

Learning to Tessellate: Point Cloud Generation via Recursive Spectral Partitioning

Monan Sun^{1,2,4} , Bangzhen Liu³  , Huaidong Zhang¹  , and Shengfeng He⁴ 

¹ South China University of Technology, China

² University of Chinese Academy of Sciences, China

³ City University of Hong Kong, Hong Kong SAR, China

⁴ Singapore Management University, Singapore

mnsun94@gmail.com, bangzliu@cityu.edu.hk

huaidongz@scut.edu.cn, shengfenghe@smu.edu.sg

<https://huggingface.co/Mo-nan/PointRSP>

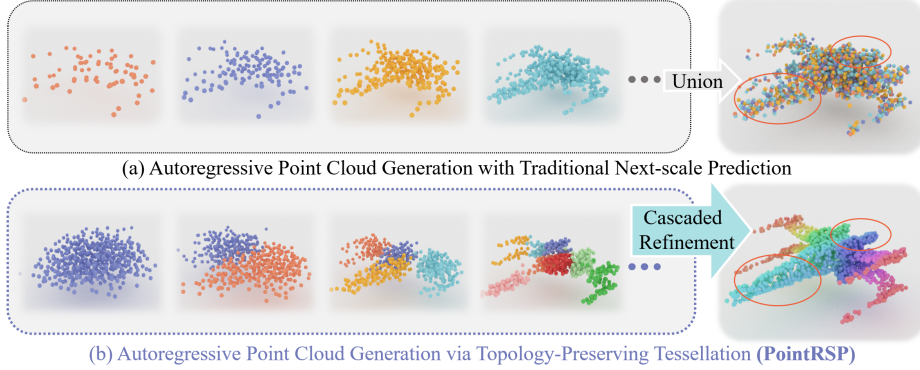


Fig. 1: Comparison between PointRSP and conventional fragmented autoregressive models. Traditional methods (a) generate shapes by merging independently generated subsets, whereas ours (b) follows a topology-preserving tessellation process.

Abstract. Autoregressive models have emerged as an effective paradigm for point cloud generation. However, most existing approaches rely on heuristic tokenization strategies, such as spatial sorting or stochastic downsampling, which often disrupt intrinsic point cloud topology and weaken the structural coherence of the generated shapes. In this paper, we present PointRSP, an autoregressive framework that reformulates point cloud generation as a topology-preserving tessellation process via recursive spectral partitioning. Instead of constructing token sequences heuristically, we introduce a topology-aware partitioning autoencoder that decomposes an unstructured point cloud into a non-balanced binary tree through a hybrid recursive spectral partitioning

 Corresponding authors.

strategy. This hierarchical representation provides a deterministic geometric blueprint that preserves topological relationships while capturing multiscale structural dependencies within a quantized latent space. To synthesize shapes in this space, we propose a dual-stream cascaded generator that jointly models structural evolution and feature synthesis. In addition, we design a geometry-calibrated positional encoding mechanism that anchors latent embeddings using multi-scale structural centers, which stabilizes cascaded generation during the early stages of structural formation. Extensive experiments show that PointRSP achieves state-of-the-art performance in generation quality and diversity, demonstrating strong generalization across complex 3D topologies.

Keywords: Point Cloud Generation · Autoregressive Generative

1 Introduction

Synthesizing high-quality point clouds is pivotal for modern 3D intelligent systems, facilitating the modeling of complex 3D distributions and driving advancements in embodied perception [30, 33] and reconstruction [7, 35]. While diffusion-based models have emerged as the prevailing paradigm [15, 20, 24, 26, 29, 41, 42], their reliance on holistic denoising within a global geometric context often precludes the preservation of fine-grained local features, resulting in artifacts that violate the underlying topological structures.

Autoregressive (AR) models [6, 12, 22, 23, 34] offer an alternative by explicitly modeling long-range geometric dependencies through sequential prediction. Despite their efficiency, effectively tokenizing point clouds while respecting their intrinsic geometric topology remains a great challenge. Unlike regular grids, point clouds are unstructured and lack a canonical ordering aligned with the underlying shape manifold. Existing schemes typically rely on heuristic discretization, such as spatial coordinate sorting [36] or space-filling curves [4], which often fail to capture manifold continuity, resulting in incoherent local details and fragmented structural outputs. Therefore, we posit that it is more natural to view point cloud tokenization as a structural decomposition problem, where the model learns to implicitly tessellate 3D space in a manner consistent with topology.

Recently, next-scale prediction AR models have advanced this domain [17, 18, 40] by tokenizing point clouds along an axis of increasing geometric detail (Figure 1(a)). While this paradigm alleviates the restrictive unidirectional dependencies of fixed linear orderings, the scale definitions remain largely heuristic and stochastic. Consequently, these frameworks struggle to preserve stable spatial or structural correspondence across varying levels of detail. To establish robust structural priors, a topology-aware strategy is required to partition the geometric structure rather than relying on heuristic downsampling.

In this context, local geometric relations are critical for learning expressive 3D representations. This principle has been validated by graph-based point cloud understanding architectures such as DGCNN [32]. Since these models successfully learn from local topology via dynamically constructed neighborhood graphs,

a compelling question naturally arises: can we develop an AR generative model that tokenizes 3D space by explicitly preserving graph connectivity? By leveraging graph clustering, we can explicitly decompose a global point cloud into a series of locally connected groups, preserving the structural integrity throughout the generation process.

Motivated by the above insights, we introduce PointRSP, an autoregressive generation framework that reformulates point cloud generation as a topology-preserving tessellation process driven by recursive spectral partitioning. PointRSP consists of two components: a Topology-Aware Partitioning Autoencoder and a Dual-Stream Autoregressive Generator. To construct a robust geometric blueprint, the autoencoder represents point clouds as nearest neighbor topology graphs and applies a hybrid recursive spectral partitioning strategy. This strategy combines a greedy merging stage to capture local geometric details and a top-down spectral bisection stage to preserve global manifold topology. By tracking these sequential partition decisions, we linearize the unstructured point set into a hierarchical, non-balanced binary tree. This explicit tree structure acts as a deterministic prior, directly guiding the structural quantization of local point features to construct a multiscale, topology-aware latent space.

