Novelty Assessment Report

Paper: Unified Cross-Scale 3D Generation and Understanding via Autoregressive Modeling
PDF URL: https://openreview.net/pdf?id=mnI8CFj2WP
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

3D structure modeling is essential across scales, enabling applications from fluid simulation and 3D reconstruction to protein folding and molecular docking. Yet, despite shared 3D spatial patterns, current approaches remain fragmented, with models narrowly specialized for specific domains and unable to generalize across tasks or scales. We propose Uni-3DAR, a unified autoregressive framework for cross-scale 3D generation and understanding. At its core is a coarse-to-fine tokenizer based on octree data structures, which compresses diverse 3D structures into compact 1D token sequences. We further propose a two-level subtree compression strategy, which reduces the octree token sequence by up to 8x. To address the challenge of dynamically varying token positions introduced by compression, we introduce a masked next-token prediction strategy that ensures accurate positional modeling, significantly boosting model performance. Extensive experiments across multiple 3D generation and understanding tasks, including small molecules, proteins, polymers, crystals, and macroscopic 3D objects, validate its effectiveness and versatility. Notably, Uni-3DAR surpasses previous state-of-the-art diffusion models by a substantial margin, achieving up to 256\% relative improvement while delivering inference speeds up to 21.8x faster.

Core Task Landscape

This paper addresses: **Cross-Scale 3D Structure Generation and Understanding**
A total of **50 papers** were analyzed and organized into a taxonomy with **30 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Autoregressive and Hierarchical 3D Generation**
- **Part-Aware and Compositional 3D Generation**
- **Large-Scale and Multi-Modal 3D Generation**
- **3D Reconstruction and Understanding**
- **Point Cloud Processing and Multi-Scale Feature Learning**
- **Mesh Processing and Multi-Scale Denoising**
- **Application-Specific Cross-Scale Methods**
- **Cross-Scale Physical and Material Systems**
- **Cross-Scale Biological and Medical Imaging**

Complete Taxonomy Tree

- Cross-Scale 3D Structure Generation and Understanding Survey Taxonomy
- Autoregressive and Hierarchical 3D Generation
  - Multi-Scale Tokenization and Autoregressive Modeling ★ (5 papers)
  - [0] Unified Cross-Scale 3D Generation and Understanding via Autoregressive Modeling (Anon et al., 2026) View paper
  - [1] SAR3D: Autoregressive 3D object generation and understanding via multi-scale 3D VQVAE (Yong-wei Chen, 2025) View paper
  - [18] 3D-WAG: Hierarchical Wavelet-Guided Autoregressive Generation for High-Fidelity 3D Shapes (Medi, 2024) View paper
  - [43] Octree Transformer: Autoregressive 3D Shape Generation on Hierarchically Structured Sequences (Moritz Ibing, 2021) View paper
  - [47] OctGPT: Octree-based Multiscale Autoregressive Models for 3D Shape Generation (Siâ□□Tong Wei, 2025) View paper
  - Hierarchical Voxel and Octree-Based Generation (2 papers)
  - [3] Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies (Xuanchi Ren, 2024) View paper
  - [6] HierOctFusion: Multi-scale Octree-based 3D Shape Generation via Part-Whole-Hierarchy Message Passing (Gao Xin-jie, 2025) View paper
  - Hierarchical Latent Space and Set-Structured Generation (2 papers)
  - [17] Setvae: Learning hierarchical composition for generative modeling of set-structured data (Jinwoo Kim, 2021) View paper
  - [19] LION: Latent Point Diffusion Models for 3D Shape Generation (Zeng, 2022) View paper
- Part-Aware and Compositional 3D Generation
  - Part-Aware Mesh and Structure Generation (2 papers)
  - [8] PartCrafter: Structured 3D Mesh Generation via Compositional Latent Diffusion Transformers (Lin Yu-Chen, 2025) View paper
  - [9] Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion (Yang, 2025) View paper
  - Sphere-Guided and Spatial Prior-Based Generation (1 papers)
  - [2] SP-GAN: Sphere-guided 3D shape generation and manipulation (Li RuiHui, 2021) View paper
- Large-Scale and Multi-Modal 3D Generation
  - Large-Scale Generative Models for 3D Assets (1 papers)
  - [4] CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets (Long-Wen Zhang, 2024) View paper
  - Text-Conditioned and Hierarchical 3D Generation (2 papers)

- [28] HyperSDFusion: Bridging Hierarchical Structures in Language and Geometry for Enhanced 3D Text2Shape Generation (Leng Zhiying, 2024) View paper
- [37] MuSeLLM: SDF Generation and Understanding via Multi-Scale Tokenization with Position-Aware Guidance (Ding Tianwei, 2025) View paper
- Scene-Level and Hierarchical 3D Generation (2 papers)
- [14] SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation With Fine-Grained Geometry (Lin Gao, 2023) View paper
- [20] HiScene: creating hierarchical 3D scenes with isometric view generation (Wen-qi Dong, 2025) View paper

