

Compositional Generative Model of Unbounded 4D Cities

Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu

Abstract—3D scene generation has garnered growing attention in recent years and has made significant progress. Generating 4D cities is more challenging than 3D scenes due to the presence of structurally complex, visually diverse objects like buildings and vehicles, and heightened human sensitivity to distortions in urban environments. To tackle these issues, we propose CityDreamer4D, a compositional generative model specifically tailored for generating unbounded 4D cities. Our main insights are 1) 4D city generation should separate dynamic objects (*e.g.*, vehicles) from static scenes (*e.g.*, buildings and roads), and 2) all objects in the 4D scene should be composed of different types of neural fields for buildings, vehicles, and background stuff. Specifically, we propose Traffic Scenario Generator and Unbounded Layout Generator to produce dynamic traffic scenarios and static city layouts using a highly compact BEV representation. Objects in 4D cities are generated by combining stuff-oriented and instance-oriented neural fields for background stuff, buildings, and vehicles. To suit the distinct characteristics of background stuff and instances, the neural fields employ customized generative hash grids and periodic positional embeddings as scene parameterizations. Furthermore, we offer a comprehensive suite of datasets for city generation, including OSM, GoogleEarth, and CityTopia. The OSM dataset provides a variety of real-world city layouts, while the Google Earth and CityTopia datasets deliver large-scale, high-quality city imagery complete with 3D instance annotations. Leveraging its compositional design, CityDreamer4D supports a range of downstream applications, such as instance editing, city stylization, and urban simulation, while delivering state-of-the-art performance in generating realistic 4D cities.

Index Terms—City Generation, 4D Generation, Generative Models, NeRF

I. INTRODUCTION

AMID the rise of the metaverse, 3D and 4D asset generation has garnered significant attention. Notable progress has been made in generating 3D objects [1]–[3], avatars [4]–[6], and scenes [7]–[9], as well as 4D objects [10], [11] and avatars [12]–[14]. Cities, as one of the most essential assets, are widely used in diverse applications such as urban planning, environmental simulations, and game asset development. Therefore, the challenge of making 3D/4D city development accessible to a wider audience, including artists, researchers, and players, becomes both significant and impactful.

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221-0012, MOE-T2EP20223-0002), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). (Corresponding author: Ziwei Liu.)

The authors are with S-Lab, Nanyang Technological University, Singapore 637335 (email: haozhe.xie@ntu.edu.sg; zhaoxi001@ntu.edu.sg; fangzhou.hong@ntu.edu.sg; ziwei.liu@ntu.edu.sg)

Project page is available at <https://haozhexie.com/project/city-dreamer-4d>

In recent years, notable advancements have been made in scene generation. Video-based methods [15]–[17] generate 3D scenes by producing videos conditioned on input images, but they cannot guarantee temporal consistency. Outpainting-based methods [18]–[20] generate 3D scenes through continuous outpainting on RGB and depth images, but they lack a compact scene representation, resulting in scenes that are typically small in scale. PCG-based methods [21]–[23] create unbounded cities by integrating large language models (LLMs) with procedural content generation (PCG), but the diversity of the generated cities is constrained by the 3D assets employed. 3D-aware-GAN-based methods, represented by GANCraft [24] and SceneDreamer [7], use volumetric neural rendering to generate images within a 3D scene, leveraging 3D coordinates and corresponding semantic labels. These methods show promising results in generating 3D natural scenes by leveraging pseudo-ground-truth images generated by SPADE [25]. InfiniCity [26] follows a similar pipeline for 3D city generation but it is more complex than 3D natural scenes due to the greater appearance variation in buildings and vehicles, unlike the relatively consistent appearance of objects with the same semantic label in natural scenes. This variation reduces the quality of generated buildings and vehicles when all instances within their respective classes are assigned the same semantic label. Generating 4D scenes poses greater challenges than 3D scenes, as existing methods [27]–[30] either fail to ensure temporal consistency or are confined to tiny scales.

To address these problems, we propose CityDreamer4D, a compositional generative model designed for unbounded 4D cities. As shown in Fig. 1, the unbounded 4D city generation framework separates dynamic objects from static scenes. Static scenes are defined by the city layout from Unbounded Layout Generator, arranging elements like roads, highways, vegetation, and buildings, with the capability to extrapolate to unbounded areas. Dynamic objects, such as vehicles, are defined by traffic scenarios generated by Traffic Scenario Generator, which determines their spatial positioning on high-definition (HD) maps derived from city layouts. Unlike existing methods that use a single module for all objects, CityDreamer4D divides the generation process into three distinct modules: Building Instance Generator for buildings, Vehicle Instance Generator for vehicles, and City Background Generator for background stuff. These generators leverage a highly compact bird’s-eye-view (BEV) scene representation to ensure efficiency and scalability. The scene parameterization is designed to address the unique characteristics of background stuff and instances: background stuff often features similar

appearances with irregular textures, while buildings and vehicles display diverse appearances with regular periodic patterns. To handle these variations, we use generative hash grids for the background and apply periodic positional encodings to each instance. We also place buildings in an object-centric coordinate space and vehicles in an object-canonical coordinate space, using specialized methods designed to capture their compact shapes. Compositor combines the rendered background stuff with the building and vehicle instances to create a unified image.

To improve the realism of our generated cities, we construct a suite of datasets, including OSM, GoogleEarth, and CityTopia. The OSM dataset, sourced from OpenStreetMap [31], includes semantic maps and height fields for 80 cities worldwide, covering over 6,000 km². The semantic maps indicate the locations of roads, buildings, urban greenery, and water bodies, while the height fields primarily represent building heights. The GoogleEarth dataset is a real-world dataset collected using Google Earth Studio [32], featuring 400 drone-view orbit trajectories over New York City. It includes 24,000 real-world city images, with 3D semantic annotations for all classes and 3D instance annotations for buildings. The CityTopia dataset is a high-quality synthetic dataset spanning 11 cities generated with 3D assets from the Unreal Engine 5 City Sample project [33]. It offers 37,500 high-fidelity street-view and drone-view images, featuring precise 2D and 3D semantic annotations for all classes, along with 3D instance annotations for buildings and vehicles.

The contributions are summarized as follows:

- We propose CityDreamer4D, the first generative model for unbounded 4D cities that disentangles dynamic objects from static scenes and enables instance editing, city stylization, and urban simulation.
- We introduce stuff-oriented and instance-oriented neural fields to generate background stuff and instances (buildings and vehicles) in 4D scenes, effectively capturing their diversity.
- We create comprehensive datasets for city generation, using OSM for realistic layouts and Google Earth and CityTopia for detailed city visuals with 3D semantic and instance annotations.
- The proposed CityDreamer4D demonstrates superior capability in generating unbounded, diverse 4D cities and enables instance-level editing within them.

A preliminary version of this work, named CityDreamer, has been published in CVPR 2024 [34]. We make several extensions in this work compared to the preliminary version. **1)** We evolve CityDreamer into CityDreamer4D, enabling 4D city generation through Traffic Scenario Generator and Vehicle Instance Generator, effectively separating dynamic objects from static scenes. **2)** We enhance the highly compact BEV representation by incorporating an additional bottom-up height map, enabling the representation of hollow structures in cities, such as highways. **3)** We propose Traffic Scenario Generator, which creates HD maps from city layouts to produce realistic traffic scenarios with vehicles in unbounded cities. **4)** We introduce Vehicle Instance Generator, designed to generate vehicle instances within cities using a novel scene

parameterization method grounded on the canonical feature space. **5)** We build the CityTopia dataset, offering nearly 40k high-quality street-view and drone-view images with both 2D and 3D semantic and instance annotations.

II. RELATED WORKS

A. 3D-aware GANs

Building on the recent success of 2D GANs [35], [36], various approaches have been introduced to generate 3D content using GANs as well. The core idea is to represent the generated scenes using a 3D representation and apply rendering techniques to produce images from various viewpoints, enabling image-level adversarial learning [37]. Early methods use explicit shapes like voxels [38]–[40], meshes [41], and 3D primitives [42] to render images from different viewpoints. However, their limited expressiveness and efficiency hinder the synthesis of complex scenes and photorealistic details. NeRF [43], known for producing high-fidelity novel view synthesis, are introduced to 3D-aware generative models. Yet, the high computational cost of NeRF-based GANs restricts earlier attempts [44]–[47] from generating high-quality images. To address this, many follow-up works [48]–[52] avoid rendering NeRFs at high resolution by applying 2D super-resolution on low-resolution feature maps, though this compromises 3D consistency. To maintain strict 3D consistency, newer approaches shift to sparser 3D representations, like sparse voxels [53], radiance manifolds [54], and multi-plane images [55], enabling direct high-resolution rendering. Nevertheless, most of these methods are trained on curated datasets for bounded scenes, such as human faces [56], [57], human bodies [58], [59], and objects [60], [61].

B. 3D Scene Generation

Unlike advanced 2D generative models that mainly focus on individual categories or familiar objects, generating scene-level content is more challenging due to the vast diversity and complexity of scenes [62]. Earlier methods [15], [16] generate scenes by synthesizing videos, but they lack 3D awareness and fail to ensure 3D consistency. Semantic image synthesis approaches [24], [63] have shown promising results in generating scene-level content by conditioning on pixel-wise dense correspondences, like semantic segmentation or depth maps. Several techniques [18]–[20] generate 3D scenes by performing inpainting and outpainting on RGB images or feature maps, though most can only interpolate or extrapolate a limited distance from the input views and lack true generative capabilities. Significant research has investigated procedural content generation (PCG) for creating natural [21], [64], indoor [65]–[67], and urban scenes [23], [68]–[70], but the diversity of the generated scenes is limited by the 3D assets used. Recent methods [7], [9], [26] achieve 3D-consistent scenes at an infinite scale through unbounded layout extrapolation. Other approaches [71]–[73] focus on indoor scene synthesis, relying on costly 3D datasets [74], [75] or CAD object retrieval [76]–[78].

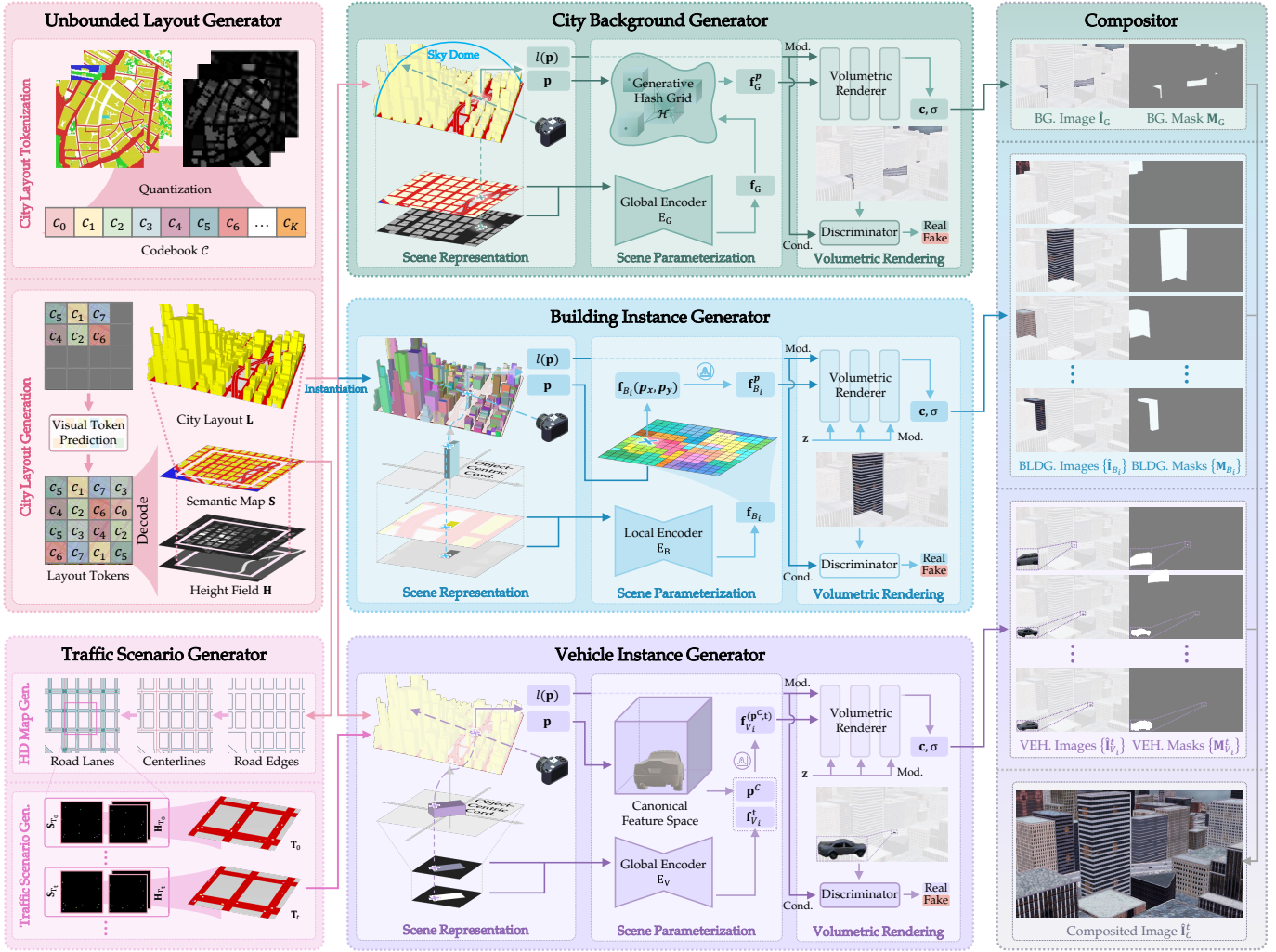


Fig. 1. **Overview of CityDreamer4D.** 4D city generation comprises static and dynamic scenes, conditioned on city layout \mathbf{L} and time-varying traffic scenario \mathbf{T}_t , generated by the Unbounded Layout and Traffic Scenario Generators, respectively. City Background Generator uses \mathbf{L} to create background images $\hat{\mathbf{I}}_G$ for stuff like roads, vegetation, and the sky, while Building Instance Generator renders the buildings $\{\hat{\mathbf{I}}_{B_i}\}$ within the city. Using \mathbf{T}_t , Vehicle Instance Generator generates vehicles $\{\hat{\mathbf{I}}_{V_i}^t\}$ at time step t . Finally, Compositor combines the rendered background, buildings, and vehicles into a unified and coherent image $\hat{\mathbf{I}}_C^t$. “Gen.”, “Mod.”, “Cond.”, “BG.”, “BLDG.”, and “VEH.” denote “Generation”, “Modulation”, “Condition”, “Background”, “Building”, and “Vehicle”, respectively.