To synthesize point clouds within this latent space, our Dual-Stream Generator employs a cascaded architecture where structural priors explicitly dictate content synthesis. At each generation step, a structural predictor models the binary branching patterns of the tessellation tree, while a feature predictor generates residual coordinates conditioned on this evolving structure. To overcome the cold-start instability inherent in early cascaded generation, where reconstructed coordinate cues are highly unreliable, we propose a Geometry-Calibrated Positional Encoding mechanism. This module anchors early latent embeddings by calibrating them against multiscale geometric centers derived directly from the generated structural hierarchy. By tightly intertwining structural prediction with feature synthesis, PointRSP achieves stable, interpretable, and topologically faithful 3D generation.

Our contributions are summarized as fourfold:

1. We reformulate point cloud tokenization as a principled structural decomposition. By framing generation as a manifold-consistent tessellation, the generative sequence is governed by intrinsic geometric topology, mitigating the structural fragmentation caused by heuristic ordering strategies.
2. We propose a topology-aware partitioning autoencoder that constructs a hierarchical binary tree using hybrid recursive spectral partitioning, capturing multiscale geometric dependencies within a quantized latent space.
3. We introduce a dual-stream cascaded generator that jointly models structural evolution and feature synthesis, together with a geometry-calibrated positional encoding that stabilizes early-stage generation.
4. Extensive experiments show that PointRSP achieves state-of-the-art performance in generation quality and diversity, while faithful to 3D topologies of complex 3D shapes.

2 Related Work

Point Cloud Generation. Existing point cloud generation methods can be broadly categorized into continuous distribution models and discrete autoregressive frameworks. Early continuous approaches, such as PointFlow [38], employ normalizing flows to model latent geometric distributions. More recently, diffusion models have become the dominant paradigm [13, 15, 20, 26, 29, 42], generating shapes by iteratively reversing a noise corruption process in continuous or latent spaces. While these methods achieve high fidelity, the sparsity of point clouds combined with high-dimensional sampling trajectories results in substantial computational cost. Moreover, their reliance on global denoising contexts may limit the preservation of fine-grained geometric topology.

Autoregressive models provide a more efficient alternative by flattening unordered point sets into discrete sequences for step-wise prediction. A key challenge is defining a meaningful ordering for inherently permutation-invariant data. Prior work addresses this through heuristic linearization strategies, including voxel-based spatial traversal [36], canonical patch serialization [6], and space-filling curves [4]. Although these approaches reduce certain directional dependencies, the imposed orderings often disrupt intrinsic manifold continuity, leading to structural fragmentation. Recent next-scale prediction methods, such as PointNSP [17], extend visual autoregressive modeling [25] by progressively predicting subsets of points at increasing resolutions. This strategy improves stability compared with strict linear prediction. However, stochastic or heuristic subset sampling can weaken geometric correspondence across scales. In contrast, PointRSP reformulates the autoregressive sequence through a deterministic structural hierarchy derived from recursive spectral partitioning, enabling topology-consistent generation across the entire process.

Visual Autoregressive Generation Models. Visual autoregressive models synthesize complex outputs by iteratively generating discrete components. These approaches are typically divided into sequential and cascaded paradigms. Sequential models [5, 8] represent images as collections of tokens or patches generated in a fixed order. Cascaded models [11, 19] instead adopt a hierarchical coarse-to-fine strategy, first producing low-resolution representations and progressively refining them to higher resolutions. Recent work further explores factorizing generation targets through token granularity [2, 9] or frequency-domain representations [39] to improve efficiency.

Despite the efficiency advantages of cascaded generation, most autoregressive methods for 3D point clouds [4, 6, 36], including subset-based approaches such as PointNSP [17], largely follow the sequential paradigm. In these frameworks, token groups across stages remain discrete and independent, imposing strong constraints on sampling strategies. In contrast, PointRSP introduces a cascaded autoregressive framework guided by recursive spectral partitioning. The resulting topology-aware hierarchical sequence preserves structural dependencies across stages, improving both generation efficiency and geometric fidelity.

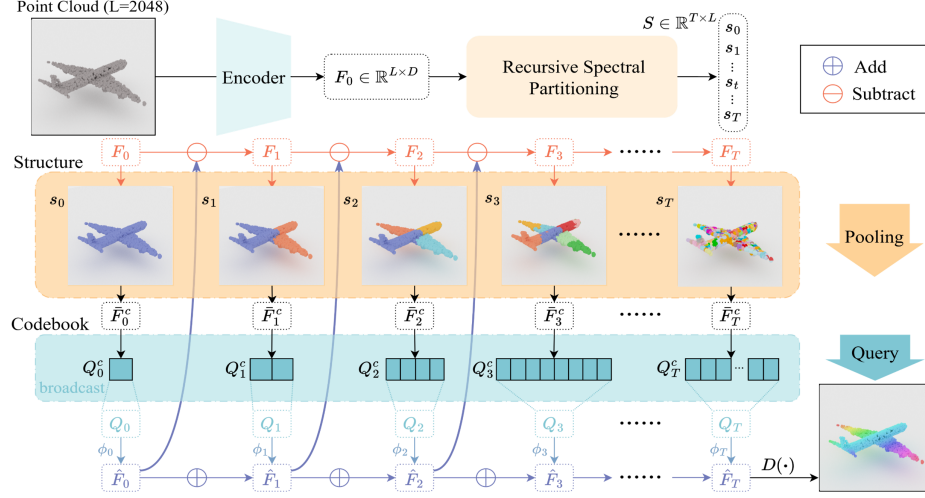


Fig. 2: Overview of the Topology-Aware Partitioning Autoencoder. The framework extracts a hierarchical feature sequence, applying a topology-preserving tessellated quantization via a recursive spectral partitioning strategy. By dynamically quantizing features within disjoint geometric subsets dictated by a deterministic structural blueprint, the model effectively preserves intrinsic topology during reconstruction.

3 Method

Standard autoregressive frameworks [23, 27] synthesize 3D shapes by decoding discrete tokens according to a predefined, heuristic ordering. However, such artificial linearizations fundamentally disrupt the intrinsic manifold continuity of unstructured point clouds. To address this, we propose PointRSP, which reformulates point cloud tokenization and generation as a topology-preserving tessellation process driven by recursive spectral partitioning. Accordingly, Section 3.1 details the Topology-Aware Partitioning Autoencoder, which decomposes the point cloud into a non-balanced binary tree to serve as a deterministic geometric blueprint for latent space quantization. Section 3.2 introduces the Dual-Stream Autoregressive Generator, which employs a geometry-calibrated cascaded architecture to jointly model structural evolution and continuous feature synthesis.

