Octree-MambaDiffuser: State-Space Diffusion with Octree Guidance for 3D Generation

1st Suhang Qian
School Of Computer Software
Tianjin University
Tianjin, China
qsh@tju.edu.cn

2nd Chao Xu*
School Of Computer Software
Tianjin University
Tianjin, China
xuchao@tju.edu.cn

3rd Yushi Li*
Department of Intelligent Science
Xi'an Jiaotong-Liverpool University
Suzhou, China
Yushi.Li@xjtlu.edu.cn

4th Chengtao Ji
Department of Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
Chengtao.Ji@xjtlu.edu.cn

5th Xiaobo Jin
Department of Intelligent Science
Xi'an Jiaotong-Liverpool University
Suzhou, China
Xiaobo.Jin@xjtlu.edu.cn

6th Xuanmo Zhang
Tianjin Second High School
Tianjin, China
13821826336@163.com

***Abstract*—Diffusion model-based approaches for 3D shape generation have advanced significantly. However, existing methods often fail to preserve both fine-grained geometric details and long-range structural coherence in complex shapes. While state space models (SSMs) can improve geometric consistency, directly integrating them with diffusion architecture often introduces causal constraints and limited local perception. To address these issues, we propose Octree-MambaDiffuser, a hierarchical framework that integrates octree-guided spatial partitioning with state-space diffusion. At first, we bridge the structural gap between 1D SSMs and 3D point clouds through a z-order serialization of octree representations. This provides a spatially coherent 1D sequence that better aligns with SSMs' autoregressive generation assumptions and improves long-range dependency modeling compared to naive traversal orders. The sequential data is then effectively processed by our scale-flexible MambaBlock which is enhanced with a Feature Conditioning Module (FCM). Acting as a critical stabilizer, the FCM pre-processes the input feature to reconcile Mamba's discrete-state dynamics with the continuous demands of the diffusion process. By applying a simple yet strategic sequence of pre-norm Layer Normalization and SiLU activation, the module ensures stable gradient propagation throughout the denoising trajectory. This not only primes the underlying state-space mechanism to model the evolving shape but also unlocks its ability to capture subtle geometric details. The subsequent dual-branch Mamba dynamically balances local feature with global structural context. Extensive experiments demonstrate that Octree-MambaDiffuser generates high-fidelity and diverse 3D shapes, outperforming existing methods in both visual quality and structural integrity.**

***Index Terms*—octree, state space model, diffusion model, 3D generation**

I. Introduction

As a fundamental generative task, 3D shape synthesis involves learning a mapping to the high-dimensional, non-Euclidean manifold of valid geometries. The objective is to produce novel, structurally sound 3D asset. While early generative models like GANs [1]–[3] and VAEs [4] laid important groundwork, diffusion models [5], [6] have recently become the dominant paradigm. This shift was catalyzed by their breakthrough success in 2D image synthesis [7], [8] and is sustained by their key advantages for 3D generation, including superior training stability, a principled likelihood-based framework that ensures better mode coverage, and a progressive denoising process well-suited for constructing coherent 3D structures. Early 3D diffusion models that operated directly in dense voxel or point cloud spaces faced significant challenges in computational efficiency and output fidelity due to the high dimensionality of 3D data. The adoption of latent diffusion strategies—inspired by breakthroughs in 2D generative modeling [8]—has enabled high-resolution 3D generation by performing diffusion in a compressed, learned latent space. This key insight has driven a series of powerful 3D diffusion architectures. For instance, Diffusion-SDF [9] employs a 3D VAE to compress Signed Distance Functions (SDFs) into a continuous latent space, while SDFusion [5] uses a 3D VQ-VAE to learn a discrete latent representation for SDFs; both train diffusion models in their respective latent

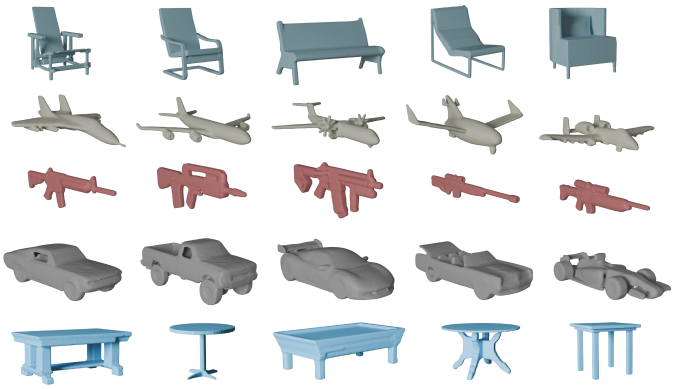

Fig. 1. High-quality and diverse results generated by our model.

* Corresponding authors.

spaces to generate high-fidelity 3D shapes. Similarly, 3D-LDM [10] applies latent diffusion to neural implicit representations by operating on shape-specific latent codes. These approaches significantly improve geometric quality, detail preservation, and scalability compared to earlier direct-domain methods. Beyond unconditional generation, conditional 3D synthesis has advanced through methods like DreamFusion [11] and Magic3D [12], which leverage pre-trained 2D diffusion models as priors via Score Distillation Sampling (SDS) to enable text-to-3D generation.