C. 4D Scene Generation

In recent years, representations like D-NeRF [79] and Deformable 3D Gaussians [80] have been proposed for 4D object and human generation. However, 4D scene generation remains in its early stages, as existing representations are not designed for large-scale scene generation. Mainstream approaches typically formulate it as 4D occupancy generation [81], [82] and distillation from video diffusion [27]–[30]. However, these methods lack compact representations, restricting the scale of the generated scenes.

III. METHOD

As illustrated in Figure 1, CityDreamer4D decouples unbounded 4D city generation into static scene generation and dynamic object generation. For static scene generation, Unbounded Layout Generator (Section III-A) creates an arbitrarily large city layout \mathbf{L} . City Background Generator (Section III-C) then produces the background image $\hat{\mathbf{I}}_G$ along with

its corresponding mask \mathbf{M}_G . Following this, Building Instance Generator (Section III-D) generates images for building instances $\{\hat{\mathbf{I}}_{B_i}\}_{i=1}^{n_B}$ and their respective masks $\{\mathbf{M}_{B_i}\}_{i=1}^{n_B}$, where n_B is the number of building instances. For dynamic object generation, the traffic generator (Section III-B) first creates the traffic scenario \mathbf{T}_t for time step t . Then, Vehicle Instance Generator (Section III-E) produces images for vehicle instances $\{\hat{\mathbf{I}}_{V_i}^t\}_{i=1}^{n_V}$ and their corresponding masks $\{\mathbf{M}_{V_i}^t\}_{i=1}^{n_V}$ based on \mathbf{T}_t , where n_V denotes the number of vehicles. Finally, Compositor (Section III-F) merges the rendered background, building instances, and vehicle instances into a cohesive image $\hat{\mathbf{I}}_C^t$ for time step t .

A. Unbounded Layout Generator

City Layout Representation. The city layout defines the locations of static 3D objects within the city, which are grouped into categories such as roads, highways, buildings, vegetation, water areas, and others. Additionally, a null class

is included to represent empty spaces in the 3D volume. The city layout in CityDreamer4D, represented as a 3D volume \mathbf{L} , is constructed by extruding pixels from the semantic map \mathbf{S}_L according to their corresponding values in the height field $\mathbf{H}_L = \{\mathbf{H}_L^{\text{BU}}, \mathbf{H}_L^{\text{TD}}\}$, where \mathbf{H}_L^{BU} and \mathbf{H}_L^{TD} represent the bottom-up heights and the top-down heights, respectively. Specifically, the value of \mathbf{L} at (i, j, k) is defined as

$$\mathbf{L}(i, j, k) = \begin{cases} \mathbf{S}_L(i, j) & \text{if } \mathbf{H}_L^{\text{BU}}(i, j) \leq k \leq \mathbf{H}_L^{\text{TD}}(i, j) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where 0 denotes empty spaces in the 3D volumes.

City Layout Generation. Obtaining unbounded city layouts is translated into generating extendable semantic maps and height fields. To achieve this, we design Unbounded Layout Generator based on MaskGIT [83], which naturally supports inpainting and extrapolation. Specifically, we leverage VQVAE [84] to tokenize patches of semantic maps and height fields, encoding them into a discrete latent space with a codebook $\mathcal{C} = \{c_k \mid c_k \in \mathbb{R}^{d_C}\}_{k=1}^{d_K}$. During inference, the layout tokens are generated autoregressively, and the VQVAE decoder reconstructs a pair of semantic map \mathbf{S}_L and height field \mathbf{H}_L . Since VQVAE produces fixed-size outputs, we perform image extrapolation to create arbitrarily large layouts. This involves using a sliding window with a 25% overlap to iteratively predict local layout tokens at each step.

Loss Functions. The VQVAE handles the generation of the height field and semantic map as separate tasks, optimized with L1 Loss and Cross-Entropy Loss \mathcal{E} , respectively. To enhance the sharpness of the height field near building edges, we incorporate an additional Smoothness Loss \mathcal{S} [85]

$$\ell_{\text{VQ}} = \lambda_{\text{R}} \|\hat{\mathbf{H}}_L^p - \mathbf{H}_L^p\| + \lambda_{\text{S}} \mathcal{S}(\hat{\mathbf{H}}_L^p, \mathbf{H}_L^p) + \lambda_{\text{E}} \mathcal{E}(\hat{\mathbf{S}}_L^p, \mathbf{S}_L^p) \quad (2)$$

where $\hat{\mathbf{H}}_L^p$ and $\hat{\mathbf{S}}_L^p$ denote the generated height field and semantic map patches, respectively. \mathbf{H}_L^p and \mathbf{S}_L^p are the corresponding ground truth. MaskGIT's autoregressive transformer is optimized with a reweighted ELBO loss [86].

B. Traffic Scenario Generator

Traffic Scenario Representation. The city layout \mathbf{L} defines the static elements of the unbounded city, while the dynamic aspects are captured by the traffic scenario, represented as $\mathcal{T} = \{\mathbf{T}_t\}_{t=1}^{n_T}$, where n_T represents the number of frames. Similar to the city layout \mathbf{L} , \mathbf{T}_t is likewise derived from the semantic map \mathbf{S}_{T_t} and the height field $\mathbf{H}_{T_t} = \{\mathbf{H}_{T_t}^{\text{BU}}, \mathbf{H}_{T_t}^{\text{TD}}\}$, where the semantic map specifies the positions of dynamic objects, and the height field defines their elevations. Specifically, the value of \mathbf{T}_t at (i, j, k) is

$$\mathbf{T}_t(i, j, k) = \begin{cases} \mathbf{S}_{T_t}(i, j) & \text{if } \mathbf{H}_{T_t}^{\text{BU}}(i, j) \leq k \leq \mathbf{H}_{T_t}^{\text{TD}}(i, j) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where 0 denotes empty spaces in the 3D volumes.

Traffic Scenario Generation. The generation of traffic scenario \mathcal{T} is conceptualized as the frame-by-frame production of semantic maps $\mathbf{S}_T = \{\mathbf{S}_{T_t}\}_{t=1}^{n_T}$ and height fields $\mathbf{H}_T = \{\mathbf{H}_{T_t}\}_{t=1}^{n_T}$. To guarantee realistic and continuous placement of dynamic objects, a high-definition (HD) map is derived from

the city layout \mathbf{L} . Unlike the city layout, which only specifies the positions of roads and highways, the HD map includes details about lanes, intersections, and traffic signals. Using the generated HD map, an off-the-shelf model [87] determines the per-frame bounding boxes of dynamic objects. The corresponding semantic map and height field are generated based on the bounding boxes.

HD Map Generation. In HD maps, we adopt the entity definitions from the Waymo Motion dataset [88], which include road edges, road lanes, road lines, stop signs, and traffic lights. *Road Edges*, representing the boundaries of roads, are generated by applying Canny edge detection [89] to \mathbf{S}_L and converting the continuous edges into a graph structure using vectorization, which involves detecting corner points and connecting them sequentially.

Road Lanes, representing the centerlines of lanes where vehicles can travel, are derived by skeletonizing [90] \mathbf{S}_L to extract road structures and identifying intersections where multiple edges connect. The image is then converted into road centerline graphs using graph-based traversal. The number and positions of the lanes are determined based on road width, and lanes at intersections are connected using Bézier curves.

Road Lines, such as solid single white or solid double yellow, are generated according to the positions and attributes of the road lanes.

Stop Signs and *Traffic Lights* are positioned at the intersections, where multiple road lanes converge.

C. City Background Generator

Scene Representation. Following SceneDreamer [7], we adopt a bird's-eye-view (BEV) representation for its efficiency and expressiveness, particularly suited for unbounded scenes. Unlike GANCraft [24] and InfiniCity [26], which parameterize features at voxel corners, our BEV representation uses a feature-free 3D volume constructed from a height field and a semantic map, as described in Equation 1. Specifically, we extract a local window of resolution $N_G^H \times N_G^W \times N_G^D$ from the city layout \mathbf{L} . This local window \mathbf{L}^G is generated using the corresponding height field \mathbf{H}_L^G and semantic map \mathbf{S}_L^G .

Scene Parameterization. To achieve generalizable 3D representation learning across various scenes and align content with 3D semantics, it is necessary to parameterize the scene representation into a latent space, making adversarial learning easier. For background stuff, we adopt the generative neural hash grid [7] to learn generalizable features across scenes by modeling the hyperspace beyond 3D space. Specifically, we first encode the local scene $(\mathbf{H}_L^G, \mathbf{S}_L^G)$ using the global encoder E_G to produce the compact scene-level feature $\mathbf{f}_G \in \mathbb{R}^{d_G}$.

$$\mathbf{f}_G = E_G(\mathbf{H}_L^G, \mathbf{S}_L^G) \quad (4)$$

Using a learnable neural hash function \mathcal{H} , the indexed feature $\mathbf{f}_G^{\mathbf{p}}$ at the 3D position $\mathbf{p} \in \mathbb{R}^3$ is derived by mapping \mathbf{p} and \mathbf{f}_G into a hyperspace, specifically $\mathbb{R}^{3+d_G} \rightarrow \mathbb{R}^{N_G^C}$.

$$\mathbf{f}_G^{\mathbf{p}} = \mathcal{H}(\mathbf{p}, \mathbf{f}_G) = \left(\bigoplus_{i=1}^{d_G} f_G^i \pi^i \bigoplus_{j=1}^3 p^j \pi^j \right) \bmod N_E \quad (5)$$

where \oplus represents the bitwise XOR operation, while π^i and π^j are distinct large prime numbers. To capture multi-scale features, we construct N_H^L levels of multi-resolution hash grids. Each level contains up to N_E entries, with N_G^C denoting the number of channels in each feature vector.

Volumetric Rendering. In the perspective camera model, every pixel in the image is associated with a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v}$, which originates at the projection center \mathbf{o} and extends along the direction \mathbf{v} . The pixel value $C(\mathbf{r})$ is then computed as an integral along this ray.

$$C(\mathbf{r}) = \int_0^\infty A(t) \mathbf{c}(\mathbf{f}_G^{\mathbf{r}(t)}, l(\mathbf{r}(t))) \boldsymbol{\sigma}(\mathbf{f}_G^{\mathbf{r}(t)}) dt \quad (6)$$

where $A(t) = \exp\left(-\int_0^t \sigma(\mathbf{f}_G^{\mathbf{r}(s)}), ds\right)$ represents the accumulated transmittance. $l(\mathbf{p})$ denotes the semantic label at the 3D position \mathbf{p} . The symbols \mathbf{c} and $\boldsymbol{\sigma}$ correspond to the color and volume density, respectively.

Loss Function. City Background Generator is optimized with a hybrid objective that combines reconstruction loss and adversarial loss. In particular, it uses an L1 loss, a perceptual loss \mathcal{P} [91], and a GAN loss \mathcal{G} [92] as part of this objective.

$$\ell_G = \lambda_G^{L1} \|\hat{\mathbf{I}}_G - \mathbf{I}_G\| + \lambda_G^{\mathcal{P}} \mathcal{P}(\hat{\mathbf{I}}_G, \mathbf{I}_G) + \lambda_G^{\mathcal{G}} \mathcal{G}(\hat{\mathbf{I}}_G, \mathbf{S}_G) \quad (7)$$

where \mathbf{I}_G represents the ground truth background image, while \mathbf{S}_G corresponds to the perspective-view semantic map obtained by accumulating semantic labels sampled from \mathbf{L}^G along each ray. The weights for the three losses are denoted by λ_G^{L1} , $\lambda_G^{\mathcal{P}}$, and $\lambda_G^{\mathcal{G}}$. Note that ℓ_G is only applied to pixels whose semantic labels are classified as background stuff.

D. Building Instance Generator

Scene Representation. Building Instance Generator also employs the BEV scene representation. It extracts a local window \mathbf{L}^{B_i} from the city layout \mathbf{L} with dimensions $N_B^H \times N_B^W \times N_B^D$. This window is centered around the 2D coordinates $(c_x^{B_i}, c_y^{B_i})$ of the building instance B_i . The height field and semantic map used to construct \mathbf{L}^{B_i} are represented as $\mathbf{H}_L^{B_i}$ and $\mathbf{S}_L^{B_i}$, respectively. Since all buildings share the same semantic label in \mathbf{S}_L , we perform building instantiation by detecting connected components. Notably, real-world building facades and roofs exhibit distinct visual distributions. To capture this, we assign different semantic labels to the facade and roof of each building instance B_i in \mathbf{L}^{B_i} , with the roof assigned to the top-most voxel layer. All other building instances are excluded from \mathbf{L}^{B_i} by assigning them a value of 0.

