3D-WAG: Hierarchical Wavelet-Guided Autoregressive Generation for High-Fidelity 3D Shapes

Tejaswini Medi[1]
tejaswini.medi@uni-mannheim.de
Arianna Rampini[2]
arianna.rampini@autodesk.com
Pradyumna Reddy[2]
chinthala.pradyumna.reddy@autodesk.com
Pradeep Kumar Jayaraman[2]
pradeep.kumar.jayaraman@autodesk.com
Margret Keuper[1,3]
keuper@uni-mannheim.de

[1] Chair for Machine Learning
University of Mannheim
Germany

[2] Autodesk AI Lab

[3] MPI for Informatics
Saarland Informatics Campus
Germany

Abstract

Autoregressive (AR) models excel in language and image generation, but their role in 3D generation faces high computational cost and resolution challenges. Existing 3D AR methods, using voxel grids or implicit representations, produce long, redundant token sequences, limiting high-fidelity 3D shape generation and incurring high inference cost. To address these issues, we introduce 3D-WAG, a novel autoregressive framework employing compact wavelet-based hierarchical representations for efficient and expressive 3D shape generation. By representing the shapes in the wavelet domain, 3D-WAG captures coarse to fine geometric details as multi-scale discrete token maps, using a 3D vector-quantized variational autoencoder (VQVAE), enabling efficient AR modeling and detailed shape understanding. Unlike conventional next-token prediction, 3D-WAG formulates 3D shape generation as a next-scale token map prediction problem, achieving a faster inference time of 1.15 seconds per sample on a single NVIDIA H100 GPU, which is 15 times faster than the state-of-the-art diffusion-based 3D generation model UDiFF. Furthermore, 3D-WAG supports unconditional, class-conditional, and text-conditional shape generation. Experimental results on standard 3D benchmarks, including ShapeNet and DeepFashion3D, show that 3D-WAG outperforms state-of-the-art methods on metrics such as Minimum Matching Distance (MMD) and Coverage (COV), generating high-quality 3D shapes that accurately represent real-world data distributions.

1 Introduction

The transformative power of autoregressive (AR) models, exemplified by the GPT series [1, 4, 58, 60, 61] and subsequent advancements [2, 3, 15, 22, 35, 73, 74, 76, 77], has revolutionized natural language processing. AR models have also shown significant promise in image generation, with pioneering works such as VQGAN [91] and DALL-E [63], and

Figure 1: 3D-WAG overview. We propose an autoregressive approach to 3D shape generation, based on the "next-scale" prediction paradigm. *Left:* Our hierarchical, multi-scale wavelet encoding enables the efficient generation of highly detailed surfaces, outperforming prior models (*e.g.*, UDiFF [104], in the red box) for comparable latent code dimensionality, and showcasing generalization in conditional 3D generation tasks, like text-to-3D. *Right:* Unconditional generation of diverse, high-fidelity shapes, supporting varied implicit spatial representations using our approach.

more recent developments [37, 65, 90], showcasing their potential. Notably, recent studies [42, 75, 93] have demonstrated that AR models, through advancements in image tokenization and training data augmentation, can surpass diffusion models in image generation tasks. This growing body of research underscores the versatility and efficacy of autoregressive modeling, hinting at its vast potential for exploration in diverse domains.

However, the full potential of AR models in 3D shape generation remains largely untapped. While several works have explored AR models for 3D synthesis [11, 54, 56, 67, 72, 88], the breakthroughs that have enabled AR models to outperform diffusion models in other domains have yet to effectively translate to 3D. Existing 3D AR methods, which sequentially predict "next tokens" to model geometric structures, primarily operate on voxels, triangles, or point coordinates. These approaches face two fundamental challenges: firstly, generating high-resolution 3D shapes requires predicting prohibitively long sequences of fine-grained tokens, leading to immense computational costs; secondly, a focus on token-by-token prediction often compromises global geometric structure. This leads to limited detail with slow predictions, resulting in incoherent shapes. Despite high inference costs[30, 95], diffusion-based models have thus become the dominant choice for 3D shape generation [12, 25, 26, 38, 66, 69].

To address the limitations of autoregressive (AR) models in 3D shape generation, we propose 3D Hierarchical Wavelet-Guided Autoregressive Generation (3D-WAG), a novel framework leveraging compact hierarchical wavelet representations, which are recently also being used in diffusion-based generation models [68, 104]. Wavelet representation allows for efficient capture of multi-scale shape information from coarse to fine, utilizing low and high frequency details of wavelets, enabling direct modeling of shapes and avoiding 2D-to-3D translation artifacts in generation. Our two-stage training paradigm first employs a 3D VQVAE to learn a quantized latent space of multi-scale wavelet features effectively. Second, inspired by VAR [75], we introduce a "token map" prediction paradigm, wherein our AR model predicts entire multi-scale feature maps hierarchically, capturing both global and local details. Wavelet inversion, following the reconstruction of the wavelet representation, yields the final 3D shape at the original implicit resolution. 3D-WAG offers a compelling, efficient alternative to diffusion models, integrating wavelet representations and hierarchical

AR modeling for high-fidelity, controllable, and real-time shape generation.
Our key contributions are:

- ✓ Extending recent advancements in autoregressive generation to direct 3D shape generation with the proposed **3D-WAG** framework, introducing a next-scale prediction strategy for detailed and efficient 3D shape generation.
- ✓ Leveraging compact **wavelet representations** for scalable autoregressive modeling of high-resolution 3D shapes, significantly reducing computational overhead and enabling the use of larger AR architectures.
- ✓ Providing comprehensive experimental validation, demonstrating that **3D-WAG** outperforms existing 3D generative models, including diffusion-based methods, while providing strong generalization in categorical and text-conditioned shape generation.