- 3D Reconstruction and Understanding
  - Multi-View and Diffusion-Based 3D Reconstruction (2 papers)
  - [12] PSHuman: Photorealistic Single-image 3D Human Reconstruction using Cross-Scale Multiview Diffusion and Explicit Remeshing (Peng Li, 2024) View paper
  - [21] FreeSplat: Generalizable 3D Gaussian Splatting Towards Free-View Synthesis of Indoor Scenes (Wang Yun-song, 2024) View paper
  - CAD Model Reconstruction and Sequence-Based Modeling (1 papers)
  - [16] Brep2Seq: a dataset and hierarchical deep learning network for reconstruction and generation of computer-aided design models (Shuming Zhang, 2024) View paper
  - Multi-Scale Representation and Uncertainty Estimation (1 papers)
  - [27] Variational Multi-scale Representation for Estimating Uncertainty in 3D Gaussian Splatting (Yiu-Ming Cheung, 2024) View paper
  - NeRF-Based Multi-Scale and Wavelet Representations (2 papers)
  - [45] New view synthesis via multiscale-depth and transformers (Xue Jiao Chen, 2024) View paper
  - [48] Trinerflet: A wavelet based triplane nerf representation (Rajaei Khatib, 2024) View paper
- Point Cloud Processing and Multi-Scale Feature Learning
  - Multi-Scale Geometry-Aware Attention for Point Clouds (2 papers)
  - [29] Multi-scale Geometry-aware Transformer for 3D Point Cloud Classification (Wei Xian, 2023) View paper
  - [36] Multi-scale Geometry-aware Self-Attention for 3D Point Cloud Classification (Cui Heng, 2023) View paper
  - Cross-Level and Cross-Scale Point Cloud Networks (1 papers)
  - [44] 3CROSSNet: Cross-level cross-scale cross-attention network for point cloud representation (Xian-Feng Han, 2022) View paper
  - Hierarchical Shape Correspondence and Matching (1 papers)
  - [39] 3D-TRANS: 3D Hierarchical Transformer for Shape Correspondence Learning (Hao Huang, 2024) View paper
  - Adversarial Robustness and Geometric Structure Preservation (1 papers)
  - [23] Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack (Daizong Liu, 2024) View paper
- Mesh Processing and Multi-Scale Denoising (1 papers)
  - [35] ResGEM: Multi-Scale Graph Embedding Network for Residual Mesh Denoising (Ziqi Zhou, 2024) View paper
- Application-Specific Cross-Scale Methods
  - Autonomous Driving and Multi-Modal Fusion (2 papers)
  - [5] Drivedreamer: Towards real-world-drive world models for autonomous driving (Xiaofeng Wang, 2024) View paper
  - [31] Interactive multi-scale fusion of 2D and 3D features for multi-object vehicle tracking (Guangming Wang, 2023) View paper
  - Robotics and Language-Embedded 3D Understanding (1 papers)
  - [42] Language embedded radiance fields for zero-shot task-oriented grasping (Sharma, 2023) View paper
  - Multi-Modal Spatial Environment Understanding (2 papers)
  - [11] Multi-modal and multi-scale spatial environment understanding for immersive visual text-to-speech (Liu Rui, 2025) View paper
  - [34] Advancing fine-grained visual understanding with multi-scale alignment in multi-modal models (Wang Wei, 2025) View paper
  - Object Detection and Cross-Scale Feature Mining (1 papers)
  - [7] CIDNet: Cross-Scale Interference Mining Detection Network for underwater object detection (Gaoli Zhao, 2025) View paper
  - 3D Lane Detection and Hierarchical Shape Matching (1 papers)
  - [46] Flexible 3D Lane Detection by Hierarchical Shape Matching (Zhihao Guan, 2023) View paper
  - Protein and Molecular Structure Generation (1 papers)
  - [49] ProtComposer: Compositional Protein Structure Generation with 3D Ellipsoids (StÃ¤rk, 2025) View paper
- Cross-Scale Physical and Material Systems
  - Multi-Scale Material Characterization and Reconstruction (2 papers)
  - [25] Multi-scale reconstruction of porous media from low-resolution core images using conditional generative adversarial networks (Yongfei Yang, 2022) View paper
  - [26] A supervised learning-assisted multi-scale study for thermal and mechanical behavior of porous Silica (Ali Khalvandi, 2024) View paper
  - Cross-Scale Fabrication and Hierarchical Manufacturing (5 papers)
  - [13] Melt electrowriting enabled 3D liquid crystal elastomer structures for cross-scale actuators and temperature field sensors (Xueming Feng, 2024) View paper
  - [30] 3D Printing of Hierarchical Structures Made of Inorganic Silicon-Rich Glass Featuring Self-Forming Nanogratings (Po-Han Huang, 2024) View paper
  - [38] 3D-Printed functionally graded PCL-HA scaffolds with Multi-Scale porosity (Hatice Kubra Bilgili, 2025) View paper
  - [41] Low-temperature-field-assisted fabrication of cross-scale tissue engineering scaffolds (Jing Ye, 2024) View paper
  - [50] Aluminium surface work hardening enables multi-scale 3D lithography (Lang Wang, 2025) View paper
  - Cross-Scale Physical Simulation and Modeling (2 papers)
  - [10] Global cross-scale simulation and experiment of supercritical CO2 boiler tube wall temperature based on bidirectional fluid-thermal coupling (Xuan Wang, 2025) View paper
  - [15] A cross-scale numerical framework of PEM water electrolyzer with pore-scale transport properties of PTL (Jianlin Fu, 2025) View paper
  - Hierarchical Material Design and Electromagnetic Applications (2 papers)
  - [22] Unraveling Crossâ□□Scale Fluorination Mechanisms in Nonâ□□Fullerene Acceptors for Highâ□□Efficiency Organic Photovoltaics (Xingpeng Liu, 2025) View paper
  - [40] Dimensional engineering of hierarchical nanopagodas for customizing crossâ□□scale magnetic coupling networks to enhance electromagnetic wave absorption (Longjun Rao, 2023) View paper