Scene Parameterization. Unlike City Background Generator, Building Instance Generator employs a distinct scene parameterization, encoding the local scene $(\mathbf{H}_L^{B_i}, \mathbf{S}_L^{B_i})$ with \mathbf{E}_B to produce pixel-level features \mathbf{f}_{B_i} of resolution $N_B^H \times N_B^W \times N_B^C$.

$$\mathbf{f}_{B_i} = \mathbf{E}_B(\mathbf{H}_L^{B_i}, \mathbf{S}_L^{B_i}) \quad (8)$$

For a 3D position $\mathbf{p} = (p_x, p_y, p_z)$, the corresponding feature $\mathbf{f}_{B_i}^{\mathbf{p}}$ is obtained as

$$\mathbf{f}_{B_i}^{\mathbf{p}} = \mathcal{O}(\text{Concat}(\mathbf{f}_{B_i}(p_x, p_y, p_z))) \quad (9)$$

where $\text{Concat}(\cdot)$ denotes the concatenation operation. $\mathbf{f}_{B_i}(p_x, p_y) \in \mathbb{R}^{N_B^C}$ represents the feature vector corresponding to the coordinates (p_x, p_y) . $\mathcal{O}(\cdot)$ refers to the positional encoding function adopted in the standard NeRF [43].

$$\mathcal{O}(\mathbf{x}) = \{\sin(2^i \pi \mathbf{x}), \cos(2^i \pi \mathbf{x})\}_{i=0}^{N_P^L-1} \quad (10)$$

Note that $\mathcal{O}(\cdot)$ is applied separately to each element of the feature \mathbf{x} , with the values normalized to the range $[-1, 1]$.

Volumetric Rendering. Unlike the volumetric rendering approach used in City Background Generator, Building Instance Generator incorporates a style code \mathbf{z} to capture the variability in building appearances. The pixel value $C(\mathbf{r})$ is computed through an integration process.

$$C(\mathbf{r}) = \int_0^\infty A(t) \mathbf{c}(\mathbf{f}_{B_i}^{\mathbf{r}(t)}, \mathbf{z}, l(\mathbf{r}(t))) \boldsymbol{\sigma}(\mathbf{f}_{B_i}^{\mathbf{r}(t)}) dt \quad (11)$$

where $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v} - [c_x^{B_i}, c_y^{B_i}, 0]^T$, which is employed to center the buildings within their local coordinate system.

Loss Function. The training of Building Instance Generator relies solely on the GAN loss \mathcal{G} , formulated as

$$\ell_B = \mathcal{G}(\hat{\mathbf{I}}_{B_i}, \mathbf{S}_{B_i}) \quad (12)$$

where \mathbf{S}_{B_i} represents the semantic map of the building instance B_i in perspective view, generated similarly to \mathbf{S}_G . Note that ℓ_B is only applied to pixels with semantic labels corresponding to the building instance.

E. Vehical Instance Generator

Scene Representation. Vehicle Instance Generator, like Building Instance Generator, leverages the BEV scene representation. It extracts a local window $\mathbf{T}_t^{V_i}$ from the traffic scenario \mathbf{T}_t , with dimensions $N_V^H \times N_V^W \times N_V^D$, to generate the vehicle instances within the scene. This window is centered around the 2D coordinates $(c_x^{V_i}, c_y^{V_i})$ of the vehicle instance V_i . The height field and semantic map used to construct $\mathbf{T}_t^{V_i}$ are represented as $\mathbf{H}_{T_t}^{V_i}$ and $\mathbf{S}_{T_t}^{V_i}$, respectively. Unlike buildings, vehicle instances are instantiated during the generation of the traffic scenario. Instances other than V_i are removed from $\mathbf{T}_t^{V_i}$ by assigning them a value of 0.

Scene Parameterization. Compared to building instances, vehicle instances demonstrate greater structural regularity, closely tied to their relative positions. For instance, within the same vehicle, the front, rear, and body exhibit distinct appearances, yet these structural features remain consistent across different vehicles. Building on this observation, we propose a scene parameterization method based on the canonical feature space. Given a 3D position $\mathbf{p} = (p_x, p_y, p_z)$, the canonicalized point \mathbf{p}^C is

$$\mathbf{p}^C = \mathbf{R} \left(\mathbf{p} - [c_x^{V_i}, c_y^{V_i}, c_z^{V_i}]^T \right) \quad (13)$$

where $c_x^{V_i}$, $c_y^{V_i}$, $c_z^{V_i}$ represent the center coordinates of the vehicle V_i along the X, Y, and Z axes, respectively. \mathbf{R} is the rotation matrix used to normalize the 3D point into the canonical feature space.

$$\mathbf{R} = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta \cos \gamma & \cos \theta \cos \gamma & \sin \gamma \\ \sin \theta \sin \gamma & -\cos \theta \sin \gamma & \cos \gamma \end{bmatrix} \quad (14)$$

TABLE I
COMPARISON OF STATISTICS AND PROPERTIES: GOOGLE EARTH, CITYTOPIA, AND PREVIOUS DATASETS. ONLY ANNOTATED IMAGES ARE COUNTED. "EXT." STANDS FOR "EXTENDABLE", INDICATING WHETHER THE DATASET CAN BE EASILY EXPANDED FOLLOWING THE CURRENT DATA GENERATION PIPELINE. "3DM.", "SEM.", AND "INST." REFER TO "3D MODEL", "SEMANTIC", AND "INSTANCE", RESPECTIVELY.

Dataset	#Images ($\times 10^3$)	#Cities	Area (km^2)	Source	Ext.	3DM.	Lighting		View Type		Dense Annotations			
							Day	Night	Street	Aerial	2D Sem.	2D Inst.	3D Sem.	3D Inst.
KITTI [93]	0.2	1	-	Real	✗	✗	✓	✗	✓	✗	✓	✓	✗	✗
Cityscapes [94]	25	50	-	Real	✗	✗	✓	✗	✓	✗	✓	✓	✗	✗
AeroScapes [95]	3.2	-	-	Real	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗
nuScenes [96]	93	2	-	Real	✗	✗	✓	✗	✓	✗	✓	✓	✗	✗
GTA-V [97]	25	-	-	Synthetic	✗	✗	✓	✗	✓	✗	✓	✗	✗	✗
SYNTIA [98]	213	1	-	Synthetic	✗	✗	✓	✗	✓	✓	✓	✓	✗	✗
VEIS [99]	61	-	-	Synthetic	✗	✗	✓	✗	✓	✓	✓	✓	✗	✗
MatrixCity [100]	519	2	28	Synthetic	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
HoliCity [101]	6.3	1	20	Real	✗	CAD	✓	✗	✓	✗	✓	✓	✗	✗
KITTI-360 [102]	78	1	-	Real	✗	CAD	✓	✗	✓	✗	✓	✓	✗	✗
UrbanScene3D [103]	6.1 [†]	-	3 [†]	Real	✗	Mesh	✓	✗	✗	✓	✗	✗	✗	✓
GoogleEarth	24	1	25	Real	✓	Voxel	✓	✗	✗	✓	✓	✓	✓	✓
CityTopia	37.5	11	36	Synthetic	✓	Voxel	✓	✓	✓	✓	✓	✓	✓	✓

[†] Only the real-world image subset is counted for this dataset.

where $\theta \in (-180^\circ, 180^\circ]$ denotes the yaw angle, indicating the vehicle's heading in the XY-plane relative to the $-y$ -axis, while $\gamma \in (-90^\circ, 90^\circ)$ represents the pitch angle, with positive or negative values indicating upward or downward tilt relative to the XY-plane. The feature $\mathbf{f}_{V_i}^{(\mathbf{p}^C, t)}$ corresponding to the vehicle V_i at time step t for \mathbf{p}^C is derived as

$$\mathbf{f}_{V_i}^{(\mathbf{p}^C, t)} = \mathcal{O}(\text{Concat}(\mathbf{f}_{V_i}^t, \mathbf{p}^C)) \quad (15)$$

where $\mathbf{f}_{V_i}^t \in \mathbb{R}^{d_V}$ is the features extracted from the local scene $(\mathbf{H}_{T_i}^{V_i}, \mathbf{S}_{T_i}^{V_i})$ using the global encoder E_V .

$$\mathbf{f}_{V_i}^t = E_V(\mathbf{H}_{T_i}^{V_i}, \mathbf{S}_{T_i}^{V_i}) \quad (16)$$

Volumetric Rendering. The volumetric rendering mirrors Building Instance Generator, using a style code \mathbf{z} to represent the variability in vehicle appearances. The pixel value $C(\mathbf{r})$ is calculated through an integration process as described in Equation 11. The camera ray $\mathbf{r}(t)$ is normalized to the canonical feature space following Equation 13.

Loss Function. Vehicle Instance Generator is optimized with a hybrid objective that integrates reconstruction and adversarial objectives. Specifically, the training process incorporates an L_1 loss, a perceptual loss \mathcal{P} , and a GAN loss \mathcal{G} to balance fidelity and realism.

$$\ell_V = \lambda_V^{L_1} \|\hat{\mathbf{I}}_{V_i}^t - \mathbf{I}_{V_i}^t\| + \lambda_V^{\mathcal{P}} \mathcal{P}(\hat{\mathbf{I}}_{V_i}^t, \mathbf{I}_{V_i}^t) + \lambda_V^{\mathcal{G}} \mathcal{G}(\hat{\mathbf{I}}_{V_i}^t, \mathbf{S}_{V_i}^t) \quad (17)$$

where $\mathbf{I}_{V_i}^t$ denotes the ground truth image of the vehicle instance V_i at time step t , while $\mathbf{S}_{V_i}^t$ is the corresponding perspective-view semantic map, generated in a manner similar to \mathbf{S}_G . The weights of the three losses are represented as $\lambda_V^{L_1}$, $\lambda_V^{\mathcal{P}}$, and $\lambda_V^{\mathcal{G}}$. Note that ℓ_V is applied exclusively to pixels with semantic labels belonging to the vehicle instance.

F. Compositor

As there are no ground truth images available for the outputs generated by City Background Generator, Building Instance

Generator, and Vehicle Instance Generator, training neural networks to combine these images becomes challenging. Consequently, Compositor merges the generated images and their corresponding masks into one unified image.

$$\mathbf{I}_C^t = \hat{\mathbf{I}}_G \mathbf{M}_G + \sum_{i=1}^{n_B} \hat{\mathbf{I}}_{B_i} \mathbf{M}_{B_i} + \sum_{i=1}^{n_V} \hat{\mathbf{I}}_{V_i}^t \mathbf{M}_{V_i}^t \quad (18)$$

IV. DATASETS

A. OSM Dataset

The OSM dataset, collected from OpenStreetMap [31], includes rasterized semantic maps and height fields for 80 cities across the globe, covering more than 6,000 km^2 . In the rasterization step, vector data is transformed into images by converting longitude and latitude coordinates into the EPSG:3857 coordinate system at zoom level 18, which gives a resolution of approximately 0.597 meters per pixel. As shown in Fig. 2, The segmentation maps use different colors to indicate various elements: red for roads, yellow for buildings, green for urban greenery, cyan for construction areas, and blue for water bodies. The height fields mainly capture building elevations, based on OpenStreetMap data. The heights for roads are set to 4, water bodies at 0, and urban greenery is assigned random heights, generated using Perlin noise [104] within a range of 8 to 16 meters.

B. GoogleEarth Dataset

CityDreamer4D generates each building instance in the city separately to handle the diversity of buildings, which requires dense 3D instance annotations. As shown in Table I, no existing dataset provides both dense 3D semantic and instance annotations. To address this, we automatically generate dense 3D semantic and building instance annotations for the GoogleEarth dataset by geographically aligning Google Earth and OpenStreetMap using latitude and longitude.

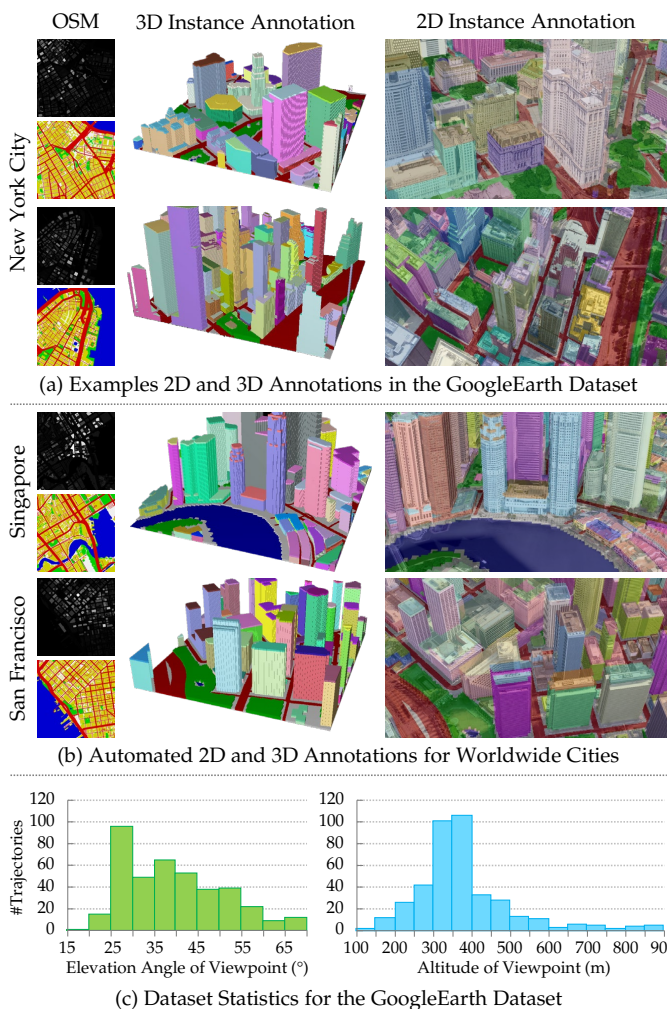


Fig. 2. **Overview of the OSM and GoogleEarth Datasets.** (a) Examples of the 2D and 3D annotations in the GoogleEarth dataset, which can be automatically generated using the OSM dataset. (b) The automatic annotation pipeline can be readily adapted for worldwide cities. (c) The dataset statistics highlight the diverse perspectives in the GoogleEarth dataset.