2 Related Work

Neural networks have revolutionized 3D applications, driving significant progress in shape modeling, scene reconstruction, and virtual simulation [24, 28, 40, 41, 48, 49, 81, 82, 86, 96, 99, 101, 102, 103]. We build upon the established efficacy of neural implicit functions.

2.1 3D Implicit Representations.

Neural Implicit Functions (NIFs) have become indispensable for 3D representation, impacting diverse applications from surface reconstruction [52, 59] to novel view synthesis [53]. Their continuous representation enables high-fidelity rendering and geometry modeling, with advancements ranging from MLP-based SDFs and occupancy fields [52, 59] to neural unsigned distance fields (UDFs) for non-watertight surfaces [8, 13, 43, 44, 79, 98, 100]. However, direct generation with NIFs suffers from a fundamental trade-off: high fidelity necessitates dense sampling and complex networks, leading to significant computational bottlenecks, particularly in AR settings. Recent research leverages compact wavelet-based representations, exploiting signal processing domain knowledge for efficient shape modeling [16, 26, 68, 104]. These representations offer reduced computational effort and maintain detailed latent space information, mitigating compression losses.

In this work, we use compact wavelet-based hierarchical representation that efficiently represents the implicit surface information using lower spatial resolution and increased frequency channels. This transformation reduces the computational burden of 3D convolutions while preserving geometric details, enabling usage of deeper models without parameter overhead (See Table 6 in Appendix). Inherent multi-scale analysis of wavelets aligns perfectly with hierarchical AR modeling, facilitating efficient progressive generation from coarse to fine geometries, a significant advantage over usual direct NIF generation.

2.2 3D Generative Models.

Early 3D generation methods extended image generation techniques like GANs [18], VAEs [31], and flow-based models [32] to 3D, producing point clouds [5, 27, 39, 57, 89] and voxel grids [70, 83]. Recently, diffusion models have emerged as the dominant approach [12, 25, 26, 38, 66, 69, 87]. These models often leverage implicit representations like signed distance fields (SDFs) [14, 19, 25, 26, 34] or occupancy fields [94] for high-fidelity results, with meshes extracted using marching cubes [45]. To enhance training efficiency, techniques such as Diffusion-SDF [14] and 3D-LDM [55] employ VAEs to transform shapes into latent codes. However, diffusion models face challenges related to training stability and slow inference, limiting their practical applicability [21, 30, 71].

Concurrently, autoregressive (AR) models have shown remarkable progress in 2D image generation [17, 64, 78], and more recently in 3D tasks [11, 54, 56, 72, 80, 88]. A few works explore 3D generation from 2D inputs using AR models; however, they often rely on highly optimized inputs or externally computed 2D feature mappings, limiting real-world applicability [9, 62, 84, 85]. This motivates the need for direct 3D generation using AR methods. Generating high-quality, high-resolution 3D shapes is challenging due to complex representation learning. ShapeFormer [88] adopts a sparse voxel representation quantizing occupied grid cells but uses a fixed row-major order. In contrast, AutoSDF [54] avoids sequential ordering but operates over full volumetric grids, incurring high compute costs. These efforts highlight the trade-off between ordering constraints and efficiency in AR-based 3D generation. For neural implicit functions (NIFs), we propose a paradigm shift from traditional next-token prediction to "next-scale" token map prediction. Conventional AR models struggle with long sequences in dense 3D settings [11, 36, 50, 88]. We address this by using large language models (LLMs) such as GPT-2 [61] to perform both unconditional and conditional generation via simple condition concatenation, enabling faster inference and improved control.

3 Wavelet Autoregressive Shape Generation

Our 3D-WAG employs a wavelet-based representation to compactly represent high-resolution shapes with minimal loss of detail (Sec. 3.1). We follow a two-stage training approach for 3D shape generation. First, a 3D VQVAE autoencoder is trained to encode the wavelet coefficients of 3D shapes into discrete, multi-scale token maps for autoregressive modeling (Sec. 3.2). Subsequently, we use a GPT-2 [61]-like transformer-based autoregressive (AR) model to progressively generate these multi-scale token maps, predicting the next higher-scale token map conditioned on all previous ones (Sec. 3.3). An overview of our complete 3D-WAG framework is provided in Figure 2.

3.1 Compact Wavelet Representations

Modeling high-resolution 3D data in AR settings presents several challenges compared to their 2D image counterparts. The additional third spatial dimension exponentially increases the number of input variables that the model must learn, resulting in a substantial increase in network parameters and memory-intensive feature maps. This added computational overhead on GPUs limits the resolution of 3D shapes that can be modeled and significantly prolongs training and inference times. Moreover, scaling generative models to 3D introduces complex data-handling challenges: the storage and processing of 3D data incur substantially higher I/O costs, such as longer data loading times during each training iteration. Recent methods [25, 26, 68] have leveraged wavelet-based compression techniques to generate high-resolution 3D shapes with diffusion models. Building on these approaches, we adopt a compact wavelet-based spatial frequency representation for efficient signed distance field (SDF) generation in our autoregressive framework.