- Cross-Scale Biological and Medical Imaging
  - Whole-Organ Cross-Scale Pathological Analysis (1 papers)
  - [24] Generation of a Highâ□□Precision Whole Liver Panorama and Crossâ□□Scale 3D Pathological Analysis for Hepatic Fibrosis (Xiaochuan Zhang, 2025) View paper
  - Single-Cell Cross-Scale Tracing and Tumor Analysis (1 papers)
  - [33] Cross-scale tracing of nanoparticles and tumors at the single-cell level using the whole-lung atlas (Zeying Cao, 2023) View paper
  - Oceanic and Environmental Cross-Scale Structure Analysis (1 papers)
  - [32] Observed 3D Structure, Generation, and Dissipation of Oceanic Mesoscale Eddies in the South China Sea (Zhiwei Zhang, 2016) View paper

Narrative

Core task: cross-scale 3D structure generation and understanding. This field addresses the challenge of representing and synthesizing three-dimensional data across multiple levels of detail, from fine geometric features to large-scale scene layouts. The taxonomy reveals a rich landscape organized around several complementary perspectives. Autoregressive and hierarchical generation methods leverage multi-scale tokenization and sequential modeling to build structures progressively, often using octree-based representations (e.g., Octree Transformer[43], OctGPT[47]) or hierarchical fusion strategies (HierOctFusion[6]). Part-aware and compositional approaches emphasize decomposing objects into meaningful components (PartCrafter[8], Omnipart[9]), while large-scale and multi-modal branches integrate diverse data sources—such as text, images, and sensor inputs—to generate expansive environments (DriveDreamer[5], CLAY[4]). Reconstruction and understanding methods focus on inferring 3D structure from observations, point cloud processing tackles efficient multi-scale feature learning, and mesh processing addresses denoising across resolutions. Application-specific branches span physical material systems, biological imaging, and domain-tailored workflows, reflecting the breadth of cross-scale challenges.

Within this ecosystem, a particularly active line of work centers on autoregressive and hierarchical tokenization strategies that encode 3D data at multiple resolutions for efficient generation. Unified CrossScale 3D[0] sits squarely in this branch, emphasizing multi-scale tokenization and autoregressive modeling to handle varying levels of geometric detail. It shares conceptual ground with SAR3D[1], which also adopts autoregressive frameworks for 3D synthesis, and with Xcube[3], another recent effort exploring hierarchical representations. Compared to these neighbors, Unified CrossScale 3D[0] appears to pursue a more integrated treatment of scale transitions within a single generative pipeline, whereas works like 3D-WAG[18] and LION[19] may prioritize different trade-offs between expressiveness and computational efficiency. Across the field, open questions persist around balancing fine-grained fidelity with scalability, integrating part-level semantics into hierarchical models, and bridging the gap between purely geometric methods and multi-modal, application-driven systems.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. SAR3D: Autoregressive 3D object generation and understanding via multi-scale 3D VQVAE

Authors: Yong-wei Chen, Yushi Lan, Yongwei Chen, Shangchen Zhou, Tengfei Wang, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Autoregressive models have demonstrated remarkable success across various fields, from large language models (LLMs) to large multimodal models (LMMs) and 2D content generation, moving closer to artificial general intelligence (AGI). Despite these advances, applying autoregres-sive approaches to 3D object generation and understanding remains largely unexplored. This paper introduces Scale AutoRegressive 3D (SAR3D), a novel framework that leverages a multi-scale 3D vector-quantized variational aut...

Relationship Analysis

Both papers belong to the Multi-Scale Tokenization and Autoregressive Modeling category, employing hierarchical representations for 3D generation. They overlap in using multi-scale vector quantization (VQ-VAE) and autoregressive transformers for 3D structure generation, with both addressing macroscopic 3D objects. However, the original paper (Uni-3DAR) uses octree-based tokenization with 2-level subtree compression and masked next-token prediction for cross-scale generation spanning microscopic to macroscopic domains, while the candidate paper (SAR3D) focuses on multi-scale triplane VQ-VAE with next-scale prediction specifically for macroscopic 3D object generation and understanding, without the octree structure or cross-scale unification.