In parallel, state space models like Mamba [13]–[15] have emerged as a powerful and efficient alternative for sequence modeling, which challenges the dominance of Transformers. Their linear complexity and selective long-range reasoning are highly attractive for scaling 3D tasks. Yet, a direct application to 3D diffusion is nontrivial due to a fundamental architectural mismatch. Mamba's 1D, causal sequential processing conflicts with the non-sequential, spatial nature of 3D data, imposing two key limitations: it cannot efficiently model multi-directional relationships, and its token-by-token focus inherently lacks the flexible local perception required for high-resolution detail.

To overcome the limitations of causal SSMs in 3D generation, we propose Octree-MambaDiffuser, a hierarchical diffusion framework that bridges the spatial awareness of 3D data with the efficiency of Mamba. Specifically, we first introduce z-order serialization of octree representations, transforming the hierarchical 3D structure into a spatially coherent 1D sequence. This critical alignment step preserves local spatial neighborhoods in the serialized order and enables the subsequent SSM to efficiently model long-range geometric dependencies. In other words, it addresses the structural mismatch between 1D sequences and 3D data. Second, to enable stable cooperation between Mamba's discrete dynamics and the continuous diffusion process, we propose a Feature Conditioning Module (FCM). The module conditions the input features via a pre-norm and SiLU activation, ensuring stable gradient flow and adapting the state-space mechanism to the iterative denoising task. Finally, a dual-branch architecture synergistically combines a Bidirectional Mamba2 core for global sequence modeling with a 3D Temporal Swin Attention branch for local feature refinement. This integration guarantees that the model dynamically balances structural coherence with detailed surface fidelity. As shown in Fig. 1, our model is able to produce high-fidelity and descriptive 3D shapes. In summary, our main contributions are as follows:

- We present Octree-MambaDiffuser, a hierarchical architecture that successfully adapts Mamba to 3D shape generation. By combining octree-based spatial partitioning with a state-space diffusion process, it overcomes the fundamental mismatch between 1D sequence modeling and 3D data.
- We design a scale-flexible MambaBlock, which integrates two key components to overcome the specific limitations of Mamba in 3D diffusion: a Feature Conditioning Module (FCM) to stabilize the discrete-continuous process transition, and a dual-branch architecture to dynamically balance global structure and local detail. This design ensures robust training and enables high-fidelity 3D generation.
- We validate our framework through comprehensive evaluations. Octree-MambaDiffuser outperforms existing methods across multiple metrics, demonstrating a marked improvement in the generation of high-fidelity, structurally coherent, and diverse 3D shapes.

II. Related Work

Early 3D generative models were built upon GANs and VAEs. For instance, IM-GAN [16] learns implicit fields via adversarial training, while SDF-StyleGAN [17] combines StyleGAN with signed distance functions for structured shape synthesis. Although expressive, such approaches frequently face instability and mode collapse. To improve scalability, sparse representations such as octrees [18] and hash grids [19] model only occupied regions, circumventing the high memory cost of dense voxel grids.

More recently, diffusion models have become a dominant paradigm for 3D generation, since they offer greater training stability and sample diversity. Initial efforts like RenderDiffusion [20] train 3D-aware image diffusion models with multi-view consistency enforced via differentiable rendering. Later methods operate directly in 3D: Point-Voxel Diffusion (PVD) [21] performs diffusion on point clouds; DiT-3D [22] extends Vision Transformers to voxelized point clouds; and SPAGHETTI [23] generates part-aware implicit surfaces via skeletal control. For implicit fields, SDF-Diffusion [24] applies diffusion-based super-resolution to coarse SDFs, while both Diffusion-SDF [9] and SDFusion [5] adopt latent diffusion strategies, using continuous and discrete latent spaces respectively, to model high-resolution SDFs. Beyond surfaces, DiffRF [25] applies direct diffusion to explicit radiance fields, and NFD [26] leverages triplane parameterizations with 2D diffusion. Closest to our work, OctFusion [6] introduces an octree-structured latent diffusion framework that enables adaptive-resolution, hierarchical generation.

However, nearly all existing methods rely on self-attention or dense convolutional backbones, whose quadratic complexity limits their ability to capture long-range dependencies in large-scale or deeply hierarchical 3D data. In contrast, we propose a linear-complexity state-space diffusion model that leverages structured octree tokenization and dual-path modeling to jointly capture global coherence and local geometric detail.

III. Method

A. Overall Framework

As shown in Fig. 2, Octree-MambaDiffuser consists of two main stages: representation learning and diffusion-based generation. A variational autoencoder (VAE) learns to encode a 3D shape into a hierarchical octree of latent features, which can be decoded into a continuous signed distance field (SDF). Then, we flatten the octree into a 1D sequence via z-order serialization and feed it into a hierarchical state-space diffusion