3.1 Topology-Aware Partitioning Autoencoder

Starting from an input point cloud $X = \{x_i\}_{i=1}^L$ with L points, where $x_i \in \mathbb{R}^3$ is the coordinate in the Euclidean space. We first employ a sparse convolutional encoder [37] to obtain latent point feature $F_0 \in \mathbb{R}^{L \times D}$, where D denotes the latent dimension. The tessellation of the point features follows a cascaded design, where we establish T scale levels to produce a coarse-to-fine global shape feature hierarchy $\mathbf{F} = \{F_t\}_{t=0}^T$. Figure 2 illustrates the overall autoencoding and quantization pipeline. Through our designed *topology-preserving tessellated*

quantization, at each level t , every token $f(i, t) \in F_t$ is explicitly associated with a structural label $s(i, t)$. These labels collectively form the topology annotation S via a hybrid recursive spectral partitioning strategy, which acts as the deterministic geometric blueprint for the subsequent training.

At scale t , the structural label S_t assigns each token to a disjoint geometry subset identified by a unique label $c \in \{1, \dots, C_t\}$. The C_t here denotes the possible number of subset at scale t . We then pool the token features within each subset to obtain structure-level representations:

$$F_t^c = \{f(i, t) \mid s(i, t) = c\}, \quad \bar{F}_t^c = \frac{1}{|F_t^c|} \sum_{f \in F_t^c} f, \quad (1)$$

where $|F_t^c|$ denotes the cardinality of the subset with label c . Subsequently, the pooled feature $\bar{F}_t^c$ serves as the atomic quantization target. This representation is mapped to a discrete code via a nearest-neighbor search within a learnable codebook $\mathbb{Z}$. The selected discrete code is subsequently broadcast back to all tokens sharing the label c , yielding a quantized sequence $Q_t \in \mathbb{R}^{L \times D}$. When $t > 0$, we proceed the residual features at scale $t-1$ to its next scale for stable learning of the quantized features [25]. To encourage the progressive refinement of residual geometric details at finer scales, we process the quantized embeddings Q_t through a lightweight mapping network $\phi_t(\cdot)$ to generate the reconstructed features $\hat{F}_t$. The final 3D shape is reconstructed by aggregating these quantized features across all T levels and decoding the accumulated representation via a MLP decoder D :

$$\hat{F}_t = \phi_t(Q_t), \quad \hat{X} = D \left(\sum_{t=0}^T \hat{F}_t \right). \quad (2)$$

Topology-Preserving Tessellated Quantization. Existing hierarchical quantizers, such as NVG [31], initialize a binary sequence matrix $S^{(2)} \in \{0, 1\}^{L \times T}$ to record token partition decisions. Here the partition decision of token i at scale t is denoted as $s^{(2)}(i, t)$. Structural labels are obtained by converting these binary codes into decimal indices via operation $\text{Dec}(\cdot)$:

$$\text{Dec}(s^{(2)}(i, j)) = \sum_{j=0}^t s^{(2)}(i, j) 2^{t-j}. \quad (3)$$

This formulation preserves cross-scale hierarchical relations but implicitly enforces balanced spatial bisections through KNN-style clustering, restricting the number of structural labels at stage t to exactly 2^t . While acceptable for dense 2D grids, uniform splitting is highly restrictive for unstructured point clouds, where geometry is naturally unbalanced across different branches and scales. Forced uniform bisection frequently blurs topological boundaries and erroneously merges structurally unrelated regions (see Figure 3(a)).

To accurately respect intrinsic geometric relations, we propose a recursive spectral partitioning strategy based on graph-Laplacian spectral clustering [3, 28]. Because token-level spectral partitioning across the entire point cloud is

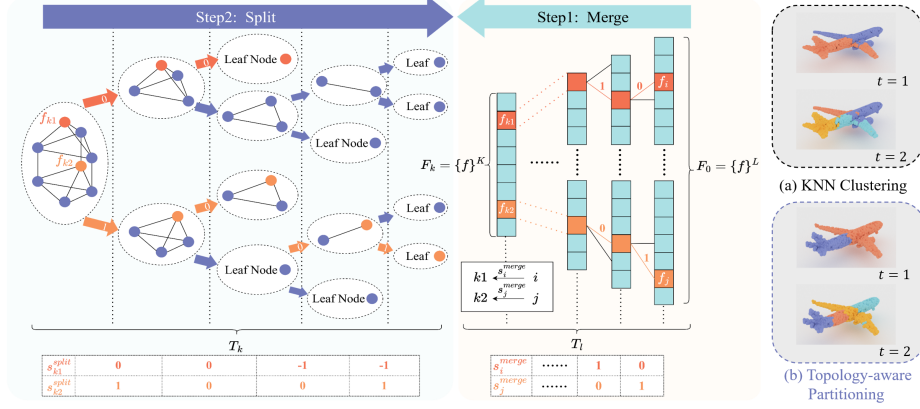


Fig. 3: Visualization of the recursive spectral partitioning strategy. The geometric blueprint is constructed through a hybrid approach: a bottom-up greedy merge stage (T_l) captures fine-grained local details, while a top-down spectral split stage (T_k) preserves global topology. Compared to (a) standard KNN-style balanced clustering, (b) our approach explicitly accommodates the non-balanced nature of 3D geometries to maintain strict boundary integrity.

computationally expensive and can yield misaligned structural sequences, we introduce a hierarchical hybrid strategy. As illustrated in Figure 3, we decompose the quantization blueprint into a bottom-up greedy merge stage with T_l steps and a top-down spectral split stage with T_k steps, satisfying $T_l + T_k = T$.

In the initial merge stage, we start from the raw feature sequence F_0 and iteratively merge the most similar token pairs based on their local feature distances. This coarse-scale greedy merging captures fine-grained local details and reduces the sequence length from L to K , bounded by the condition $\log_2 L - \log_2 K = T_l$. We record these bottom-up binary decisions at each step as $S^{\text{merge}} \in \{0, 1\}^{L \times T_l}$ and extract the pooled features F_k to serve as computationally efficient nodes for the subsequent global stage.

