MagicCity: Geometry-Aware 3D City Generation from Satellite Imagery with Multi-View Consistency

Xingbo Yao[1,*] Xuanmin Wang[3,*] Hao Wu[1,2,*] Chengliang Ping[1] Doudou Zhang[1] Hui Xiong[1,2,†]
[1]Hong Kong University of Science and Technology (Guangzhou)
[2]Hong Kong University of Science and Technology
[3]Tianjin University

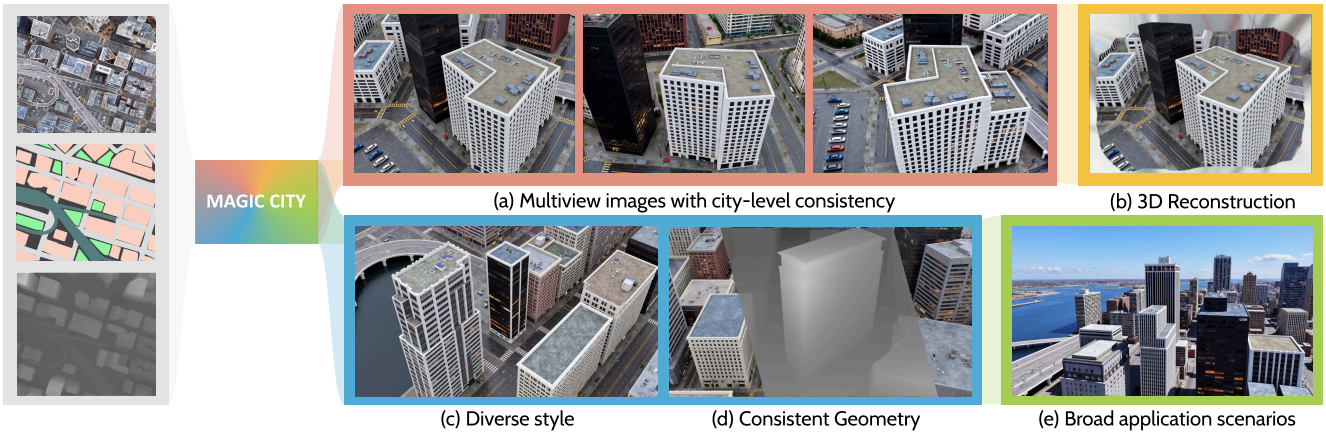

Figure 1. **Overview of MagicCity.** Given satellite input, MagicCity generates (a) multi-view images with *city-level consistency*. (b) These images are then fed into a *robust reconstruction* pipeline to generate a 3D city. Our approach achieves (c) *diverse style* generation while maintaining (d) *geometric consistency* across views. (e) The proposed method has *broad application scenarios*, including 3D modeling assets for games, urban simulation, and more.

Abstract

Directly generating 3D cities from satellite imagery opens up new possibilities for gaming and mapping services. However, this task remains challenging due to the limited information in satellite views, making it difficult for existing methods to achieve both photorealistic textures and geometric accuracy. To address these challenges, we propose MagicCity, a novel large-scale generative model for photorealistic 3D city generation with geometric consistency. Given a satellite image, our framework first extracts 3D geometric information and encodes it alongside textural features using a dual encoder. These features then guide a multi-branch diffusion model to generate city-scale, geometrically consistent multi-view images. To further enhance texture consistency across different viewpoints, we propose an Inter-Frame Cross Attention mechanism that enables feature sharing across different frames. Additionally, we incorporate a Hierarchical Geometric-Aware Module and a Consistency Evaluator to improve overall scene consistency. Finally, the generated images are fed into our robust 3D reconstruction pipeline to produce high-visual quality and geometrically consistent 3D cities. Moreover, we contribute CityVista, a high-quality dataset comprising 500 3D city scenes along with corresponding multi-view images and satellite imagery to advance research in 3D city generation. Experimental results demonstrate that MagicCity surpasses state-of-the-art methods in both geometric consistency and visual quality. Our project page: https://github.com/YaoXingbo/MagicCity

1. Introduction

3D city generation is driving innovation in gaming [16, 45], urban simulation [32, 33], and mapping services [14, 46]. Directly generating 3D cities from satellite imagery offers an efficient way to transform real-world environments into detailed digital twins. This approach not only frees designers from tedious manual tasks but also preserves the authenticity of the real city. However, the generated 3D cities often suffer from low texture quality or geometric inconsistencies. Satellite imagery offers only a top-down view and is missing crucial details such as building facades, street-level

*Equal contribution. † Corresponding author.

features, and architectural styles.

Existing approaches to city generation can generally be divided into two main categories: **geometry prior-based** methods [6, 19, 21, 22, 38–40] and **image prior-based** methods [3, 5, 8, 9, 12, 17, 34–36, 41]. Geometry prior-based methods first construct 3D city geometry from satellite imagery by extracting semantic segmentation maps and depth maps, then apply generative models for surface texture synthesis. This approach preserves geometric accuracy but struggles to produce high-quality textures for large scenes. As a result, they can only generate limited style variety and are often limited to smaller urban areas. Image-prior-based methods leverage recent advances in diffusion models to offer a promising alternative for city scene generation. They can produce high-fidelity textures using video diffusion or multi-view diffusion models. However, these frame-by-frame approaches lack 3D geometric constraints, causing geometric inconsistencies across multi-views. This problem becomes worse in city-scale scenes, where complex architecture and spatial relationships makes it difficult to maintain consistency.