---

2. 3D-WAG: Hierarchical Wavelet-Guided Autoregressive Generation for High-Fidelity 3D Shapes

Authors: Medi, Tejaswini, Rampini, Arianna, Tejaswini Medi, et al. (14 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

Autoregressive (AR) models have achieved remarkable success in natural language and image generation, but their application to 3D shape modeling remains largely unexplored. Unlike diffusion models, AR models enable more efficient and controllable generation with faster inference times, making them especially suitable for data-intensive domains. Traditional 3D generative models using AR approaches often rely on ``next-token"predictions at the voxel or point level. While effective for certain appl...

Relationship Analysis

Both papers belong to the Multi-Scale Tokenization and Autoregressive Modeling category, employing hierarchical representations for 3D generation. While Uni-3DAR uses octree-based tokenization with masked next-token prediction for cross-scale structures (molecules to macroscopic objects), 3D-WAG employs wavelet-guided multi-scale token maps with next-scale prediction specifically for implicit distance fields of 3D shapes. The key difference is that Uni-3DAR focuses on unified cross-scale generation and understanding across diverse domains using octree compression, whereas 3D-WAG specializes in high-fidelity shape generation using wavelet decomposition for implicit representations.

---

3. Octree Transformer: Autoregressive 3D Shape Generation on Hierarchically Structured Sequences

Authors: Moritz Ibing, Gregor Kobsik, Leif Kobbelt, L. Kobbelt | **Year/Venue**: 2021 • 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | **URL**: View paper

Abstract

Autoregressive models have proven to be very powerful in NLP text generation tasks and lately have gained popularity for image generation as well. However, they have seen limited use for the synthesis of 3D shapes so far. This is mainly due to the lack of a straightforward way to linearize 3D data as well as to scaling problems with the length of the resulting sequences when describing complex shapes. In this work we address both of these problems. We use octrees as a compact hierarchical shape ...

Relationship Analysis

Both papers belong to the Multi-Scale Tokenization and Autoregressive Modeling category, employing octree-based hierarchical tokenization for autoregressive 3D generation. They overlap in using octree structures to compress 3D data into sequences and applying

autoregressive transformers for generation. However, the original paper (Uni-3DAR) introduces a unified cross-scale framework spanning microscopic to macroscopic domains with masked next-token prediction and two-level subtree compression, while the candidate paper (Octree Transformer) focuses specifically on macroscopic 3D shape generation with a different compression strategy using strided convolutions and does not address cross-scale unification or microscopic structures.

---

4. OctGPT: Octree-based Multiscale Autoregressive Models for 3D Shape Generation

Authors: Siâ□□Tong Wei, Rui-Huan Wang, Si-Tong Wei, Chuan-Zhi Zhou, Baoquan Chen, et al. (6 authors total) | **Year/Venue**: 2025 • Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers | **URL**: View paper

Abstract

Autoregressive models have achieved remarkable success across various domains, yet their performance in 3D shape generation lags significantly behind that of diffusion models. In this paper, we introduce OctGPT, a novel multiscale autoregressive model for 3D shape generation that dramatically improves the efficiency and performance of prior 3D autoregressive approaches, while rivaling or surpassing state-of-the-art diffusion models. Our method employs a serialized octree representation to effici...

Relationship Analysis

Both papers belong to the Multi-Scale Tokenization and Autoregressive Modeling category, employing octree-based hierarchical representations for autoregressive 3D generation. They overlap in using octree tokenization to compress 3D structures into sequences and applying autoregressive transformers for generation tasks. However, the original paper (Uni-3DAR) focuses on cross-scale unification across molecular to macroscopic domains with a two-level subtree compression and masked next-token prediction strategy, while OctGPT specializes in high-resolution macroscopic 3D shape generation (up to 1024^3) with VQVAE-based binary tokens and token-parallel generation schemes for computational efficiency.

Contributions Analysis

Overall novelty summary. Uni-3DAR proposes a unified autoregressive framework for cross-scale 3D generation and understanding, employing an octree-based tokenizer with two-level subtree compression and masked next-token prediction. The paper resides in the 'Multi-Scale Tokenization and Autoregressive Modeling' leaf, which contains five papers total, including SAR3D, Xcube, and two others. This leaf represents a moderately active research direction within the broader autoregressive and hierarchical generation branch, focusing specifically on tokenization strategies and sequential prediction for 3D synthesis across varying resolutions.

The taxonomy reveals that Uni-3DAR's leaf sits within a larger autoregressive and hierarchical generation subtopic, which also includes neighboring leaves on hierarchical voxel/octree-based generation and hierarchical latent space methods. Adjacent branches address part-aware compositional generation, large-scale multi-modal synthesis, and reconstruction/understanding tasks. The scope note for this leaf emphasizes multi-scale vector quantization and hierarchical latent codes, while excluding diffusion-based or GAN-based methods without autoregressive components. Uni-3DAR's cross-domain ambitions (molecules to macroscopic objects) distinguish it from sibling papers that may target narrower application scopes or single-scale regimes.

Among thirty candidates examined, the contribution-level analysis found limited prior work overlap. The unified autoregressive framework (Contribution A) examined ten candidates with zero refutable matches, suggesting relative novelty in bridging multiple domains under one model. The octree tokenizer with two-level compression (Contribution B) examined ten candidates and identified one refutable match, indicating some precedent for hierarchical octree tokenization but potentially novel compression strategies. The masked next-token prediction for dynamic positions (Contribution C) also examined ten candidates with zero refutations, hinting at a less-explored technique within this limited search scope.