Image Collection. The GoogleEarth dataset, collected from Google Earth Studio [32], includes 400 orbit trajectories over the New York City, totaling 24,000 images at a 960x540 resolution. As shown in Fig. 2c, orbit radii range from 125 to 813 meters, with altitudes from 112 to 884 meters. Google Earth Studio also provides camera intrinsic and extrinsic parameters for each image.

2D and 3D Annotation. The 3D annotations can be generated by: 1) performing connected components detection on the OSM semantic map to create the instance map for buildings, while keeping the labels for background stuff unchanged, and 2) generating 3D volumes by extruding the pixels in the instance map based on height values from the OSM dataset. The dense 3D annotations can be used to create 2D annotations by projecting the 3D volumes onto images, leveraging the camera parameters from Google Earth Studio. Fig. 2a shows the 2D and 3D instance annotations in the GoogleEarth dataset, highlighting the efficiency of automated data annotation. Fig. 2b shows how the automated annotation pipeline can be applied to cities worldwide.

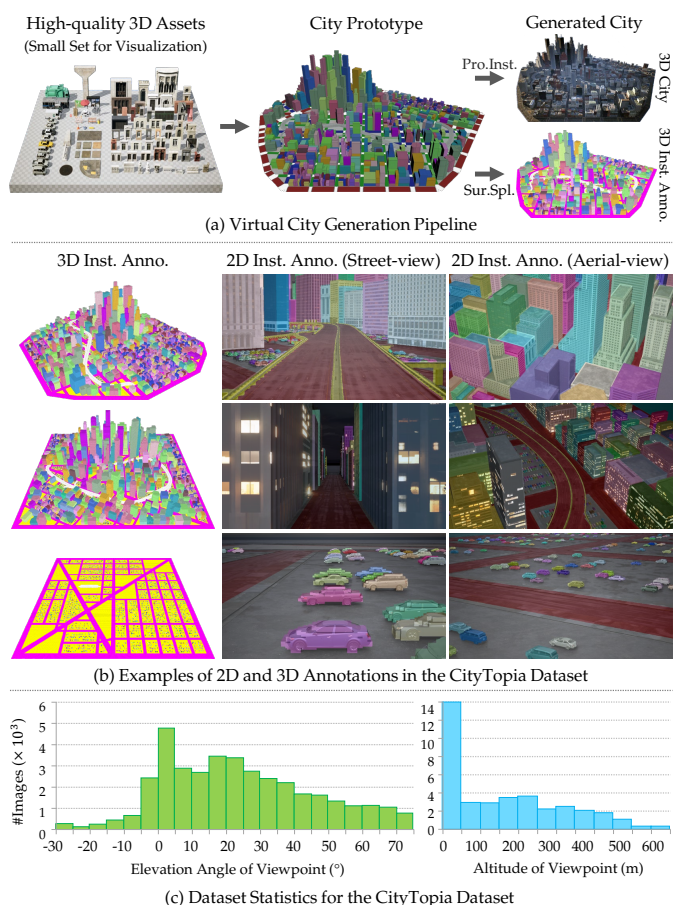


Fig. 3. **Overview of the CityTopia Dataset.** (a) The virtual city generation pipeline. “Pro.Inst.,” “Sur.Spl,” and “3D Inst. Anno.” denote “Prototype Instantiation,” “Surface Sampling,” and “3D Instance Annotation”, respectively. (b) Examples of 2D and 3D annotations in the CityTopia dataset are shown from both daytime and nighttime street-view and aerial-view perspectives, automatically generated during virtual city generation. (c) The dataset statistics highlight the diverse perspectives in both street and aerial views.

C. CityTopia Dataset

The GoogleEarth dataset provides images with dense 3D semantic and instance annotations but faces three challenges: 1) it lacks street-view images due to suboptimal 3D reconstructions near ground level in Google Earth Studio [32]; 2) its annotations, sourced from OpenStreetMap [31], have some imprecision due to differing data sources; and 3) elevated structures like highways remain unannotated due to missing height data in OpenStreetMap. To address these challenges, we construct the CityTopia dataset, featuring precise 3D dense annotations on high-fidelity day and night images from both street and aerial views. As shown in Table I, it is the largest dataset to date, offering unparalleled scene diversity and detailed annotations for urban cities.

Virtual City Generation. To build the CityTopia dataset, we design 11 virtual cities in Houdini and Unreal Engine, generating 3D annotations and realistic images with controlled lighting conditions. As illustrated in Fig. 3a, we use a diverse, high-quality set of approximately 5,000 3D assets from the CitySample project [33] to procedurally generate a city pro-

TABLE II

QUANTITATIVE COMPARISON. THE BEST VALUES ARE HIGHLIGHTED IN BOLD. NOTE THAT INFINICITY IS NOT INCLUDED IN THIS COMPARISON AS IT IS NOT OPEN-SOURCED.

Methods	GoogleEarth					CityTopia				
	FID ↓	KID ↓	VBench ↑	DE ↓	CE ↓	FID ↓	KID ↓	VBench ↑	DE ↓	CE ↓
SGAM [105]	277.6	0.358	0.691	0.575	239.2	330.1	0.284	0.690	0.571	233.5
PersistentNature [106]	123.8	0.109	0.706	0.326	86.37	235.3	0.215	0.713	0.428	127.3
SceneDreamer [7]	232.2	0.204	0.781	0.153	0.186	195.1	0.126	0.708	0.185	0.162
DreamScene4D [107]	-	-	-	-	-	288.2	0.136	0.715	0.199	0.146
DimensionX [108]	206.9	0.182	0.805	-	-	171.4	0.070	0.815	-	-
CityDreamer4D (Ours)	96.83	0.096	0.834	0.138	0.060	88.48	0.049	0.825	0.150	0.063

prototype in Houdini¹. This city prototype stores the 6D poses of all 3D assets within the city. Through surface sampling, we can assign each 3D point a semantic and instance label, and by instantiating the city prototype in Unreal Engine², we produce a fully generated virtual city.

Image Collection. Once the virtual city is instantiated in Unreal Engine, camera trajectories are set to generate 3,000 images for cities with buildings and 7,500 for a vehicle-only city. Daytime and nighttime scenes are rendered for each trajectory, with sunlight removed to help the network more easily learn lighting consistency during the generation process. To avoid Moiré effects, each image is sampled 8x spatially and 32x temporally during rendering. As shown in Fig. 3c, the CityTopia dataset provides a wider range of viewpoints, shown by its broader elevation angles compared to the GoogleEarth dataset, as well as more street-level perspectives, evidenced by the large number of images taken at near-zero altitude.

2D and 3D Annotation. Since the precise 3D annotations are natively generated from the virtual city pipeline, once the camera poses are set in Unreal Engine, 2D annotations are produced by projecting the 3D annotations using the given camera poses. Fig. 3b highlights the perfect alignment of 2D and 3D instance annotations with both street-view and aerial-view images in the CityTopia dataset. The last row features a vehicle-only scene, enhancing vehicle generation learning. The accurate vehicle annotations demonstrate the effectiveness of the pipeline, which can be scaled by adding more 3D assets.

V. EXPERIMENTS

A. Evaluation Protocols

We evaluate our method by generating 1,024 unique city layouts, each with 20 variations created by randomly sampling the style code \mathbf{z} . For each variation, images are rendered at a resolution of 960×540 pixels using randomized camera trajectories. Frames from these renderings are randomly selected for evaluation, depending on the specific metrics used. The evaluation metrics are as follows.

FID and KID. Fréchet Inception Distance (FID) [109] and Kernel Inception Distance (KID) [110] measure image quality. FID and KID are calculated between 15,000 generated frames and 15,000 randomly sampled images from datasets.

VBench. VBench [111] provides a comprehensive evaluation of video generative models, considering dimensions such as background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. The VBench score is computed from 150 videos, each consisting of 100 frames rendered at 16 FPS.

Depth Error (DE). To assess 3D geometry, DE is evaluated following EG3D [52]. A pretrained model [112] generates pseudo ground truth depth maps by accumulating density σ . DE is calculated as the L2 distance between the normalized depth maps, evaluated on 100 frames per method.

Camera Error (CE). CE measures multi-view consistency, following SceneDreamer [7]. CE is computed on a static 3D scene by comparing the inferred camera trajectory with the one estimated by COLMAP [113]. This metric is calculated on 600 frames rendered from an orbit trajectory and is defined as the scale-invariant normalized L2 distance between the generated and reconstructed camera poses.

B. Implementation Details

Hyperparameters

Unbounded Layout Generator. The codebook size d_K is set to 512, with each code having a dimension d_C of 512. Height field and semantic map patches are cropped to 512×512 and compressed by a factor of 16. The loss weights are $\lambda_R = 10$, $\lambda_S = 10$, and $\lambda_E = 1$.

City Background Generator. For the GoogleEarth dataset, the local window resolutions are set to $N_G^H = 1536$, $N_G^W = 1536$, and $N_G^D = 640$. For the CityTopia dataset, they are set to $N_G^H = 3072$, $N_G^W = 3072$, and $N_G^D = 2560$. The dimension of scene-level features d_G is 2. For the generative hash grid, $N_H^L = 16$, $N_E = 2^{19}$, and $N_G^C = 8$. The prime numbers used in Equation 5 are $\pi^1 = 1$, $\pi^2 = 2654435761$, $\pi^3 = 805459861$, $\pi^4 = 3674653429$, and $\pi^5 = 2097192037$. The loss function weights are set to $\lambda_G^{L1} = 10$, $\lambda_G^P = 10$, and $\lambda_G^G = 0.5$.

Building Instance Generator. For the GoogleEarth dataset, the local window resolutions are set to $N_B^H = 672$, $N_B^W = 672$, and $N_B^D = 640$. For the CityTopia dataset, these values are $N_B^H = 768$, $N_B^W = 768$, and $N_B^D = 2560$. The pixel-level features have 63 channels ($N_B^C = 63$), and the dimension N_P^L is set to 10.

Vehicle Instance Generator. The dimension of scene-level features d_V is 2. The local window resolutions are set to

¹<https://www.sidefx.com>

²<https://www.unrealengine.com>

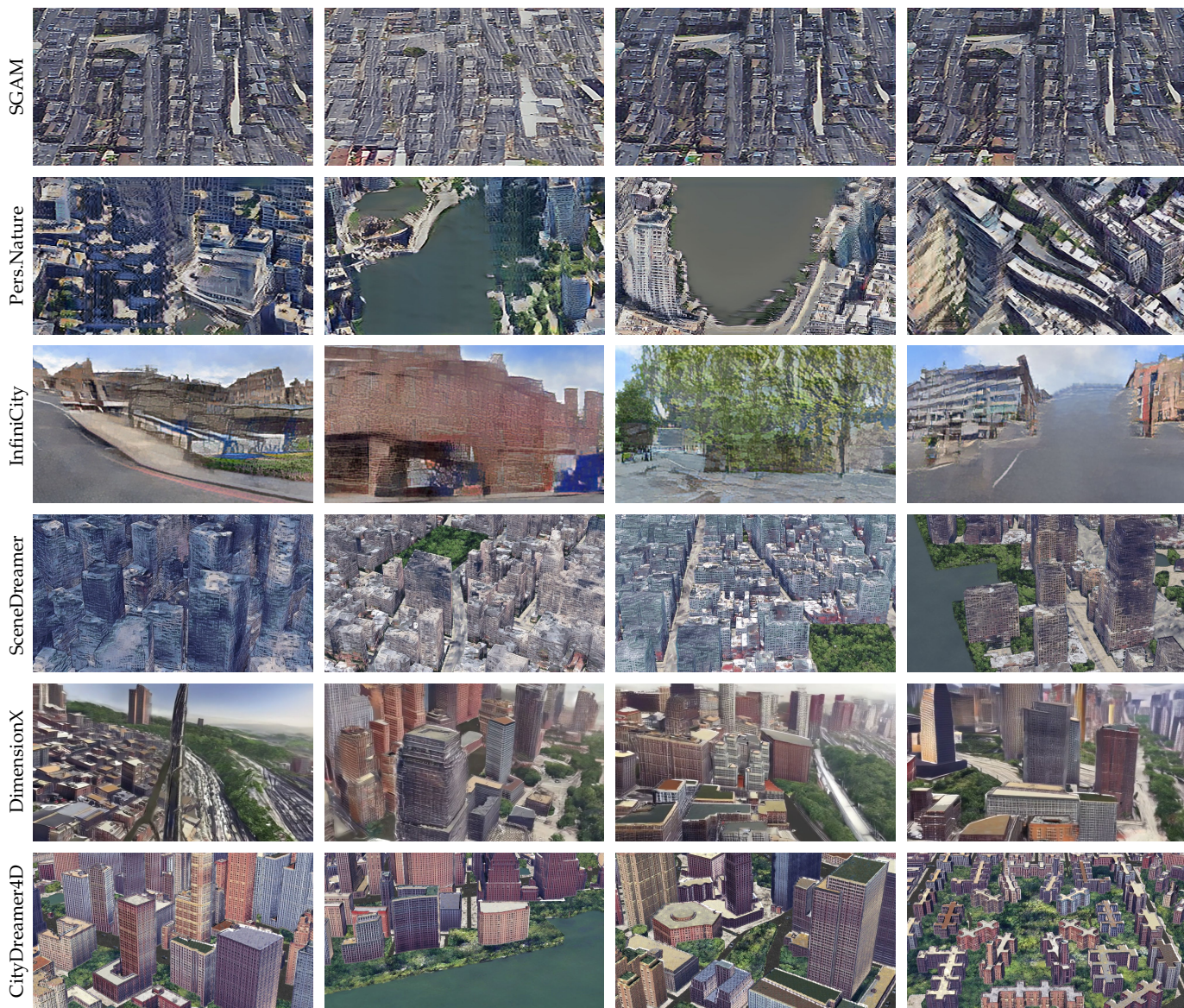


Fig. 4. **Qualitative Comparison on Google Earth.** For SceneDreamer [7] and CityDreamer4D, vehicles are generated using models trained on CityTopia due to the lack of semantic annotations for vehicles in Google Earth. For DimensionX [108], the initial frame is provided by CityDreamer4D. The visual results of InfiniCity [26], provided by the authors, have been zoomed in for better viewing. “Pers.Nature” stands for “PersistentNature” [106].