To address the aforementioned challenges, we propose **MagicCity**, an innovative framework that integrates the strengths of both geometry prior-based and image prior-based methods. Our method leverages geometric and textural priors from satellite imagery as controls to guide our city-scale multi-view generative model in producing *scene-level, view-consistent* images. These images are then fed into a robust reconstruction pipeline to generate 3D cities with high-fidelity textures and consistent geometry. As shown in Figure 2, we first extract CLIP-based texture features and construct 3D geometry from semantic and depth maps, which are jointly encoded into embeddings by our Dual Encoder (DE). The embeddings guide our City-Scale Multi-View Diffusion (CMD) model to generate consistent multiple views. We adopt a progressive generation strategy where we first generate the key frames and then the remaining frames. Each frame is assigned a consistency score to quantify its cross-view consistency. Finally, these images are used to optimize a robust 3D Gaussian Splatting process, where their consistency scores guide the optimization of Gaussian point colors across views. To support city generation, we introduce the CityVista dataset, which contains 500 city scenes with paired satellite and multi-view images. Experiments show that MagicCity outperforms state-of-the-art approaches. Our method generates more *photorealistic* city scenes while maintaining *strict geometric consistency* across multiple views.

The key contributions are summarized as:

- We introduce MagicCity, a novel framework to generate photorealistic 3D cities from satellite imagery while maintaining scene-level geometric consistency.
- We propose a city-scale multi-view diffusion model that generates 3D-consistent images by incorporating explicit geometric constraints.
- We develop a robust 3D Gaussian Splatting strategy for synthesizing detailed 3D reconstructions from generated multi-view images.
- We present CityVista, a novel dataset consisting of 500 high-quality city scenes with paired with multi-view images and satellite images, to support research in 3D city generation.

2. Related Works

Existing 3D city generation methods can be broadly categorized into geometry prior-based and image prior-based approaches.

Geometry Prior-based Methods. These methods utilize semantic maps and depth information from satellite images as geometric references to construct 3D geometric structures, followed by the generation of surface textures using 3D-native generation models. InfiniCity [22] pioneered this direction by constructing octree-based voxels and applying SPADE-based neural rendering. They efficiently expanding the octree representation to support large-scale scenes while ensuring spatially consistent textures for scalable and editable 3D city generation. CityDreamer [38] further refined this approach by decomposing the scene into foreground buildings and background roads, separately training a 3D Generative Adversarial Network (GAN) on Google Earth data for each component. This method significantly enhances the visual fidelity of generated buildings. Based on that, GaussianCity [39] integrates a Gaussian Splatting framework, improving computational efficiency and enabling real-time unbounded city generation. More recent methods, such as Sat2Scene [21], focus on generating city street blocks from initial 3D structures extracted from satellite images. Followed by employing a 3D diffusion model to color sparse point clouds and a 2D diffusion model to synthesize skies. Similarly, GeoSpecific [40] predicts ground-view images at given geolocations by incorporating comprehensive satellite information, achieving a significant resolution boost.

Despite these advancements, these methods share common limitations in both style diversity and texture quality. In contrast, our method achieves superior visual fidelity and style diversity.

Image Prior-based Methods. Recent advances in image and video diffusion models have opened new possibilities for 3D content generation. Early works focused on object-level generation through multi-view synthesis [3, 34, 36] and view-consistent modeling [12, 13, 41], but struggled with larger scene synthesis. More recent approaches have attempted to address scene-level generation. CAT3D [9] proposes a two-stage framework that first generates consistent novel views through multi-view diffusion,

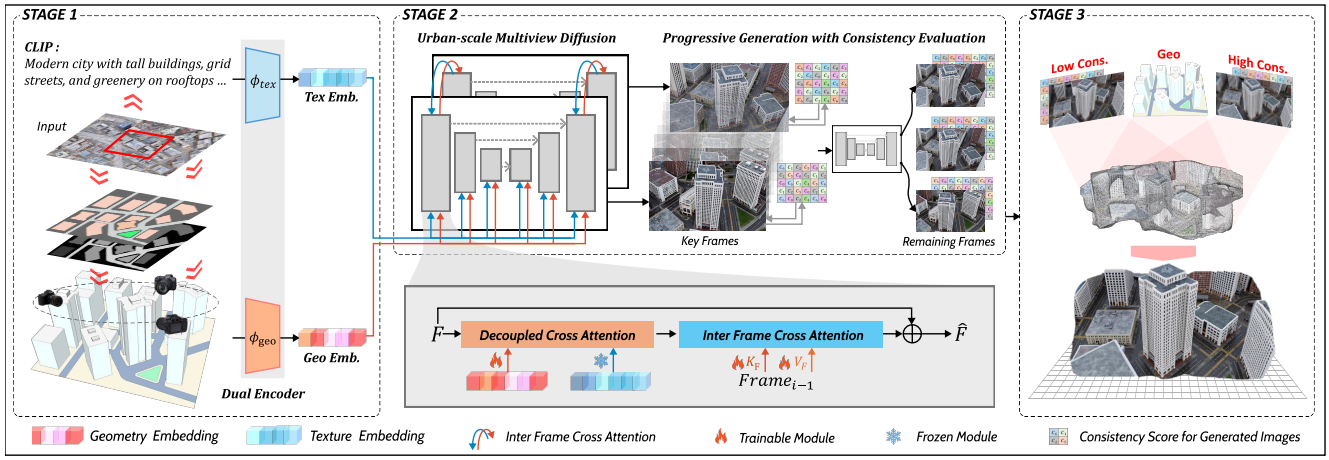

Figure 2. **Framework of MagicCity.** (a) Given a satellite image as input, we first predict its texture descriptors from CLIP *and* 3D structure from depth/segmentation maps and then encode these features through our **Dual Encoder**. (b) These encoded features are injected into the **City-scale Multi-view Diffusion** model to generate photorealistic multi-view images with *city-scale structural consistency*. Subsequently, we implement instance-level feature matching to evaluate consistency scores across generated images. (c) Finally, the generated multi-view images are fed into a robust 3D Gaussian Splatting pipeline, where instance-level consistency scores guide the color initialization and **adaptive optimization** of 3D Gaussian points across views.