During the spectral split stage, we construct a nearest-neighbor graph Laplacian over the K merged tokens and perform recursive spectral bipartitioning for $t = 0, \dots, T_k - 1$. At each step, we compute the Fiedler vector of the graph Laplacian to capture the dominant geometric variation. Let $v(k, t)$ denote the Fiedler vector value for token k at step t . Structural split labels are deterministically assigned based on the sign of $v(k, t)$. Because this topology-driven partitioning yields a non-balanced binary tree, tokens reach terminal leaf states at variable depths. Let l_k denote the effective split sequence length of the k -th merged token. We define the global split horizon as $T_k = \max\{l_k\}$ for $k = 1, \dots, K$. The

split labels S^{split} are assigned as:

$$s^{\text{split}}(k, t) = \begin{cases} 0, & v(k, t) \leq 0, \quad t < l_k, \\ 1, & v(k, t) > 0, \quad t < l_k, \\ -1, & \text{otherwise.} \end{cases} \quad (4)$$

Finally, we establish the global topology-preserving structural blueprint $S^{(2)}$ by integrating the binary labels from both the merging and splitting stages. An alignment function $\text{map}(\cdot)$ traces each original point i through the merge hierarchy to its corresponding aggregated token k in the split stage. This alignment guarantees structural consistency across the entire generation process:

$$s^{(2)}(i, t) = \begin{cases} s^{\text{split}}(k, t), & t < l_k, \quad k = \text{map}(i, S^{\text{merge}}), \\ s^{\text{merge}}(i, t - l_k), & l_k \leq t < l_k + T_l, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Training Loss. The Topology-Aware Partitioning Autoencoder is optimized end-to-end using a weighted combination of three objectives: a Smooth- ℓ_1 loss L_{L1}^{smooth} [10] between the original points X and the reconstructed $\hat{X}'$, a Chamfer loss L_{cmf} to enforce global shape-level geometric fidelity, and an improved quantization loss L_{vq} with better codebook usage [16] to ensure latent feature consistency between the initial features F_0 and the quantized outputs $\hat{F}$:

$$\mathcal{L}_{\text{total}} = \lambda_{L1} L_{L1}^{\text{smooth}}(X, \hat{X}) + \lambda_{\text{cmf}} L_{\text{cmf}}(X, \hat{X}) + \lambda_{\text{vq}} L_{\text{vq}}(F_0, \hat{F}_T). \quad (6)$$

3.2 Dual-Stream Autoregressive Generator

Standard autoregressive 3D generators typically encode structural dependencies implicitly via manually crafted attention masks. In contrast, PointRSP adopts a topology-aware dual-stream cascaded paradigm (see Figure 4) comprising two explicit modules: a structural predictor that models inter- and intra-scale geometry, and a feature predictor that performs residual feature synthesis conditioned on the predicted structure. By promoting structural prior from an implicit constraint to an interpretable, recursively updated representation, the generation process is directly guided by the evolving geometric blueprint.

Let $\tilde{F}_t$ denote the generated feature representation at scale t , and let $\tilde{F}_{:t}$ represent the accumulation of generated features from resolution levels 1 to t . Prior to generation, the initial feature sequence at $t = 1$ is initialized as a zero vector, and the corresponding structural sequence is initialized by sampling from a standard Gaussian distribution. During generation, the scale index t iterates from 1 to T . Each cascaded step executes three core operations: computing *geometry-calibrated positional embeddings*, predicting next-level features conditioned on the current structure, and updating the structural representation. Both predictors are built upon self-attention mechanisms and utilize Rotary Position

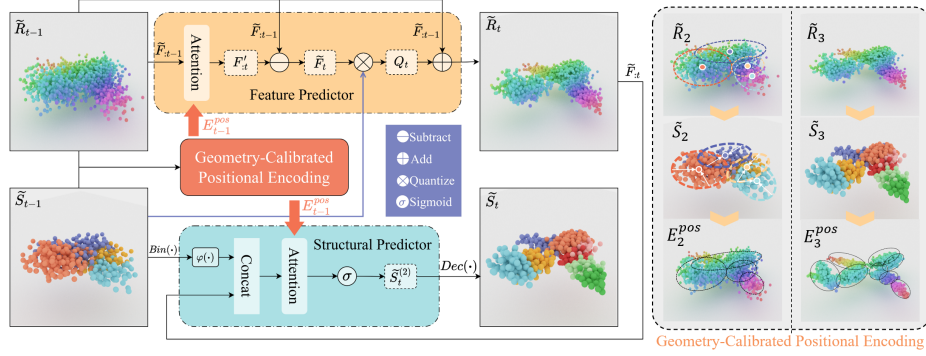


Fig. 4: Overview of the cascaded generation from scale $t-1$ to t . The dual-stream generator employs a structural predictor and a feature predictor, stabilized by a geometry-calibrated positional encoding. The right panel illustrates the dynamic construction of topology-aware positional embeddings to resolve early-stage cold-start instability.

Embedding (RoPE) [21]. First, the feature representation $\tilde{F}_t$ is predicted from the accumulated history $\tilde{F}_{:t-1}$ using the feature attention module Attn^f :

$$\tilde{F}_t = \text{Attn}^f(\tilde{F}_{:t-1}) - \tilde{F}_{:t-1}. \quad (7)$$

These features are then quantized under the guidance of the preceding structural sequence $\tilde{S}_{t-1}$. We employ a MLP-based projection layer to associate the pooled features with codebook entries, ensuring stable gradient propagation. The accumulated feature sequence for the next step is computed as:

$$\tilde{F}_{:t} = \text{Quantize}(\tilde{F}_t, \tilde{S}_{t-1}) + \tilde{F}_{:t-1}. \quad (8)$$

To predict the subsequent structural sequence, we concatenate the accumulated features $\tilde{F}_{:t}$ with the prior structural labels $\tilde{S}_{t-1}$, enabling the generated feature content to iteratively refine the structural blueprint. A lightweight projection φ ensures dimensional compatibility prior to processing by the structural attention module Attn^s :

$$\tilde{S}_t^{(2)} = \sigma \left(\text{Attn}^s \left(\left[\tilde{F}_{:t}, \varphi(\text{Bin}(\tilde{S}_{t-1})) \right] \right) \right) > 0, \quad (9)$$

where $\text{Bin}(\cdot)$ denotes the inverse process of Eq. 3, $\sigma(\cdot)$ is the sigmoid activation function to squash the output into probabilistic binary decisions that define the structural path of each node. After T recursive steps, the final accumulated feature sequence $\tilde{F}_{:T}$ is decoded into the generated 3D shape.

