WorldFlow3D: Flowing Through 3D Distributions for Unbounded World Generation

Amogh Joshi[1*], Julian Ost[1*], Felix Heide[1, 2]

[1]Princeton University [2]Torc Robotics

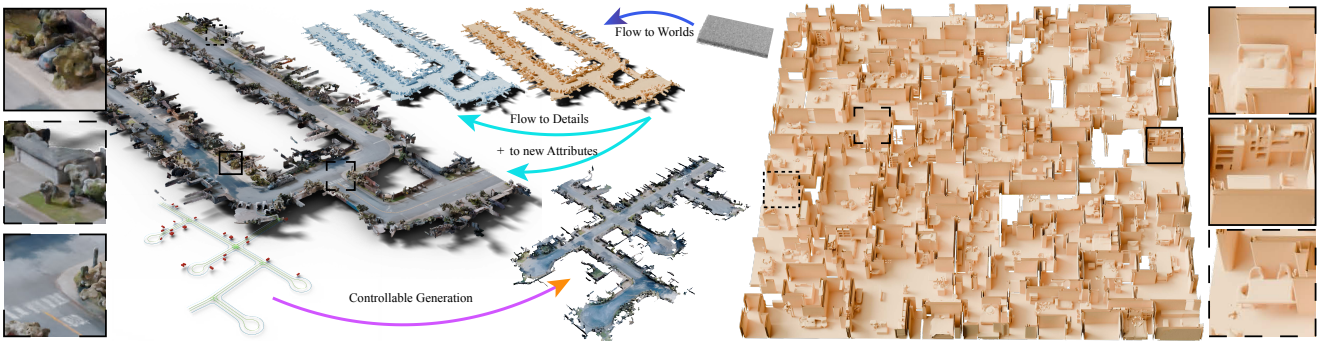


Figure 1. **WorldFlow3D** is a novel method for the generation of unbounded 3D worlds. We show the capabilities of WorldFlow3D for the generation of large-scale outdoor and indoor scenes, with insets showing learned distributions of fine geometric detail and realistic texture.

Abstract

Unbounded 3D world generation is emerging as a foundational task for scene modeling in computer vision, graphics, and robotics. In this work, we present WorldFlow3D, a novel method capable of generating unbounded 3D worlds. Building upon a foundational property of flow matching – namely, defining a path of transport between two data distributions – we model 3D generation more generally as a problem of flowing through 3D data distributions, not limited to conditional denoising. We find that our latent-free flow approach generates causal and accurate 3D structure, and can use this as an intermediate distribution to guide the generation of more complex structure and high-quality texture – all while converging more rapidly than existing methods. We enable controllability over generated scenes with vectorized scene layout conditions for geometric structure control and visual texture control through scene attributes. We confirm the effectiveness of WorldFlow3D on both real outdoor driving scenes and synthetic indoor scenes, validating cross-domain generalizability and high-quality generation on real data distributions. We confirm favorable scene generation fidelity over approaches in all tested settings for unbounded scene generation. For more, see `https://light.princeton.edu/worldflow3d`.

*Equal contribution.

1. Introduction

Developing spatially intelligent systems in large-scale environments has long been a central pursuit in computer vision and robotics. A core display of intelligence in this context is the ability to synthesize and reason over realistic 3D models of the real world. This implicitly demonstrates coherent world understanding and processing of spatial relationships centered around visual and geometric causality. A growing body of recent work has enabled high-quality *3D reconstructions* from real-world scene captures [1, 33]. Recent learned neural scene representations are capable of producing both implicit [2, 25] and explicit [15, 16] 3D models from images. Scene modeling by reconstruction, however, is fundamentally constrained by a reliance on real data – naturally translating into a need for purely generative approaches for producing unlimited data.

Modern *3D generation* approaches [38, 43] have shown great success in object-level generation, with high fidelity in both structure and visual texture. However, modeling large-scale, realistic 3D *scenes* requires a distinct level of 3D scene understanding, consisting of objects within a broader spatial domain and environmental context. Procedural modeling methods are capable of producing theoretically unbounded scenes [19, 29, 30, 36], but their hand-crafted rule-based approach comes at the cost of not only photorealism in texture but also realism in structure. Real-world environments, in contrast, exhibit vast diversity in scale, struc-

ture, and appearance. Some works have shown the ability to model large synthetic environments [18, 24, 42], but real-world, open-world scenes are more complex. Open-world outdoor scenes, such as driving scenes [37], are fundamentally structurally sparse – preventing existing unbounded approaches from translating to such environments. More recent scene-focused 3D generation approaches use hierarchical 3D latent diffusion [17, 24, 42, 46], but are either constrained to a specific data distribution [24], prohibiting generalizability, or are fixed in spatial extent [31]. Therefore, as summarized in Table 1, a method capable of producing unbounded scenes with high-fidelity geometry and texture, *and* allowing full controllability across domains, remains an open challenge.

We introduce WorldFlow3D, a novel approach for generating unbounded 3D worlds with full controllability. WorldFlow3D is built on a foundational property of flow matching [6, 20] – defining a path of transport between *any* two data distributions. Building upon this, we reformulate 3D generation not as a problem of progressive hierarchical conditional denoising but as flowing through sequential 3D data distributions. As such, WorldFlow3D directly produces volumetric scene representations successively from noise, through coarse structure, and into detailed, causal geometry and high-fidelity texture – all modeled as transport through data distributions. Our approach allows for latent-free, purely volumetric generative models, breaking from the standard autoencode $\rightarrow$ generate paradigm of existing methods [24, 31, 43].

We validate WorldFlow3D for unbounded scene generations across real, open-world outdoor driving scenes [37] and synthetic indoor rooms [12] – confirming that our method obtains high quality across multiple distinct data distributions. We allow for explicit 3D controllability through vectorized scene layouts for structure and scene attributes for texture. Our flow matching formulation also enables rapid, latent-free training convergence even on complex 3D data distributions, and efficient inference for generating large-scale worlds. We introduce an extension to existing schedulers by aligning predicted flow fields across smaller chunks at inference time, unlocking truly unbounded scene generation (limited only by compute) without visible border artifacts. We measure major improvements upon all existing tested methods, across multiple data domains. Additionally, visual analysis and a blind user study confirms the significance of our results qualitatively.

We summarize our contributions as follows:

- We introduce a novel 3D world generation method that formulates 3D generation as flow matching across 3D data distributions.
- Our proposed method allows for latent-free generation of scenes with (a) unbounded spatial extent, (b) high-quality geometric structure and realistic visual texture, (c) full controllability over scene layout and visual attributes.
- We validate our method with large-scale generations across domains, confirming favorable 3D geometric and texture fidelity in all experiments.

2. Related Work

3D Object Generation and Procedural Scene Generation. Recent advances in 3D object generation have demonstrated remarkable capabilities in synthesizing high-quality textured 3D assets. Object generation methods [38, 43], and more generally diffusion-based [5, 8, 11, 13, 38–40] and transformer-based [7, 34, 41] methods, have recently shown that generative priors learned from large-scale datasets allow for producing realistic object-level geometry and appearance with explicit control. These methods establish a foundation for generative 3D modeling, but remain limited to isolated objects or bounded spatial contexts. As such, they cannot directly scale to complex scene-level synthesis involving multiple entities, spatial layouts, and environmental context.

Scene Generation. Early examples of simulated worlds have been crafted as manual assets [9], thus enabling large-scale experimentation, composition with dynamic actors, and replayable evaluation of perception models. However, hand-crafted 3D design is prohibitively expensive and time-consuming. Procedural modeling approaches [19, 29, 30, 36] have been proposed to resolve this bottleneck, but at the cost of photorealism and variability, both critical aspects of simulation efficacy. Some approaches have built on existing 3D object-centric approaches and integrate multiple components together to create pipelines for block-wise 3D world construction [4, 10], yet these are limited by individual component cohesiveness and broadly lack real 3D awareness.

	XCube [31]	BlockFusion [42]	LidarDM [46]	LT3SD [24]	InfiniCube [23]	WorldGrow [18]	XScene [44]	**Ours**
Unbounded	✗	✓	✗	✓	✓	✓	✓	✓
Controllable	✗	✗	✗	✗	✓	✗	✗	✓
Cross-Domain	✓	✓	✗	✗	✗	✗	✗	✓
Texture	✗	✗	✗	✗	✓	✓	✓	✓

Table 1. **Summary of recent 3D scene generation methods.** Ours is the only approach satisfying all desirable criteria.

Hierarchical Latent Scene Generation. Building upon object-level generative models, recent works [14, 23, 24, 27, 31, 32, 42] have extended 3D generation to scene-level synthesis for both indoor and outdoor environments.

XCube [31] sets a benchmark for 3D generation quality via a multi-resolution sparse voxel hierarchy, while SCube [32] and InfiniCube [23] introduce controllability and texture modeling. Nevertheless, they remain limited in spatial extent or geometric fidelity. BlockFusion [42] represents scenes as latent tri-planes and performs spatial extrapolation for larger-scale outpainting, while LT3SD [12] introduces a latent tree-structured representation for patch-wise geometry synthesis over expansive environment. However, LT3SD is explicitly confined to dense indoor scenes and neither are capable of appearance modeling. In the outdoor domain, LidarDM [46] generates LiDAR via underlying 3D scene modeling, yet remains limited in fidelity and scope. WoVoGen [22] and XScene [44] explore joint voxel-based occupancy and image generation, but struggle with geometry-texture alignment at scale. The very recent LSD-3D [27] produces high-quality scene textures, but depends on the above methods to supply proxy geometry.

Overall, prior works are limited in some combination of fidelity, texture synthesis, spatial extent, or controllability. We propose a novel controllable and unbounded 3D generation method that generalizes across domains, situated among recent works in Table 1.

3. WorldFlow3D

In this section, we introduce WorldFlow3D, a novel formulation of the hierarchical 3D generation problem via flow matching. We formulate our method as transport between hierarchical distributions via flow models in Sec. 3.1. We propose to *directly generate* volumetric distributions of scene surfaces in Sec. 3.2, where we revisit autoencoder-free generation departing from latent diffusion models. Finally, we describe how WorldFlow3D can be used to perform controllable (3.4) and unbounded (3.3) 3D world generation through conditional and inference-time flow field alignment across smaller chunks.