followed by 3D reconstruction. DimensionX [35] introduces a spatial-temporal decomposition strategy for video generation and establishes 3D scenes through multi-loop refinement. DreamScene [17] employs Formation Pattern Sampling and progressive camera strategies to generate 3D-consistent scenes. However, these methods are limited to simple scenarios like single buildings or natural scenes, and often produce results with noticeable synthetic artifacts. While works like MagicDrive [8] and Streetscape [5] demonstrate impressive progress in generating high-quality street-view videos of urban environments, their frame-by-frame generation approach leads to geometric inconsistencies across views, resulting in distorted 3D reconstruction that deviates from real-world structure.

In contrast, our method incorporates geometric constraints into the multi-view generation process and designs specific strategies to ensure city-level geometric consistency across the generated views.

3. Methods

Our MagicCity follows a novel 3D-consistent framework for 3D city generation. First, we initialize the scene structure by combining depth maps and semantic maps while extracting texture information from CLIP. These two modalities are encoded into feature embeddings via our Dual Encoder (Section 3.1). Next, these features are fed into our City-scale Multi-view Diffusion model to generate 3D-consistent images. We employ a progressive generation strategy that first synthesizes critical key frames, followed by the remaining frames. Each generated image is assigned with a consistency score to quantify its cross-view consistency (Section 3.2). Finally, the generated multi-view frames undergo a robust 3D Gaussian Splatting process. During this stage, the initial 3D structure from the first stage initializes Gaussian point positions, while consistency scores guide both color initialization and iterative optimization of Gaussian points using the generated images.(Section 3.3).

3.1. Scene Initialization and Dual Encoding

Motivation. Prior works excel at generating photorealistic videos of city scenes, but they struggle to maintain geometric consistency across different views. To address this, we leverage satellite data to initialize the 3D scene geometry, followed by encoding both the geometric structure and satellite texture into embeddings using our designed Dual Encoder. These embeddings serve as constraints to ensure consistency in multi-view generation. The details are as follows:

3D Structure Generation. As shown in Figure 2, we first process the satellite image I to obtain instance segmentation S and depth estimation D. The 3D volume $V \in \mathbb{R}^{H\times W\times D}$ is then constructed by lifting each pixel (i, j) from S to its corresponding 3D position according to D, where each voxel $v(x, y, z)$ stores its instance ID and semantic class.

Dual Encoder. After obtaining the 3D volume from satellite imagery, we design a Dual Encoder that processes geometry and texture information into features separately:

Texture Branch: We employ CLIP [28] to extract textual descriptions T from I, capturing architectural characteristics such as building styles, materials, and regional features.

Through the proposed encoder ϕ_{tex}, these semantic tokens $T \in \mathbb{R}^K$ are transformed into texture features $f_{tex} \in \mathbb{R}^D$ that guide consistent texture generation across views.

Geometric Branch: To encode geometric properties, we represent the 3D volume using view-specific features. Specifically, we render the 3D volume V from the input trajectory, which consists of N camera positions $\{v_i\}_{i=1}^N$. The rendering process can be formulated as:

$$\{S_i, D_i\}_{i=1}^N = R(V, \{v_i\}_{i=1}^N) \tag{1}$$

where R denotes the rendering function, v_i represents the i-th camera parameters, S_i and D_i are the corresponding instance segmentation and depth maps. Notably, if no predefined camera inputs are available, the viewpoints are uniformly sampled along a circular trajectory. For each viewpoint i, we concatenate S_i, D_i with camera parameters v_i into $F_i \in \mathbb{R}^{H\times W\times 3}$. These features are then processed through the proposed geometric encoder ϕ_{geo} to obtain per-view geometric features $f_{geo}^i \in \mathbb{R}^D$ that ensure structural consistency during generation.

3.2. City-scale Multi-view Generation with Consistency Evaluation

Motivation. City-scale multi-view generation is more challenging than object-level tasks due to the extensive spatial range and complex interactions between multiple buildings and roads [5, 38]. To address these challenges, we propose the City-scale Multi-view Diffusion (CMD) that generates consistent multiple views through a multi-branch diffusion architecture. As shown in Figure 2, each branch follows the LDM framework [30] but incorporates three key modifications for large-scale scene generation: (1) Dual embeddings injection at each UNet block, (2) Inter-Frame Consistency Attention (IFCA) between branches, and (3) Hierarchical Geometric Aware module in middle layers. Furthermore, we adopt a progressive generation strategy with consistency evaluation, where key frames are first generated and evaluated by our proposed consistency evaluator before the parallel generation of the remaining frames. The details are as follows:

Network Architecture. Our CMD model first generates multiple key frames in parallel. Taking the i-th frame as an example, let F_i denote its feature map and $\{F_r\}$ represent features of other frames. Within each block, we inject dual-encoded constraints through decoupled cross attention:

$$\tilde{F}_i = F_i + \mathcal{A}(F_i, f_{tex}) + \mathcal{A}(F_i, f_{geo}^i) \tag{2}$$

where $\mathcal{A}(\cdot,\cdot)$ represents the cross-attention operation that enables dynamic feature selection from texture and geometric constraints, respectively. Inspired by IP-Adapter [43], we adopt an asymmetric training strategy: the geometric constraint branch is trainable, while the texture constraint branch remains frozen. This design stabilizes training by preserving structural priors while allowing the model to adaptively learn texture variations.