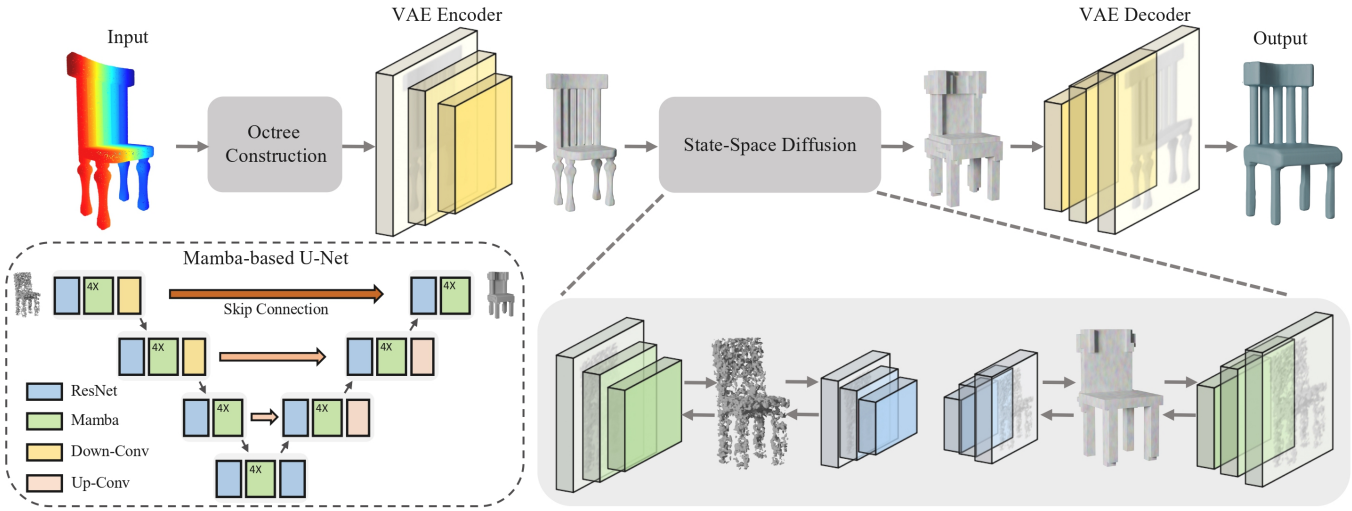

Fig. 2. The overall architecture of Octree-MambaDiffuser comprises two main stages. First, a variational autoencoder (VAE) is trained to obtain octree-structured latent embeddings from 3D shapes; these embeddings are decoded into continuous signed distance fields (SDFs) for geometric reconstruction. Second, a hierarchical state-space diffusion model iteratively generates both the hierarchical octree topology and its associated latent features. The bottom-left subplot details the architecture of our core Mamba-based U-Net.

model, which iteratively denoises the sequence to jointly generate the octree structure and its latent features.

B. Octree-based Latent Representation

In the first stage, we encode 3D shapes using a hierarchical octree VAE built upon dual octree graph networks [18]. This representation learning stage serves as a geometric tokenizer, transforming raw, irregular inputs, such as point clouds, into a discrete, multi-resolution latent description. The encoder processes an input octree, constructed from the input shape, and produces a set of latent features at its leaf nodes. The decoder then reconstructs a continuous signed distance field (SDF) from these features via MLPs, with global feature fusion handled by the MPU module [27]. This design avoids reliance on an explicit SDF formulation while preserving high-frequency geometric details. Importantly, the octree organizes shape information across scales, and the dual graph network explicitly encodes neighborhood and hierarchical relationships. This embeds essential structural priors into the learned latent space.

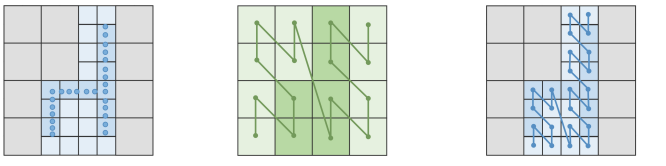

Fig. 3. Illustration of the octree structure and z-order traversal, using a 2D chair shape for clarity. (a) Input point cloud and its octree decomposition. Node states are color-coded: darker shades represent occupied nodes (containing points), lighter shades denote empty nodes, and gray indicates nodes not present at the depicted depth. (b, c) z-order curves traversing the octree nodes at depths 2 and 3, respectively.

To bridge this spatial-hierarchical representation with sequential generation, we serialize the octree into a $1D$ token sequence using z-order curve traversal as illustrated in Fig. 3. This space-filling curve preserves spatial locality and implicitly encodes hierarchical relationships, which guarantees that proximate 3D regions remain adjacent in the sequence. The resulting linearized order performs a vital modal translation. It transforms the structured octree into a format compatible with sequential likelihood models. Each node becomes a discrete token, establishing the "language" in which shapes will be generated, where 3D coherence is encoded as sequential context.

C. State-Space Diffusion

The prevailing Transformer-based 3D diffusion models face a fundamental dilemma: while global attention ensures structural coherence, its quadratic complexity is computationally prohibitive at 3D scales. Conversely, local attention reduces cost but often fragments holistic shape understanding. To resolve this trade-off, we replace the Transformer with a Mamba-based selective state space model (SSM). Mamba maintains linear-time complexity in sequence length while preserving global modeling capability, thereby achieving both computational efficiency and structural consistency.

Formally, the Mamba SSM [28] maps an input signal $\mathbf{x}_t$ to a latent state $\mathbf{h}_t$ and output $\mathbf{y}_t$ via a linear ordinary differential equation:

$$\begin{aligned} \mathbf{h}_t' &= \mathbf{A}\mathbf{h}_t + \mathbf{B}\mathbf{x}_t, \\ \mathbf{y}_t &= \mathbf{C}\mathbf{h}_t + \mathbf{D}\mathbf{x}_t, \end{aligned} \tag{1}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are learnable system matrices. This formulation is particularly advantageous for 3D data, as it captures long-range dependencies without incurring the quadratic cost of attention.

However, naively integrating Mamba into a diffusion framework introduces two key limitations: its inherent causal modeling bias and restricted local receptive field. To adapt Mamba for 3D geometric generation, we therefore propose a tailored variant which emphasizes structural coherence over sequential causality. We integrate it as a plug-and-play replacement for attention in the diffusion U-Net.