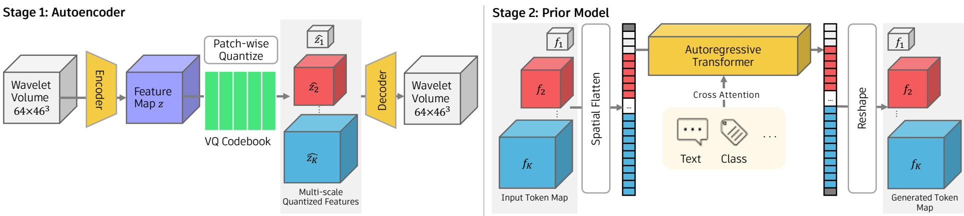

Figure 2: Overview of our architecture. We follow a two-stage training approach that is standard with latent generative models. In Stage 1, we train a vector-quantized autoencoder (VQ-VAE) on the wavelet volumes with multi-scale patch-wise quantization [75] in the latent space giving us quantized feature maps $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_K$. In Stage 2, the multi-scale VQ-VAE codebook indices in the form of token maps $f_1, f_2, \ldots, f_K$ are flattened and learned with an autoregressive decoder-only Transformer prior model, enabling next-scale generation. During inference, the generated token maps are reshaped and mapped to obtain multi-scale quantized features by using codebook. Then the obtained features are converted into wavelet volumes by the decoder from Stage 1. The wavelet volumes can be further converted into implicit functions using wavelet inversion.

We first decompose the implicit truncated signed distance field (TSDF) of a 3D shape into multi-scale wavelet coefficients using biorthogonal wavelet transformation, following the previous work [26]. Their selective retention of informative high-frequency detail coefficients ensures fine shape detail preservation with memory efficiency, achieving 99.64 percent IoU concerning input TSDF representation. This enables to use of 3D convolutions with an increased number of channels to learn the shape. Extending beyond watertight meshes, we utilize unsigned distance functions (UDFs) for non-watertight surfaces using the DeepFashion3D dataset. Furthermore, inspired by UDiff [104], we utilize a self-reconstruction process to optimize the parameters of the wavelet filter during both the decomposition and reconstruction phases. Once these wavelet parameters are learned, we use them to create a compact representation of the UDFs consisting of coarse and fine coefficient volumes. We use both of these compact wavelet volumes to represent the 3D shape and showcase the capability of the method in handling the compact wavelet volume input data using our 3D-WAG.

In Table 1, we compare the computational burden, measured in multiply–accumulate operations (MACs), for a single 3D convolution applied to different shape representations using the same kernel. The results show that the wavelet-based representation is significantly more efficient than the implicit representation at the same voxel resolution. Additional details on the compact wavelet input are provided in Appendix 6.

Table 1: **MACs of 3D convolutions on implicit and wavelet representations.** Wavelet representation reduces MACs by $\sim 3\times$.

Representation	Ch.	Filter	MACs
Implicit (1×256^3)	1	3^3	453M
Wavelet (64×46^3)	64	3^3	168M

3.2 Multi-Scale Wavelet Tokenization

To facilitate generative autoregressive (AR) modeling via next-token prediction, it is imperative to first translate each data instance into a sequence of discrete tokens by tokenization. Subsequently, a specific one-dimensional (1D) ordering of these tokens must be established, enabling the model to learn a unidirectional predictive pattern. Predicting each token conditioned on its predecessors allows for coherent and context-aware generation. However, for the non-sequential data, such as 3D shapes, the direct flattening of spatial features into linear sequences for autoregressive modeling can disrupt the inherent spatial structure. This naïve flattening approach weakens the natural correlations between the distant features, effectively disregarding the original spatial relationships.

To mitigate these limitations, we formulate 3D shape generation as a hierarchical feature prediction task [75], wherein the model estimates wavelet token maps instead of individual tokens at progressively higher resolutions. To enable this multi-scale generation process, we train an autoencoder, 3D VQVAE that encodes each 3D shape into K discrete token maps spanning varying resolutions or scales. This multi-scale discretization allows for the representation of shapes at different levels of detail, ranging from coarse structural information in low-resolution token maps to fine geometric details in high-resolution ones.

For each shape S, an autoencoder 3D VQVAE learns to compress continuous wavelet coefficients W into multi-scale discrete token maps $F = (f_1, f_2, \dots, f_K)$, where f_1 and f_K are the coarsest and finest scales, respectively. Using a 3D convolutional VAE with a modified multi-scale quantization layer, inspired by [36, 75], we quantize each f_k conditioned on $(f_1, f_2, \dots, f_{k-1})$ with a shared codebook Z. This enables progressive, coherent multi-scale reconstruction, preserving details for autoregressive generation. The multi-scale encoding and reconstruction steps of 3D VQVAE are detailed in stage 1 in Figure 2.