$N_V^H = 32$, $N_V^W = 32$, and $N_V^D = 32$. The loss function weights are assigned as $\lambda_V^{L1} = 10$, $\lambda_V^P = 10$, and $\lambda_V^G = 0.5$.

Training Details

Unbounded Layout Generator. The VQVAE model is trained over 1,250,000 iterations using a batch size of 16, an Adam optimizer with $\beta = (0.5, 0.9)$, and a learning rate of 7.2×10^{-5} . The autoregressive transformer is trained for 250,000 iterations with a batch size of 80, an Adam optimizer with $\beta = (0.9, 0.999)$, and a learning rate of 2×10^{-4} .

Stuff and Instance Generators. The City Background Generator, Building Instance Generator, and Vehicle Instance Generator are trained with an Adam optimizer, using $\beta = (0, 0.999)$ and a learning rate of 10^{-4} . The discriminators use the same optimizer settings with a learning rate of 10^{-5} . Training runs for 298,500 iterations with a batch size of 8, and images are randomly cropped to 192×192 resolution.

C. Main Results

Comparison Methods. We compare CityDreamer4D against several state-of-the-art methods, including SGAM [105], PersistentNature [106], SceneDreamer [7], and InfiniCity [26]. Since no method exists for 4D scene generation, we use DreamScene4D [107] for 4D novel view synthesis and DimensionX [108] for 4D video generation as competitive baselines. To ensure a fair comparison, all methods, except for InfiniCity and DimensionX, are retrained using their released code on the GoogleEarth and CityTopia datasets. Since SceneDreamer cannot generate city layouts or traffic scenarios, their inputs are supplied by Unbounded Layout Generator and Traffic Scenario Generator. Additionally, because the GoogleEarth dataset lacks annotations for dynamic objects, vehicles are generated using models trained on the CityTopia dataset to support 4D generation.

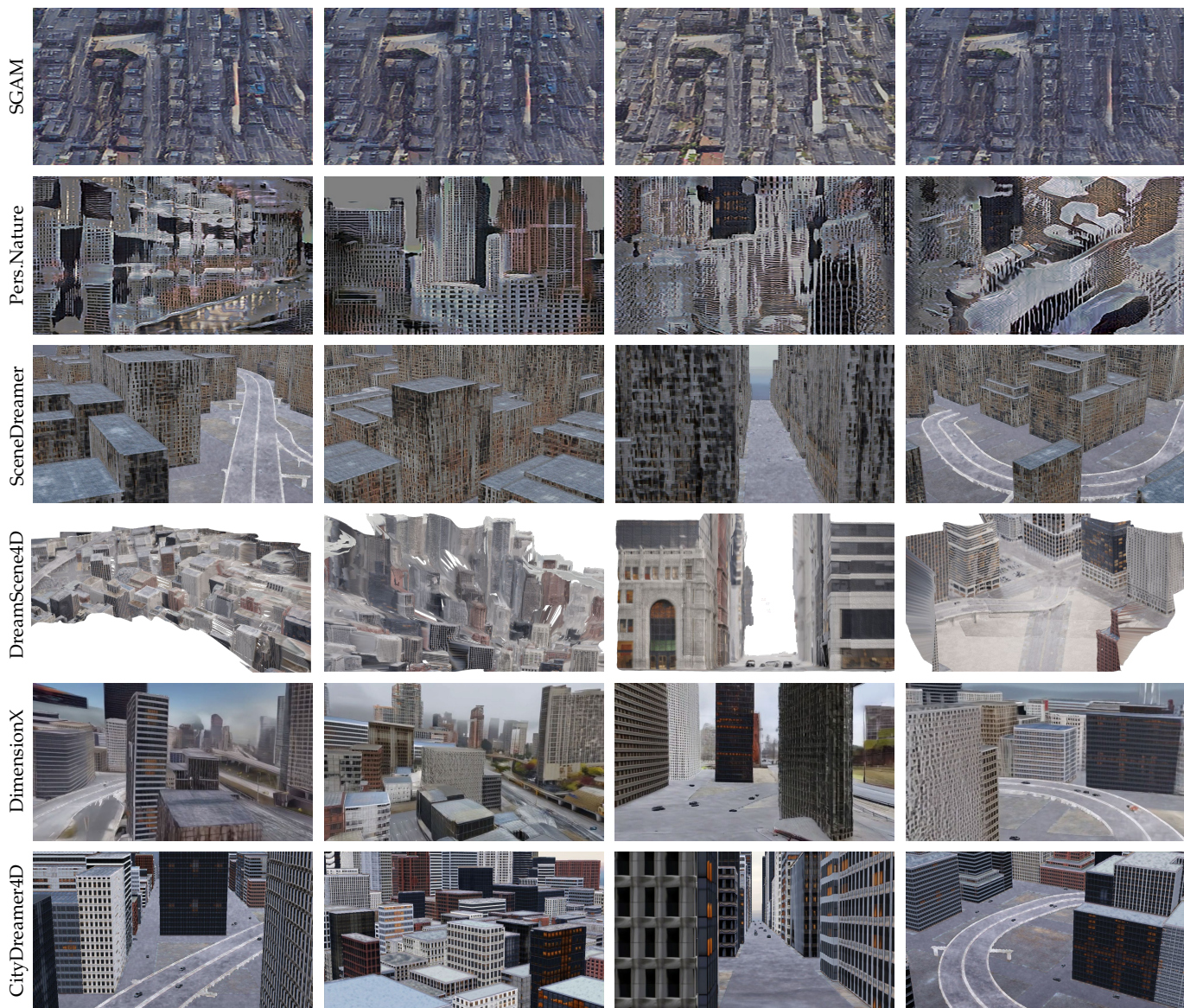


Fig. 5. **Qualitative Comparison on CityTopia.** The initial frame for DimensionX and the input frames for DreamScene4D are chosen from the dataset. “Pers.Nature” refers to “PersistentNature” [106].

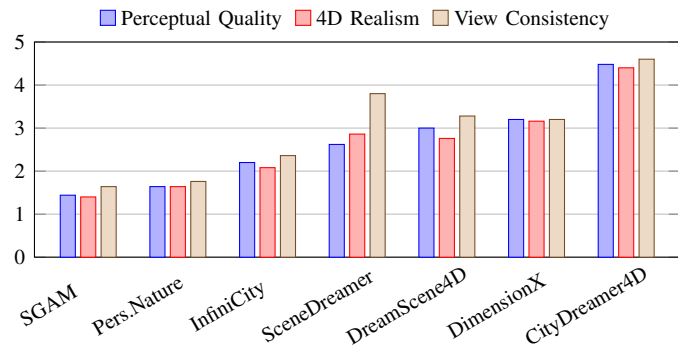


Fig. 6. **User Study on 4D City Generation.** All scores are in the range of 5, with 5 indicating the best. “Pers.Nature” refers to “PersistentNature” [106].

Qualitative Comparison. Fig. 4 and 5 present qualitative comparisons with the baseline methods on the GoogleEarth and CityTopia datasets, respectively. SGAM faces difficulties in generating realistic results and maintaining multi-view consistency due to the inherent challenges of extrapolating views for complex 4D cities. PersistentNature, which adopts a tri-plane representation, also struggles to produce realistic renderings. Both InfiniCity and SceneDreamer use BEV maps as their scene representation, but they still experience significant structural distortions in instance-level objects, such as buildings and vehicles, because all instances are assigned the same semantic label. DreamScene4D cannot directly generate 4D scenes but transforms monocular videos into 4D scenes by decoupling dynamic objects from the background, yet it struggles to reconstruct their 3D shapes. During the generation of orbit 4D videos, DimensionX exhibited severe distortions and failed to maintain multi-view consistency in the results.

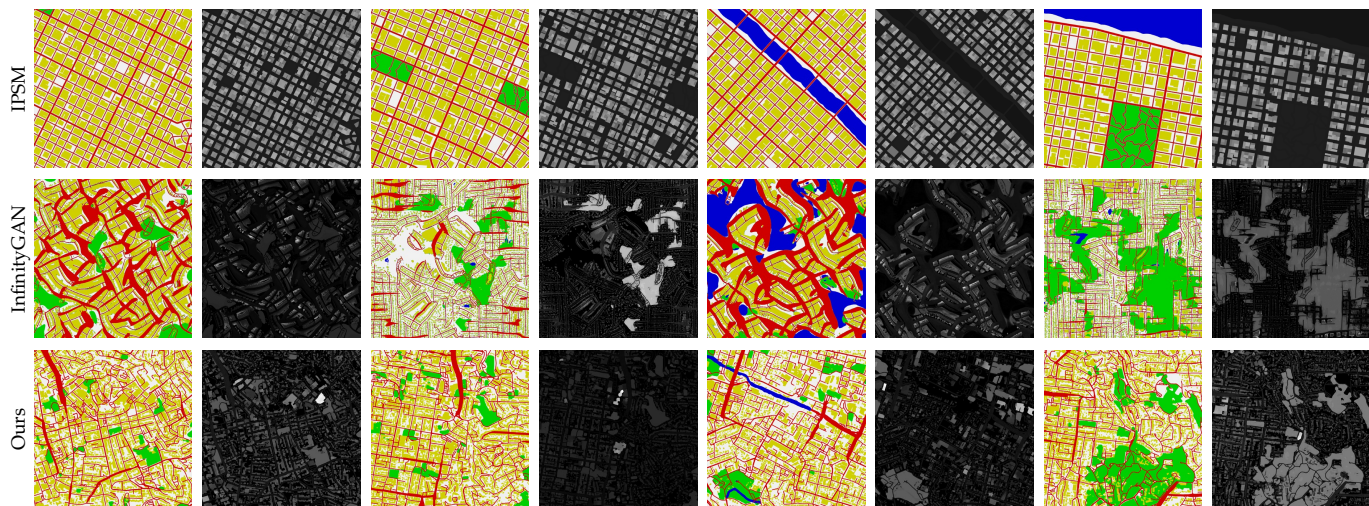


Fig. 7. **Qualitative Comparison of City Layout Generators.** The height map values are normalized to a range of $[0, 1]$ by dividing each value by the maximum value within the map.

TABLE III

QUANTITATIVE COMPARISON OF UNBOUNDED LAYOUT GENERATOR (ULG). THE BEST VALUES ARE HIGHLIGHTED IN BOLD. THE GENERATED IMAGES ARE CENTRALLY CROPPED TO A SIZE OF 4096×4096 .

Methods	FID ↓	KID ↓
IPSM [114]	321.47	0.502
InfinityGAN [115]	183.14	0.288
ULG (Ours)	124.45	0.123

In comparison, the proposed CityDreamer4D generates more realistic and diverse results compared to all the baselines³.

Quantitative Comparison. Table II shows the quantitative metrics, where CityDreamer4D outperforms the baselines in FID, KID, and VBench, highlighting its motion smoothness, dynamic degree, and aesthetic quality. Additionally, CityDreamer4D achieves the lowest DE and CE errors, demonstrating accurate 3D geometry, view consistency, and photorealistic image generation.

User Study. To better evaluate the multi-view consistency and quality of unbounded 4D city generation, we perform a user study following CityDreamer’s protocol [34]. In this survey, 25 volunteers rate each generated city on three aspects: 1) perceptual quality, 2) 4D realism, and 3) view consistency. Ratings are on a scale of 1 to 5, with 5 being the highest. As shown in Fig. 6, the proposed CityDreamer4D outperforms the baselines by a significant margin.

D. Ablation Studies

Effectiveness of Unbounded Layout Generator. Unbounded Layout Generator (ULG) is essential for producing “unbounded” city layouts. To demonstrate the effectiveness of ULG, we evaluate its performance against InfinityGAN [115], which is utilized in InfiniCity, alongside the rule-based city layout generation technique, IPSM [114]. Following InfiniCity [26], we use FID and KID to quantitatively evaluate the

³More results can be found on our project page.