To further maintain consistency across views, we propose the Inter-Frame Consistency Attention (IFCA) module that operates across multiple frames. The inter-frame feature aggregation is computed as:

$$M_i = \sum_{l\neq i} \text{softmax}(W_q\tilde{F}_i \cdot W_k\tilde{F}_r^T) \cdot W_v\tilde{F}_r \tag{3}$$

where $W_Q\tilde{F}_i$ acts as the query, encoding the content of the current frame i, while $W_K\tilde{F}_r$ and $W_V\tilde{F}_r$ serve as the key and value from a neighboring frame r. The dot product $W_q\tilde{F}_i \cdot W_k\tilde{F}_r^T$ measures the similarity between corresponding regions across frames, and the softmax operation determines how much information from frame r should be transferred to frame i. By aggregating relevant features from adjacent frames, IFCA reduces temporal flickering and improves coherence in the generated views.

Additionally, we introduce a Hierarchical Geometric-Aware Attention (HGA) module in the middle blocks with dual attention paths: a scene-level path (4 heads $\times$ 128 dim) for global layout and an object-level path (8 heads $\times$ 64 dim) for local structural details.

The training objective is defined as:

$$L := \mathbb{E}_{\{z_t^i\},\varepsilon_i,t,c}\left[\sum_{i=1}^N \|\varepsilon^i - \varepsilon_\theta^i(\{z_t^i\}, t, \tau_\theta(c)\|_2^2\right] \tag{4}$$

where $\{z_t^i\}$ denotes the noisy latents of N views at timestep t, $\varepsilon^i \sim \mathcal{N}(0,1)$ is the sampled noise, and $\tau_\theta(c)$ encodes our consistency constrains into diffusion condition.

Progressive Generation with Consistency Evaluation. Similar to CAT3D [9], we adopt an efficient progressive generation strategy that first synthesizes key frames at anchor viewpoints and then generates the remaining views. Unlike previous works, a consistency evaluator is introduced to compute the consistency score of each image. The image generation process is iteratively refined until it meets a predefined threshold. Consistency scores are computed based on instance-level feature matching. Specifically, for each instance i in frame k, its consistency score C_i^k is computed as:

$$C_i^k = \frac{1}{|V_i|}\sum_{j\in V_i} \cos(f_i^k, f_i^j) \tag{5}$$

where V_i denotes the set of views containing instance i, f_i^k represents the average DINO [27] features of instance i in frame k, and $\cos(\cdot,\cdot)$ computes cosine similarity. The overall consistency score for frame k is calculated as:

$$C^k = \sum_i w_i C_i^k, \quad w_i = \frac{A_i}{\sum_j A_j} \tag{6}$$

where A_i is the pixel area of instance i, serving as a natural weight for the instance's contribution to the overall consistency. We proceed to generate intermediate views only when the consistency scores of all key frames exceed a threshold τ. Additionally, all remaining images are also assigned a consistency score to evaluate their cross-view consistency.

3.3. Consistency Score-guided 3D Reconstruction

Motivation. Our multi-view generation framework produces high-quality views that maintain *scene-level geometric consistency*. However, even state-of-the-art video generation models [2, 42, 44, 47] cannot ensure perfect *pixel-level texture consistency*. 3D Gaussian Splatting (GS) [15] relies on feature matching for point cloud initialization and refines Gaussian points through iterative multi-view rendering, pixel-level inconsistencies in the input views can lead to sparse or failed initialization and noticeable artifacts during optimization. To address this, we propose a robust 3D Gaussian splatting pipeline that leverages the consistency score to guide the reconstruction process.

Point Cloud Initialization. Point initialization in GS requires accurate 3D point positions and colors. For a 3D point p, its position is initialized using the 3D structure from Section 3.1. For color, the RGB value may vary across different views containing p. To address this, we use the consistency scores and place more weight on views with higher consistency scores during the color initialization process. Formally, the initial color of point c_p is computed as follows:

$$c_p = \frac{\sum_k (C_i^k \cdot c_k)}{\sum_k C_i^k} \tag{7}$$

where c_k is the color from view k, and C_i^k is the consistency score of instance i in view k.

Adaptive Optimization Strategy. Similarly, we use the consistency scores of multi-view images to guide the optimization of Gaussian points. Views with higher consistency scores are given more weight during the optimization process. Formally, for each point p, the reconstruction loss is weighted as:

$$L_p = L_{\text{render}}^p \cdot C_p \tag{8}$$

where L_{render}^p is the standard rendering loss for point p, and C_p is the consistency score of the current view. This allows us to reconstruct high-quality 3D cities from the generated views.

4. Experiments

4.1. Dataset

CityVista Dataset. To the best of our knowledge, there is currently no dataset that provides high-quality 3D cities with paired multi-view images and satellite imagery. To bridge this gap, we introduce CityVista dataset, comprising 500 city scenes from two sources: (1) Since the Matrixcity dataset [20] provides high-quality drone imagery of urban environments, we perform reconstruction and split them into 300 distinct city scenes; (2) We collect approximately 2,000 3D assets from Sketchfab[*] and CitySample[†], normalize their scales, and assemble them into 200 synthetic city scenes using methods proposed by CityCraft [7].