Based on the top-thirty semantic matches and citation expansion, Uni-3DAR appears to occupy a moderately novel position, particularly in its cross-domain unification and compression strategy. However, the analysis does not cover exhaustive literature beyond these candidates, and the single refutable match for the tokenizer suggests some overlap with existing octree-based methods. The framework's claimed versatility across molecular and macroscopic scales remains a distinguishing feature within the examined scope, though broader validation would require deeper exploration of domain-specific prior work.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Unified autoregressive framework for cross-scale 3D generation and understanding

Description: The authors introduce Uni-3DAR, a single autoregressive model that handles both 3D generation and understanding tasks across multiple scales, from microscopic structures like molecules and proteins to macroscopic 3D objects. This framework unifies previously fragmented domain-specific approaches into one architecture.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Polygen: An autoregressive generative model of 3d meshes

URL: View paper

Brief Assessment

PolyGen[53] focuses exclusively on 3D mesh generation for macroscopic objects using autoregressive modeling. It does not address cross-scale modeling (molecules to macroscopic objects) or unified generation-understanding tasks that are central to the original paper's contribution.

---

2. Autoregressive models in vision: A survey

URL: View paper

Brief Assessment

Autoregressive Vision Survey[56] is a broad survey paper covering autoregressive models across various vision domains. It does not present a specific unified framework for cross-scale 3D generation and understanding, but rather reviews existing approaches in the field.

---

3. Autosdf: Shape priors for 3d completion, reconstruction and generation

URL: View paper

Brief Assessment

AutoSDF[55] focuses on a non-sequential autoregressive prior for 3D shape completion, reconstruction, and generation using discretized latent representations. It does not address cross-scale modeling (molecules to macroscopic objects) or unified generation-understanding tasks as claimed in the original paper.

---

4. Autopartgen: Autogressive 3d part generation and discovery

URL: View paper

Brief Assessment

AutoPartGen[52] focuses on compositional 3D part generation using autoregressive modeling for object decomposition, not on cross-scale unified frameworks spanning molecules to macroscopic objects.

---

5. Bamm: Bidirectional autoregressive motion model

URL: View paper

Brief Assessment

BAMM[54] focuses exclusively on human motion generation from text, not cross-scale 3D structures. It does not address molecular, crystal, protein, or macroscopic object modeling that the original paper claims as novel.

---

6. HERMES: A Unified Self-Driving World Model for Simultaneous 3D Scene Understanding and Generation

URL: View paper

Brief Assessment

HERMES[57] focuses on driving world models for autonomous driving scenarios using BEV representations, not cross-scale 3D generation spanning molecules to macroscopic objects. The domains and technical approaches are fundamentally different.

---

7. Pushing auto-regressive models for 3d shape generation at capacity and scalability

URL: View paper

Brief Assessment

Autoregressive 3D Capacity[51] focuses exclusively on 3D shape generation at macroscopic scales using autoregressive models, without addressing cross-scale unification or understanding tasks that span from microscopic (molecules, proteins) to macroscopic structures.

---

8. OctGPT: Octree-based Multiscale Autoregressive Models for 3D Shape Generation

URL: View paper

Brief Assessment

OctGPT[47] focuses exclusively on 3D shape generation tasks without addressing understanding tasks (e.g., property prediction, binding site prediction). It does not claim to unify generation and understanding within a single framework.

---

9. Autoregressive 3d shape generation via canonical mapping

URL: View paper

Brief Assessment

Canonical Mapping 3D[59] focuses exclusively on point cloud generation for macroscopic 3D objects (ShapeNet categories) using canonical sphere mapping. It does not address cross-scale modeling (molecules to objects) or unified generation-understanding tasks.

---

10. Show-o2: Improved Native Unified Multimodal Models

URL: View paper

Brief Assessment

Show-o2[58] focuses on multimodal models for text, images, and videos using autoregressive modeling and flow matching in a 2D/temporal domain. The original paper addresses 3D spatial structures (molecules, proteins, crystals, 3D objects) using octree-based tokenization for cross-scale 3D generation and understanding, which is fundamentally different from Show-o2's focus on 2D visual content and temporal sequences.

Contribution 2: Coarse-to-fine octree-based tokenizer with two-level subtree compression

Description: The authors develop a hierarchical tokenization method using octree data structures to efficiently compress sparse 3D structures into 1D sequences. They introduce a two-level subtree compression that merges parent-child nodes into single tokens, achieving up to 8x reduction in sequence length while maintaining lossless representation.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mem4Nav: Boosting Vision-and-Language Navigation in Urban Environments with a Hierarchical Spatial-Cognition Long-Short Memory System

URL: View paper

Brief Assessment

Mem4Nav[65] uses octree structures for spatial indexing in navigation tasks, not for tokenization or compression of 3D structures into 1D sequences for autoregressive modeling as in the original paper.