Our denoising diffusion model operates directly on the serialized latent sequences generated by the octree-based z-order traversal. Following the established denoising diffusion framework [7], we define a forward noising process as a Markov chain that progressively corrupts a clean octree latent sequence over a series of timesteps:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ with $\alpha_t \in (0, 1)$ forming a monotonically decreasing noise schedule. The reverse process is learned by a neural network $\mathcal{F}_\theta(\mathbf{x}_t, t)$ that predicts the denoised sequence $\mathbf{x}_0$ from $\mathbf{x}_t$.

To bridge the gap between Mamba's discrete-state dynamics and the continuous diffusion process, we introduce a Feature Conditioning Module (FCM) as the initial stage of our Mamba block. The FCM stabilizes gradient propagation during iterative denoising through a simple yet effective pre-normalization and activation step. Given the input token sequence $\mathbf{x}_t$, the module first applies Layer Normalization $\mathcal{LN}$, followed by a SiLU activation, producing the conditioned features:

$$\tilde{\mathbf{x}}_{t-1} = \mathcal{LN}(\mathbf{x}_{t-1}) \cdot \sigma(\mathcal{LN}(\mathbf{x}_{t-1})), \tag{3}$$

where σ denotes the sigmoid function within the SiLU activation. This preprocessing ensures that features entering the state-space mechanism are both normalized and non-linearly transformed, mitigating numerical instability caused by fluctuating gradients. By introducing this conditioning early, the FCM enhances the model's adaptability to evolving latent representations across timesteps. This also enables robust cooperation between the discrete Mamba framework and the continuous denoising objective.

Relying on the stable signal produced by our proposed Feature Conditioning Module (FCM), our Mamba block adopts a dual-branch architecture, inspired by LinGen [29], to model both global structure and local detail. The first branch, termed the MA-branch, ensures global structural coherence through a sequence of transformations: a Rotary Major Scan (RMS) that reorganizes the 1D token sequence by cyclically alternating row-, column-, and depth-major scan orders across layers to preserve spatial adjacency; a Bidirectional Mamba2 (BM2) module that captures multi-directional, long-range dependencies along all three spatial axes within the scanned sequence; and a final Reverse RMS (Rev-RMS) step that restores the original token ordering, stabilizing gradient flow and ensuring compatibility with downstream layers. This branch thereby maintains global structural relationships across the entire octree hierarchy.

The second 3D Temporal Swin Attention (3D-TESA) branch operates directly on the feature map $\tilde{\mathbf{x}}_{t-1}$. Different from the Temporal Swin Attention (TESA) originally proposed for video [29], we adapt this module to 3D geometry by reinterpreting temporal shifts as spatial window shifts, yielding a purely spatial attention module tailored for hierarchical geometry. This variant of TESA efficiently captures local geometric features, such as edges and contours, through local self-attention within shifted windows, enabling cross-window communication and avoiding isolated local representations.

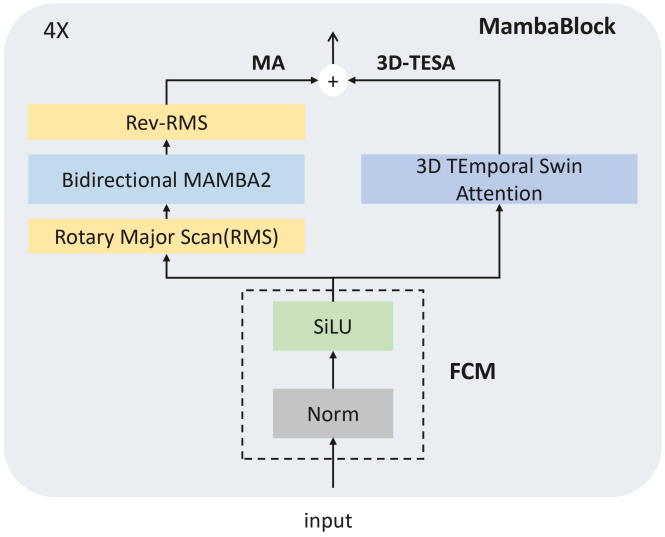

Fig. 4. Architecture of the proposed Mamba block. To effectively adapt Mamba for the diffusion framework, we introduce a two-branch module built on state-space scanning to capture long-range dependencies, integrated with a FCM module for non-linear feature enhancement.

The outputs of both branches are summed to fuse global structural awareness with local detail refinement:

$$\mathbf{x}^{(k)} = \mathcal{M}(\mathbf{x}^{(k-1)}) + \mathcal{T}\left(\mathbf{x}^{(k-1)}\right), \quad (k = 1, 2, 3, 4).$$

where $\mathcal{M}$ and $\mathcal{T}$ denote the MA and 3D-TESA branches respectively. Importantly, $\mathbf{x}^{(0)} = \tilde{\mathbf{x}}$ and this Mamba block is applied recursively four times in a cyclic scanning manner. This multi-scan design progressively aggregates features from all three dimensions, enabling the model to build a comprehensive geometric understanding. By integrating z-order serialization with our adapted MambaBlock, our approach resolves the inherent conflict between 1D sequential SSMs and 3D spatial hierarchies, achieving both high-fidelity detail preservation and long-range structural consistency in generated shapes. Fig. 4 exhibits the internal architecture of our Mamba block.