Geometry-Calibrated Positional Encoding. Unlike 2D images equipped with regular grid indices, unstructured point clouds lack dense positional priors. Furthermore, collapsing 3D coordinates into a single 1D index space [17, 20] severely distorts spatial proximity and undermines local geometric consistency.

To address this, we formulate the positional embedding at scale t as the concatenation of a structural term E_t^s and a feature-based term E_t^{pos} . The structural

term is directly derived from the discrete identifiers in $\tilde{S}_t$, reducing embedding collisions by providing unique topological cues. For the feature-based term, we utilize the intermediate reconstructed coordinates $\tilde{X}_{t-1} \in \mathbb{R}^{L \times 3}$ as geometric anchors, which are decoded from the accumulated features via $\tilde{X}_{t-1} = D(\tilde{F}_{:t-1})$.

However, cascaded generation suffers from a ‘‘cold-start’’ instability. In early stages, the reconstructed coordinates are highly unreliable, making direct coordinate-based positional encoding prone to severe noise. To resolve this, we propose a geometry-calibrated positional encoding that anchors early features using structure-induced geometric centers. Given the predicted structure labels $\tilde{S}_t$, we compute group centers $P_t \in \mathbb{R}^{M_t \times 3}$, where M_t is the number of distinct structural groups. For a token i with reconstructed coordinates $x(i, t) \in \tilde{X}_t$, its geometry-calibrated positional embedding $e(i, t) \in E_t^{\text{pos}}$ is defined as:

$$e(i, t) = \begin{cases} \alpha \cdot x(i, t) + (1 - \alpha) \sum_{j=0}^{t-1} w_j \cdot P(\text{map}(i, \tilde{S}_t), j), & t < T_k, \\ x(i, t), & t \geq T_k. \end{cases} \quad (10)$$

Here, $\text{map}(i, \tilde{S}_t)$ returns the structural group index of token i . The hyperparameter α controls the reliance on the raw reconstructed coordinates. To prioritize finer geometric details, the weight w_j for the historical geometric center at scale j increases exponentially ($w_{j+1} = 2w_j$), constrained by $\sum_{j=0}^{t-1} w_j = 1$. When $t < T_k$, the representation is robustly anchored by these multi-scale structural centers. Once generation reaches the split horizon ($t \geq T_k$), the encoding transitions exclusively to $x(i, t)$ since it is deemed geometrically stable.

Training Loss. During the generation phase, both the feature predictor and the structural predictor are supervised end-to-end using a Mean Squared Error (MSE) objective. We measure the deviation between the generated features $\tilde{F}_{:t}$ and structures $\tilde{S}_t^{(2)}$ against their corresponding quantized ground-truth representations extracted by the autoencoder:

$$\mathcal{L}_{\text{Gen}} = \frac{1}{T} \sum_{t=1}^T L_{\text{mse}}(\hat{F}_t, \tilde{F}_{:t}) + \frac{1}{T_k} \sum_{t=1}^{T_k} L_{\text{mse}}(S_t^{(2)}, \tilde{S}_t^{(2)}). \quad (11)$$

Crucially, we restrict structural supervision exclusively to the spectral split stage ($t = 1, \dots, T_k$). This targeted supervision prevents the model from overfitting to local variations and stabilizes the training of the structural predictor during the critical early stages of cascaded generation.

4 Experiments

4.1 Settings

Dataset. Following established benchmarking protocols, we evaluate PointRSP on the ShapeNetV2 dataset, utilizing the preprocessing pipeline introduced by PointFlow [38]. We focus on the three most structurally diverse and widely

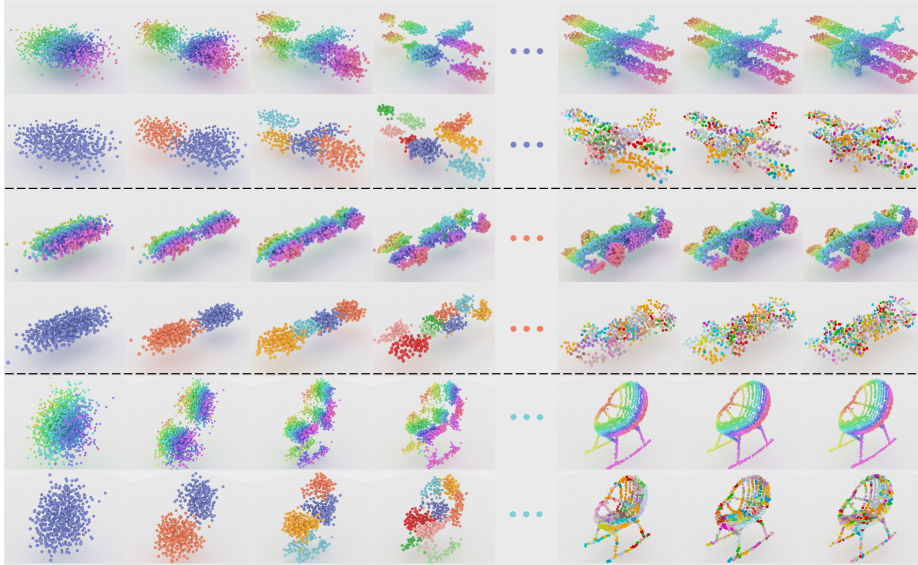


Fig. 5: Qualitative visualization of the stage-wise cascaded generation process across the airplane, car, and chair categories. As the generation progresses from coarse to fine scales, the synthesized points (top row) are progressively anchored and refined by the evolving topology-aware geometric blueprint (bottom row), culminating in highly coherent and structurally precise 3D shapes.

evaluated categories: airplanes, chairs, and cars. Each 3D shape is globally normalized within the range of $[-1.0, 1.0]$ and uniformly sampled to 2,048 points. We adhere to the train/test splits in TIGER [20], which comprise 2,832/405, 4,612/662, and 2,458/352 for airplanes, chairs, and cars, respectively. All experiments strictly follow the same preprocessing strategy and data splits to ensure fair and consistent comparisons with the compared methods.

Evaluation Metrics. Consistent with the prior state-of-the-art models (e.g., PVD [42] and LION [26]), we adopt 1-nearest neighbor accuracy (1-NNA) [14] as our primary quantitative metric. 1-NNA rigorously assesses both the generation quality and the geometric diversity of the synthesized point clouds. A score approaching 50% indicates that the generated distribution is virtually indistinguishable from the ground-truth reference distribution [38]. To construct the 1-NN distance matrix, we compute two standard point cloud similarity measures: Chamfer Distance (CD) and Earth Mover’s Distance (EMD). Furthermore, we employ Maximum Mean Discrepancy (MMD) to evaluate the global alignment between the generated and reference sets. In our ablation studies, we specifically report the standard deviation (Std.) of the MMD computed across batches. This batch-wise variance serves as a direct quantitative indicator of generation stability, allowing us to rigorously validate the efficacy of our proposed geometry-calibrated positional encoding in solving the "cold-start" issue.