3.2.1 Multi-Scale Encoding

The multi-scale vector quantization encoding scheme progressively encodes an input shape representation W across multiple resolutions, utilizing residual connections to capture details at varying scales (Stage 1 of Figure 2). Given a wavelet input W, an encoder E generates an initial feature map z. This map is then iteratively quantized over K steps, each corresponding to a predefined resolution (h_k, w_k, d_k). At each scale k, the current residual tensor z_k^{res} is interpolated to match (h_k, w_k, d_k). At the finest scale K, z_K^{res} is initialized with z. The interpolated tensors are quantized using the method from [65, 75] to obtain token maps f_k from codebook Z, which stores learned feature vectors with indices. These maps retrieve feature vectors from Z, creating a quantized representation $\hat{z}_k{}^{res}$. At scale k, $\hat{z}_k{}^{res}$ serves as an estimate of z_k^{res}. Subsequently, $\hat{z}_k{}^{res}$ is upsampled to the original scale. A residual is computed by subtracting $\mathrm{C}_k(\hat{z}_k{}^{res})$ from z_k^{res}, yielding z_{k-1}^{res}. This process is repeated for all scales. Here, C_k is a convolutional layer at scale k, mitigating information loss during upscaling. Notably, convolutional layers are omitted during downsampling of z. This recursive process generates a sequence of multi-scale tokens F, encapsulating hierarchical and multi-resolution information from the wavelet encodings.

3.2.2 Multi-Scale Reconstruction

The multi-scale reconstruction scheme reverses the encoding process, reconstructing a wavelet representation from the multi-scale token maps generated during encoding. Given a sequence of multi-scale token maps F, the scheme iteratively reconstructs a feature map $\hat{z}^{recon}$ by decoding information at progressively finer scales. The process begins by initializing $\hat{z}^{recon}$ to zero. For each scale k in the sequence of K steps, we retrieve the quantized representation f_k from F. Then, uses a codebook (Z) lookup to obtain a latent representation z_k^{recon}, corresponding to the token map f_k. This latent representation z_k^{recon} is interpolated to match the current resolution (h_k, w_k, d_k). A residual connection is employed, where the current estimate $\hat{z}^{recon}$ is updated by adding $C_k(z_k^{recon})$. Once all scales are processed, a decoder D is applied to $\hat{z}^{recon}$ to reconstruct the final wavelet volume $\hat{W}$.

The autoencoder is trained by minimizing a composite loss that encourages accurate reconstruction, latent consistency, and fine-detail preservation. The objective includes three terms: (i) a reconstruction loss between the original wavelet representation W and its reconstruction $\hat{W}$, (ii) a commitment loss that encourages the encoder to map consistently to discrete embeddings across K scales, and (iii) a detail-aware loss detailed in Sec. 3.2.3 focusing on high-frequency coefficients. The overall training objective is:

$$L = \lambda_{\text{recon}} \|W - \hat{W}\|_2 + \lambda_{\text{commit}} \sum_{k=1}^{K} \|z_k^{\text{res}} - \hat{z}_k^{\text{res}}\|_2 + \lambda_{\text{detail}} L_{\text{detail}}, \tag{1}$$

where λ_{recon}, λ_{commit}, and λ_{detail} are weighting hyperparameters.

3.2.3 Detail Loss

To improve the fidelity of 3D shape reconstruction, especially around fine-grained geometric features we utilize a **detail-aware loss** that focuses specifically on high-frequency components of the wavelet representation. This loss selectively penalizes reconstruction errors at spatial locations identified as containing meaningful high-frequency information. Let $\mathcal{M}_b \subseteq \Omega$ denote the set of valid high-frequency spatial indices for sample b, determined by a binary mask, and let C be the number of channels. The detail loss measures squared error between the predicted and ground truth wavelet coefficients at these high-frequency locations, averaged across channels and samples in a batch of size B:

$$L_{\text{detail}} = \frac{1}{B} \sum_{b=1}^{B} \left[\frac{1}{C} \sum_{c=1}^{C} \left(\frac{1}{|\mathcal{M}_b|} \sum_{i \in \mathcal{M}_b} \left(W_{b,i}^{(c)} - \hat{W}_{b,i}^{(c)} \right)^2 \right) \right], \tag{2}$$

where $W_{b,i}^{(c)}$ and $\hat{W}_{b,i}^{(c)}$ denote the ground truth and predicted wavelet coefficients at spatial index i, channel c, and sample b. By explicitly encouraging accurate reconstruction of high-frequency details, 3DWAG ensures that the model captures fine surface structure more precisely, enabling high-fidelity modeling of wavelet-encoded 3D data.

3.3 Next-Scale Token Map Prediction using AR

The multi-scale token map feature encodings, learned during the first stage of training, are used to train an autoregressive shape generation model. The weights of 3D VQVAE during the second stage of training are frozen. In the second stage, the goal of the AR Transformer model is to generate token maps $\{f_1, f_2, \ldots, f_K\}$ using a "next-scale prediction" approach (Stage 2 part of Figure 2). Starting with a random initial token map, the model generates subsequent token maps one step at a time, each at progressively higher resolutions. This process continues until it reaches f_K, which matches the resolution of the original feature map. To accomplish this, we utilize a GPT-2 [61] style decoder-only Transformer model. The AR likelihood is defined as: $p(f_1, f_2, \ldots, f_K) = \prod_{i=1}^{K} p(f_i | f_{<i})$,where each f_i is conditioned on all preceding token maps $f_{<i}$. It is important to note that, at each step, all token maps at the **current** scale are estimated simultaneously rather than sequentially. For class-conditional synthesis, we use a class embedding as the start token, denoted by $[s]$, which also serves as