Preliminaries. Continuous normalizing flows (CNFs) [6] were originally introduced to train ordinary differential equations with black-box solvers, modeling their underlying vector field v_t end-to-end with deep neural networks. CNFs model the continuous-time flow ϕ_t over $t = \{0...T\}$ between two d-dimensional distributions p_0 and p_1 using a deep neural network $f_\theta(x_t, t)$ with trainable parameters θ for any sample $x \in \mathbb{R}^d$. More recently, conditional flow matching (CFM) [20] methods train CNFs for optimal transport between two distributions using linear solvers. CFMs only require samples x from the underlying data distributions p_0 and p_1, and are trained to regress the underlying conditional vector field v_t given sample x_1 of fixed conditional probability paths with

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_0),p_t(x|x_0)} \| f_\theta\left(x,t\right) - v_t\left(x|x_0\right) \|^2, \quad (1)$$

extending the scope of this approach.

While probabilistic modeling of differential equations was also proposed for diffusion models, with stochastic differential equations [35], CFM has been investigated as a more efficient way to model paths between a Gaussian distribution $p_0 \sim \mathcal{N}(0, I)$ and a data distribution p_1 such as images or 3D data. Note that CFM generalizes well to arbitrary, non-diffusion probability paths such as optimal transport between *any* two data distributions. We subsequently leverage this property to propose a novel formulation of hierarchical generation.

3.1. Flowing Through Hierarchical Data Distributions

For WorldFlow3D, we define hierarchical generation as a sequence of distributions over progressively richer scene representations — concretely, from coarse geometry to fine geometry to full appearance — where each transition between adjacent levels corresponds to an independent learned flow. Fig. 2 depicts our approach, with separate paths indicating distinct hierarchies. We assign a data distribution p_i to each such level $i \in [0, N]$, where adjacent levels differ in fidelity and attribute composition, and therefore have distinct dimensions $d_i \geq d_{i-1}$. We train an independent velocity field $f_{\theta,i}$ at each level i to model the optimal transport of a sample $\mathbf{x}$, with the following rectified flow [21] objective as

$$\begin{aligned} &\mathcal{L}_{\text{CFM}}(\theta_i) = \\ &\mathbb{E}_{t\sim\mathcal{U}[i-1,i],x\sim p_{i-1},\, x_i\sim p_i} \| f_{\theta,i}(x_t,t) - (x_i - x_{i-1}) \|^2. \end{aligned} \quad (2)$$

As finer-level attributes may introduce additional attributes, e.g., RGB color, we accomodate them at lower level target distributions p_i, by assuming Gaussian distributions over unknown source dimensions such that

$$\begin{aligned} x_{i-1} &= \hat{x}_{i-1} \oplus \tilde{x}_{i-1}, \\ &\text{with } \tilde{x}_{i-1} \sim \mathcal{N}(0, I) \in \mathbb{R}^{(d_i - d_{i-1})} \\ &\text{and } \hat{x}_{i-1} \sim p_{i-1}, \in \mathbb{R}^{d_{i-1}}. \end{aligned} \quad (3)$$

3.2. 3D Scene Generation

We employ the hierarchical formulation introduced in Sec. 3.1 for 3D scene generation, where faithful synthesis requires resolving structure simultaneously at multiple spatial scales, and both the global spatial layout and local surface detail are necessary.

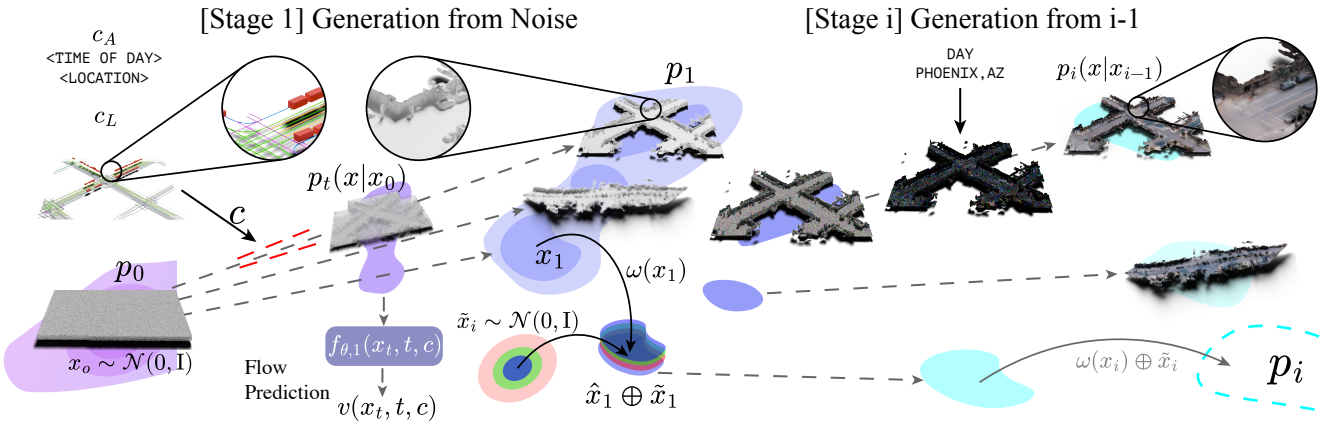


Figure 2. **WorldFlow3D** decomposes generation into a sequence of independent flows over progressively richer representations — transporting from noise, through coarse geometry into fine geometry, and visual appearance (Sec. 3.1) . All flows operate directly in raw volumetric space (Sec. 3.2), enabling a latent-free, hierarchical scene generation procedure. Generation is controlled by a vectorized geometric layout and discrete scene attributes, giving consistent structural and semantic control at every level (Sec. 3.4).

Latent-Free 3D Scene Representation. While 3D data comes in many forms (meshes, point clouds, signed distance functions), volumetric representations in particular can represent fine-grained geometric structure at a discretized voxel level, making the learning of 3D distributions p_i tractable without complex compression. Motivated by their simplicity, we choose truncated unsigned distance fields (TUDFs) to represent the surface of the scene at its zero-level set.

Generating 3D scenes that are both geometrically detailed and spatially coherent requires representations that can express structure at multiple scales of resolution and richness. Each scene sample $\mathbf{x}_i \in \mathbb{R}^{l_i \times c_i}$, is a raw volumetric tensor of shape $l_i = X \times Y \times Z$ and attribute channels c_i. Coarser levels operate at higher metric size $s_i \leq s_{i-1}$ of each individual voxel, allowing us to capture broad geometric structure. Finer levels refine detail at lower s_i and may introduce additional volumetric attributes. At each level i, x_i is composed of a subset of volumetric attributes defined over a voxel grid at resolution l_i: a TUDF $\mathcal{D}_i \in [0, \tau]$. Each voxel stores the unsigned distance to the nearest surface, truncated at τ; and optionally a sparse attribute volume $\mathcal{C}_i$ at the surface-set defined by $D_i\,(x, y, z) < \tau$. τ is the same across levels, and only varies by hierarchy.

We instantiate each sample from a distribution p_i directly over all volumetric scene representations, without a latent intermediate. Our method *does not* require a latent vector produced by a latent autoencoder. We thus achieve higher training and inference efficiency, eliminating the two-stage learning approach common in latent generation. As we find in 4.2, velocity prediction between *geometry* distributions benefits from unmediated access to geometric structure at every point along the trajectory, avoiding the representational bottleneck and reconstruction error introduced by a learned compression. We also note that latent-space flows remain fully compatible with the framework and may be incorporated at any stage where compression is warranted; direct volumetric generation is simply the more natural choice when the data space is tractable.

Flow Through the 3D Scene Hierarchy. In our method, the hierarchical scene representation defines a structured sequence of distributions that the generative process traverses. At the coarsest level, $\mathbf{v}_{\theta^{(0)}}$ transports Gaussian noise samples $x_0 \sim \mathcal{N}(0, \mathrm{I})$ to the data distribution over coarse geometry p_1. At each subsequent level i, a function ω_{i-1} may — depending on the target distribution — (1) spatially upsample existing attributes in l_{i-1} with voxel size s_{i-1} to the target tensor of shape l_i with voxel size s_i, (2) inject additive noise to prevent mode collapse in low data regimes

$$\omega_{i-1}(x_{i-1}) = \uparrow_r (x_{i-1} + \varepsilon), \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \in \mathbb{R}^{d_{i-1}}, r = \frac{s_{i-1}}{s_i}, \tag{4}$$

or (3) append independent noise channels as specified in Eq. (3) for any new attributes introduced at level i, that is

$$x_{i-1}^{(d_i)} = \omega_{i-1}(x_{i-1}^{(d_{i-1})}) \oplus \tilde{x}_{i-1}. \tag{5}$$

Each learned $f_{\theta,i}$ then independently models $v_{\theta,i}(x_t^{d_i}, t)$, as described in Eq. 2.

3.3. Unbounded World Synthesis with Chunk-Aware Velocity Averaging

Unbounded — or even large-scale in general — 3D scene generation is not obtainable in a single inference pass, due

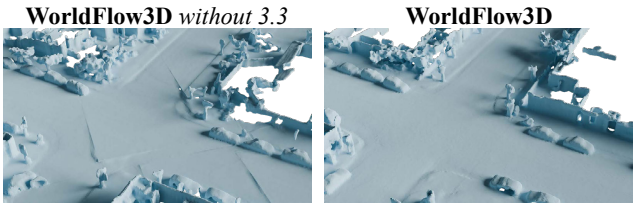


Figure 3. **Feather weighted velocity averaging** in overlapping chunk regions significantly improves the generated geometry for unbounded generations, as shown above.

to compute limits set by available technology. As a result, we partition scenes into overlapping chunks $\{\Omega_k\}$, and generate individual Ω_k in each inference pass. Naïve sequential outpainting and masking [26] with these chunks, however, results in artfiacts (see Fig. 3). We therefore generate the full volume x_i by integrating all chunks through the flow matching ODE at the same time. At each timestep t, the local volume $x|_{\Omega_k}$ is extracted for all chunks and passed through the flow model f_θ to obtain the per-chunk velocity $v(x|_{\Omega_k}, t, c_k)$, where c_k represents the local layout conditioning and global attributes. All chunks at t are then combined into the global velocity field $\bar{v}(t)$ via a spatially varying feather-weighted average

$$\bar{v}(\mathbf{s}, t) = \frac{\sum_{k:\,\mathbf{s}\in\Omega_k} \gamma_k(\mathbf{s})\; f_\theta\left(\mathbf{x}\big|_{\Omega_k}(t),\; t,\; c_k\right)\,[\mathbf{s}]}{\sum_{k:\,\mathbf{s}\in\Omega_i} \gamma_k(\mathbf{s})}. \quad (6)$$

Here, $\gamma_k(s)$ represents the feather weight at global location $\mathbf{s} \in \mathbb{R}^l$, which ramps linearly from a small value at chunk borders to 1 at the center of each chunk. This smoothly blends adjacent chunks and reduces to simple single-chunk prediction in non-overlapping regions. The full volume is then advanced with a standard Euler integration step $x_{t+1} = x_t - \Delta t\ \bar{v}$. In practice, we keep global conditioning, samples x_t, and computed local velocity fields $v(x|_{\Omega_k}, t, c_k)$ in CPU memory and only transfer to GPU memory for the forward pass of f_θ, enabling the generation of theoretically infinite scenes, constrained only by compute resources.