Table 1: Quantitative evaluation of single-category generation on ShapeNetV2 for airplane, car, and chair, respectively. The results are reported with 1-NNA (%) under both CD and EMD metrics. The best and the second-best results are highlighted in **bold** and underline, respectively.

Model	Type	Airplane		Chair		Car		Mean	Mean
		CD	EMD	CD	EMD	CD	EMD	CD	EMD
1-GAN [1]	GAN	87.30	93.95	68.58	83.84	66.49	88.78	74.12	88.86
PointFlow [38]	Normalizing Flow	75.68	70.74	62.84	60.57	58.10	56.25	65.54	62.52
DPM [15]	Diffusion	76.42	86.91	60.05	74.77	68.89	79.97	68.45	80.55
PVD [42]	Diffusion	73.82	64.81	56.26	53.32	54.55	53.83	61.54	57.32
LION [26]	Diffusion	<u>72.99</u>	64.21	55.67	53.82	53.47	<u>53.21</u>	61.75	57.59
TIGER [20]	Diffusion	73.02	<u>64.10</u>	55.15	53.18	<u>53.21</u>	53.95	60.46	57.08
PointGrow [23]	Autoregressive	82.20	78.54	63.14	61.87	67.56	65.89	70.96	68.77
CanonicalVAE [6]	Autoregressive	80.15	76.27	62.78	61.05	63.23	61.56	68.72	66.29
PointGPT [4]	Autoregressive	74.85	65.61	57.24	55.01	55.91	54.24	63.44	62.24
PointNSP [17]	Autoregressive	73.39	64.58	55.60	54.17	53.66	54.26	60.88	57.67
PointRSP(Ours)	Autoregressive	70.68	63.94	53.88	53.96	51.47	52.68	58.67	56.86

4.2 Single-category Generation

Quantitative Evaluation. To rigorously evaluate PointRSP, we benchmark our framework against a comprehensive suite of recent state-of-the-art point cloud generation models. The baselines span both diffusion-based paradigms, including LION [26] and TIGER [20], and autoregressive approaches, such as PointGrow [23], CanonicalVAE [6], PointGPT [4], and PointNSP [17]. To ensure a fair comparison under the subset-scale paradigm, we instantiate the PointNSP baseline by adopting its official quantization-based VAE reconstruction module coupled with a lightweight global self-attention generator. As shown in Table 1, PointRSP consistently achieves state-of-the-art 1-NNA across all evaluated categories. This superiority over fragmented autoregressive models validates our motivation that heuristic spatial linearizations and stochastic downsampling inherently disrupt manifold continuity. Furthermore, our substantial improvements in the 1-NNA EMD metric, which is notoriously sensitive to local point density and fine-grained topological details, confirm the efficacy of our non-balanced binary tree representation. Unlike standard bisection strategies that force uniform spatial splits, our hybrid recursive spectral partitioning dynamically adapts to the intrinsic complexity of the 3D manifold, strictly preserving part boundaries. All experiments are conducted under a subset-scale configuration consistent with the shared transformer backbone setting used in our framework.

Qualitative Evaluation. Figure 5 visualizes the cascaded generation process of PointRSP across different scales, explicitly rendering the intermediate point clouds alongside their corresponding structural labels. The visualization confirms that generated points consistently anchor around their predicted structural centers. As the geometric blueprint undergoes recursive spectral partitioning across progressive scales, the topological structure becomes systematically refined. This coarse-to-fine fidelity directly highlights the stability and geometric awareness inherent to our dual-stream cascaded design. Figure 6 presents the visually comparison between PointRSP and representative competitors. Compared to exist-

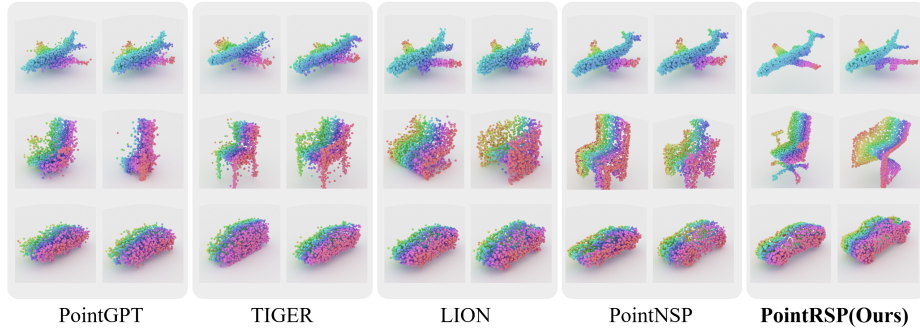


Fig. 6: Qualitative comparison between PointRSP and representative state-of-the-art baselines across single-category generation. Compared to existing autoregressive and diffusion-based methods, PointRSP explicitly preserves intrinsic manifold continuity, resulting in synthesized 3D shapes with strictly maintained structural boundaries, fewer spatial outliers, and a highly coherent global geometry.

ing methods, PointRSP synthesizes cleaner shapes characterized by fewer spatial outliers, strictly preserved structural boundaries, and a highly coherent global manifold. More visualizations are provided in the supplementary material.

4.3 Multi-class Generation and Point Upsampling

Multi-Category Generation. Beyond the single-category setting, we investigate the scalability and expressiveness of PointRSP within a comprehensive multi-category generation paradigm. Following standard benchmarking protocols, we train the models with class conditioning across all 55 ShapeNet object categories. This highly challenging setting requires the generator to simultaneously capture shared cross-category geometric regularities and preserve distinct, class-specific topological characteristics within a unified latent space.

As detailed in the left forms of Table 2, PointRSP demonstrates exceptional generalization capability across these heterogeneous distributions. Specifically, under rigorous CD- and EMD-based evaluations, our framework achieves state-of-the-art performance in four of the experiments and preserves highly competitive results in the remaining two, demonstrating the superiority of PointRSP in models highly diverse 3D topologies without suffering from the manifold collapse.