the conditioning factor for network AdaLN. Additionally, we normalize the queries and keys to unit vectors before the attention mechanism, enhancing training stability in our generative model. The cross-entropy loss L_{CE} is employed for AR shape generation:

$$L_{\mathrm{CE}} = -\sum_{i=1}^{K} \log p(f_i | f_{<i}), \tag{3}$$

where $p(f_i|f_{<i})$ is the predicted probability of the i-th token map f_i given all previous token maps $f_{<i}$. This loss function encourages the model to assign high probabilities to the correct token map f_i at each step, ensuring detail oriented autoregressive generation by minimizing the negative log-likelihood of the true token maps. During autoregressive inference, multi-scale token maps are sampled progressively, capturing spatial details at increasing resolutions. This sequential process enables the model to coherently generate complex shapes, refining details at each scale to produce a realistic output. These token maps are transformed into a wavelet reconstruction $\hat{W}$ using the **Reconstruction** algorithm presented in Sec. 3.2.

The wavelet representation is inverted using an inverse wavelet transform to generate an implicit representation of the shape. Additional design choices and implementation details of shape retrieval from the wavelet representation are given in Appendix 7, 8, 9. We also note that, in the 3D case, generating a volume token sequence $\{f_1, f_2, \ldots, f_K\}$ with a conventional self-attention Transformer requires $\mathcal{O}(n^3)$ autoregressive steps and incurs a computational cost of $\mathcal{O}(n^9)$. By employing the 3D-WAG approach, the time complexity for an $n \times n \times n$ latent representation is substantially reduced to $\mathcal{O}(n^6)$, see Appendix 10 for the proof.

4 Experiments

4.1 Dataset and Metrics

We evaluate our method on DeepFashion3D [105] (1,798 garments, 563 items, 10 categories) and ShapeNet [6] (51,300 models, 55 categories), using 80/20 train/test splits. Shapes are represented as signed (ShapeNet) or unsigned (DeepFashion3D) distance fields at 256^3 resolution. Unconditional generation performance is assessed via Coverage (COV), Minimum Matching Distance (MMD), and 1-Nearest Neighbor Classifier Accuracy (1-NNA), generating 1000 (DeepFashion3D) and 2000 (ShapeNet) samples with 2048 points sampled from the surfaces. For ShapeNet, quantitative metrics are reported on airplane and chair categories, following [25, 34], while qualitative evaluations utilize the entire dataset. We argue that our model can also be used on other 3D datasets; however, due to limited computing resources, we restrict our evaluations to the mentioned datasets. In Appendix 13, we further provide an ablation on our design choices.

4.2 Results on ShapeNet

We present our results on ShapeNet dataset in Table 2. We compare our 3D-WAG model to previous methods, including UDiFF [104], IM-GAN [10], Voxel-GAN [33], PointDiff [46], SPAGHETTI [20], WaveGen [25], and SALAD [34]. We also evaluate our approach in comparison to the autoregressive approach ImAM [47], which uses token-by-token prediction to generate shapes. Our model improves over these previous models.

Specifically, we obtain improved COV and MMD across the Chair and Airplane categories. The high COV reflects the model's ability to generate diverse, realistic shapes, while

Table 2: **Shape generation performance and time on ShapeNet.** Inference time is in seconds on A6000. MMD-CD and MMD-EMD scores are scaled by 10^3 and 10^2, respectively.

Method	Chair COV ↑ CD	Chair COV ↑ EMD	Chair MMD ↓ CD	Chair MMD ↓ EMD	Chair 1-NNA ↓ CD	Chair 1-NNA ↓ EMD	Airplane COV ↑ CD	Airplane COV ↑ EMD	Airplane MMD ↓ CD	Airplane MMD ↓ EMD	Airplane 1-NNA ↓ CD	Airplane 1-NNA ↓ EMD	Time (s)
IM-GAN [10]	57.01	55.23	12.02	14.76	62.15	63.78	62.21	63.10	3.298	8.479	76.50	76.29	2.98
Voxel-GAN [33]	44.29	39.89	15.35	17.59	80.74	81.62	39.01	39.78	5.875	11.53	93.60	92.98	2.75
PointDiff [46]	51.88	56.11	12.92	16.35	62.02	63.95	60.55	62.71	3.615	9.620	74.98	72.63	–
SPAGHETTI [20]	50.11	50.73	15.02	16.05	72.82	69.93	57.30	59.12	4.342	9.021	79.84	79.33	27.4
SALAD (Global) [34]	50.13	49.12	11.93	14.35	63.25	61.73	55.42	59.85	3.891	9.073	82.61	80.87	33.2
SALAD [34]	56.91	55.65	11.89	14.50	**58.23**	**58.90**	63.65	65.82	3.721	8.329	**74.35**	**71.52**	38.0
WaveGen [25]	50.12	50.67	12.31	14.47	65.58	63.39	61.40	59.57	3.605	8.075	76.23	73.25	15.4
ImAM [47]	54.38	55.71	11.90	13.97	62.43	62.11	65.82	65.41	3.248	7.900	77.12	79.13	14.2
UDiff [104]	53.02	56.51	11.82	14.19	66.42	63.91	65.23	64.22	3.229	7.879	74.85	79.42	39.0
Ours	**57.30**	**58.41**	**10.90**	**13.06**	68.10	66.31	**66.80**	**65.93**	**2.900**	**7.543**	82.10	80.92	**2.50**