---

2. GAEM: Graph-driven Attention-based Entropy Model for LiDAR Point Cloud Compression

URL: View paper

Brief Assessment

GAEM[64] focuses on entropy coding for LiDAR point cloud compression using graph attention mechanisms on octree structures, not on tokenization for autoregressive generation. The candidate addresses compression efficiency through entropy models rather than developing a hierarchical tokenization method for converting 3D structures into discrete token sequences for generative modeling.

---

3. Octattention: Octree-based large-scale contexts model for point cloud compression

URL: View paper

Brief Assessment

OctAttention[60] focuses on point cloud compression using octree structures for entropy modeling and arithmetic coding, not on tokenization for autoregressive 3D generation tasks.

---

4. Learning-based Lossless Event Data Compression

URL: View paper

Brief Assessment

Lossless Event Compression[68] applies octree partitioning to event camera data compression, not to 3D structure tokenization for autoregressive generation. The technical domains and objectives differ fundamentally.

---

5. TopNet: Transformer-Efficient Occupancy Prediction Network for Octree-Structured Point Cloud Geometry Compression

URL: View paper

Brief Assessment

TopNet[67] focuses on point cloud geometry compression using octree structures for efficient encoding, not on general 3D structure generation or autoregressive modeling across diverse domains as in the original paper.

---

6. Uni-3dar: Unified 3d generation and understanding via autoregression on compressed spatial tokens

URL: View paper

Brief Assessment

Uni-3DAR[63] uses octree tokenization with 2-level subtree compression for 3D structures, but focuses on unified cross-scale generation and understanding tasks. The specific two-level subtree compression strategy (merging parent and 8 children into a single 8-bit token) appears to be a distinct implementation detail not directly comparable to the original paper's approach.

---

7. VoxelContext-Net: An Octree based Framework for Point Cloud Compression

URL: View paper

Brief Assessment

VoxelContext-Net[66] focuses on point cloud compression using octree structures with voxel context for entropy coding, not on tokenization for autoregressive 3D generation across diverse molecular and macroscopic structures.

---

8. OctGPT: Octree-based Multiscale Autoregressive Models for 3D Shape Generation

URL: View paper

Prior Art Analysis

OctGPT[47] demonstrates that octree-based tokenization with hierarchical compression for 3D structures was previously proposed. Both papers use octree structures to compress 3D shapes into sequences, employ coarse-to-fine hierarchical representations, and use binary tokens for fine-grained details. OctGPT[47] explicitly describes using 'serialized octree representation to efficiently capture the hierarchical and spatial structures of 3d shapes' where 'coarse geometry is encoded via octree structures, while fine-grained details are represented by binary tokens generated using a vector quantized variational autoencoder (vqvae)'. This directly parallels the original paper's claimed contribution of octree-based tokenization with hierarchical compression.

Evidence

Evidence 1 - **Rationale**: Both papers claim octree-based compression as a core innovation. OctGPT[47] describes its 'serialized octree representation' for 'compact multiscale binary sequences', which directly corresponds to the original paper's 'coarse-to-fine tokenizer based on octree data structures' that 'compresses diverse 3d structures into compact 1d token sequences'. - **Original**: at its core is a coarse-to-fine tokenizer based on octree data structures, which compresses diverse 3d structures into compact 1d token sequences. we further propose a two-level subtree compression strategy, which reduces the octree token sequence by up to 8x. - **Candidate**: octgpt, a novel multiscale autoregressive model for 3d shape generation that dramatically improves the efficiency and performance of prior 3d autoregressive approaches, while rivaling or surpassing state-of-the-art diffusion models. our method employs a serialized octree representation to efficientl...

---

9. Octsqueeze: Octree-structured entropy model for lidar compression

URL: View paper

Brief Assessment

OctSqueeze[61] focuses on LiDAR compression using octrees for entropy modeling, not on tokenization for autoregressive 3D generation. The candidate does not address hierarchical tokenization or sequence length reduction for generative modeling.

---

10. OG-Mapping: Octree-based Structured 3D Gaussians for Online Dense Mapping

URL: View paper

Brief Assessment

OG-Mapping[62] uses octrees for 3D Gaussian representations in dense mapping, not for tokenization or sequence compression. The candidate focuses on online RGB-D mapping with spatial scene representation, while the original paper develops hierarchical tokenization for autoregressive 3D generation across multiple scales.

Contribution 3: Masked next-token prediction strategy for dynamic token positions

Description: The authors propose a novel training strategy that duplicates tokens with masked placeholders to handle the challenge of unpredictable token positions in sparse 3D structures. This method enables the model to predict token content while being conditioned on correct positional information, maintaining causal attention flow without complex sampling schemes.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Dynamic Token Masking in Spiking Neural Network

URL: View paper

Brief Assessment

Dynamic Token Masking[76] addresses token masking in Spiking Neural Networks for energy efficiency in vision tasks, not autoregressive sequence generation with dynamic spatial positions in 3D structures.

---

2. Medical Referring Image Segmentation via Next-Token Mask Prediction

URL: View paper

Brief Assessment

NextToken Mask Prediction[77] focuses on medical image segmentation with next-k token prediction and contrastive learning strategies, not on handling dynamic token positions in sparse 3D structures through masked placeholders as in the original paper.