3.4. Controllable 3D Generation

Our method allows for explicit control over geometric structure and visual texture, a crucial requirement for usability of generated scenes. We provide control through a geometric layout c_L represented as a vectorized primitive — polylines defining structural boundaries and bounding boxes defining object extents — and discrete scene attributes c_A encoding scene-level visual descriptors such as environment type and lighting conditions, see Fig. 2. Typical forms of c_L are room layouts or maps, while c_A spans from natural text to discrete categories as presented in this work.

3D World Control. Maintaining c_L in vectorized form makes it resolution-agnostic and generalizable across map formats. At each level i, it is voxelized on-the-fly into $c_{L,i} \in \mathbb{R}^{s_i \times K}$, where each of the K channels encodes a distinct semantic class of the boundary or object type. This allows for a single layout specification to condition generation consistently across all levels without reprocessing, and decouples the control representation from the spatial resolution of the generator. Scene attributes $c_A \in \mathcal{A}$ are encoded as a compact embedding $c_{A,i}$ and injected globally. In practice, we use discrete environment tags; however, c_A may encode any arbitrary scene-level descriptor.

3.5. Scale-Space Embeddings and Losses

Each velocity field $v_i(x|c_L, c_A, x_{i-1})$ is represented by $f_{\theta,i}(x_t, t, c_L, c_A)$ as a 3D UNet with residual blocks and self-attention at multiple spatial scales. The intermediate sample x_t is concatenated channel-wise with the layout volume $c_L^{(i)}$, and the scene attributes $c_A^{(i)}$ broadcast spatially to $l_i \times \mathcal{A}$, providing the model with direct spatial access to prior-level structure, layout, and scene-level descriptors. Each residual block applies FiLM conditioning [28] via a conditioning embedding $e^{(i)}(t)$ formed by summing independent learned embeddings of the timestep t, and a global layout summary $\phi_L(c_L^{(i)})$ as

$$e^{(i)}(t) = \phi_t(t) + \phi_L(c_L^{(i)}), \quad (7)$$

where ϕ_t and ϕ_L are small learned encoders used to predict the scale and shift parameters of each residual block.

3.6. Implementation Details

The coarsest model in each sequential flow hierarchy, translating from $p^{(0)} := \mathcal{N}(0, 1) \longmapsto p^{(1)}$, is trained for up to 1 day across 2 NVIDIA H100 GPUs (empirically, we observe saturation in quality at this point), while all subsequent flow models $p^{i-1} \longmapsto p^i$ are trained for 12 hours on the same infrastructure. We use the AdamW optimizer with a learning rate of 2×10^{-6}. We provide additional details on flow sequence modeling in the Supplementary Material.

4. Assessment

We evaluate the effectiveness of our method via comparisons to existing generative methods on both indoor rooms and outdoor driving environments, and both real and synthetic data distributions.

Datasets. We evaluate on three data distributions, using the Waymo Open Dataset [37] as our data distribution for open-world 3D driving scenes. and the 3D-FRONT dataset [12] for synthetic, indoor worlds. For all scenes, we construct volumetric TUDFs and sparse color volumes

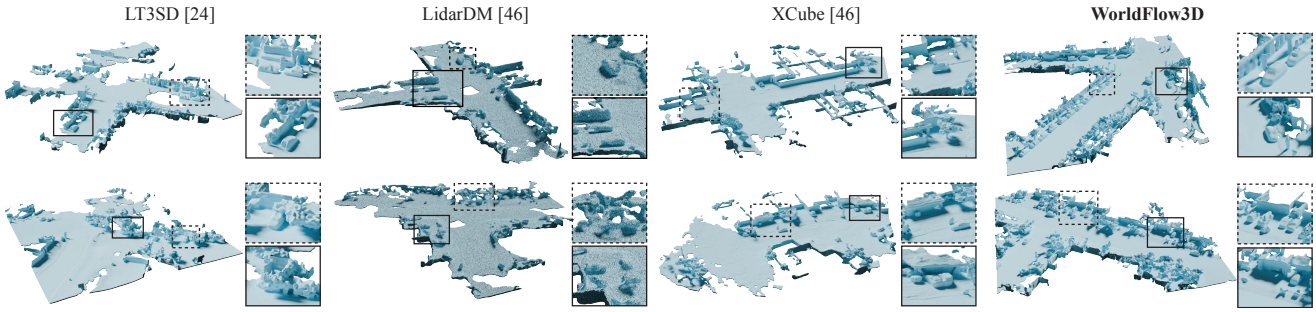


Figure 4. **Qualitative comparison on outdoor scene generation** with WorldFlow3D and baseline methods trained on the Waymo [37] dataset. We showcase scenes generated at moderate scales, and closer-up views of specific details including buildings and vehicles. We obtain high-quality, realistic geometry and smooth surfaces with a good amount of detail, as viewed from coherent building structure, smooth road surfaces, and distinct vehicle geometry.

which are used as training data. For outdoor scenes, we use a hierarchy with $\{s_1, s_2\} = 0.4m, 0.2m$ and $\tau = 1m$, and for indoor scenes, we use $\{s_1, s_2\} = 0.044m, 0.022m$ and $\tau = 0.1m$. We provide further detail on our data processing and parameter choices in the Supplementary Material.

Baselines. We compare against five recent methods across our target domains. For outdoor generation on Waymo, we evaluate against (a) XCube [31], a hierarchical voxel latent diffusion model which set a benchmark on 3D quality, and (b) LidarDM [46], using the intermediate 3D scene generation branch. For indoor generation on 3D-Front, we compare against (c) BlockFusion [42], which generates scenes via latent triplane-based spatial outpainting, and (d) WorldGrow [18], a sequential framework for unbounded 3D indoor scene synthesis. For all datasets, we compare to (e) LT3SD [24], a latent tree-structured patch diffusion model which we re-train on Waymo and 3D-FRONT. We use official checkpoints for all other baselines, and omit InfiniCube [23] and XScene [44] due to unavailable implementation and task mismatch respectively. Further details, including our evaluation procedure and inference procedure for baselines, are provided in the Supplementary Material.

Evaluation Metrics. We evaluate generation quality using five complementary metrics: Coverage (COV), Minimum Matching Distance (MMD), 1-Nearest Neighbour Accuracy (1-NNA), Jensen-Shannon Divergence (JSD), and Fréchet Distance ($\text{FD}_{\text{Concerto}}$). COV measures diversity as the fraction of reference scenes $r \in \mathcal{R}$ matched by at least one generated sample; MMD measures fidelity as $\frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \min_{g \in \mathcal{G}} d(r, g)$; and 1-NNA is a leave-one-out classifier over $\mathcal{G} \cup \mathcal{R}$, where an accuracy of 50% indicates statistically indistinguishable distributions. COV, MMD, and 1-NNA are each computed under both Chamfer Distance (CD) and Earth Mover's Distance (EMD) as the underlying similarity measure $d(\cdot, \cdot)$. JSD measures spatial overlap by voxelizing both sets into a shared occupancy grid and computing $\frac{1}{2}\text{KL}(P\|M) + \frac{1}{2}\text{KL}(Q\|M)$, where $M = \frac{1}{2}(P + Q)$. $\text{FD}_{\text{Concerto}}$ computes the Fréchet Distance between per-scene embeddings extracted from Concerto [45], a large-scale 3D foundation model pretrained on real-world point clouds, capturing high-level semantic and structural similarity beyond what point-distance metrics can express. All metrics are computed over $N = 5{,}000$ surface points sampled per scene chunk, and we select $1,000$ scene chunks per method. Extended evaluation information and per-dataset sampling details are provided in the Supplementary Material.

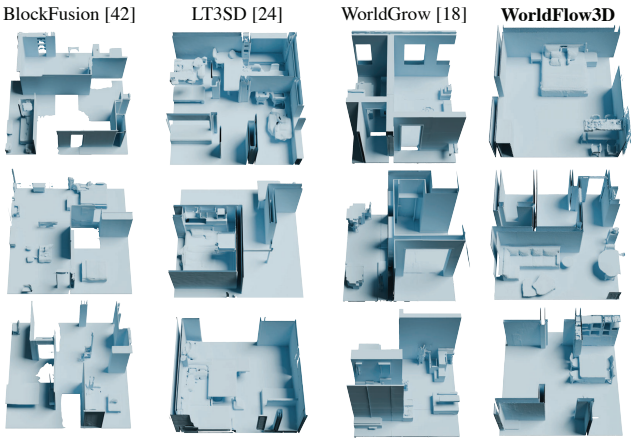


Figure 5. **Qualitative comparison on indoor scene generation** with WorldFlow3D and baseline methods trained on the 3D-FRONT [12] dataset. We showcase generations of regions including (potentially multiple) rooms with various objects. Our generations are high-fidelity and contain smooth surfaces and realistic geometry.

4.1. 3D Generation Results

We conduct a quantitative evaluation of the 3D generation quality of our method and competing baseline approaches. In Tab. 2, we provide results for outdoor scene generation on the Waymo dataset, both uncondition-

Table 2. **Quantitative Evaluation for Outdoor 3D Scene Generation** on the Waymo Open Dataset [37] for WorldFlow3D and existing approaches. We show results for unconditional generation in the first section and conditional in the second. The best results for each metric are in **bold**; second-best are underlined. We evaluate distribution coverage and alignment (COV, JSD), generation fidelity (MMD and 1-NNA), and feature-based distributional distance (FD_C). We report metrics on *large-scale* scene sizes of $96m \times 96m$.