For all 500 scenes, we render images along diverse camera trajectories including orbital paths (radius: 50–500m, altitude: 50–200m), forward-facing circles, and exploratory spline paths. Each scene contains 60 high-resolution (1920×1080) images with corresponding satellite views. Moreover, we render satellite viewpoints for every scene to provide paired aerial imagery. Finally, we provide comprehensive annotations for all multi-view images, including instance segmentation for buildings, roads, vegetation, and water bodies, as well as the paired depth maps. The final dataset contains 30,000 views spanning diverse architectural styles and city layouts.

OSM Dataset. Following CityDreamer [38], we also utilize the OpenStreetMap (OSM) dataset[‡] for additional testing and evaluation. The OSM dataset contains satellite imagery from 80 cities worldwide, along with corresponding semantic maps providing five-category classification (roads, buildings, green lands, construction sites, and water areas) and height fields derived from OSM data. This large-scale dataset serves as an extended evaluation benchmark for our method.

4.2. Implementation Details

Our generative model builds upon the Latent Diffusion Model (LDM) framework [30]. Following Section 3.2, we adapt it for city generation. For training, the dataset is resized to 1280×720 resolution, which improves training efficiency while maintaining good visual quality. We use the AdamW optimizer with a learning rate of $1e^{-5}$ on 4 A800 GPUs. We first fine-tuned the model on single-view images. Then, we froze the backbone and trained the CMD blocks using multi-view images. The consistency threshold τ of our consistency evaluation process is set to 0.85, and images are iteratively generated until this threshold is satisfied. For robust 3D reconstruction, we apply a $0.1\times$ standard learning rate for low-confidence regions during optimization.

4.3. Evaluation Protocols

We randomly select 500 satellite images from the test set for evaluation. To ensure fair testing, we apply the same preprocessing method in Sec. 3.1 to obtain semantic maps and

[*] https://sketchfab.com

[†] https://www.unrealengine.com/marketplace/en-US/product/city-sample

[‡] https://openstreetmap.org

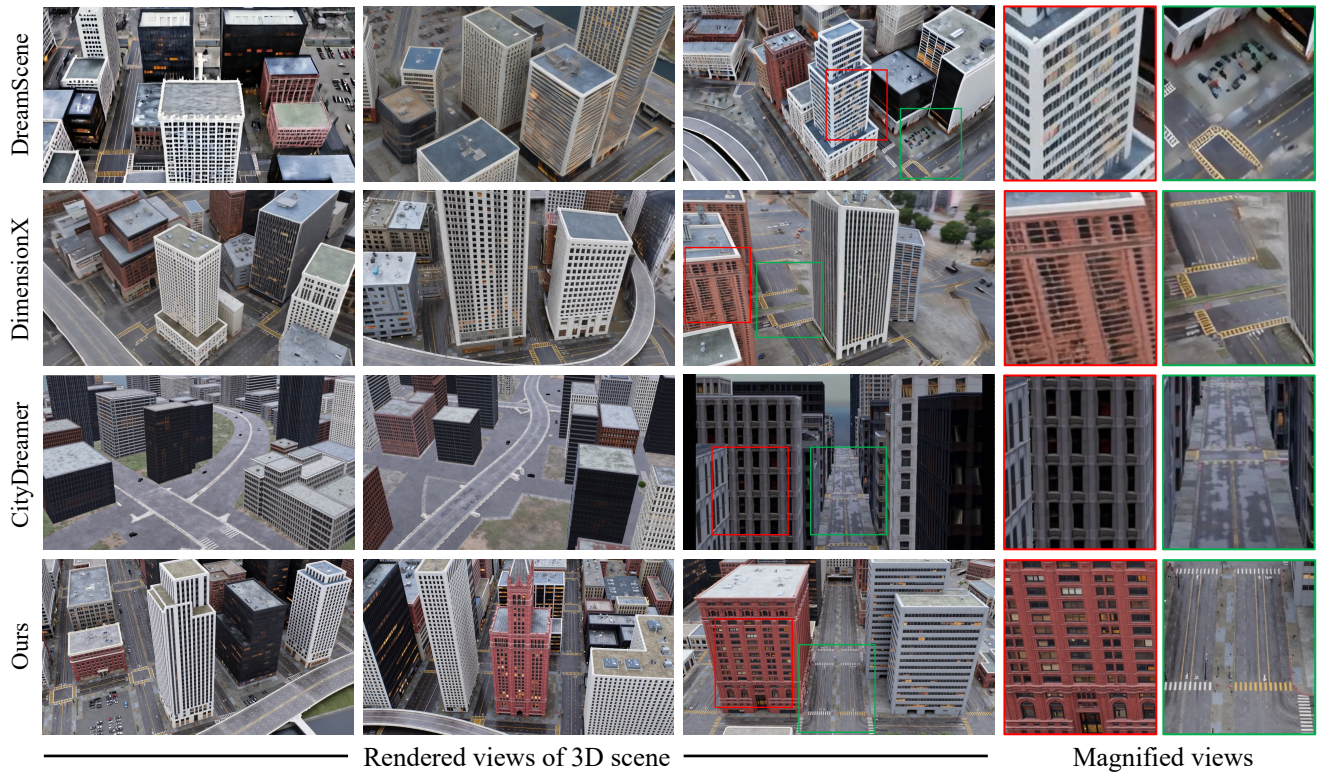

Figure 3. **Qualitative comparison.** Our method produces higher-quality 3D cities with better consistency compared with the baselines. We strongly recommend **zooming in** to examine the detailed differences.