D. Loss Functions

We employ separate loss functions for training the two main components of our framework: the VAE-based geometry representation and the diffusion-based generative model.

1) VAE Training Objective: The VAE is optimized using a composite loss that balances geometric reconstruction accuracy ($\mathcal{L}_s$) with octree constraint ($\mathcal{L}_o$) and latent distribution regularization ($\mathcal{L}_k$):

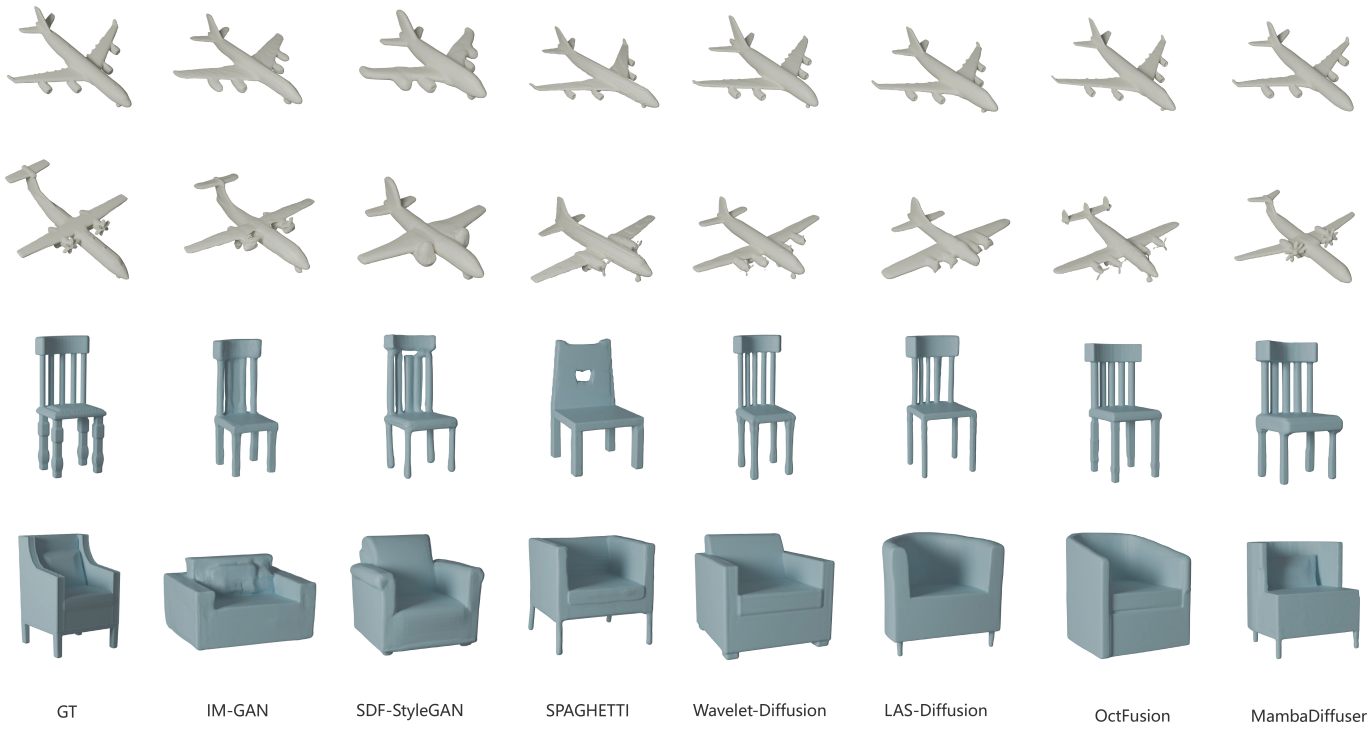

Fig. 5. Visual comparison of different generative models. Our method consistently generates shapes with both accurate global structure and finely resolved details, outperforming other approaches that tend to compromise one for the other.

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{s}} + \mathcal{L}_{\text{o}} + \lambda_{\text{k}}\mathcal{L}_{\text{k}} \tag{4}$$

where λ_k is the hyperparameter applied to control the weight of $\mathcal{L}_k$. In practice, we set λ_k as 0.1.

SDF Reconstruction Loss ($\mathcal{L}_\mathbf{s}$) enforces accurate implicit surface reconstruction by matching both predicted SDF values and their spatial gradients to ground truth. For a set of points $\mathcal{Q}$ uniformly sampled in 3D space:

$$\mathcal{L}_{\text{s}} = \frac{1}{|\mathcal{Q}|}\sum_{\mathbf{x}\in\mathcal{Q}}\left(\lambda_s||F(\mathbf{x}) - G(\mathbf{x})||_2^2 + ||\nabla F(\mathbf{x}) - \nabla G(\mathbf{x})||_2^2\right), \tag{5}$$

where $F(\mathbf{x})$ is the predicted SDF, $G(\mathbf{x})$ is the ground-truth SDF, ∇ denotes the spatial gradient, and $\lambda_s = 200$ balances the relative importance of SDF values versus gradient accuracy.