---

3. Mask-predict: Parallel decoding of conditional masked language models

URL: View paper

Brief Assessment

Mask-Predict[74] addresses masked prediction in machine translation with bi-directional attention for parallel decoding, not the specific challenge of dynamic token positions in sparse 3D octree structures that the original paper tackles.

---

4. Uni-3dar: Unified 3d generation and understanding via autoregression on compressed spatial tokens

URL: View paper

Brief Assessment

Uni-3DAR[63] proposes masked next-token prediction to handle dynamic token positions in octree sequences. However, this is applied specifically to their octree-based 3D structure tokenization framework, which differs from the original paper's context and application domain.

---

5. Csi-LLM: A Novel Downlink Channel Prediction Method Aligned with LLM Pre-Training

URL: View paper

Brief Assessment

CSI-LLM[72] focuses on channel prediction in wireless communication using LLM alignment strategies. It does not address the challenge of dynamic token positions in sparse 3D structures or propose masked next-token prediction for handling unpredictable spatial token positions in octree-based representations.

---

6. Dyset: A dynamic masked self-distillation approach for robust trajectory prediction

URL: View paper

Brief Assessment

Dyset[73] focuses on masked trajectory prediction for autonomous driving with dynamic token sampling based on informativeness, not on handling unpredictable token positions in sparse 3D structures through masked placeholders as in the original paper.

---

7. Denoising token prediction in masked autoregressive models

URL: View paper

Brief Assessment

Denoising Token Prediction[69] focuses on masked autoregressive models for image generation with denoising heads, not on handling dynamic token positions in sparse 3D structures as in the original paper's octree-based approach.

---

8. Context-aware Rotary Position Embedding

URL: View paper

Brief Assessment

Rotary Position Embedding[70] focuses on context-aware positional encodings for transformers, not on handling dynamic token positions in sparse 3D structures through masked prediction strategies.

---

9. FlexTok: Resampling Images into 1D Token Sequences of Flexible Length

URL: View paper

Brief Assessment

FlexTok[71] addresses variable-length token sequences in image generation using nested dropout and rectified flow decoders, not masked next-token prediction. The candidate's approach to handling dynamic sequences differs fundamentally from the original paper's masked prediction strategy for 3D spatial structures.

---

10. Mst: Masked self-supervised transformer for visual representation

URL: View paper

Brief Assessment

MST[75] focuses on masked token prediction for vision transformers in self-supervised learning, not on handling dynamic token positions in sparse 3D autoregressive generation. The masking strategies serve fundamentally different purposes in different domains.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Uni-3dar: Unified 3d generation and understanding via autoregression on compressed spatial tokens

Detected in: Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Unified Cross-Scale 3D Generation and Understanding via Autoregressive Modeling View paper
- [1] SAR3D: Autoregressive 3D object generation and understanding via multi-scale 3D VQVAE View paper
- [2] SP-GAN: Sphere-guided 3D shape generation and manipulation View paper
- [3] Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies View paper
- [4] CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets View paper
- [5] Drivedreamer: Towards real-world-drive world models for autonomous driving View paper
- [6] HierOctFusion: Multi-scale Octree-based 3D Shape Generation via Part-Whole-Hierarchy Message Passing View paper
- [7] CIDNet: Cross-Scale Interference Mining Detection Network for underwater object detection View paper
- [8] PartCrafter: Structured 3D Mesh Generation via Compositional Latent Diffusion Transformers View paper
- [9] Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion View paper
- [10] Global cross-scale simulation and experiment of supercritical CO2 boiler tube wall temperature based on bidirectional fluid-thermal coupling View paper
- [11] Multi-modal and multi-scale spatial environment understanding for immersive visual text-to-speech View paper
- [12] PSHuman: Photorealistic Single-image 3D Human Reconstruction using Cross-Scale Multiview Diffusion and Explicit Remeshing View paper
- [13] Melt electrowriting enabled 3D liquid crystal elastomer structures for cross-scale actuators and temperature field sensors View paper