Method	FID ↓	KID ↓	NIQE ↓	BRISQUE ↓	DE ↓	CE ↓
CityDreamer [38]	155.390	$0.251^{\pm 0.012}$	$8.632^{\pm 0.709}$	$86.773^{\pm 11.492}$	0.157	0.083
DimensionX [35]	126.890	$0.175^{\pm 0.006}$	$6.595^{\pm 0.425}$	$70.207^{\pm 7.157}$	-	-
DreamScene [17]	104.627	$0.125^{\pm 0.002}$	$6.018^{\pm 0.671}$	$30.311^{\pm 11.732}$	0.223	0.371
Ours	**86.096**	$\mathbf{0.087}^{\pm 0.001}$	$\mathbf{4.553}^{\pm 0.412}$	$\mathbf{28.018}^{\pm 5.634}$	**0.137**	**0.072**

Table 1. **Quantitative comparison**. MagicCity outperforms the baselines across all metrics, including visual quality and geometric consistency. Lower values indicate better performance.

Dataset	Images	Resolution	View	3D mod.	Anno.
KITTI [10]	15k	1242 ×375	street	×	sem.
OmniCity [18]	6k	512 ×512	street/sate.	×	ins./plane
GoogleEarth [38]	24k	960 ×540	drone	×	sem./ins.
HoliCity [49]	108k	512 ×512	street	✓	sem./plane
UrbanScene3D [23]	6.1k	800 ×450	drone	✓	ins.
Ours	30k	1920 ×1080	drone/sate.	✓	sem./ins.

Table 2. **Comparison of MagicCity with city-related datasets.** "sate.", "sem.", "plane", and "ins." denote satellite, semantic segmentation, plane segmentation, and instance segmentation respectively. "Anno.", "3D mod." represents the annotation types and the availability of 3D models respectively.

depth maps for all approaches. For each scene, after generating 3D cities using all methods, we render 60 frames with a resolution of 1280×720 under diverse camera trajectories (i.e., circular, spiral). Subsequently, we randomly select frames for visualization and quantitative evaluation.

Image Quality Metrics. Following previous works [21, 38, 48], we employ Fréchet Inception Distance (FID) [11] and Kernel Inception Distance (KID) [1] to assess the distribution similarity between generated and real city scenes. FID compares the mean and covariance of features extracted from an Inception network, while KID uses a polynomial kernel to compare distributions. These metrics are computed between 10k randomly sampled generated views and 10k real street-view images from the dataset. Additionally, we use no-reference image quality metrics

BRISQUE [25] and NIQE [26] to evaluate the naturalness and perceptual quality of generated views. BRISQUE measures naturalness by analyzing spatial features, and NIQE evaluates image naturalness based on statistical properties. We evaluate BRISQUE and NIQE on 10k randomly sampled rendered images.

Geometric Consistency Metrics. Following CityDreamer [38], we evaluate geometric consistency using two metrics. Camera Error (CE) measures the scale-invariant L_2 distance between reconstructed and ground-truth camera poses, quantifying the difference between the inference camera trajectory and the estimated trajectory from COLMAP [31]. Depth Error (DE) evaluates 3D geometry accuracy by computing the normalized L_2 distance between predicted and pseudo ground-truth depth maps. Inspired by EG3D [4], we generate pseudo ground-truth using a pretrained depth estimation model [29]. Both the predicted and reference depths are obtained by applying the same model to rendered RGB images. To eliminate scale ambiguity, the depth maps are normalized to zero mean and unit variance before computing DE.

4.4. Main Results

Comparison Methods. We evaluate our approach against the recent state-of-the-art methods: DreamScene [17], CityDreamer [38], and DimensionX [35]. For fair comparison, all methods are retrained on the CityVista dataset except DimensionX, whose code has not been available.

Qualitative Comparison. Figure 3 shows the render results from diffusion models, both DreamScene, CityDreamer, and our method demonstrate strong *geometric consistency* due to their geometry-aware generation processes. In contrast, DimensionX, which relies on controllable video diffusion without explicit geometric constraints, exhibits noticeable distortions in building structures and road layouts.

Regarding *visual quality*, our method produces realistic textures for both building facades and ground details. The magnified view demonstrates that our texture details outperform the comparison methods. In contrast, CityDreamer shows limited style diversity in building colors and road textures. DimensionX generates scenes with greater style variation than CityDreamer but exhibits significant distortion and defects in the details, as shown in the magnified view. Additionally, DreamScene produces results with pronounced synthetic artifacts despite also generating diverse styles.

Quantitative Comparison. Table 1 presents the quantitative evaluation results. Our method demonstrates significant improvements in distribution metrics (FID and KID), indicating better alignment with real city scenes. This is further supported by our results shown in Figure 3. The superior performance in no-reference metrics (NIQE and BRISQUE) validates the enhanced perceptual quality of our results. Furthermore, our method achieves the lowest DE and CE scores, confirming our ability in maintaining geometric accuracy and cross-view consistency.

4.5. Ablation Study

We further conduct ablative experiments to validate the effectiveness of the three components utilized in our method.

Effectiveness of DE. The Dual Encoder (DE) encodes 3D geometric features and textural features into embeddings, and injects them into our generative model through cross-attention mechanisms. We compare our approach with two baselines: removing the encoder entirely and directly concatenating the two embeddings instead of using cross-attention. All other components remain unchanged during retraining. Results are shown in Table 3, where our DE achieves better performance across all metrics. This is especially evident on geometric metrics DE and CE, demonstrating the effectiveness of our approach in preserving geometric accuracy.