TABLE IV

QUANTITATIVE COMPARISON OF BUILDING INSTANCE GENERATOR VARIANTS. THE BEST VALUES ARE HIGHLIGHTED IN BOLD. NOTE THAT “INST.” AND “POS.ENC.” REFER TO “INSTANCE LABELS” AND “POSITIONAL ENCODING”, WHILE “G” AND “L” DENOTE “GLOBAL ENCODER” AND “LOCAL ENCODER”, RESPECTIVELY.

BIG	Inst.	Encoder		Pos.Enc.		Evaluation Metrics			
		G	L	Hash	SinCos	FID ↓	KID ↓	DE ↓	CE ↓
✗	✗	-	-	-	-	195.1	0.126	0.185	0.162
✓	✗	✗	✓	✗	✓	167.8	0.094	0.157	0.087
✓	✓	✓	✗	✓	✗	196.8	0.124	0.165	0.159
✓	✓	✓	✗	✗	✓	197.9	0.132	0.162	0.152
✓	✓	✗	✓	✓	✗	182.3	0.111	0.155	0.092
✓	✓	✗	✓	✗	✓	88.48	0.049	0.150	0.063

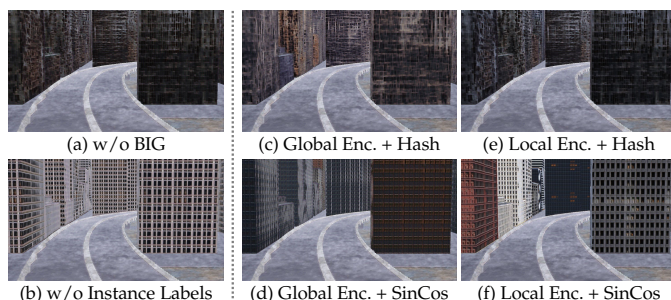


Fig. 8. **Qualitative Comparison of Building Instance Generator (BIG) Variants.** (a) and (b) illustrate the effects of removing BIG and instance labels, respectively. (c)–(f) present the results of various scene parameterizations. Note that “Enc.” is an abbreviation for “Encoder”.

quality of the generated layouts. As illustrated in Table III, ULG achieves the best results in terms of all metrics compared to IPSM and InfinityGAN. The qualitative results shown in Fig. 7 also demonstrate the high quality and diversity of the proposed method.

Effectiveness of Building Instance Generator. We highlight the essential role of Building Instance Generator (BIG)

TABLE V
QUANTITATIVE COMPARISON OF VEHICLE INSTANCE GENERATOR VARIANTS. ALL METRICS ARE COMPUTED ON THE VEHICLE-ONLY CITY FROM THE CITYTOPIA DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD. NOTE THAT “CAN.” AND “POS.ENC.” REFER TO “CANONICALIZATION” AND “POSITIONAL ENCODING”, WHILE “G” AND “L” DENOTE “GLOBAL ENCODER” AND “LOCAL ENCODER”, RESPECTIVELY.

VIG	Can.	Encoder		Pos.Enc.		Evaluation Metrics			
		G	L	Hash	SinCos	FID ↓	KID ↓	DE ↓	CE ↓
✗	✗	-	-	-	-	419.3	0.576	0.364	1.276
✓	✗	✓	✗	✗	✓	273.4	0.530	0.289	0.966
✓	✓	✓	✗	✓	✗	229.2	0.428	0.259	0.989
✓	✓	✓	✗	✗	✓	142.3	0.276	0.202	0.824
✓	✓	✗	✓	✓	✗	273.4	0.521	0.265	0.997
✓	✓	✗	✓	✗	✓	200.5	0.403	0.332	1.117

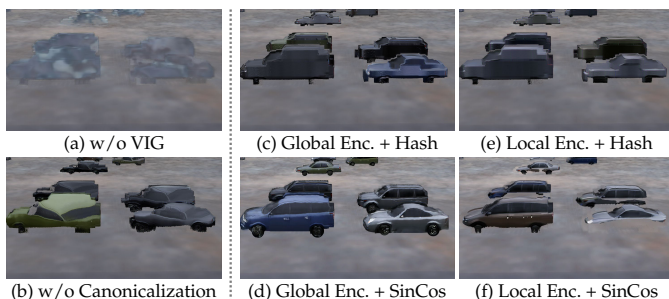


Fig. 9. **Qualitative Comparison of Vehicle Instance Generator (VIG) Variants.** (a) and (b) illustrate the effects of removing VIG and canonicalization, respectively. (c)–(f) present the results of various scene parameterizations. Note that “Enc.” is an abbreviation for “Encoder”.

in achieving successful unbounded 4D city generation. To validate its effectiveness, we perform an ablation study for BIG. We first compare BIG with two alternative designs: (1) Removing BIG from CityDreamer4D, effectively reverting the model to SceneDreamer, and (2) Generating all buildings simultaneously using BIG without incorporating instance labels. As shown in the first two rows of Table IV and Fig. 8a-b, both alternative designs result in significant degradation in generation quality, underscoring the importance of BIG and instance labels. Scene parameterization directly impacts the quality of 4D city generation. BIG uses vanilla SinCos positional encoding with pixel-wise features from the local encoder. To demonstrate the effectiveness of the scene parameterization in BIG, we compare BIG with other alternative scene parameterization designs. Scene parameterization plays a critical role in the quality of 4D city generation. BIG leverages vanilla SinCos positional encoding combined with pixel-wise features from the local encoder. To evaluate the effectiveness of BIG’s scene parameterization, we compare it with alternative designs. As shown in the last four rows of Table IV and Fig. 8c-f, using generative hash grid positional encoding results in distorted building façades, while Global Encoders with SinCos encoding introduce repetitive façade patterns. These comparisons emphasize the significance of BIG’s well-designed parameterization in achieving realistic and varied results.

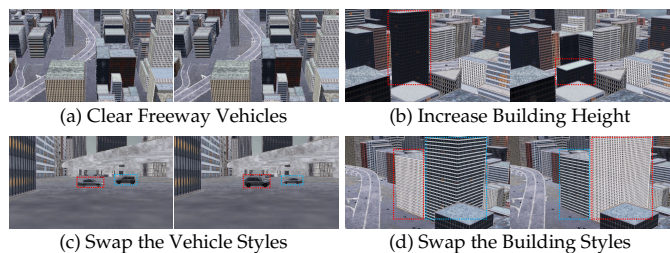


Fig. 10. **Localized Editing on the Generated Cities.** (a) and (c) show vehicle editing results, while (b) and (d) present building editing results.

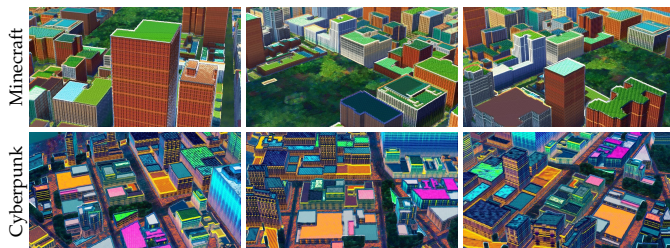


Fig. 11. **Text-driven City Stylization with ControlNet.** The multi-view consistency is preserved in stylized Minecraft and Cyberpunk cities.

Effectiveness of Vehicle Instance Generator. Vehicle Instance Generator (VIG) plays a critical role in generating vehicles within 4D cities. To validate its effectiveness, we conduct an ablation study on VIG. We compare it with two alternative designs: (1) Removing VIG from CityDreamer4D and treating vehicles as background stuff, allowing City Background Generator to handle their generation, and (2) Generating vehicles without canonicalization, meaning they are not produced in a canonical feature space. As shown in the first two rows of Table V and Fig. 9a-b, both alternative designs lead to severe distortions in the generated results, highlighting the importance of VIG and canonicalization. Scene parameterization is equally critical in VIG. To validate this, we compare different scene parameterization designs within VIG. Currently, VIG uses vanilla SinCos positional encoding combined with global-level features from the global encoder. In the canonical feature space, combining global-level features with 3D coordinates allows the network to better share features across different vehicles, facilitating better convergence. As shown in the last row of Table V and Fig. 9f, using a local encoder with SinCos positional encoding, as in BIG, makes learning more challenging, resulting in incomplete vehicle shapes. Similarly, using generative hash grid in VIG leads to structural distortions by complicating the network’s ability to associate texture features with 3D coordinates, as illustrated in Fig. 9c and 9e as well as the 3rd and 5th rows of Table V.

E. Applications

Urban Simulator. CityDreamer4D can be a powerful urban simulator, capable of generating realistic 4D urban scenes with dynamic objects and detailed environments. Unlike traditional simulators such as CARLA [68], which are limited to predefined, bounded areas, this method supports unbounded urban scenes, creating vast, seamless cityscapes. Furthermore,

TABLE VI

VISUAL LANGUAGE NAVIGATION (VLN) RESULTS IN GENERATED 4D CITIES. METRICS INCLUDE: PL (PATH LENGTH, SCALED BY 1/10, IN METERS), SR (SUCCESS RATE, %), SPL (SUCCESS RATE WEIGHTED BY NORMALIZED INVERSE PATH LENGTH), AND RT (RESET TIMES). NOTE THAT EACH TRAJECTORY IS EXECUTED FIVE TIMES AND IS CONSIDERED A FAILURE IF ALL ATTEMPTS RESULT IN RESETS.

Methods	#Param (B)	PL↓	SR↑	SPL↑	RT↓
Human (5 participants)	-	20.73	100.0	85.87	0.00
Gemini 2.5 Pro [117]	-	9.32	12.40	4.43	0.45
GPT-4o [118]	-	8.97	36.00	17.32	0.11
SAIL-VL 1.6 [119]	8.33	14.56	23.40	7.63	0.28
Ovis2 [120]	8.94	13.96	17.00	5.01	0.35
Qwen2.5-VL [121]	8.29	5.01	15.00	7.01	0.37
Ola [122]	8.88	9.15	18.00	8.30	0.32
InternVL3 [123]	7.94	9.02	25.60	12.66	0.23

it can generate both street-view and aerial-view perspectives, providing a richer variety of scenarios for applications like autonomous driving, urban planning, and virtual reality.

Localized Editing. Benefiting from the compositional architecture, CityDreamer4D allows for localized editing on building and vehicle instances. In Fig. 10a and 10c, vehicle positions and styles can be independently modified without affecting other scene elements. Similarly, as shown in Fig. 10b and 10d, building appearances adapt seamlessly to varying heights while maintaining a consistent style. This capability facilitates customized scene refinement in post-production.

City Stylization. The generated cities can be seamlessly restyled by leveraging ControlNet [116], fine-tuning pretrained models on images created with ControlNet conditioned on HED edges. Fig. 11 shows examples of city styles such as Minecraft and Cyberpunk. These results maintain multiview consistency, enabled by the proposed scene representation and parameterization in CityDreamer4D.

Visual Language Navigation. To evaluate the practicality of our generated 4D urban environments, we conduct experiments on Visual Language Navigation, where an embodied agent navigates based on natural language instructions. Specifically, the agent operates within scenes generated by CityDreamer4D, pretrained on the CityTopia dataset. We manually annotate a test set of 100 instruction-trajectory pairs, each guiding the agent to a distinct landmark in the generated scenes. Following the protocol in GRUtopia [124], the agent receives its current image observation and a language prompt as input to a vision-language model (VLM), which selects one of 12 discrete actions: move forward or diagonally (2/4/6 meters), turn left/right (45°), or stop. This process continues iteratively until the model outputs “stop”. We additionally collect a human performance reference from 5 volunteers, who complete the full test set following the same navigation protocol as the VLMs. We adopt the zero-shot evaluation setting from GRUtopia, using pre-trained VLMs without task-specific fine-tuning. We evaluate the latest state-of-the-art VLMs, including open-source models such as SAIL-VL 1.6 [119], Ovis2 [120], Qwen2.5-VL [121], Ola [122], and InternVL3 [123], as well as closed-source models like Gemini 2.5 Pro [117] and GPT-4o [118]. Navigation performance is evaluated using four

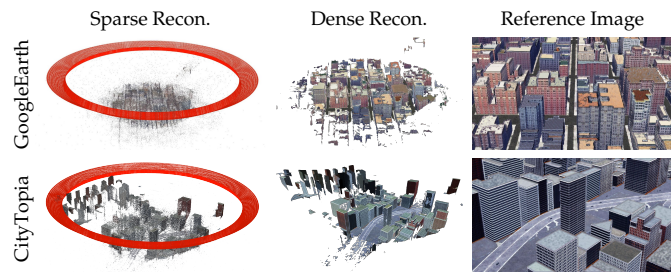


Fig. 12. **COLMAP Reconstruction of 600-frame Orbital Videos.** The red ring shows the camera positions, and the clear point clouds demonstrate CityDreamer4D’s consistent rendering. Note that “Recon.” stands for “Reconstruction.”

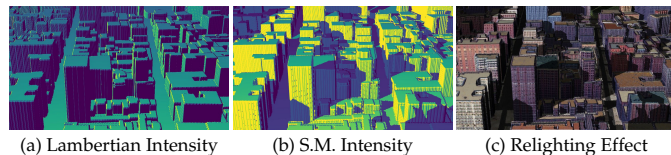


Fig. 13. **Directional Light Relighting Effect.** (a) and (b) show the lighting intensity. (c) illustrates the relighting effect. Note that “S.M.” denotes “Shadow Mapping”.



Fig. 14. **Pedestrians in the generated 4D cities.** The three consecutive frames illustrate a group of pedestrians crossing the street.