Point Cloud Upsampling. Furthermore, we evaluate PointRSP on the downstream task of point cloud upsampling to validate its capacity for fine-grained geometric refinement. By conditioning the generator on a sparse input, we apply an upsampling factor of $2\times$, increasing the spatial resolution from 1,024 to 2,048 points. As shown in the right of Table 2, PointRSP explicitly outperforms all competing baseline approaches across every single evaluated category under both distance metrics. These definitive results indicate that our topology-preserving cascaded design not only excels in unconditional and class-conditional global generation but also seamlessly transfers to local geometric refinement. By an-

Table 2: Results of multi-class generation (left) and upsampling (right) on Airplane, Chair, and Car. Both are reported with 1-NNA (%) under CD and EMD metrics. The best and the second-best results are highlighted in **bold** and underline, respectively.

Model	Airplane		Chair		Car		Model	Airplane		Chair		Car	
	CD↓	EMD↓	CD↓	EMD↓	CD↓	EMD↓		CD↓	EMD↓	CD↓	EMD↓	CD↓	EMD↓
LION	86.30	77.04	66.50	63.83	<u>64.52</u>	54.21	LION	70.41	<u>59.65</u>	53.98	54.33	57.14	47.56
TIGER	83.54	81.55	<u>57.34</u>	61.45	65.79	57.24	TIGER	71.65	59.94	<u>52.80</u>	<u>52.98</u>	57.90	<u>48.01</u>
PointGPT	94.94	91.73	71.83	79.00	89.35	87.22	PointGPT	72.11	60.12	53.75	53.21	57.26	47.85
PointNSP	<u>78.95</u>	<u>68.84</u>	58.79	55.10	69.97	52.89	PointNSP	<u>69.86</u>	59.68	53.79	53.84	<u>56.69</u>	46.93
Ours	76.52	68.79	57.11	<u>56.07</u>	58.39	<u>53.51</u>	Ours	67.29	59.09	52.05	52.63	54.95	48.63

Table 3: Ablation study on the single-category Car setting. MMD with its batch-wise Std. are reported to measure distribution alignment and stability. TPA and GCPE denote the Topology-Aware Partitioning Autoencoder and Geometry-Calibrated Positional Encoding, respectively. Cas./Struc. denote Cascaded and Structured processing.

Model	Components				Results ↓			
	Cas.	Struc.	TPA	GCPE	Acc.CD _{NNA}	Std.CD _{MMD}	Acc.EMD _{NNA}	Std.EMD _{MMD}
Base					53.66%	0.357×10^{-3}	54.26%	1.027×10^{-3}
V1	✓				53.02%	0.343×10^{-3}	53.84%	0.923×10^{-3}
V2	✓	✓			52.75%	0.318×10^{-3}	53.65%	0.713×10^{-3}
V3	✓	✓	✓		51.89%	0.304×10^{-3}	53.01%	0.636×10^{-3}
Full	✓	✓	✓	✓	51.47%	0.294×10^{-3}	52.68%	0.506×10^{-3}

choring generation to a deterministic structural blueprint, PointRSP accurately predicts highly localized structural details from coarse inputs.

Extensive visualizations for both multi-category generation and upsampling are provided in the supplementary material.

4.4 Ablation Study

To thoroughly quantify the effects of each proposed components, we conduct comprehensive ablation studies under the single-category Car generation setting, as detailed in Table 3. Following the description in Section 4.2, we adopt the quantization-based VAE prototype of PointNSP [17] together with coupled with a lightweight global self-attention generator as our baseline.

Building upon the baseline, we first replace the standard fragmented generator with the cascaded generation framework (V1). Without the structural prior, the cascaded generation takes the low-resolution prediction as condition to generate high-resolution features. As evidenced by the reduced batch-wise standard deviation in MMD, the cascaded paradigm significantly stabilizes the generation process while preserving competitive fidelity. Next, we integrate a structural geometric prior. Emulating the NVG [31] approach (V2), we explicitly construct a full, balanced binary tree using the KNN clustering strategy. The results confirm that anchoring generation to a structural prior drastically mitigates the random fluctuations inherently caused by heuristic token orderings.

However, forced uniform bisection remains suboptimal for 3D manifolds. By upgrading to our proposed Topology-Aware Partitioning Autoencoder driven by hybrid recursive spectral partitioning (V3), we explicitly accommodate the naturally non-balanced topological distribution of 3D point clouds. This topology-preserving blueprint prevents boundary blurring and leads to sharp improvements in 1-NNA. Finally, the full PointRSP framework integrates Geometry-Calibrated Positional Encoding, which explicitly resolves cold-start instability by anchoring early-stage latent features with multi-scale structural geometric centers, yielding our best generative performance.

Extensive hyperparameter analyses, including the effective spectral split horizon T_k utilized in the TPA split stage and the positional encoding reconstruction weight α , are provided in the supplementary material.

5 Conclusion

In this paper, we presented PointRSP, an autoregressive framework designed to reformulate 3D point cloud generation as a topology-preserving tessellation process. Our model employs a novel hybrid recursive spectral partitioning autoencoder to decompose point clouds into non-balanced binary trees, allowing the autoencoder to capture multiscale geometric dependencies within a quantized latent space. We further proposed a dual-stream generator to jointly model structural evolution and feature synthesis, supported by a geometry-calibrated positional encoding mechanism to stabilize early-stage generation. Extensive experiments demonstrate that PointRSP achieves state-of-the-art generation quality and diversity while exhibiting robust generalization on preserving topology across diverse 3D shapes.

Limitations and Future Directions. While PointRSP achieves promising results, it is currently limited to single-modality generation on object-centric shapes of standard densities, and the scalability of its recursive spectral partitioning mechanism to ultra-dense point sets or large-scale scene-level data has not been fully validated. Our future work will focus on incorporating multimodal conditional inputs such as text or images, and further optimizing the architecture to support the effective generation of large-scale, high-density point clouds.