Effectiveness of CMD. Our City-Scale Multi-view Diffusion (CMD) model incorporates three key components: Inter-Frame Consistency Attention (IFCA) for enabling block-level cross-attention between different diffusion branches during parallel multi-view image generation. Hierarchical Geometric Aware attention (HGA) for capturing global-local relationships in the middle layers of the UNet, and Consistency Evaluation strategy that calculates a consistency score for each image, continuing the iterative generation process only until a threshold is reached. We evaluate the contribution of each component by removing them individually while keeping other parts unchanged. As shown in Table 4, removing each component leads to degraded performance, particularly in geometric consistency metrics (DE and CE). This demonstrates the effectiveness of our proposed approach in maintaining geometric consistency and enhancing texture quality.

Comparison on 3D Reconstruction. To validate the effectiveness of our Consistency Score-guided reconstruction strategy (CSR), we compare it with ReconFusion [37], a state-of-the-art method for robust 3D reconstruction from generated images. As shown in Figure 4, our method achieves more accurate 3D geometry with fewer artifacts, particularly in regions with pixel-level color inconsistencies. All results are visualized using 3D Gaussian Splatting (3DGS).

4.6. Multi-view Generation

To further demonstrate the capability in generating large-scale consistent multi-view of our MagicCity, we conduct additional experiments on 360° scene synthesis. We compare our approach with recent state-of-the-art methods in multi-view generation (Wonder3D [24]), novel view syn-

Method	FID↓	NIQE↓	DE↓	CE↓
Baseline (w/o all)	95.23	6.124	0.315	0.152
Direct Concat	88.12	5.235	0.189	0.092
Ours (DE)	**86.09**	**4.552**	**0.137**	**0.071**

Table 3. **Ablation study on the effectiveness of Dual Encoder.** The best values are highlighted in blod. Note that "w/o all" denotes the removal of both geometric and textural embeddings from our encoder.

Method	FID↓	NIQE↓	DE↓	CE↓
w/o IFCA	89.23	4.892	0.186	0.098
w/o HGA	87.56	4.763	0.159	0.085
w/o CE	88.12	4.835	0.192	0.102
Ours	**86.09**	**4.552**	**0.137**	**0.071**

Table 4. **Ablation study on the effectiveness of CMD.** The best values are highlighted in blod. Note that "IFCA" denotes Inter-Frame Consistency Attention, "HGA" represents the Hierarchical Geometric Aware module, and "CE" refers to Consistency Evaluation.

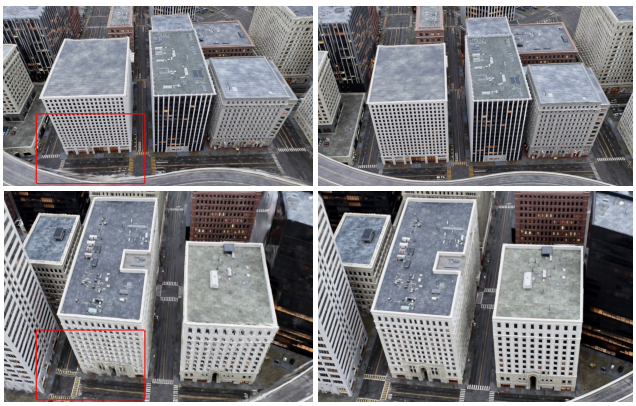

Figure 4. **Comparison on 3D reconstruction results.** Our method (CSR) produces more robust 3D reconstructions with fewer artifacts compared to the baseline ReconFusion. Note that the input views are in drone-view perspective, which are generated from our multi-view diffusion model.

thesis (ViewCraft[44]), and camera-controlled video generation (DimensionX[35]). To ensure consistent input across all methods, we randomly select from drone-view perspectives in our dataset. For the baseline methods, we used the default parameters from their open-source implementations.

As shown in Figure 5, given the same input view, our method maintains superior scene-level consistency across the novel viewpoints. The generated views demonstrate *accurate building structures* and *consistent texture patterns*

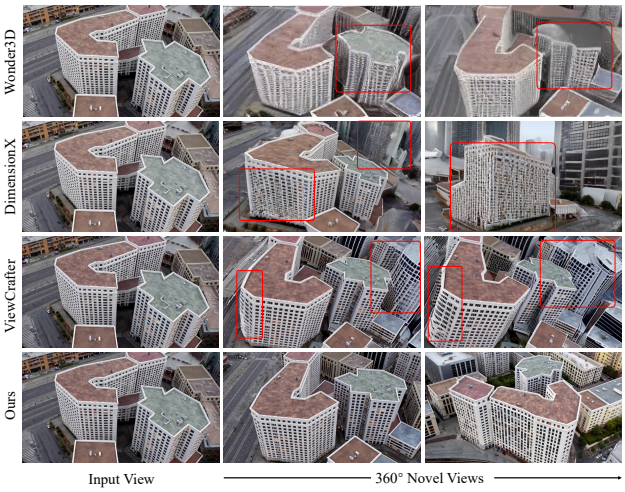

Figure 5. **Multi-view generation comparison.** Our method maintains consistent structure and texture across viewpoints, outperforming the baseline methods.

throughout the sequence. In contrast, baseline methods show notable limitations in the regions marked by the red boxes. Specifically, Wonder3D fails to preserve building proportions and exhibits significant distortion of building geometry in challenging viewpoints. DimensionX performs relatively well in maintaining building geometry but struggles with large-angle changes, where facades become distorted. In ViewCrafter, buildings also become distorted with camera rotation, resulting in the loss of architectural features across different angles.