- [14] SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation With Fine-Grained Geometry View paper
- [15] A cross-scale numerical framework of PEM water electrolyzer with pore-scale transport properties of PTL View paper
- [16] Brep2Seq: a dataset and hierarchical deep learning network for reconstruction and generation of computer-aided design models View paper
- [17] Setvae: Learning hierarchical composition for generative modeling of set-structured data View paper
- [18] 3D-WAG: Hierarchical Wavelet-Guided Autoregressive Generation for High-Fidelity 3D Shapes View paper
- [19] LION: Latent Point Diffusion Models for 3D Shape Generation View paper
- [20] HiScene: creating hierarchical 3D scenes with isometric view generation View paper
- [21] FreeSplat: Generalizable 3D Gaussian Splatting Towards Free-View Synthesis of Indoor Scenes View paper
- [22] Unraveling Crossâ□□Scale Fluorination Mechanisms in Nonâ□□Fullerene Acceptors for Highâ□□Efficiency Organic Photovoltaics View paper
- [23] Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack View paper
- [24] Generation of a Highâ□□Precision Whole Liver Panorama and Crossâ□□Scale 3D Pathological Analysis for Hepatic Fibrosis View paper
- [25] Multi-scale reconstruction of porous media from low-resolution core images using conditional generative adversarial networks View paper
- [26] A supervised learning-assisted multi-scale study for thermal and mechanical behavior of porous Silica View paper
- [27] Variational Multi-scale Representation for Estimating Uncertainty in 3D Gaussian Splatting View paper
- [28] HyperSDFusion: Bridging Hierarchical Structures in Language and Geometry for Enhanced 3D Text2Shape Generation View paper
- [29] Multi-scale Geometry-aware Transformer for 3D Point Cloud Classification View paper
- [30] 3D Printing of Hierarchical Structures Made of Inorganic Silicon-Rich Glass Featuring Self-Forming Nanogratings View paper
- [31] Interactive multi-scale fusion of 2D and 3D features for multi-object vehicle tracking View paper
- [32] Observed 3D Structure, Generation, and Dissipation of Oceanic Mesoscale Eddies in the South China Sea View paper
- [33] Cross-scale tracing of nanoparticles and tumors at the single-cell level using the whole-lung atlas View paper
- [34] Advancing fine-grained visual understanding with multi-scale alignment in multi-modal models View paper
- [35] ResGEM: Multi-Scale Graph Embedding Network for Residual Mesh Denoising View paper
- [36] Multi-scale Geometry-aware Self-Attention for 3D Point Cloud Classification View paper
- [37] MuSeLLM: SDF Generation and Understanding via Multi-Scale Tokenization with Position-Aware Guidance View paper
- [38] 3D-Printed functionally graded PCL-HA scaffolds with Multi-Scale porosity View paper
- [39] 3D-TRANS: 3D Hierarchical Transformer for Shape Correspondence Learning View paper
- [40] Dimensional engineering of hierarchical nanopagodas for customizing crossâ□□scale magnetic coupling networks to enhance electromagnetic wave absorption View paper
- [41] Low-temperature-field-assisted fabrication of cross-scale tissue engineering scaffolds View paper
- [42] Language embedded radiance fields for zero-shot task-oriented grasping View paper
- [43] Octree Transformer: Autoregressive 3D Shape Generation on Hierarchically Structured Sequences View paper
- [44] 3CROSSNet: Cross-level cross-scale cross-attention network for point cloud representation View paper
- [45] New view synthesis via multiscale-depth and transformers View paper
- [46] Flexible 3D Lane Detection by Hierarchical Shape Matching View paper
- [47] OctGPT: Octree-based Multiscale Autoregressive Models for 3D Shape Generation View paper
- [48] Trinerflet: A wavelet based triplane nerf representation View paper
- [49] ProtComposer: Compositional Protein Structure Generation with 3D Ellipsoids View paper
- [50] Aluminium surface work hardening enables multi-scale 3D lithography View paper
- [51] Pushing auto-regressive models for 3d shape generation at capacity and scalability View paper
- [52] Autopartgen: Autogressive 3d part generation and discovery View paper
- [53] Polygen: An autoregressive generative model of 3d meshes View paper
- [54] Bamm: Bidirectional autoregressive motion model View paper
- [55] Autosdf: Shape priors for 3d completion, reconstruction and generation View paper
- [56] Autoregressive models in vision: A survey View paper
- [57] HERMES: A Unified Self-Driving World Model for Simultaneous 3D Scene Understanding and Generation View paper
- [58] Show-o2: Improved Native Unified Multimodal Models View paper
- [59] Autoregressive 3d shape generation via canonical mapping View paper
- [60] Octattention: Octree-based large-scale contexts model for point cloud compression View paper
- [61] Octsqueeze: Octree-structured entropy model for lidar compression View paper
- [62] OG-Mapping: Octree-based Structured 3D Gaussians for Online Dense Mapping View paper
- [63] Uni-3dar: Unified 3d generation and understanding via autoregression on compressed spatial tokens View paper
- [64] GAEM: Graph-driven Attention-based Entropy Model for LiDAR Point Cloud Compression View paper
- [65] Mem4Nav: Boosting Vision-and-Language Navigation in Urban Environments with a Hierarchical Spatial-Cognition Long-Short Memory System View paper
- [66] VoxelContext-Net: An Octree based Framework for Point Cloud Compression View paper
- [67] TopNet: Transformer-Efficient Occupancy Prediction Network for Octree-Structured Point Cloud Geometry Compression View paper
- [68] Learning-based Lossless Event Data Compression View paper
- [69] Denoising token prediction in masked autoregressive models View paper
- [70] Context-aware Rotary Position Embedding View paper
- [71] FlexTok: Resampling Images into 1D Token Sequences of Flexible Length View paper
- [72] Csi-LLM: A Novel Downlink Channel Prediction Method Aligned with LLM Pre-Training View paper
- [73] Dyset: A dynamic masked self-distillation approach for robust trajectory prediction View paper
- [74] Mask-predict: Parallel decoding of conditional masked language models View paper
- [75] Mst: Masked self-supervised transformer for visual representation View paper
- [76] Dynamic Token Masking in Spiking Neural Network View paper
- [77] Medical Referring Image Segmentation via Next-Token Mask Prediction View paper