheng, Y., Fan, H., Ding, J., Feng, J., Chen, J., Tian, L., Li, Y.: Urbanworld: An urban world model for 3d city generation. arXiv preprint arXiv:2407.11965 (2024)
28. Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., Nießner, M.: Meshgpt: Generating triangle meshes with decoder-only transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19615–19625 (2024)
29. Siddiqui, Y., Thies, J., Ma, F., Shan, Q., Nießner, M., Dai, A.: Retrievalfuse: Neural 3d scene reconstruction with a database. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12568–12577 (2021)
30. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)

31. Wu, Y., Zhang, K., Qian, J., Xie, J., Yang, J.: Text2lidar: Text-guided lidar point cloud generation via equirectangular transformer. In: European Conference on Computer Vision. pp. 291–310. Springer (2024)
32. Wu, Z., Li, Y., Yan, H., Shang, T., Sun, W., Wang, S., Cui, R., Liu, W., Sato, H., Li, H., et al.: Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. ACM Transactions on Graphics (ToG) **43**(4), 1–17 (2024)
33. Yang, Y., Sun, F.Y., Weihs, L., VanderBilt, E., Herrasti, A., Han, W., Wu, J., Haber, N., Krishna, R., Liu, L., et al.: Holodeck: Language guided generation of 3d embodied ai environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16227–16237 (2024)
34. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5916–5926 (2025)
35. Yu, H.X., Duan, H., Hur, J., Sargent, K., Rubinstein, M., Freeman, W.T., Cole, F., Sun, D., Snavely, N., Wu, J., et al.: Wonderjourney: Going from anywhere to everywhere. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6658–6667 (2024)
36. Zhai, G., Örnek, E.P., Wu, S.C., Di, Y., Tombari, F., Navab, N., Busam, B.: Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. Advances in Neural Information Processing Systems **36**, 30026–30038 (2023)
37. Zheng, H., Liang, Y.: Sseditor: Controllable mask-to-scene generation with diffusion model. arXiv preprint arXiv:2411.12290 (2024)
38. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. arXiv preprint arXiv:2310.06773 (2023)
39. Zhou, S., Fan, Z., Xu, D., Chang, H., Chari, P., Bharadwaj, T., You, S., Wang, Z., Kadambi, A.: Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In: European Conference on Computer Vision. pp. 324–342. Springer (2024)