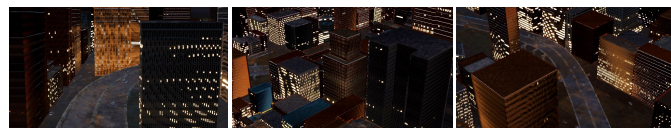


Fig. 15. **Night-view Generation Results.** Despite achieving realistic effects, managing global illumination in the generated scenes remains a challenge.

standard metrics: success rate (SR), path length (PL), success rate weighted by normalized inverse path length (SPL), and the number of resets due to occlusions (RT). As shown in Table VI, VLMs struggle with spatial reasoning in 4D cities, as indicated by low SR and SPL scores. GPT-4o performs best, followed by InternVL3. For comparison, the human reference achieves an SR of 100.0% and an SPL of 85.87%, indicating a substantial gap between human and model performance. The average human path length is slightly higher than that of VLMs, mainly because the latter often stop before actually reaching the destination. These results, along with recent findings [124], [125], underscore the difficulty of grounding spatial instructions in complex urban environments and suggest that generated 4D cities can serve as valuable benchmarks for evaluating the spatial reasoning and navigation capabilities of VLMs.

F. Discussions

View Consistency. To demonstrate CityDreamer4D’s multi-view consistent renderings, we use COLMAP [113] for structure-from-motion and dense reconstruction on orbital videos generated using models trained on the GoogleEarth and CityTopia datasets. The video sequence comprises 600 frames at a resolution of 960×540 , captured from a circular camera trajectory orbiting the scene at a fixed height, with the camera focused on the center. Reconstruction is performed solely from the images, without specifying camera parameters. As illustrated in Fig. 12, the estimated camera poses closely align with the sampled trajectory, and the resulting point cloud is both dense and well-defined.

Relighting. In CityDreamer4D, the generation of background stuff and instances is deliberately decoupled, offering two key benefits: (1) Simplified learning for building instances, vehicle instances, and background stuff, and (2) Enabling localized editing of building and vehicle instances. This approach can be viewed as an inverse rendering process, where CityDreamer4D generates the albedo, normals, and depth of urban scenes. Lighting and shading effects are then computed based on the given lighting conditions. As shown in Fig. 13, the shading effects are divided into two components: Lambertian shading and shadow mapping. Lambertian shading accounts for the light direction and surface normal, resulting in uniform lighting across all directions, as shown in Fig. 13a. Shadow mapping considers light visibility, enabling the simulation of shadows and occlusion from other objects in the scene, as shown in Fig. 13b. The final relighting effects, with the camera positioned on the left side of the scene, are presented in Fig. 13c.

Diverse Agents Support. To explore the potential for supporting more diverse agents in CityDreamer4D, we conduct a preliminary experiment integrating pedestrians into the generated scenes. We use MoMask [126] to synthesize motion, retarget it to 3D human avatars, and render the animated pedestrians into our generated 4D environments. As shown in Figure 14, the resulting animation demonstrates coherent pedestrian behavior such as street crossing. This highlights the feasibility of extending our framework to support richer multi-agent simulations beyond vehicles.

Limitations. Despite the realistic generation results, CityDreamer4D has some limitations. 1) During the inference process, buildings and vehicles are generated individually, leading to a slightly higher computational cost. 2) The current implementation does not account for global illumination and reflections, which are essential for realistic night scenes. As illustrated in Fig. 15, the emitted light from buildings and vehicles does not illuminate the surrounding environment, limiting the realism of the generated cities under such conditions.

VI. CONCLUSION

In this paper, we introduce CityDreamer4D, a generative model tailored for unbounded 4D city generation. Our method simplifies the process by decoupling dynamic objects from static scenes, enabling greater flexibility and realism driven by dynamic traffic scenarios and static city layouts. Objects in the

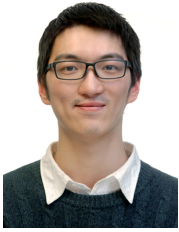
4D cities are generated using a composition of stuff-oriented and instance-oriented neural fields for background stuff, buildings, and vehicles. Additionally, we construct a comprehensive suite of datasets, including OSM, GoogleEarth, and CityTopia, which provide real-world city layouts and cityscapes with high-quality 3D annotations. CityDreamer4D achieves state-of-the-art performance in generating large-scale, realistic 4D cities with instance-level editing, leveraging its compositional design to capture urban diversity and unlock new opportunities for research and practical applications in urban simulation.

REFERENCES

- [1] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, “Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images,” *IJCV*, vol. 128, no. 12, pp. 2919–2935, 2020.
- [2] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, “DreamGaussian: Generative Gaussian splatting for efficient 3D content creation,” in *ICLR*, 2024.
- [3] Z. Chen, J. Tang, Y. Dong, Z. Cao, F. Hong, Y. Lan, T. Wang, H. Xie, T. Wu, S. Saito, L. Pan, D. Lin, and Z. Liu, “3DTopia-XL: High-quality 3D PBR asset generation via primitive diffusion,” in *CVPR*, 2025.
- [4] F. Hong, Z. Chen, Y. Lan, L. Pan, and Z. Liu, “EVA3D: compositional 3D human generation from 2D image collections,” in *ICLR*, 2023.
- [5] Z. Chen, F. Hong, H. Mei, G. Wang, L. Yang, and Z. Liu, “PrimDiffusion: Volumetric primitives diffusion for 3D human generation,” in *NeurIPS*, 2023.
- [6] X. Liu, X. Zhan, J. Tang, Y. Shan, G. Zeng, D. Lin, X. Liu, and Z. Liu, “HumanGaussian: Text-driven 3D human generation with Gaussian splatting,” in *CVPR*, 2024.
- [7] Z. Chen, G. Wang, and Z. Liu, “SceneDreamer: Unbounded 3D scene generation from 2D image collections,” *IEEE TPAMI*, vol. 45, no. 12, pp. 15 562–15 576, 2023.
- [8] Z. Wu, Y. Li, H. Yan, T. Shang, W. Sun, S. Wang, R. Cui, W. Liu, H. Sato, H. Li, and P. Ji, “BlockFusion: Expandable 3D scene generation using latent tri-plane extrapolation,” *ACM TOG*, vol. 43, no. 4, pp. 43:1–43:17, 2024.
- [9] H. Xie, Z. Chen, F. Hong, and Z. Liu, “Generative Gaussian splatting for unbounded 3D city generation,” in *CVPR*, 2025.
- [10] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao, “Consistent4D: Consistent 360° dynamic object generation from monocular video,” in *ICLR*, 2024.
- [11] J. Ren, K. Xie, A. Mirzaei, H. Liang, X. Zeng, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, and H. Ling, “L4GM: large 4D Gaussian reconstruction model,” in *NeurIPS*, 2024.
- [12] Y. Ma, Z. Lin, J. Ji, Y. Fan, X. Sun, and R. Ji, “X-Oscar: A progressive framework for high-quality text-guided 3D animatable avatar generation,” in *ICML*, 2024.
- [13] Z. Chai, C. Tang, Y. Wong, and M. S. Kankanhalli, “STAR: skeleton-aware text-based 4D avatar generation with in-network motion retargeting,” *arXiv 2406.04629*, 2024.
- [14] X. Guo, M. Zhang, H. Xie, C. Gu, and Z. Liu, “CrowdMoGen: Zero-shot text-driven collective motion generation,” *arXiv 2407.06188*, 2024.
- [15] A. Liu, A. Makadia, R. Tucker, N. Snavely, V. Jampani, and A. Kanazawa, “Infinite Nature: Perpetual view generation of natural scenes from a single image,” in *ICCV*, 2021.
- [16] Z. Li, Q. Wang, N. Snavely, and A. Kanazawa, “InfiniteNature-Zero: Learning perpetual view generation of natural scenes from single images,” in *ECCV*, 2022.
- [17] B. Deng, R. Tucker, Z. Li, L. J. Guibas, N. Snavely, and G. Wetzstein, “Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion,” in *SIGGRAPH*, 2024.
- [18] Y. Liang, X. Yang, J. Lin, H. Li, X. Xu, and Y. Chen, “LucidDreamer: Towards high-fidelity text-to-3D generation via interval score matching,” in *CVPR*, 2024.
- [19] J. Shriram, A. Trevithick, L. Liu, and R. Ramamoorthi, “RealmDreamer: Text-driven 3D scene generation with inpainting and depth diffusion,” in *3DV*, 2025.
- [20] H. Yu, H. Duan, C. Herrmann, W. T. Freeman, and J. Wu, “WonderWorld: Interactive 3D scene generation from a single image,” in *CVPR*, 2025.
- [21] M. Zhou, J. Hou, C. Luo, Y. Wang, Z. Zhang, and J. Peng, “SceneX: Procedural controllable large-scale scene generation via large-language models,” in *AAAI*, 2025.

- [22] J. Deng, W. Chai, J. Huang, Z. Zhao, Q. Huang, M. Gao, J. Guo, S. Hao, W. Hu, J. Hwang, X. Li, and G. Wang, "CityCraft: A real crafter for 3D city generation," *arXiv 2406.04983*, 2024.
- [23] S. Zhang, M. Zhou, Y. Wang, C. Luo, R. Wang, Y. Li, X. Yin, Z. Zhang, and J. Peng, "CityX: Controllable procedural content generation for unbounded 3D cities," *arXiv 2407.17572*, 2024.
- [24] Z. Hao, A. Mallya, S. J. Belongie, and M. Liu, "GANcraft: Unsupervised 3D neural rendering of minecraft worlds," in *ICCV*, 2021.
- [25] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, 2019.
- [26] C. H. Lin, H. Lee, W. Menapace, M. Chai, A. Siarohin, M. Yang, and S. Tulyakov, "InfiniCity: Infinite-scale city synthesis," in *ICCV*, 2023.
- [27] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. J. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell, "4D-fy: Text-to-4D generation using hybrid score distillation sampling," in *CVPR*, 2024.
- [28] Y. Zheng, X. Li, K. Nagano, S. Liu, O. Hilliges, and S. D. Mello, "A unified approach for text-and image-guided 4D scene generation," in *CVPR*, 2024.
- [29] D. Xu, H. Liang, N. P. Bhatt, H. Hu, H. Liang, K. N. Plataniotis, and Z. Wang, "Comp4D: Llm-guided compositional 4D scene generation," *arXiv 2403.16993*, 2024.
- [30] H. Yu, C. Wang, P. Zhuang, W. Menapace, A. Siarohin, J. Cao, L. A. Jeni, S. Tulyakov, and H. Lee, "4Real: Towards photorealistic 4D scene generation via video diffusion models," in *NeurIPS*, 2024.
- [31] <https://openstreetmap.org>.
- [32] <https://earth.google.com/studio>.
- [33] <https://www.unrealengine.com/marketplace/en-US/product/city-sample>.
- [34] H. Xie, Z. Chen, F. Hong, and Z. Liu, "CityDreamer: Compositional generative model of unbounded 3D cities," in *CVPR*, 2024.
- [35] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE TPAMI*, vol. 43, no. 12, pp. 4217–4228, 2021.
- [36] A. Melnik, M. Miasayedenkau, D. Makarovets, D. Pirstuk, E. Akbulut, D. Holzmann, T. Renusch, G. Reichert, and H. J. Ritter, "Face generation and editing with StyleGAN: A survey," *IEEE TPAMI*, vol. 46, no. 5, pp. 3557–3576, 2024.
- [37] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.
- [38] M. Gadelha, S. Maji, and R. Wang, "3D shape induction from 2D views of multiple objects," in *3DV*, 2017.
- [39] P. Henzler, N. J. Mitra, and T. Ritschel, "Escaping plato's cave: 3D shape from adversarial rendering," in *ICCV*, 2019.
- [40] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y. Yang, "Hologan: Unsupervised learning of 3d representations from natural images," in *ICCV*, 2019.
- [41] A. Szabó, G. Meishvili, and P. Favaro, "Unsupervised generative 3D shape learning from natural images," *arXiv 1910.00287*, 2019.
- [42] Y. Liao, K. Schwarz, L. M. Mescheder, and A. Geiger, "Towards unsupervised learning of generative models for 3D controllable image synthesis," in *CVPR*, 2020.
- [43] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.
- [44] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "GRAF: generative radiance fields for 3D-aware image synthesis," in *NeurIPS*, 2020.
- [45] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "PiGAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis," in *CVPR*, 2021.
- [46] T. DeVries, M. Á. Bautista, N. Srivastava, G. W. Taylor, and J. M. Susskind, "Unconstrained scene generation with locally conditioned radiance fields," in *ICCV*, 2021.
- [47] X. Xu, X. Pan, D. Lin, and B. Dai, "Generative occupancy fields for 3d surface-aware image synthesis," in *NeurIPS*, 2021.
- [48] M. Niemeyer and A. Geiger, "GIRAFFE: representing scenes as compositional generative neural feature fields," in *CVPR*, 2021.
- [49] J. Gu, L. Liu, P. Wang, and C. Theobalt, "StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis," in *ICLR*, 2022.
- [50] Y. Xue, Y. Li, K. K. Singh, and Y. J. Lee, "GIRAFFE HD: A high-resolution 3D-aware generative model," in *CVPR*, 2022.
- [51] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman, "StyleSDF: High-resolution 3D-consistent image and geometry generation," in *CVPR*, 2022.
- [52] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3d generative adversarial networks," in *CVPR*, 2022.
- [53] K. Schwarz, A. Sauer, M. Niemeyer, Y. Liao, and A. Geiger, "VoxGRAF: Fast 3D-aware image synthesis with sparse voxel grids," in *NeurIPS*, 2022.
- [54] Y. Deng, J. Yang, J. Xiang, and X. Tong, "GRAM: generative radiance manifolds for 3D-aware image generation," in *CVPR*, 2022.
- [55] X. Zhao, F. Ma, D. Güera, Z. Ren, A. G. Schwing, and A. Colburn, "Generative multiplane images: Making a 2D GAN 3D-aware," in *ECCV*, 2022.
- [56] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.
- [57] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3d face prediction," in *CVPR*, 2020.
- [58] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE TPAMI*, vol. 36, no. 7, pp. 1325–1339, 2014.
- [59] Z. Cai, D. Ren, A. Zeng, Z. Lin, T. Yu, W. Wang, X. Fan, Y. Gao, Y. Yu, L. Pan, F. Hong, M. Zhang, C. C. Loy, L. Yang, and Z. Liu, "HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling," in *ECCV*, 2022.
- [60] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. Qian, D. Lin, and Z. Liu, "OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation," in *CVPR*, 2023.
- [61] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3D objects," in *CVPR*, 2023.
- [62] B. Wen, H. Xie, Z. Chen, F. Hong, and Z. Liu, "3d scene generation: A survey," *arXiv 2505.05474*, 2025.
- [63] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. P. Xing, "Multimodal image synthesis and editing: The generative AI era," *IEEE TPAMI*, vol. 45, no. 12, pp. 15098–15119, 2023.
- [64] A. Raistrick, L. Lipson, Z. Ma, L. Mei, M. Wang, Y. Zuo, K. Kayan, H. Wen, B. Han, Y. Wang, A. Newell, H. Law, A. Goyal, K. Yang, and J. Deng, "Infinite photorealistic worlds using procedural generation," in *CVPR*, 2023.
- [65] M. Li, A. G. Patil, K. Xu, S. Chaudhuri, O. Khan, A. Shamir, C. Tu, B. Chen, D. Cohen-Or, and H. R. Zhang, "GRAINS: generative recursive autoencoders for indoor scenes," *ACM TOG*, vol. 38, no. 2, pp. 12:1–12:16, 2019.
- [66] H. Fu, B. Cai, L. Gao, L. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, and H. Zhang, "3D-FRONT: 3D furnished rooms with layouts and semantics," in *ICCV*, 2021.
- [67] A. Raistrick, L. Mei, K. Kayan, D. Yan, Y. Zuo, B. Han, H. Wen, M. Parakh, S. Alexandropoulos, L. Lipson, Z. Ma, and J. Deng, "Infinigen Indoors: Photorealistic indoor scenes using procedural generation," in *CVPR*, 2024.
- [68] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, and V. Koltun, "CARLA: an open urban driving simulator," in *CoRL*, 2017.
- [69] S. I. Nikolenko, "Synthetic data for deep learning," *arXiv 1909.11512*, 2019.
- [70] Y. Li, X. Ran, L. Xu, T. Lu, M. Yu, Z. Wang, Y. Xiangli, D. Lin, and B. Dai, "Proc-GS: Procedural building generation for city assembly with 3D Gaussians," in *CVPR Workshops*, 2025.
- [71] Y. Jiang, H. S. Koppula, and A. Saxena, "Modeling 3D environments through hidden human context," *IEEE TPAMI*, vol. 38, no. 10, pp. 2040–2053, 2016.
- [72] D. Paschalidou, A. Kar, M. Shugrina, K. Kreis, A. Geiger, and S. Fidler, "ATISS: autoregressive transformers for indoor scene synthesis," in *NeurIPS*, 2021.
- [73] L. Gao, J. Sun, K. Mo, Y. Lai, L. J. Guibas, and J. Yang, "SceneHGN: Hierarchical graph networks for 3D indoor scene generation with fine-grained geometry," *IEEE TPAMI*, vol. 45, no. 7, pp. 8902–8919, 2023.
- [74] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *CVPR*, 2017.
- [75] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Y. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. De Nardi, M. Goesele, S. Lovegrove, and R. A. Newcombe,