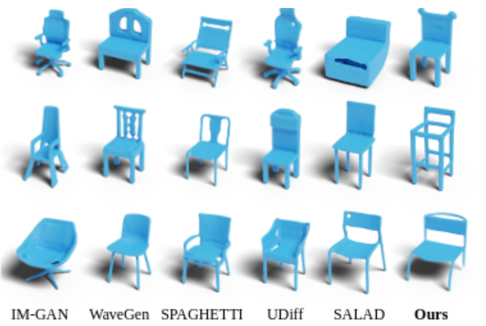

Figure 3: 3D chair generation results on ShapeNet, comparing IM-GAN, WaveGen, SPAGHETTI, UDiff, and SALAD, with our 3D-WAG on the far right. 3D-WAG shows improved structural coherence, finer details, and a wider diversity of generated designs, capturing both simple and complex geometries more effectively than other methods.

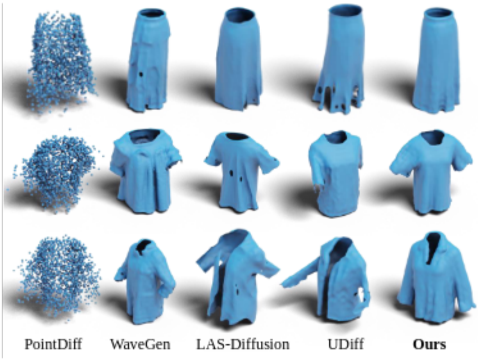

Figure 4: 3D garment generation results on the DeepFashion3D dataset [105], comparing PointDiff, WaveGen, LAS-Diffusion, UDiff, and our method (3D-WAG) on the far right. 3D-WAG produces more coherent and realistic garment shapes, capturing intricate structural details such as folds, sleeves, and openings with higher fidelity than other methods.

the low MMD indicates strong fidelity to the real data distribution, confirming the benefit of our autoregressive next-scale token map prediction strategy in capturing high-resolution geometric details without significant computational overhead. The slightly higher 1-NNA arises from our focus on preserving fine-grained local details, which, while enhancing geometric precision, can lead to sharper features that nearest neighbor classifiers more readily detect. This trade-off reflects our design choice to prioritize high-quality, detailed shape generation over smoothing, which may lower 1-NNA at the expense of visual and structural richness. Overall, these results demonstrate the effectiveness of 3D-WAG in jointly improving diversity, fidelity, and inference efficiency in 3D shape generation. Qualitative comparisons in Fig. 3 further highlight the superior detail and variety achieved by our method.

4.3 Results on DeepFashion3D

We compare our approach with SOTA diffusion-based generation approaches, including UDiff [104], PointDiff [46], WaveGen [25], Diffusion-SDF [14], and LAS-Diffusion [97]. Each model is evaluated with the official implementations. WaveGen [25] and UDiff [104] are also based on wavelet input representation, ensuring a fair comparison.

Table 3: **Shape generation on DeepFashion3D.** MMD-CD and MMD-EMD scores are scaled by 10^3 and 10^2, respectively.

Method	COV ↑		MMD ↓		1-NNA ↓	
	CD	EMD	CD	EMD	CD	EMD
PointDiff [46]	68.90	64.89	11.20	15.72	83.50	87.93
WaveGen [25]	62.75	52.15	15.78	17.29	93.20	94.96
Diffusion-SDF [14]	67.45	62.40	14.92	16.88	89.31	92.81
LAS-Diffusion [97]	67.82	56.43	14.75	16.79	88.95	91.68
UDiff [104]	**69.98**	68.01	11.75	14.25	**82.11**	82.47
Ours	**69.85**	**68.03**	**10.96**	**13.89**	82.01	**82.30**

Table 3 highlights the superior performance of 3D-WAG on the DeepFashion3D dataset compared to previous approaches. 3D-WAG achieves the highest Coverage (COV) scores, with 69.85 (CD) and 68.03 (EMD), indicating a broader diversity of generated shapes that better represent the dataset's range of 3D garments. Additionally, 3D-WAG excels in Minimum Matching Distance (MMD), achieving the lowest scores across both CD and EMD metrics, with values of 10.96 and 13.89, respectively, demonstrating higher fidelity to real shapes. The 1-Nearest Neighbor Accuracy (1-NNA) scores of 82.01 (CD) and 82.30 (EMD) further confirm that 3D-WAG produces realistic samples with a similarity distribution close to the real data. These results validate the effectiveness of our approach in open-surface shape generation in terms of both diversity and quality. Fig. 4 presents qualitative examples of our unconditional generation results on DeepFashion3D. More unconditional generation results, including failure cases and conditional generation, are found in Appendix 11 12 14.