Octree Structure Loss ($\mathcal{L}_\mathbf{o}$) ensures that the decoded geometry maintains structural consistency with the hierarchical octree representation. We apply binary cross-entropy loss at each depth level of the octree:

$$\mathcal{L}_o = \sum_{d\in D}\frac{1}{N_d}\sum_{o_i\in O_d} E(o_i, \tilde{o}_i) \tag{6}$$

where D is the maximum octree depth, N_d is the number of nodes at depth d, o_i is the predicted occupancy probability for the i-th node at depth d, and $\tilde{o}_i$ is its ground-truth binary occupancy.

KL Divergence Loss ($\mathcal{L}_\mathbf{k}$) regularizes the latent space by encouraging the posterior distribution $q(\mathbf{z} \mid \mathbf{x})$ to approximate a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\begin{aligned}\mathcal{L}_{\text{k}} &= \mathcal{D}_{\text{k}}\left(q(\mathbf{z} \mid \mathbf{x}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})\right) \\ &= -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right)\end{aligned} \tag{7}$$

where J is the latent dimension, and μ_j, σ_j^2 are the mean and variance of the j-th latent dimension produced by the encoder.

2) Diffusion Training Objective: The diffusion model is trained using a standard denoising objective that reconstructs clean latent codes from their noisy counterparts. Given a clean octree latent representation $\mathbf{x}_0$, we generate a noisy version $\mathbf{x}_t$ at diffusion timestep t by adding Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ according to the forward diffusion process. The model $\mathcal{F}_\theta$ learns to predict the original latent $\mathbf{x}_0$ from $\mathbf{x}_t$ and t:

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{\mathbf{x}_0,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t\sim\mathcal{U}(1,T)}||\mathcal{F}_\theta(\mathbf{x}_t, t) - \mathbf{x}_0||_2^2, \tag{8}$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ follows the standard diffusion formulation with noise schedules α_t, $\bar{\alpha}t = \prod s = 1^t\alpha_s$, and T is the total number of diffusion steps.

IV. Experiments

A. Datasets

We conduct our experiments using ShapeNetV1 [30], a large-scale repository of 3D CAD models spanning numerous object categories. From this collection, we select five categories for our study: chair, table, airplane, car, and rifle. This subset provides substantial geometric and topological diversity, including objects with thin structures (tables), hollow volumes (chairs), smooth surfaces (cars), and fine mechanical details (airplanes, rifles), making it well-suited for evaluating the robustness of 3D reconstruction and generative models.

B. Implementation Details

To ensure fair and consistent comparison with state-of-the-art methods, we adopt the same data preparation protocol as OctFusion [6]. We use their established training and testing partitions for all experiments, maintaining identical dataset splits across all evaluations. The VAE encoder processes the input octree down to depth 6, producing a 3D latent vector for each leaf node. After pre-training the VAE, we freeze its weights and train only the diffusion model. The diffusion model employs a two-stage U-Net architecture. For unconditional generation, we train separate models for each category using the AdamW optimizer [31] with a learning rate of 4×10^{-4}. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

C. Evaluation Metrics

We evaluate the visual, geometric, and statistical quality of the generated 3D shapes using a comprehensive suite of established metrics. Visual fidelity is measured via the Fréchet Inception Distance (FID) [32], computed by rendering each mesh from 20 uniformly sampled viewpoints and comparing the feature distribution of these renderings with that of ground-truth dataset renderings. Geometric and distributional quality are assessed on point-cloud representations: we sample 2,048 points from each mesh (normalized to a unit cube) and report three standard metrics, Coverage (COV), MMD, and 1-NNA [33], each computed with both Chamfer Distance (CD) and Earth Mover's Distance (EMD) as the underlying distance function. Here, COV quantifies the diversity of the generated set, MMD measures the average distance between the generated and real distributions, and 1-NNA (where 50% indicates perfect distribution matching) evaluates how well the generated distribution aligns with the true data distribution. Together, these metrics provide a multi-faceted evaluation across visual, geometric, and statistical dimensions.

D. Comparisons with SoTA Methods

Our method is comprehensively benchmarked against recent state-of-the-art generative frameworks, including GAN-based models (IM-GAN [16], SDF-StyleGAN [17]) and diffusion-based approaches (Wavelet-Diffusion [34], MeshDiffusion [35], SPAGHETTI [23], LAS-Diffusion [32]).

TABLE I
QUANTITATIVE COMPARISON WITH SoTA MODELS

Category	Method	COV(%)↑		MMD↓		1-NNA(%)↓	
		CD	EMD	CD (10^{-3})	EMD (10^{-2})	CD	EMD
chair	SDF-StyleGAN	49.41	40.56	15.25	18.76	72.16	77.62
	OctFusion	47.71	46.68	14.37	17.86	68.15	**69.07**
	MambaDiffuser	**51.26**	**47.79**	**13.93**	**17.84**	**67.69**	69.98
airplane	SDF-StyleGAN	43.14	32.76	4.80	12.63	92.71	92.83
	OctFusion	41.66	38.44	**3.99**	**10.52**	86.65	**86.40**
	MambaDiffuser	**47.96**	**40.05**	4.01	11.13	**86.22**	86.90
rifle	SDF-StyleGAN	37.26	36.21	4.19	11.48	93.05	88.53
	OctFusion	34.11	31.79	3.71	**11.03**	**83.89**	**84.84**
	MambaDiffuser	**42.74**	**36.63**	**3.65**	11.05	87.79	85.05
car	SDF-StyleGAN	35.90	30.50	5.17	11.78	95.70	94.80
	OctFusion	30.30	33.30	**4.65**	**10.93**	**86.55**	**85.70**
	MambaDiffuser	**36.80**	**38.60**	4.94	11.45	91.90	91.20
table	SDF-StyleGAN	38.95	36.60	14.83	17.16	81.87	80.32
	OctFusion	34.31	38.07	12.62	16.19	77.32	77.23
	MambaDiffuser	**44.18**	**43.95**	**12.21**	**15.84**	**69.15**	**69.48**