- “The Replica Dataset: A digital replica of indoor spaces,” *arXiv 1906.05797*, 2019.
- [76] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. A. Funkhouser, “Semantic scene completion from a single depth image,” in *CVPR*, 2017.
- [77] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. J. Maybank, and D. Tao, “3D-FUTURE: 3D furniture shape with texture,” *IJCV*, vol. 129, no. 12, pp. 3313–3337, 2021.
- [78] T. Dai, J. Wong, Y. Jiang, C. Wang, C. Gokmen, R. Zhang, J. Wu, and L. Fei-Fei, “ACDC: Automated creation of digital cousins for robust policy learning,” in *CoRL*, 2024.
- [79] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, “D-NeRF: Neural radiance fields for dynamic scenes,” in *CVPR*, 2021.
- [80] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, “Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction,” in *CVPR*, 2024.
- [81] S. Gu, W. Yin, B. Jin, X. Guo, J. Wang, H. Li, Q. Zhang, and X. Long, “DOME: Taming diffusion model into high-fidelity controllable occupancy world model,” *arXiv 2410.10429*, 2024.
- [82] H. Bian, L. Kong, H. Xie, L. Pan, Y. Qiao, and Z. Liu, “DynamicCity: Large-scale LiDAR generation from dynamic scenes,” in *ICLR*, 2025.
- [83] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, “MaskGIT: Masked generative image transformer,” in *CVPR*, 2022.
- [84] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in *NIPS*, 2017.
- [85] S. Meister, J. Hur, and S. Roth, “UnFlow: Unsupervised learning of optical flow with a bidirectional census loss,” in *AAAI*, 2018.
- [86] S. Bond-Taylor, P. Hesse, H. Sasaki, T. P. Breckon, and C. G. Willcocks, “Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes,” in *ECCV*, 2022.
- [87] L. Feng, Q. Li, Z. Peng, S. Tan, and B. Zhou, “TrafficGen: Learning to generate diverse and realistic traffic scenarios,” in *ICRA*, 2023.
- [88] S. Ettlinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov, “Large scale interactive motion forecasting for autonomous driving : The Waymo open motion dataset,” in *ICCV*, 2021.
- [89] J. F. Canny, “A computational approach to edge detection,” *IEEE TPAMI*, vol. 8, no. 6, pp. 679–698, 1986.
- [90] T. Y. Zhang and C. Y. Suen, “A fast parallel algorithm for thinning digital patterns,” *Communications of the ACM*, vol. 27, no. 3, pp. 236–239, 1984.
- [91] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *ECCV*, 2016.
- [92] J. H. Lim and J. C. Ye, “Geometric GAN,” *arXiv 1705.02894*, 2017.
- [93] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *CVPR*, 2012.
- [94] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes dataset for semantic urban scene understanding,” in *CVPR*, 2016.
- [95] I. Nigam, C. Huang, and D. Ramanan, “Ensemble knowledge transfer for semantic segmentation,” in *WACV*, 2018.
- [96] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in *CVPR*, 2020.
- [97] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, “Playing for data: Ground truth from computer games,” in *ECCV*, 2016.
- [98] G. Ros, L. Sellart, J. Materzynska, D. Vázquez, and A. M. López, “The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *CVPR*, 2016.
- [99] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, and J. M. Álvarez, “Effective use of synthetic data for urban scene semantic segmentation,” in *ECCV*, 2018.
- [100] Y. Li, L. Jiang, L. Xu, Y. Xiangli, Z. Wang, D. Lin, and B. Dai, “MatrixCity: A large-scale city dataset for city-scale neural rendering and beyond,” in *ICCV*, 2023.
- [101] Y. Zhou, J. Huang, X. Dai, L. Luo, Z. Chen, and Y. Ma, “HoliCity: A city-scale data platform for learning holistic 3D structures,” *arXiv 2008.03286*, 2020.
- [102] Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D,” *IEEE TPAMI*, vol. 45, no. 3, pp. 3292–3310, 2023.
- [103] L. Lin, Y. Liu, Y. Hu, X. Yan, K. Xie, and H. Huang, “Capturing, reconstructing, and simulating: The UrbanScene3D dataset,” in *ECCV*, 2022.
- [104] K. Perlin, “An image synthesizer,” in *SIGGRAPH*, 1985.
- [105] Y. Shen, W. Ma, and S. Wang, “SGAM: building a virtual 3D world through simultaneous generation and mapping,” in *NeurIPS*, 2022.
- [106] L. Chai, R. Tucker, Z. Li, P. Isola, and N. Snively, “Persistent Nature: A generative model of unbounded 3D worlds,” in *CVPR*, 2023.
- [107] W. Chu, L. Ke, and K. Fragkiadaki, “Dreamscene4D: Dynamic multi-object scene generation from monocular videos,” in *NeurIPS*, 2024.
- [108] W. Sun, S. Chen, F. Liu, Z. Chen, Y. Duan, J. Zhang, and Y. Wang, “DimensionX: Create any 3D and 4D scenes from a single image with controllable video diffusion,” in *ICCV*, 2025.
- [109] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *NIPS*, 2017.
- [110] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” in *ICLR*, 2018.
- [111] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, “VBench: Comprehensive benchmark suite for video generative models,” in *CVPR*, 2024.
- [112] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE TPAMI*, vol. 44, no. 3, pp. 1623–1637, 2022.
- [113] J. L. Schönberger and J. Frahm, “Structure-from-motion revisited,” in *CVPR*, 2016.
- [114] G. Chen, G. Esch, P. Wonka, P. Müller, and E. Zhang, “Interactive procedural street modeling,” *ACM TOG*, vol. 27, no. 3, p. 103, 2008.
- [115] C. H. Lin, H. Lee, Y. Cheng, S. Tulyakov, and M. Yang, “InfinityGAN: Towards infinite-pixel image synthesis,” in *ICLR*, 2022.
- [116] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *ICCV*, 2023.
- [117] R. Anil, S. Borgeaud, Y. Wu, J. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, S. Petrov, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T. P. Lillicrap, A. Lazaridou, and et al., “Gemini: A family of highly capable multimodal models,” *arXiv 2312.11805*, 2023.
- [118] OpenAI, “GPT-4o system card,” *arXiv 2410.21276*, 2024.
- [119] H. Dong, Z. Kang, W. Yin, X. Liang, C. Feng, and J. Ran, “Scalable vision language model training via high quality data curation,” in *ACL*, 2025.
- [120] S. Lu, Y. Li, Q. Chen, Z. Xu, W. Luo, K. Zhang, and H. Ye, “Ovis: Structural embedding alignment for multimodal large language model,” *arXiv 2405.20797*, 2024.
- [121] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, “Qwen2.5-VL technical report,” *arXiv 2502.13923*, 2025.
- [122] Z. Liu, Y. Dong, J. Wang, Z. Liu, W. Hu, J. Lu, and Y. Rao, “Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment,” *arXiv 2502.04328*, 2025.
- [123] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, Z. Gao, E. Cui, X. Wang, Y. Cao, Y. Liu, X. Wei, H. Zhang, H. Wang, W. Xu, H. Li, and et al., “InternVL3: exploring advanced training and test-time recipes for open-source multimodal models,” *arXiv 2504.10479*, 2025.
- [124] H. Wang, J. Chen, W. Huang, Q. Ben, T. Wang, B. Mi, T. Huang, S. Zhao, Y. Chen, S. Yang, P. Cao, W. Yu, Z. Ye, J. Li, J. Long, Z. Wang, H. Wang, Y. Zhao, Z. Tu, Y. Qiao, D. Lin, and J. Pang, “GRUtopia: dream general robots in a city at scale,” *arXiv 2407.10943*, 2024.
- [125] Y. Hong, R. Sun, B. Li, X. Yao, M. Wu, A. Chien, D. Yin, Y. N. Wu, Z. J. Wang, and K.-W. Chang, “Embodied Web Agents: Bridging physical-digital realms for integrated agent intelligence,” *arXiv 2506.15677*, 2025.
- [126] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, “MoMask: Generative masked modeling of 3D human motions,” in *CVPR*, 2024.



Haozhe Xie received his Ph.D. from the Harbin Institute of Technology, in 2021. He is currently a research fellow at MMLab@NTU, Nanyang Technological University, Singapore. Previously, he served as a senior research scientist at Tencent AI Lab from 2021 to 2023. His research interests include computer vision with a focus on 3D generation and reconstruction. He has published several papers in CVPR, ICCV, ECCV, ICLR, and IJCV, and serves as a reviewer for these journals and conferences.



Zhaoxi Chen received the bachelor's degree from Tsinghua University, in 2021. He is currently a Ph.D. student at MMLab@NTU, Nanyang Technological University, supervised by Prof. Ziwei Liu. He received the AISG PhD Fellowship in 2021. His research interests include inverse rendering and 3D generative models. He has published several papers in CVPR, ICCV, ECCV, ICLR, NeurIPS, TOG, and TPAMI. He also served as a reviewer for CVPR, ICCV, NeurIPS, TOG, and IJCV.



Fangzhou Hong received Ph.D. degree from MMLab at Nanyang Technological University, supervised by Prof. Ziwei Liu. He received a B.Eng. degree in software engineering from Tsinghua University, China, in 2020. His research interests include computer vision and deep learning. Particularly, he is interested in 3D representation learning.



Ziwei Liu is currently an Associate Professor at Nanyang Technological University, Singapore. His research revolves around computer vision, machine learning and computer graphics. He has published extensively on top-tier conferences and journals in relevant fields, including CVPR, ICCV, ECCV, NeurIPS, ICLR, SIGGRAPH, TPAMI, TOG and Nature - Machine Intelligence. He is the recipient of PAMI Mark Everingham Prize, CVPR Best Paper Award Candidate, Asian Young Scientist Fellowship, International Congress of Basic Science Frontiers of Science Award and MIT Technology Review Innovators under 35 Asia Pacific. He serves as an Area Chair of CVPR, ICCV, ECCV, NeurIPS and ICLR, as well as an Associate Editor of IJCV.