5 Conclusion

We present 3D-WAG, a novel autoregressive approach for 3D shape generation that operates directly on 3D representations. 3D-WAG transforms the traditional "next-token" prediction framework of 3D autoregressive methods to a "next-scale token map" prediction framework. The key innovation of 3D-WAG is reframing 3D autoregressive generation as a "next-scale token map" prediction task using a compact wavelet representation as input, enabling efficient generation while preserving intricate geometric details in a structured, hierarchical way. This approach is validated by 3D-WAG's outstanding performance on widely recognized benchmarks, where it achieves superior Coverage and MMD metrics compared to SOTA models. Our results show that 3D-WAG generates high-fidelity 3D shapes that are closely aligned with real data distributions, while requiring less inference time.

Future Work While this work primarily introduces a novel and competitive approach to 3D shape generation, we plan to further optimize the autoregressive generation module. Additionally, we aim to explore a broader range of applications, potentially enabling more complex 3D generation scenarios. Furthermore, we will investigate alternative conditioning mechanisms to enhance the flexibility and robustness of 3D-WAG across diverse 3D tasks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[5] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 364–381. Springer, 2020.

[6] 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.

[7] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[8] Weikai Chen, Cheng Lin, Weiyang Li, and Bo Yang. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18522–18531, 2022.

[9] Yongwei Chen, Yushi Lan, Shangchen Zhou, Tengfei Wang, and Xingang Pan. Sar3d: Autoregressive 3d object generation and understanding via multi-scale 3d vqvae. In *CVPR*, 2025.

[10] 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*, pages 5939–5948, 2019.

[11] An-Chieh Cheng, Xueting Li, Sifei Liu, Min Sun, and Ming-Hsuan Yang. Autoregressive 3d shape generation via canonical mapping. In *European Conference on Computer Vision (ECCV)*, 2022.

[12] 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*, pages 4456–4465, 2023.

[13] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020.

[14] 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*, pages 2262–2272, 2023.

[15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[16] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990.

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[18] 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*, 63(11):139–144, 2020.

[19] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023.

[20] 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)*, 41(4):1–20, 2022.

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

[22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[23] Fei Hou, Xuhui Chen, Wencheng Wang, Hong Qin, and Ying He. Robust zero level-set extraction from unsigned distance fields based on double covering. *ACM Transactions on Graphics*, 42(6), 2023.

[24] Han Huang, Yulun Wu, Junsheng Zhou, Ge Gao, Ming Gu, and Yu-Shen Liu. Neusurf: On-surface priors for neural surface reconstruction from sparse input views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[25] 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*, pages 1–9, 2022.

[26] Ka-Hei Hui, Aditya Sanghi, Arianna Rampini, Kamal Rahimi Malekshan, Zhengzhe Liu, Hooman Shayani, and Chi-Wing Fu. Make-a-shape: a ten-million-scale 3d shape model. In *Forty-first International Conference on Machine Learning*, 2024.

[27] Le Hui, Rui Xu, Jin Xie, Jianjun Qian, and Jian Yang. Progressive point cloud deconvolution generation network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 397–413. Springer, 2020.

[28] Chuan Jin, Tieru Wu, and Junsheng Zhou. Multi-grid representation with field regularization for self-supervised surface reconstruction from point clouds. *Computers & Graphics*, 2023.

[29] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[30] Tero Karras, Miika Aittala, Timo Hellsten, Samuli Laine, Jaakko Lehtinen, and Frédo Durand. Elucidating the latent space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25509–25521, 2022.

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

[32] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[33] Marian Kleineberg, Matthias Fey, and Frank Weichert. Adversarial generation of continuous implicit shape representations. *arXiv preprint arXiv:2002.00349*, 2020.

[34] Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. Salad: Part-level latent diffusion for 3d shape generation and manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14441–14451, 2023.

[35] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model, 2023.

[36] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[37] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[38] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12642–12651, 2023.

[39] Ruihui Li, Xianzhi Li, Ka-Hei Hui, and Chi-Wing Fu. Sp-gan: Sphere-guided 3d shape generation and manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–12, 2021.

[40] Shujuan Li, Junsheng Zhou, Baorui Ma, Yu-Shen Liu, and Zhizhong Han. NeAF: Learning neural angle fields for point normal estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[41] Shujuan Li, Junsheng Zhou, Baorui Ma, Yu-Shen Liu, and Zhizhong Han. Learning continuous implicit field with local distance indicator for arbitrary-scale point cloud upsampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[42] Xuantong Liu, Shaozhe Hao, Xianbiao Qi, Tianyang Hu, Jun Wang, Rong Xiao, and Yuan Yao. Elucidating the design space of language models for image generation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[43] Yu-Tao Liu, Li Wang, Jie Yang, Weikai Chen, Xiaoxu Meng, Bo Yang, and Lin Gao. Neudf: Leaning neural unsigned distance fields with volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 237–247, 2023.

[44] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, and Wenping Wang. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20834–20843, 2023.

[45] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4):163–169, 1987.

[46] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.

[47] Simian Luo, Xuelin Qian, Yanwei Fu, Yinda Zhang, Ying Tai, Zhenyu Zhang, Chengjie Wang, and Xiangyang Xue. Learning versatile 3d shape generation with improved ar models. *arXiv preprint arXiv:2303.14700*, 2023.

[48] Baorui Ma, Haoge Deng, Junsheng Zhou, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Geodream: Disentangling 2d and geometric priors for high-fidelity and consistent 3d generation. *arXiv preprint arXiv:2311.17971*, 2023.