	Method	COV ↑		MMD ↓		1-NNA ↓		JSD ↓	FD_C ↓
		CD	EMD	CD	EMD	CD	EMD		
Waymo Uncond.	XCube [31]	27.50	24.90	19.75	2.99	95.85	85.35	**0.480**	214.08
	LidarDM [46]	15.30	15.40	29.74	3.60	99.40	98.10	0.564	232.49
	LT3SD [24]	20.00	16.10	34.33	3.96	95.85	85.35	0.524	76.04
	WorldFlow3D	**33.00**	**32.70**	**16.57**	**2.81**	**89.15**	**70.15**	0.490	**74.83**
Waymo Cond.	LidarDM [46]	12.00	11.30	35.46	3.81	99.60	99.20	0.590	215.49
	WorldFlow3D	**39.70**	**35.20**	**12.44**	**2.70**	**88.60**	**81.85**	**0.483**	**80.08**

Table 3. **Quantitative Evaluation for Indoor 3D Scene Generation** on the synthetic 3D-FRONT [12] dataset. The best results for each metric are in **bold**; second-best are underlined. We report metrics on *small-scale scenes* [24, 42] of $2m \times 2m$.

	Method	COV ↑		MMD ↓		1-NNA ↓		JSD ↓	FD_C ↓
		CD	EMD	CD	EMD	CD	EMD		
3D-Front [12] Uncond.	BlockFusion [42]	29.00	28.40	0.054	0.221	91.50	90.65	0.380	165.46
	LT3SD [24]	23.40	26.20	0.056	0.223	93.00	91.60	0.230	44.97
	WorldGrow [18]	34.70	32.80	0.053	0.221	88.05	88.65	0.272	172.58
	WorldFlow3D	**38.30**	**38.10**	**0.039**	**0.195**	**74.75**	**75.80**	**0.164**	**36.45**

ally and conditionally. In Tab. 3, we compare our method for indoor scene generation to baseline results on the 3D-Front dataset. Across multiple data distributions, WorldFlow3D outperforms existing baselines in all quantitative evaluations, demonstrating not only high geometric *fidelity* but also a high degree of geometric *diversity*. We demonstrate examples of very large-scale generated scenes in Fig. 1 to supplement these numerical results, exhibiting the core elements of our method: large-scale, effectively unbounded scenes, explicit scene control, high-fidelity scene geometry, and visual attributes such as texture. In Figures 4 and 5, we ground this with further visual comparisons to the existing baselines we quantitatively compared to, supplementing our demonstration of superior quality. While competing methods demonstrate reasonable fidelity, ours attain higher levels of quality and 3D consistency. Our training objective results in broader generalizability and this is reflected in higher distribution coverage across all datasets. Furthermore, we do not compare relative to our discretized training distribution as in existing methods [18, 24], which is inherently lower resolution – but *we compare to the original, arbitrarily higher-resolution data* as our ground truth. This inherent difficulty is especially present for outdoor scene generation, as shown in the difference in evaluation metrics compared to indoor data. Nevertheless, our approach obtains reasonable distribution coverage and reasonable feature-space similarity; furthermore, visual evaluations (see 4.4) confirm the quality of our generations in their geometric structure.

Controllability Evaluation. We provide additional qualitative results which validate our method's controllability in Figure 7 – confirming, respectively, attribute control over visual texture and fine-grained geometric structure control using road map layouts. Our generated scenes not only strictly adhere to control, on the level of individual objects (such as vehicles, for road layouts), but are visually expressive for distinct textures, showcasing diversity.

4.2. Ablation Study

On a foundational level, we compare standard latent diffusion or latent flow approaches to our proposed latent-free generative method. We conduct our ablation study on the Waymo dataset. We compare our method to traditional latent-space generation approaches with `Latent Diffusion` and `Latent Flow` — using the same hierarchical structure as our main results, but incorporating a VAE at both levels and performing diffusion or flow, respectively, in latent space. We compare additionally the results of flowing from noise (conditional denoising) as opposed

Table 4. **Ablation Study** over the core contributions of our method, comparing distribution coverage (COV), geometric fidelity (MMD and 1-NNA), visual texture quality (FD_C), and training efficiency (GPU-hrs). We compare against traditional latent diffusion and latent flow approaches, followed by an ablation of our flow matching through distributions objective, and finally of our flow sequence hierarchy. We conduct our evaluation on the Waymo Open Dataset [37], and the best results are **bolded**.

Ablation	COV ↑		MMD ↓		1-NNA ↓		FD_C ↓	GPU-hrs ↓
	CD	EMD	CD	EMD	CD	EMD		
Latent Diffusion	37.75	33.75	7.426	1.801	91.38	75.63	120.87	288
Latent Flow	37.00	32.75	6.755	1.778	90.88	74.88	112.35	288
Flow from Noise	33.75	33.75	6.887	1.747	92.25	81.25	107.59	144
WorldFlow3D ($p^{(f)} = p^{(1)}$)	38.00	31.50	7.101	1.821	92.75	75.63	**89.47**	**72**
WorldFlow3D(full)	**42.50**	**37.25**	**4.942**	**1.696**	**85.00**	**70.25**	103.20	144

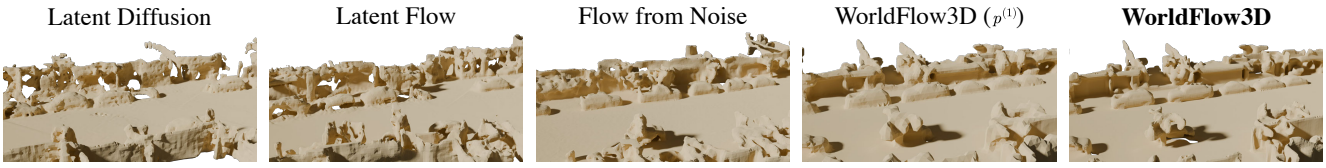


Figure 6. **Ablation of Core Components.** We provide qualitative results obtained by ablating the key components of our method. Latent diffusion and flow approaches produce structurally degenerate, non-realistic results, while flow from noise at finer distributions produces noisy outputs. Examples of this can be seen on the building walls, which are smooth with WorldFlow3D, and on vehicle details such as tires. Our hierarchical, latent-free approach is the only one that can produce high-quality, geometrically plausible results.

to flowing through distributions in `Flow From Noise`. Finally, we ablate our hierarchical structure by comparing the results of WorldFlow3D after only one distribution. The results in Figure 6 visually demonstrate that our latent-free flow through distributions is most capable of producing realistic, high-quality geometry, and a quantitative evaluation in Tab. 4 confirms this. We conduct this evaluation using smaller scene sizes than the results in Tab. 2, hence the distinct value range. This is in order to focus metric variation more specifically on geometric quality, and this is evidenced by metric variation in 1-NNA and MMD. Our approach outperforms latent-based methods and is dramatically more efficient, as we discuss in the following section.

Flowing Through Distributions vs. Conditional Denoising. The value of our approach of flowing through distributions as opposed to the traditional formulation of successive conditional denoising is most strongly confirmed in the comparison shown in Figure 6. Reducing the transport between distributions by flowing from an intermediate p^i rather than conditionally flowing from new noise allows the model to focus on generating detail rather than structure, such as the tires on the vehicles. This is also evidenced in visual color quality (see the Supplementary Material), further confirming the usefulness of our novel approach.

4.3. Training Efficiency

WorldFlow3D is at least 2× more efficient in training compared to traditional generative approaches, as a result of our latent-free paradigm and approach which minimizes transport between finer distributions. In contrast to existing methods which require multiple days of sequential autoencoder and latent generative model training [18, 24, 31], our method requires no autoencoder and can converge on complex data distributions rapidly, reaching high levels of fidelity within only a couple of hours for finer distributions p^i and requiring less than a day for full convergence (as validated quantitatively). In fact, our flow models between finer distributions converge to high quality in only 12 hours of training. In contrast, as in Tab. 4, even our comparatively efficient latent generation approaches (as conducted for our ablation) require two-stage training which takes 2× as long as our own two-level hierarchical flow. Existing baselines, furthermore, require multiple days, coming to over a week for certain methods [24]. As a result, our method is not only capable of producing higher-fidelity results, but also accomplishes this with much lower computational cost for model training.

Table 5. **User Study.** We report Bradley-Terry (BT) [3] scores with 95% bootstrap confidence intervals and overall win rates.

Method	Rank	BT Score	95% CI	Win Rate
WorldFlow3D	**1**	**0.692**	**[0.569, 0.813]**	**88%**
XCube [31]	2	0.212	[0.118, 0.314]	63%
LT3SD [24]	3	0.085	[0.047, 0.131]	43%
LidarDM [46]	4	0.011	[0.004, 0.022]	6%

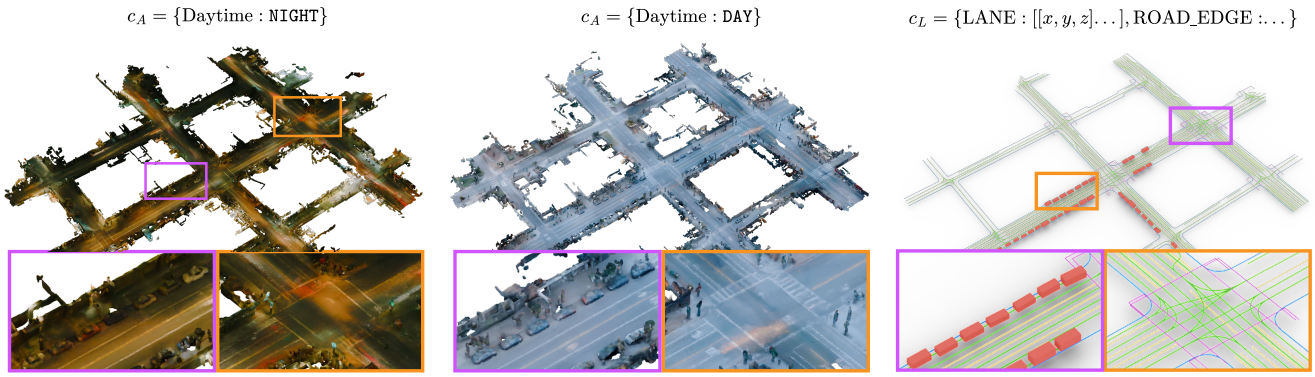


Figure 7. **Visual Texture and Controllability.** We report large-scale outdoor scenes with texture control via text attributes and geometry control via vector maps, yielding results conditioned on the same geometry to produce distinct environments.