As shown in Table I, MambaDiffuser achieves the best Coverage (COV) across all categories and underlying distance metrics (CD/EMD), demonstrating its superior ability to capture the diversity of the target data distribution. It also delivers competitive results in MMD and 1-NNA, confirming a strong balance between sample quality and distribution fidelity. Table II shows that our method attains the lowest FID scores in every category, underscoring its advantage in generating visually realistic shapes that closely match the ground-truth distribution.

TABLE II
QUANTITATIVE COMPARISON OF FID BETWEEN DIFFERENT MODELS

Method	chair	airplane	car	table	rifle
IM-GAN	63.42	74.57	141.20	51.70	103.30
SPAGHETTI	65.26	59.21	N/A	N/A	N/A
MeshDiffusion	49.01	97.81	156.21	49.71	87.96
SDF-StyleGAN	56.23	121.80	189.57	73.64	117.91
Wavelet-Diffusion	36.89	55.77	N/A	51.82	N/A
MambaDiffuser	**34.50**	**52.82**	**109.55**	**37.19**	**72.24**

Fig. 6. Hierarchical 3D shape generation results. Our model successfully synthesizes expressive local details from coarse, octree-based structural representations.

Qualitative results in Fig. 5 reveal that GAN-based methods (IM-GAN [16], SDF-StyleGAN [17]) often produce distorted geometries and artifacts, while other diffusion models (Wavelet-Diffusion [34], etc.) tend to lose fine details due to

resolution limitations. In contrast, MambaDiffuser generates more coherent, detailed, and artifact-free shapes. In Fig. 6, we additionally provide the visual results of our hierarchical generation.

E. Ablation Studies

We conduct systematic ablation studies on our Octree-MambaDiffuser model to examine the contribution of its key components and parameters. All experiments maintain identical settings and training configurations to ensure fair comparisons.

Feature Conditioning Module (FCM). As shown in the loss curves (Fig. 7), the inclusion of the FCM leads to significantly faster convergence and lower final training loss. The model with FCM (orange curve) exhibits a steeper initial descent and stabilizes at a consistently lower loss. This indicates more stable gradient flow and better conditioning for the diffusion process. In contrast, the variant without FCM (blue curve) converges more slowly and plateaus at a higher loss, demonstrating the importance of pre-normalization and SiLU activation for stable feature dynamics during iterative denoising.

Dual-Branch. The first section of Table III compares different architectural configurations. The full dual-branch design, which integrates MA branch for long-range modeling and 3D-TESA branch for local refinement—consistently outperforms single-branch variants. This synergy between global structural coherence and local detail enhancement improves generative quality and distribution matching across all metrics.

Scanning Mechanisms. As shown in Section 2 of Table III, the Rotary Major Scan (RMS) strategy surpasses conventional zigzag scanning across all evaluation metrics. By following a cyclic, multi-dimensional path, RMS better preserves spatial coherence and enables more effective bidirectional feature aggregation, confirming its advantage as an optimal scanning strategy. Section 3 of Table III examines the impact of scanning iterations. A single iteration yields the weakest performance, underscoring the necessity of multi-scan processing. While increasing iterations from 4× to 8× offers marginal gains, computational cost rises substantially. We therefore adopt 4× iterations as the default, striking a practical balance between generation quality and efficiency.

V. Conclusion

This paper introduces Octree-MambaDiffuser, a hierarchical framework for high-fidelity 3D shape generation that integrates octree spatial guidance with state-space diffusion. To reconcile the conflict between sequential state-space modeling and 3D structures, we propose a z-order serialization that linearizes the octree into a 1D sequence while preserving spatial locality. Our core MambaBlock dynamically fuses local and global features via bidirectional state-space scanning and structured attention. Experiments across five ShapeNet categories demonstrate that our method surpasses existing approaches in geometric quality, diversity, and visual fidelity.

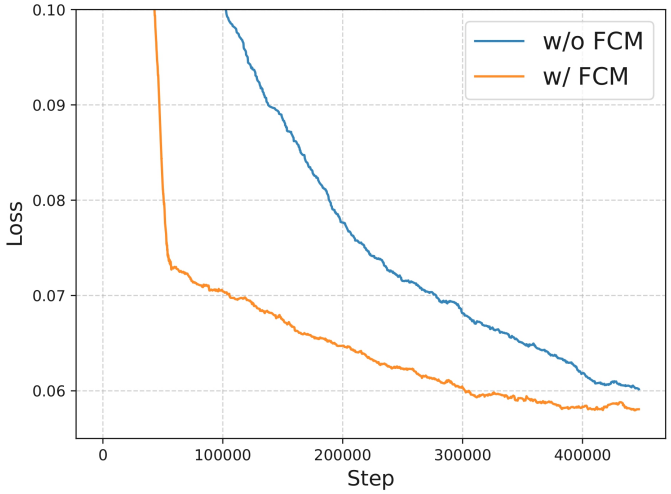

Fig. 7. Training loss curves of models with and without the Feature Conditioning Module (FCM).