[49] Baorui Ma, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Towards better gradient consistency for neural signed distance functions via level set alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17724–17734, 2023.

[50] Tejaswini Medi, Jawad Tayyub, Muhammad Sarmad, Frank Lindseth, and Margret Keuper. Fullformer: Generating shapes inside shapes. In Ullrich Köthe and Carsten Rother, editors, *Pattern Recognition*, pages 147–162, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-54605-1.

[51] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations (ICLR) 2024*, 2024.

[52] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[53] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

[54] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022.

[55] 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.

[56] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International Conference on Machine Learning*, pages 7220–7229. PMLR, 2020.

[57] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[58] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[59] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[60] Alec Radford. Improving language understanding by generative pre-training, 2018.

[61] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[62] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan T. Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2349–2359, October 2023. doi: 10.1109/ICCV51070.2023.00223.

[63] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[64] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[65] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[66] Pradyumna Reddy, Ismail Elezi, and Jiankang Deng. G3dr: Generative 3d reconstruction in imagenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9655–9665, 2024.

[67] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4209–4219, 2024.

[68] Aditya Sanghi, Aliasghar Khani, Pradyumna Reddy, Arianna Rampini, Derek Cheung, Kamal Rahimi Malekshan, Kanika Madan, and Hooman Shayani. Wavelet latent diffusion (wala): Billion-parameter 3d generative model with compact wavelet encodings, 2024. URL https://arxiv.org/abs/2411.08017.

[69] 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*, pages 20875–20886, 2023.

[70] Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, pages 87–96. PMLR, 2017.

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

[72] Yongbin Sun, Yue Wang, Ziwei Liu, Joshua Siegel, and Sanjay Sarma. Pointgrow: Autoregressively learned point cloud generation with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 61–70, 2020.

[73] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.

[74] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[75] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Oral Poster Track*, 2024.

[76] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[77] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[78] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[79] Li Wang, Jie Yang, Weikai Chen, Xiaoxu Meng, Bo Yang, Jintao Li, and Lin Gao. Hsdf: Hybrid sign and distance field for modeling surfaces with arbitrary topologies. In *Advances in Neural Information Processing Systems*, 2022.

[80] Si-Tong Wei, Rui-Huan Wang, Chuan-Zhi Zhou, Baoquan Chen, and Peng-Shuai Wang. Octgpt: Octree-based multiscale autoregressive models for 3d shape generation. In *Proceedings of the ACM SIGGRAPH 2025 Conference*, 2025.

[81] Xin Wen, Junsheng Zhou, Yu-Shen Liu, Hua Su, Zhen Dong, and Zhizhong Han. 3D shape reconstruction from 2D images with disentangled attribute flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3803–3813, 2022.

[82] Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. PMP-Net++: Point cloud completion by transformer-enhanced multistep point moving paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):852–867, 2023.

[83] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

[84] Shuang Wu, Youtian Lin, Yifei Zeng, Feihu Zhang, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In *NeurIPS 2024 Poster Session*, 2024.

[85] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

[86] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflake point deconvolution for point cloud completion and generation with skip-transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6320–6338, 2023.

[87] 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. *Computer Graphics Forum (Eurographics Symposium on Geometry Processing)*, 2025. doi: 10.1111/cgf.70198.

[88] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022.

[89] 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*, pages 4541–4550, 2019.

[90] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[91] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *International Conference on Learning Representations (ICLR) 2022*, 2022.

[92] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gzqrANCF4g.

[93] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024.

[94] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 42(4), jul 2023. ISSN 0730-0301. doi: 10.1145/3592442. URL https://doi.org/10.1145/3592442.

[95] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00355.

[96] Wenyuan Zhang, Ruofan Xing, Yunfan Zeng, Yu-Shen Liu, Kanle Shi, and Zhizhong Han. Fast learning radiance fields by shooting much fewer rays. *IEEE Transactions on Image Processing*, 2023.

[97] 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 (SIGGRAPH)*, 42(4), 2023.

[98] Junsheng Zhou, Baorui Ma, Yu-Shen Liu, Yi Fang, and Zhizhong Han. Learning consistency-aware unsigned distance functions progressively from raw point clouds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[99] Junsheng Zhou, Xin Wen, Yu-Shen Liu, Yi Fang, and Zhizhong Han. Self-supervised point cloud representation learning with occlusion auto-encoder. *arXiv e-prints*, pages arXiv–2203, 2022.

[100] Junsheng Zhou, Baorui Ma, Shujuan Li, Yu-Shen Liu, and Zhizhong Han. Learning a more continuous zero level set in unsigned distance fields through level set projection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[101] Junsheng Zhou, Baorui Ma, Wenyuan Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[102] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *International Conference on Learning Representations*, 2024.

[103] Junsheng Zhou, Xin Wen, Baorui Ma, Yu-Shen Liu, Yue Gao, Yi Fang, and Zhizhong Han. 3d-oae: Occlusion auto-encoders for self-supervised learning on point clouds. *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[104] Junsheng Zhou, Weiqi Zhang, Baorui Ma, Kanle Shi, Yu-Shen Liu, and Zhizhong Han. Udiff: Generating conditional unsigned distance fields with optimal wavelet diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21496–21506, 2024.

[105] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 512–530. Springer, 2020.