4.4. User Study

We further supplement our evaluation by conducting a two-alternative forced choice user study for outdoor scene. Participants compared pairs of extracted meshes generated by all methods in Tab. 2 and selected the one which they perceived as higher quality. We then fit a Bradley-Terry model [3] to the comparison data, and obtain a global quality score for each method in Tab. 5, estimating confidence intervals via bootstrapping with $n = 1000$. Additional pairwise win rates with binomial significance tests and a detailed study setup are provided in the Supplementary Material. Overall, users prefer the results of our method with high significance over all other baseline methods, providing further perceptual confirmation of our qualitative (Fig. 4) and quantitative (Tab. 2) results.

5. Conclusion

In this work, we revisit 3D generation and model it more generally as a problem of flowing through hierarchical distributions. In this paradigm, we present WorldFlow3D, a novel approach capable of producing unbounded 3D worlds with explicit scene control and high-quality geometry and texture. We validate WorldFlow3D across distinct data distributions, including *both* real and synthetic data, confirming our method's generalizability, fidelity, and efficiency. The generality of our flow through distributions approach opens the door to future work using flow matching to transport between even more complex distributions and scene representations, including animated 3D scenes and radiance fields, and, as such, we believe WorldFlow3D is an innovative step towards 3D world generation.

Acknowledgements

Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, a Amazon Science Research Award, and a Bosch Research Award. Felix Heide is a co-founder of Algolux (now Torc Robotics), Head of AI at Torc Robotics, and a co-founder of Cephia AI.

References

[1] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)

[2] Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19697–19705 (2023)

[3] Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)

[4] Chen, H., Liu, Y., Li, M.: Trellisworld: Training-free world generation from object generators. arXiv preprint arXiv:2510.23880 (2025)

[5] Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2416–2425 (2023)

[6] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)

[7] Chen, Y., He, T., Huang, D., Ye, W., Chen, S., Tang, J., Chen, X., Cai, Z., Yang, L., Yu, G., et al.: Meshanything: Artist-created mesh generation with autoregressive transformers. arXiv preprint arXiv:2406.10163 (2024)

[8] Chen, Z., Tang, J., Dong, Y., Cao, Z., Hong, F., Lan, Y., Wang, T., Xie, H., Wu, T., Saito, S., et al.: 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In: Proceedings of the Computer Vi-

sion and Pattern Recognition Conference. pp. 26576–26586 (2025)

[9] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)

[10] Engstler, P., Shtedritski, A., Laina, I., Rupprecht, C., Vedaldi, A.: Syncity: Training-free generation of 3d worlds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27585–27595 (2025)

[11] Erkoç, Z., Ma, F., Shan, Q., Nießner, M., Dai, A.: Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14300–14310 (2023)

[12] Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021)

[13] Gupta, A., Xiong, W., Nie, Y., Jones, I., Oğuz, B.: 3dgen: Triplane latent diffusion for textured mesh generation. arXiv preprint arXiv:2303.05371 (2023)

[14] Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7909–7920 (2023)

[15] Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: SIGGRAPH 2024 Conference Papers. Association for Computing Machinery (2024)

[16] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023)

[17] Lee, H.H., Han, Q., Chang, A.X.: Nuiscene: Exploring efficient generation of unbounded outdoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 26509–26518 (October 2025)

[18] Li, S., Yang, C., Fang, J., Yi, T., Lu, J., Cen, J., Xie, L., Shen, W., Tian, Q.: Worldgrow: Generating infinite 3d world. arXiv preprint arXiv:2510.21682 (2025)

[19] Lin, C.H., Lee, H.Y., Menapace, W., Chai, M., Siarohin, A., Yang, M.H., Tulyakov, S.: Infinicity: Infinite-scale city synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22808–22818 (2023)

[20] Lipman, Y., Chen, R., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023)

[21] Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)

[22] Lu, J., Huang, Z., Yang, Z., Zhang, J., Zhang, L.: Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In: European Conference on Computer Vision. pp. 329–345. Springer (2024)

[23] Lu, Y., Ren, X., Yang, J., Shen, T., Wu, Z., Gao, J., Wang, Y., Chen, S., Chen, M., Fidler, S., Huang, J.: Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models (2024), https://arxiv.org/abs/2412.03934

[24] Meng, Q., Li, L., Nießner, M., Dai, A.: Lt3sd: Latent trees for 3d scene diffusion (2025)

[25] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

[26] Müller, N., Schwarz, K., Rössle, B., Porzi, L., Bulò, S.R., Nießner, M., Kontschieder, P.: Multidiff: Consistent novel view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10258–10268 (2024)

[27] Ost, J., Ramazzina, A., Joshi, A., Bömer, M., Bijelic, M., Heide, F.: Lsd-3d: Large-scale 3d driving scene generation with geometry grounding. arXiv preprint arXiv:2508.19204 (2025)

[28] Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)

[29] Raistrick*, A., Kayan*, K., Mei*, L., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., Ma, Z., Deng, J.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[30] Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., et al.: Infinite photorealistic worlds using procedural generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12630–12641 (2023)

[31] Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[32] Ren, X., Lu, Y., Liang, H., Wu, J.Z., Ling, H., Chen, M., Fidler, Sanja annd Williams, F., Huang, J.: Scube:

Instant large-scale scene reconstruction using voxsplats. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)

[33] Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)

[34] Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., Nießner, M.: Meshgpt: Generating triangle meshes with decoder-only transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19615–19625 (2024)

[35] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. 9th International Conference on Learning Representations, ICLR 2021 (October 2021)

[36] Sun, C., Han, J., Deng, W., Wang, X., Qin, Z., Gould, S.: 3d-gpt: Procedural 3d modeling with large language models. In: 2025 International Conference on 3D Vision (3DV). pp. 1253–1263. IEEE (2025)

[37] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[38] Team, T.H.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation (2025)

[39] Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al.: Lion: Latent point diffusion models for 3d shape generation. Advances in Neural Information Processing Systems **35**, 10021–10039 (2022)

[40] Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4563–4573 (2023)

[41] Wei, X., Zhang, K., Bi, S., Tan, H., Luan, F., Deschaintre, V., Sunkavalli, K., Su, H., Xu, Z.: Meshlrm: Large reconstruction model for high-quality meshes. arXiv preprint arXiv:2404.12385 (2024)

[42] Wu, Z., Li, Y., Yan, H., Shang, T., Sun, W., Wang, S., Cui, R., Liu, W., Sato, H., Li, H., et al.: Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. ACM Transactions on Graphics (ToG) **43**(4), 1–17 (2024)

[43] Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)

[44] Yang, Y., Liang, A., Mei, J., Ma, Y., Liu, Y., Lee, G.H.: X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. arXiv preprint arXiv:2506.13558 (2025)

[45] Zhang, Y., Wu, X., Lao, Y., Wang, C., Tian, Z., Wang, N., Zhao, H.: Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. In: NeurIPS (2025)

[46] Zyrianov, V., Che, H., Liu, Z., Wang, S.: Lidardm: Generative lidar simulation in a generated world. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6055–6062. IEEE (2025)

WorldFlow3D: Flowing Through 3D Distributions for Unbounded World Generation (Supplementary Material)

Amogh Joshi[1*], Julian Ost[1*], Felix Heide[1, 2]

[1]Princeton University [2]Torc Robotics

This supplementary document provides additional information in support of our main manuscript. In Section 1, we describe implementation details of our method, WorldFlow3D, including our model architecture, sequential flow setup, and detail on our controllability mechanism. Section 3 includes information on our dataset setup. In Section 2, we further elaborate on our model evaluation, including a thorough explanation of our user study. Finally, in Section 4, we provide additional qualitative results and additional comparisons to demonstrate the quality of our method.

Supplementary Video. In addition to this document, we provide a separate supplementary video showcasing generated large-scale scenes with high-quality geometry and visual texture, along with further method comparisons to supplement our qualitative evaluation.

1. Additional Implementation Details

In this section, we elaborate on details of our method. This includes details of the generator architecture and control implementations, as well as a further description of our overall approach supplementing the main manuscript.

1.1. FiLM Conditioning

Every residual block in $f_{\theta,i}$, introduced in Sec. 3.1 of the main manuscript, applies FiLM conditioning [7] via a shared embedding $e^{(i)} \in \mathbb{R}^{512}$ formed as

$$e^{(i)} = \phi_t(t) + \phi_L(c_L^{(i)}). \tag{1}$$

Timestep Encoder ϕ_t. The flow-matching timestep $t \in [0, 1]$ is embedded with a sinusoidal positional encoding γ_{sin} of dimension $C = 128$, then projected via a two-layer MLP

$$\phi_t(t) = W_2\, \sigma(W_1\, \gamma_{\text{sin}}(t)), \tag{2}$$

where σ is the SiLU activation.

Layout Encoder ϕ_L. The layout volume $c_L^{(i)} \in \mathbb{R}^{M \times l_i}$ is compressed to a spatially invariant summary through a single $3\times3\times3$ convolution, followed by global 3D average pooling

$$\phi_L(c_L^{(i)}) = W_{\text{out}}\, \text{Flatten}\Big(\text{AvgPool}_{3D}\Big(\sigma(W_{\text{in}} * c_L^{(i)})\Big)\Big), \tag{3}$$

where W_{in} projects $M \to C$ channels and W_{out} maps to the embedding dimension.

Scale-shift Modulation. Within each residual block, $e^{(i)}$ is projected to per-channel scale γ and shift β, which modulate intermediate features after group normalization as

$$h \;\leftarrow\; (1 + \gamma) \odot \text{GN}(h) + \beta \tag{4}$$

via $(\gamma, \beta) = W_e\, e^{(i)}$, followed by SiLU activation and a zero-initialized $3\times3\times3$ projection.

*Equal contribution.

1.2. Constructing Layout Control

We describe our procedure for constructing our layout control, focusing on our application for outdoor driving scenes. To accomplish this, we first discretize the road and vehicle layout, represented in a vectorized format, into a voxel grid L_0 of resolution τ_0. A voxel L is considered occupied in the vehicle channel if at least 50% of its volume lies inside a static vehicle bounding box B, i.e., v is occupied if $\mathrm{Vol}(v \cap B)/\mathrm{Vol}(v) \geq 0.5$. For road channels, a voxel is occupied if it intersects a road polyline of the corresponding type. When multiple polylines exist for a road type, the occupancy is given by $O_c(v) = \max_{\ell \in \mathcal{L}_c} \mathbf{1}[v \cap \ell \neq \emptyset]$, where $\mathcal{L}_c$ denotes all polylines of channel c and $\mathbf{1}[\cdot]$ is the indicator function. The final multi-channel layout is then given by $L_i = [O_1, \ldots, O_C, O_{\text{vehicle}}]$, where R is the number of road channels and O_{vehicle} is the occupancy of the vehicle channel. To enforce consistency between layout control and geometry, we prune unsupported layout voxels by updating $L_i \leftarrow L_i \odot \mathbf{1}_{(D_i \leq \mathbf{s}_i)}$, thereby removing any occupied layout voxel that does not intersect a surface in the corresponding TUDF. Together, this produces a control signal which contains core structural scene information, and is also spatially aligned with the scene for generation. Note that the same procedure can be adopted for indoor layout control — road polylines representing walls and room division, and object bounding boxes indicating object placement for items in the rooms.