6 Acknowledgements

This research is supported by National Natural Science Foundation of China (No.62302170), Guangdong Basic and Applied Basic Research Foundation (No.2024A1515010187), the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the Singapore Ministry of Education Academic Research Fund Tier 2 (Award No. MOE-T2EP20125-0016), the Singapore Ministry of Education Academic Research Fund Tier 1 (Proposal ID: 24-SIS-SMU-015), and the Lee Kong Chian Fellowships.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: *Int. Conf. Mach. Learn.* pp. 40–49 (2018)
2. Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O.F., Amirloo, E., El-Nouby, A., Zamir, A., Dehghan, A.: Flextok: Resampling images into 1d token sequences of flexible length. In: *Int. Conf. Mach. Learn.* pp. 2241–2292 (2025)
3. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**(6), 1373–1396 (2003)
4. Chen, G., Wang, M., Yang, Y., Yu, K., Yuan, L., Yue, Y.: Pointgpt: Auto-regressively generative pre-training from point clouds. In: *Adv. Neural Inform. Process. Syst.* pp. 29667–29679 (2023)
5. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: *Int. Conf. Mach. Learn.* pp. 1691–1703 (2020)
6. Cheng, A.C., Li, X., Liu, S., Sun, M., Yang, M.H.: Autoregressive 3d shape generation via canonical mapping. In: *Eur. Conf. Comput. Vis.* pp. 89–104 (2022)
7. Cheng, Y., Su, J., Jiang, M., Liu, Y.: A novel radar point cloud generation method for robot environment perception. *IEEE Trans. Robot.* **38**(6), 3754–3773 (2022)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 12873–12883 (2021)
9. Gao, Z., Shou, M.Z.: D-ar: Diffusion via autoregressive models. *arXiv preprint arXiv:2505.23660* (2025)
10. Girshick, R.: Fast r-cnn. In: *Int. Conf. Comput. Vis.* pp. 1440–1448 (2015)
11. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.* **23**(47), 1–33 (2022)
12. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 11523–11532 (2022)
13. Liu, B., Yu, Y., Xu, X., Xu, C., Zheng, C., Yang, H., Huang, S., He, S.: Genpoly: Learning generalized and tessellated shape priors via 3d polymorphic evolving. *IEEE Trans. Pattern Anal. Mach. Intell.* pp. 1–15 (2025)
14. Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests. In: *Int. Conf. Mach. Learn. Represent.* pp. 1–12 (2017)
15. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2837–2845 (2021)
16. Luo, Z., Shi, F., Ge, Y., Yang, Y., Wang, L., Shan, Y.: Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410* (2024)
17. Meng, Z., Wang, Q., Dou, Z., Song, Z., Zhou, Z., King, I., Zhao, P.: Pointnsp: Autoregressive 3d point cloud generation with next-scale level-of-detail prediction. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 31450–31461 (2026)
18. Molodyk, P., Choi, J., Romero, D.W., Liu, M.Y., Chen, Y.: Mfm-point: Multi-scale flow matching for point cloud generation. *arXiv preprint arXiv:2511.20041* (2025)
19. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: *Adv. Neural Inform. Process. Syst.* pp. 14866–14876 (2019)
20. Ren, Z., Kim, M., Liu, F., Liu, X.: Tiger: Time-varying denoising model for 3d point cloud generation with diffusion process. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 9462–9471 (2024)

21. Shi, F., Luo, Z., Ge, Y., Yang, Y., Shan, Y., Wang, L.: Scalable image tokenization with index backpropagation quantization. In: *Int. Conf. Comput. Vis.* pp. 16037–16046 (2025)
22. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525* (2024)
23. Sun, Y., Wang, Y., Liu, Z., Siegel, J., Sarma, S.: Pointgrow: Autoregressively learned point cloud generation with self-attention. In: *IEEE Winter Conf. Appl. Comput. Vis.* pp. 61–70 (2020)
24. Tao, Y., Zhang, J., Shi, Y., Zou, D., Zhou, B.: Gsv3d: Gaussian splatting-based geometric distillation with stable video diffusion for single-image 3d object generation. In: *Int. Conf. Comput. Vis.* pp. 7751–7760 (2025)
25. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: *Adv. Neural Inform. Process. Syst.* pp. 84839–84865 (2024)
26. Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al.: Lion: Latent point diffusion models for 3d shape generation. *Adv. Neural Inform. Process. Syst.* **35**, 10021–10039 (2022)
27. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: *Adv. Neural Inform. Process. Syst.* pp. 6309–6318 (2017)
28. Von Luxburg, U.: A tutorial on spectral clustering. *Statis. Comput.* **17**(4), 395–416 (2007)
29. Wang, J., Lin, C., Liu, Y., Xu, R., Dou, Z., Long, X., Guo, H., Komura, T., Wang, W., Li, X.: Pdt: Point distribution transformation with diffusion models. In: *ACM SIGGRAPH*. pp. 1–11 (2025)
30. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 19757–19767 (2024)
31. Wang, Y., Wang, Z., Wu, Z., Tao, Q., Liao, K., Loy, C.C.: Next visual granularity generation. In: *Int. Conf. Learn. Represent.* pp. 1–14 (2026)
32. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.* **38**(5), 1–12 (2019)
33. Wang, Y., Xian, Z., Chen, F., Wang, T.H., Wang, Y., Fragkiadaki, K., Erickson, Z., Held, D., Gan, C.: Robogen: towards unleashing infinite data for automated robot learning via generative simulation. In: *Int. Conf. Mach. Learn.* pp. 51936–51983 (2024)
34. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 4840–4851 (2024)
35. Wu, Y., Zheng, W., Zuo, S., Huang, Y., Zhou, J., Lu, J.: Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. In: *Int. Conf. Comput. Vis.* pp. 26360–26370 (2025)
36. Yan, X., Lin, L., Mitra, N.J., Lischinski, D., Cohen-Or, D., Huang, H.: Shapeformer: Transformer-based shape completion via sparse representation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 6239–6249 (2022)
37. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. *Sensors* **18**(10), 3337 (2018)
38. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In: *Int. Conf. Comput. Vis.* pp. 4541–4550 (2019)

- 39. Yu, H., Luo, H., Yuan, H., Rong, Y., Zhao, F.: Frequency autoregressive image generation with continuous tokens. arXiv preprint arXiv:2503.05305 (2025)
- 40. Zhang, J., Xiong, F., Wang, G., Xu, M.: G3pt: unleash the power of autoregressive modeling in 3d generation via cross-scale querying transformer. In: Int. Jt. Conf. Artif. Intell. pp. 2350–2358 (2025)
- 41. Zheng, C., Lin, Y., Liu, B., Xu, X., Nie, Y., He, S.: Recdreamer: Consistent text-to-3d generation via uniform score distillation. In: Int. Conf. Learn. Represent. pp. 1–14 (2025)
- 42. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Int. Conf. Comput. Vis. pp. 5826–5835 (2021)