TABLE III
COMPLETE ABLATION STUDIES: ARCHITECTURAL COMPONENTS, SCANNING MECHANISMS, AND ITERATION COUNTS

Ablation Setting	COV (%) ↑		MMD ↓		1-NNA (%) ↓		FID ↓
	CD	EMD	CD (10^{-3})	EMD (10^{-2})	CD	EMD	
1. Architectural Components							
Baseline	**51.26**	**47.79**	**13.93**	**17.84**	**67.69**	**69.98**	**34.50**
w/o 3D-TESA	50.89	**47.79**	13.99	17.94	68.76	70.42	38.38
w/o MA branch	48.01	46.24	14.15	17.96	70.02	72.53	41.77
2. Scanning Mechanism							
RMS	**51.26**	**47.79**	**13.93**	**17.84**	**67.69**	**69.98**	**34.50**
Zigzag	45.50	44.47	14.54	18.08	71.28	71.79	35.42
3. Scanning Iterations							
4× iterations	51.26	47.79	**13.93**	17.84	**67.69**	**69.98**	34.50
8× iterations	**52.60**	**48.35**	14.02	**17.83**	**67.69**	70.46	**32.69**
1× iteration	46.10	43.66	15.05	18.24	70.90	72.23	38.26

Acknowledgment

This work is supported by the Tianjin Science and Technology Plan Project (Grant No. 24ZYJDSY00290) and the industrial collaboration projects (Grant Nos. RDS10120240228 and RDS10120240175).

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] Yushi Li and George Baciu, "Hsgan: Hierarchical graph learning for point cloud generation," *IEEE Transactions on Image Processing*, vol. 30, pp. 4540–4554, 2021.

[3] Yushi Li and George Baciu, "Sg-gan: Adversarial self-attention gcn for point cloud topological parts generation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 10, pp. 3499–3512, 2021.

[4] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[5] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui, "Sdfusion: Multimodal 3d shape completion, reconstruction, and generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 4456–4465.

[6] Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang, "Octfusion: Octree-based diffusion models for 3d shape generation," in *Computer Graphics Forum*. Wiley Online Library, 2025, vol. 44, p. e70198.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[8] 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*, 2022, pp. 10684–10695.

[9] Gene Chou, Yuval Bahat, and Felix Heide, "Diffusion-sdf: Conditional generative modeling of signed distance functions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2262–2272.

[10] Gimin Nam, Mariem Khlifi, Andrew Rodriguez, Alberto Tono, Linqi Zhou, and Paul Guerrero, "3d-ldm: Neural implicit 3d shape generation with latent diffusion models," *arXiv preprint arXiv:2212.00842*, 2022.

[11] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[12] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 300–309.

[13] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[14] Tri Dao and Albert Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," *arXiv preprint arXiv:2405.21060*, 2024.

[15] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al., "Jamba: A hybrid transformer-mamba language model," *arXiv preprint arXiv:2403.19887*, 2024.

[16] Zhiqin Chen and Hao Zhang, "Learning implicit fields for generative shape modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5939–5948.

[17] Xinyang Zheng, Yang Liu, Pengshuai Wang, and Xin Tong, "Sdf-stylegan: implicit sdf-based stylegan for 3d shape generation," in *Computer Graphics Forum*. Wiley Online Library, 2022, vol. 41, pp. 52–63.

[18] Peng-Shuai Wang, Yang Liu, and Xin Tong, "Dual octree graph networks for learning adaptive volumetric shape representations," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[19] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[20] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero, "Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12608–12618.

[21] Linqi Zhou, Yilun Du, and Jiajun Wu, "3d shape generation and completion through point-voxel diffusion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5826–5835.

[22] Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li, "Dit-3d: Exploring plain diffusion transformers for 3d shape generation," *Advances in neural information processing systems*, vol. 36, pp. 67960–67971, 2023.

[23] Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or, "Spaghetti: Editing implicit shapes through part aware generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–20, 2022.

[24] Jaehyeok Shim, Changwoo Kang, and Kyungdon Joo, "Diffusion-based signed distance fields for 3d shape generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20887–20897.

[25] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner, "Diffrf: Rendering-guided 3d radiance field diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4328–4338.

[26] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein, "3d neural field generation using triplane diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20875–20886.

[27] Yutaka Ohtake, Alexander Belyaev, Marc Alexa, Greg Turk, and Hans-Peter Seidel, "Multi-level partition of unity implicits," in *Acm Siggraph 2005 Courses*, pp. 173–es. 2005.

[28] Ankit Gupta, Albert Gu, and Jonathan Berant, "Diagonal state spaces are as effective as structured state spaces," *Advances in neural information processing systems*, vol. 35, pp. 22982–22994, 2022.

[29] Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, et al., "Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2578–2588.

[30] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al., "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[31] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*.

[32] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum, "Locally attentional sdf diffusion for controllable 3d shape generation," *ACM Transactions on Graphics (ToG)*, vol. 42, no. 4, pp. 1–13, 2023.

[33] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan, "Pointflow: 3d point cloud generation with continuous normalizing flows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4541–4550.

[34] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu, "Neural wavelet-domain diffusion for 3d shape generation," in *SIGGRAPH Asia 2022 conference papers*, 2022, pp. 1–9.

[35] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu, "Meshdiffusion: Score-based generative 3d mesh modeling," *arXiv preprint arXiv:2303.08133*, 2023.