1.3. Model Architecture

Each of our models $f_{\theta,i}$ in a flow sequence is a 3D UNet with residual blocks and multi-head self-attention at coarser spatial scales. The base channel width is $C = 128$ with multiplier progression $(1, 2, 4, 8)$, yielding four resolution levels with channel widths $(128, 256, 512, 1024)$. Each encoder level contains two residual blocks followed by a strided $2\times2\times2$ downsample; each decoder level contains three residual blocks followed by a $2\times$ trilinear upsample. Self-attention with 8 heads is applied at spatial downsampling factors $4\times$ and $8\times$. Table 1 provides layer specifications.

Table 1. 3D UNet architecture. **RB**: residual block with FiLM conditioning. **DS**: strided 2^3 downsample. **US**: $2\times$ trilinear upsample. $C = 128$.

Stage	Channels	Attention
Input Conv 3^3	$C_{\text{in}} \to C$	
Encoder $\ell = 0$: RB $\times2$ + DS	C	
Encoder $\ell = 1$: RB $\times2$ + DS	$2C$	
Encoder $\ell = 2$: RB $\times2$ + DS	$4C$	✓
Encoder $\ell = 3$: RB $\times2$	$8C$	✓
Bottleneck: RB + Attn + RB	$8C$	✓
Decoder $\ell = 3$: RB $\times3$ + US	$8C$	✓
Decoder $\ell = 2$: RB $\times3$ + US	$4C$	✓
Decoder $\ell = 1$: RB $\times3$ + US	$2C$	
Decoder $\ell = 0$: RB $\times3$	C	
Output: GN + SiLU + Conv† 3^3	$C \to 1$	

†Zero-initialized [3].

1.4. Training Procedure

In practice, we use a batch size of 1 for most models in order to maximize the spatial extent of each chunk size. For outdoor scenes, we use a chunk size of $256 \times 256 \times 16$ at $s_1 = 0.4m$, followed by $128 \times 128 \times 32$ at $s_2 = 0.2m$. For indoor scenes, we use a chunk size of $96 \times 96 \times 96$ at $s_1 = 0.044m$, and $64 \times 96 \times 64$ at $s_2 = 0.022m$. For training, we use a learning rate of 2×10^{-6}, although we empirically find that 5×10^{-6} produces similar results. For classifier-free guidance, we apply CFG dropout at rate p_{drop} to all conditioning signals independently during training, zeroing map, tag, and inpainting conditioning with probability p_{drop} per sample. The source sample x_1 for finer distribution flow is constructed by upsampling the coarse geometry and adding i.i.d. Gaussian noise with standard deviation $\sigma = 0.25$ to all channels, preventing the velocity field from collapsing to a trivial upsampling solution. Generator weights are maintained with an exponential moving average (EMA) using decay $\rho = 0.9999$ and a warmup schedule $\rho_n = (1 + n)/(10 + n)$ over the first updates n, after which the constant decay is applied.

2. Additional Evaluation Details

In this section, we further detail our evaluation procedure as conducted in the main manuscript, including a detailed description of our metrics, our parameter choices, and details on our user study.

2.1. Evaluation Metrics

We use coverage (COV), minimum matching distance (MMD), and 1-nearest-neighbor accuracy (1-NNA) metrics, consistent with existing works [5, 6, 8]. COV measures diversity as the fraction of reference scenes matched by at least one generated scene; MMD measures fidelity as the average distance from each reference scene to its closest generated counterpart; and 1-NNA is a leave-one-out classifier that is 50% for indistinguishable distributions and 100% for fully separable ones. These metrics are formalized as

$$\begin{aligned}
\mathrm{MMD}(S_g, S_r) &= \frac{1}{|S_r|} \sum_{Y \in S_r} \min_{X \in S_g} D(X, Y), \\
\mathrm{COV}(S_g, S_r) &= \frac{|\{\arg\min_{Y \in S_r} D(X, Y) \mid X \in S_g\}|}{|S_r|}, \\
\text{1-NNA}(S_g, S_r) &= \frac{\sum_{X \in S_g} \mathbb{1}[N_X \in S_g] + \sum_{Y \in S_r} \mathbb{1}[N_Y \in S_r]}{|S_g| + |S_r|},
\end{aligned}$$

where $D(\cdot, \cdot)$ is either the Chamfer Distance (CD) or Earth Mover's Distance (EMD), S_g and S_r are the generated and reference sets, and N_X denotes the nearest neighbor of X in the combined set $S_g \cup S_r \setminus \{X\}$.

For all point-cloud-based metrics, CD and EMD are

$$\begin{aligned}
d_{\mathrm{CD}}(A, B) &= \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} \|a - b\|^2 + \frac{1}{|B|} \sum_{b \in B} \min_{a \in A} \|a - b\|^2, \\
d_{\mathrm{EMD}}(A, B) &= \min_{\phi: A \to B} \frac{1}{|A|} \sum_{a \in A} \|a - \phi(a)\|,
\end{aligned}$$

where ϕ ranges over bijections. In practice, we approximate EMD via entropy-regularized Sinkhorn (ε=0.01, 20 iterations), restricted to the top-k nearest-neighbor candidate pairs pre-filtered by CD. To select k, we evaluate EMD on hold-out sets for $k = 10, 20, 30, 40$ and find that k=20 recovers correct values within $\epsilon = 10^{-5}$ with dramatically reduced compute time.

We also report JSD (Jensen-Shannon Divergence) as an additional signal of distributional alignment. Both sets are voxelized at δ=0.2 m within a shared bounding box into normalized occupancy distributions P and Q, that is

$$\mathrm{JSD}(P \| Q) = \frac{1}{2}\mathrm{KL}(P \| M) + \frac{1}{2}\mathrm{KL}(Q \| M), \quad M = \tfrac{1}{2}(P + Q).$$

We compute all metrics using the original ground truth meshes as the reference set, rather than pre-voxelized TUDFs as in other methods [6]. This ensures consistency across methods that use different data representations and provides a fairer evaluation, since the ground truth is at arbitrarily higher resolution than any generated voxel grid.

Feature-Based Metrics. Point-cloud distances alone cannot capture high-level structural and semantic similarity. We therefore also report $\mathrm{FD_C}$, the Fréchet Distance in the embedding space of Concerto [11], a large-scale 3D foundation model pretrained on real-world point clouds. Each scene is embedded via mean-pooling of per-point Concerto features over P=50,000 sampled surface points; coordinates are scaled by $0.2\times$ and voxelized at 0.01 m (effective resolution 0.05 m) to match Concerto's outdoor training regime, with both geometry and per-vertex color as input. Fitting multivariate Gaussians $(\mu_{\mathcal{G}}, \Sigma_{\mathcal{G}})$ and $(\mu_{\mathcal{R}}, \Sigma_{\mathcal{R}})$ to the resulting embeddings, the Fréchet Distance is

$$\mathrm{FD_C} = \|\mu_{\mathcal{G}} - \mu_{\mathcal{R}}\|^2 + \mathrm{tr}\Big(\Sigma_{\mathcal{G}} + \Sigma_{\mathcal{R}} - 2\,(\Sigma_{\mathcal{G}}\Sigma_{\mathcal{R}})^{\frac{1}{2}}\Big).$$

2.2. Evaluation Procedure

For both outdoor and indoor scenes, we evaluate by sampling $1,000$ chunks from ground truth data and generating $1,000$ chunks using our method and evaluated baselines. Outdoor scenes are evaluated with a spatial extent of $96m \times 96m \times 6m$,

as this is the minimum size across baselines, as set by LidarDM [12]; for other methods and ours, we generate chunks at size $102.4m \times 102.4m$ and crop to this reduced extent. For indoor scenes, we adopt a similar procedure, set by BlockFusion [10] which generates chunks at $2m \times 2m \times 2m$. For our ablation study, we select reduced-size chunks of $51m \times 51m \times 6m$ for outdoor scenes, in order to concentrate evaluation metrics not on broad large-scale context and instead on local geometric quality and fidelity.

2.3. User Study Details

We conduct a user study to complement our main quantitative evaluations, assessing the perceptual quality of generated geometry through a two-alternative forced choice paradigm. Fig. 2 shows each screen layout presented to participants. As a primer, participants were shown GT mesh data generated during the preprocessing described in Sec. 3, see Fig. 2(**d**). For the study, we randomly selected 10 meshes from each of the four methods [6, 8, 12] trained and quantitatively evaluated on the Waymo Open Dataset [9]. In each comparison, samples were drawn randomly from each method's pool of unused meshes. Under a blind two-alternative forced choice design, participants were asked to clearly indicate a preference for one of the two presented meshes without any additional information, see Fig. 2(**f**). Each participant completed 21 pairwise comparisons, with all methods appearing exactly 9 times outside of attention checks. The remaining 3 comparisons served as attention checks, presenting identical mesh pairs and recording response times to identify inattentive participants. One participant was discarded on this basis. We additionally recorded left/right position preferences, finding no explicit position bias across participants (48.3% vs. 51.7%, $p = 0.65$). In total, 15 participants completed the study, achieving significance ($p < 0.05$, 1) across all direct pairwise comparisons. We fit a Bradley-Terry model [1] to the collected comparisons to obtain a global quality score per method, with confidence intervals estimated via bootstrapping with $n = 1000$. The resulting ranking shows clear separation between methods: WorldFlow3D achieves the best scores by a wide margin, followed by XCube [8], then LT3SD [6], with LidarDM [12] ranking clearly last. Bradley-Terry confidence intervals are non-overlapping between most adjacent methods, confirming statistical significance; only the LT3SD vs. XCube comparison is borderline. Despite the inherent variance across generated samples, which is characteristic of open-ended generation without a fixed reconstruction target, our method, WorldFlow3D, achieves clear perceptual preference with high significance.