5. Conclusion

In this paper, we propose a large-scale generative framework for 3D city generation. Comparing to existing methods that struggle to simultaneously achieve geometric consistency and photorealistic textures, MagicCity ensures high-fidelity city scene synthesis with multi-view geometric consistency. This is achieved through our dual encoder architecture, multi-branch diffusion model, and robust reconstruction strategy. Furthermore, we introduce the CityVista dataset, comprising 500 city scenes with paired multi-view images and satellite imagery to support 3D generation research. Experimental results demonstrate that our approach surpasses state-of-the-art methods in both visual realism and structural accuracy.

Acknowledgments. This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Education Bureau of Guangzhou.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 5

[3] Emmanuelle Bourigault and Pauline Bourigault. Mvd-iff: Scalable and flexible multi-view diffusion for 3d object reconstruction from single-view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7586, 2024. 2

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 7

[5] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 4

[6] Jie Deng, Wenhao Chai, Jianshu Guo, Qixuan Huang, Wenhao Hu, Jenq-Neng Hwang, and Gaoang Wang. Citygen: Infinite and controllable 3d city layout generation. *arXiv preprint arXiv:2312.01508*, 2023. 2

[7] Jie Deng, Wenhao Chai, Junsheng Huang, Zhonghan Zhao, Qixuan Huang, Mingyan Gao, Jianshu Guo, Shengyu Hao, Wenhao Hu, Jenq-Neng Hwang, et al. Citycraft: A real crafter for 3d city generation. *arXiv preprint arXiv:2406.04983*, 2024. 5

[8] Ruiyuan Gao, Kai Chen, Enze Xie, HONG Lanqing, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 3

[9] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin Brualla, Pratul Srinivasan, Jonathan Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 37:75468–75494, 2025. 2, 4

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. The kitti vision benchmark suite. *URL http://www. cvlibs. net/datasets/kitti*, 2(5):1–13, 2015. 6

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5043–5052, 2024. 2

[13] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 2

[14] Junya Kanda, Yi He, Haoran Xie, and Kazunori Miyata. Sketch2tooncity: sketch-based city generation using neurosymbolic model. In *International Workshop on Advanced Imaging Technology (IWAIT) 2024*, pages 431–436. SPIE, 2024. 1

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 5

[16] Joon-Seok Kim, Hamdi Kavak, and Andrew Crooks. Procedural city generation beyond game development. *SIGSPATIAL Special*, 10(2):34–41, 2018. 1

[17] Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-hang Lee, and Peng Yuan Zhou. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *European Conference on Computer Vision*, pages 214–230. Springer, 2024. 2, 3, 6, 7

[18] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17397–17407, 2023. 6

[19] Weijia Li, Jun He, Junyan Ye, Huaping Zhong, Zhimeng Zheng, Zilong Huang, Dahua Lin, and Conghui He. Crossviewdiff: A cross-view diffusion model for satellite-to-street view synthesis. *arXiv preprint arXiv:2408.14765*, 2024. 2

[20] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023. 5

[21] Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, and Martin R Oswald. Sat2scene: 3d urban scene generation from satellite images with diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7150, 2024. 2, 6

[22] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22808–22818, 2023. 2

[23] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022. 6

[24] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Sin-

gle image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 7

[25] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708, 2012. 7

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[29] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 7

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 5

[31] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 7

[32] Yu Shang, Yuming Lin, Yu Zheng, Hangyu Fan, Jingtao Ding, Jie Feng, Jiansheng Chen, Li Tian, and Yong Li. Urbanworld: An urban world model for 3d city generation. *arXiv preprint arXiv:2407.11965*, 2024. 1

[33] Bingyu Shen, Boyang Li, and Walter J Scheirer. Automatic virtual 3d city generation for synthetic data collection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 161–170, 2021. 1

[34] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[35] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 3, 6, 7, 8

[36] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2

[37] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21551–21561, 2024. 7

[38] Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. Citydreamer: Compositional generative model of unbounded 3d cities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9666–9675, 2024. 2, 4, 5, 6, 7

[39] Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. Gaussiancity: Generative gaussian splatting for unbounded 3d city generation. *arXiv preprint arXiv:2406.06526*, 2024. 2

[40] Ningli Xu and Rongjun Qin. Geospecific view generation geometry-context aware high-resolution ground view inference from satellite views. In *European Conference on Computer Vision*, pages 349–366. Springer, 2024. 2

[41] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7079–7088, 2024. 2

[42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *CoRR*, 2024. 5

[43] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 4

[44] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 5, 8

[45] Shougao Zhang, Mengqi Zhou, Yuxi Wang, Chuanchen Luo, Rongyu Wang, Yiwei Li, Zhaoxiang Zhang, and Junran Peng. Cityx: Controllable procedural content generation for unbounded 3d cities. *arXiv preprint arXiv:2407.17572*, 2024. 1

[46] Fangshuo Zhou, Huaxia Li, Rui Hu, Sensen Wu, Hailin Feng, Zhenhong Du, and Liuchang Xu. Controlcity: A multimodal diffusion model based approach for accurate geospatial data generation and urban morphology analysis. *arXiv preprint arXiv:2409.17049*, 2024. 1

[47] Qiang Zhou, Shaofeng Zhang, Nianzu Yang, Ye Qian, and Hao Li. Motion control for enhanced complex action video generation. *arXiv preprint arXiv:2411.08328*, 2024. 5

[48] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision*, pages 324–342. Springer, 2024. 6

[49] Yichao Zhou, Jingwei Huang, Xili Dai, Shichen Liu, Linjie Luo, Zhili Chen, and Yi Ma. Holicity: A city-scale data platform for learning holistic 3d structures. *arXiv preprint arXiv:2008.03286*, 2020. 6