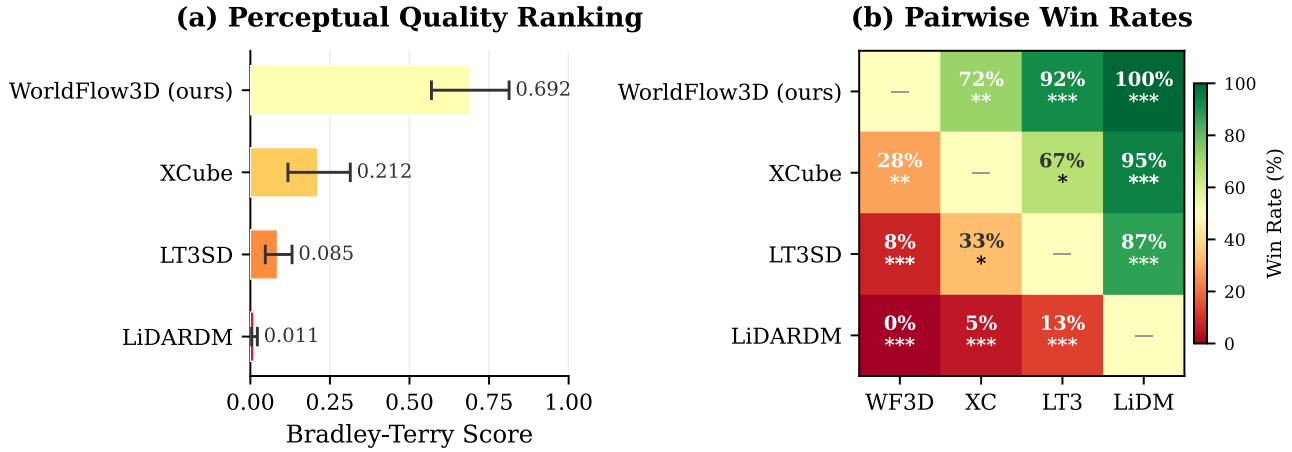
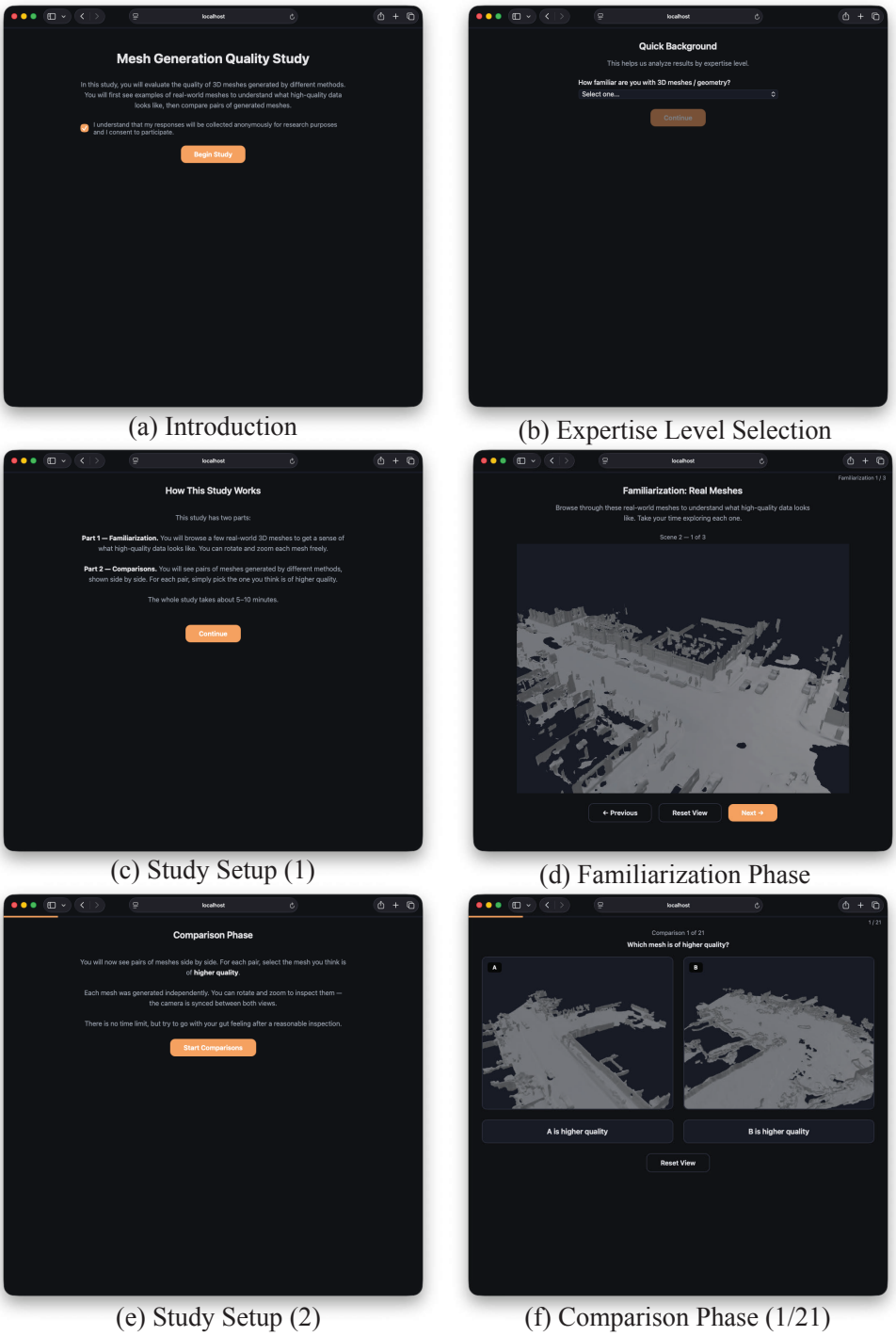


Figure 1. **User Study.** We report Bradley-Terry (BT) [1] scores with 95% bootstrap confidence intervals and overall win rates. Pairwise win rates are shown in the right table. Each cell indicates the row method's win rate against the column method. Statistical significance is denoted by $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (binomial test).

3. Dataset Preparation

For both datasets, we convert triangle meshes to truncated unsigned distance fields (TUDFs) using a fast-sweeping level set method. Exact point-to-triangle distances are computed in a narrow band around the surface and propagated to the full grid via multi-directional sweeps. Color is stored sparsely alongside the TUDF: for each voxel within a small region of the surface, color is assigned via nearest-vertex interpolation on the closest triangle. We extract a TUDF at the finest data resolution s_N for each dataset, and then deterministically compute the coarser data via dense average pooling for the distance volume and sparse, masked average pooling for the color volume.



(a) Introduction (b) Expertise Level Selection

(c) Study Setup (1) (d) Familiarization Phase

(e) Study Setup (2) (f) Comparison Phase (1/21)

Figure 2. **User Study UI.** The participants are first introduced to the overall topic (a), (c) and asked for consent with the anonymous collection of their data. (b) We additionally collect experience level. The participants are then presented with real data to familiarize themselves with the data quality they can expect (d). Ultimately, they are asked to perform comparisons for 21 meshes (f). Further details are provided in Sec. 2.

Waymo Dataset. We aggregate LiDAR sweeps across each scene sequence and back-project RGB camera images onto the accumulated point cloud to obtain colored point clouds. We then run NKSR [4] to reconstruct a watertight colored triangle

mesh per scene chunk, from which we extract the TUDF as described above.

3D-FRONT Dataset. We use the provided textured CAD meshes directly, without any reconstruction step, and apply the same TUDF extraction pipeline.

4. Additional Qualitative Results

Finally, we provide an additional set of qualitative results to supplement the evaluations provided in the main manuscript. We showcase additional large-scale scene generations, comparisons of our work with competing baselines, and furthermore introduce an additional 3D dataset and provide results (with a method comparison) to further highlight the generalizability of our method.

Outdoor Scene Patch Generation. We incorporate additional comparisons to supplement our outdoor scene patch generation evaluation in the main manuscript, showing single scene patch generations for our method and existing baselines in Figure 3.

Large-Scale Indoor Scene Generation. We supplement our indoor 3D scene generation evaluation on the synthetic 3D-Front dataset, which is included in the main manuscript, with an additional evaluation of dense, large-scale indoor scene generation for our method and competing baselines in Figure 4.

Synthetic 3D City Dataset: Large-Scale City-Scale Generation. We also provide additional qualitative results on a synthetic 3D city dataset (https://www.turbosquid.com/3d-models/city-downtown-and-suburb-1767093), composed of a set of city blocks arranged in a street grid fashion with buildings ranging from single-family houses to high-rise buildings. We also re-train LT3SD [6] using this dataset and show comparisons to it, alongside generations with our method, in Fig. 5.

Large-Scale Controllable Outdoor World Generation. Finally, in Figures 6 and 7, we showcase large-scale outdoor worlds generated by our method trained on the Waymo [9] dataset, demonstrating WorldFlow3D's ability to produce coherent, large-scale 3D worlds, and the capacity of our approach to strictly follow scene layout and visual texture control to produce diverse and consistent 3D worlds.

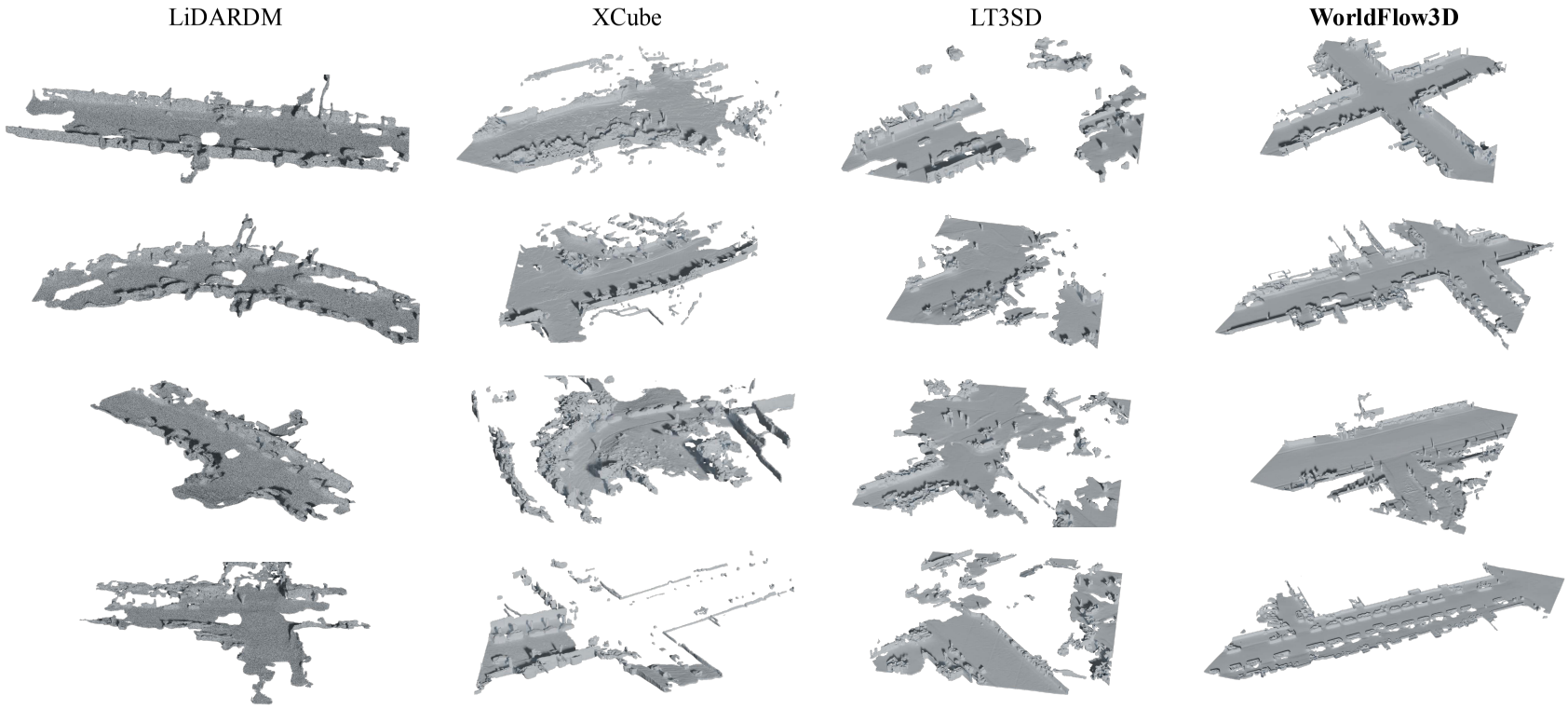


Figure 3. **Qualitative Evaluation of Outdoor Scene Patch Generation.** We demonstrate outdoor single scene patch generations on the Waymo [9] dataset from our method and baseline methods. Our patches achieve the highest quality in both scene layout and geometric structure detail, as further evidenced by this comparison.

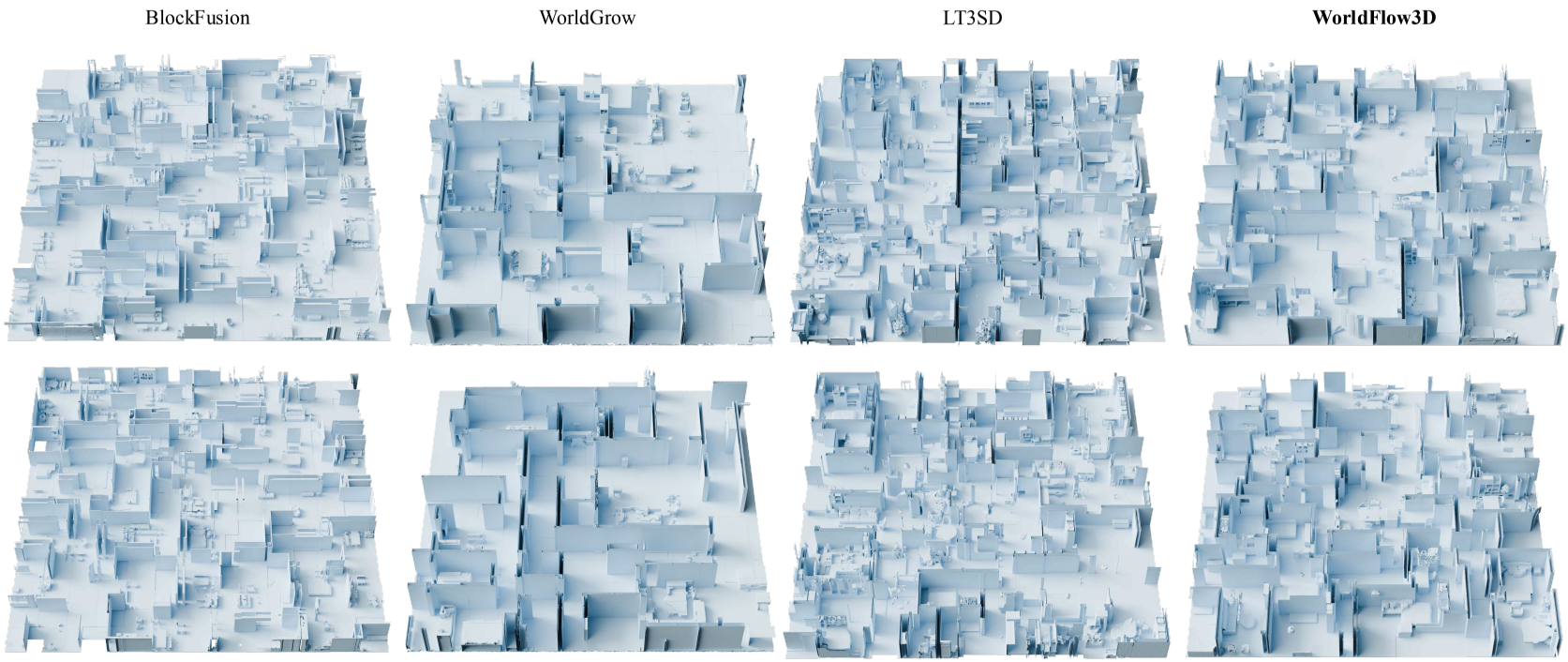


Figure 4. **Large-Scale Qualitative Scene Evaluation on the 3D-Front Dataset**. We show examples of large-scale indoor scenes generated by our method and prior baselines on the synthetic 3D-FRONT [2] dataset. We demonstrate the quality achieved by our method, as evidenced in the more coherent scene layout and structural detail of objects in various rooms, as compared to other methods.

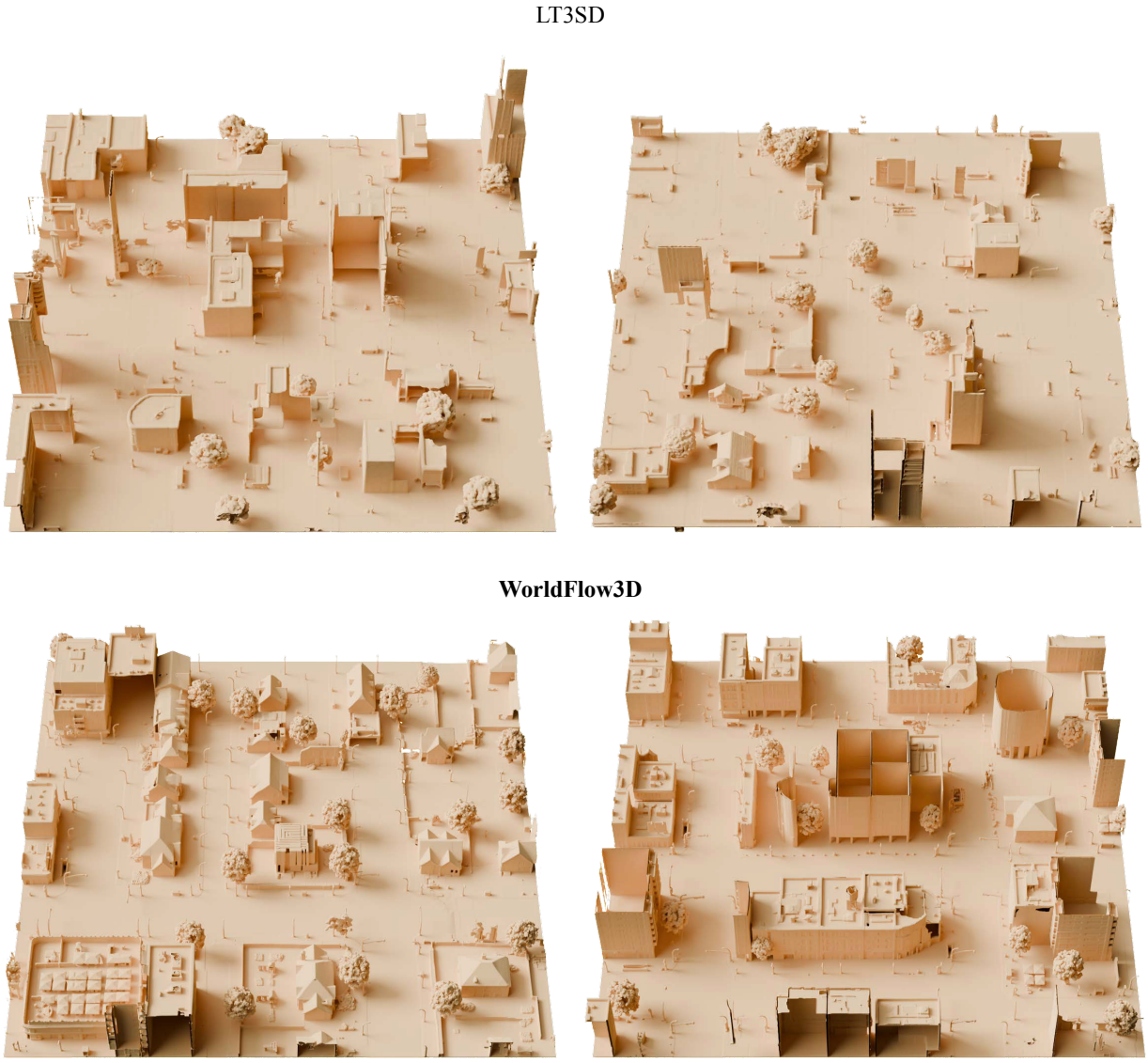


Figure 5. **Qualitative Evaluation on a Synthetic 3D City Dataset**. We provide visual results of large-scale generated scenes on the 3D-City dataset described in Section 4. We compare our method with LT3SD, a recent work which we re-train on this dataset. Our method achieves better quality in both 3D geometric structure and also overall scene layout, respecting the road structure and finer details such as the placement of streetlights.

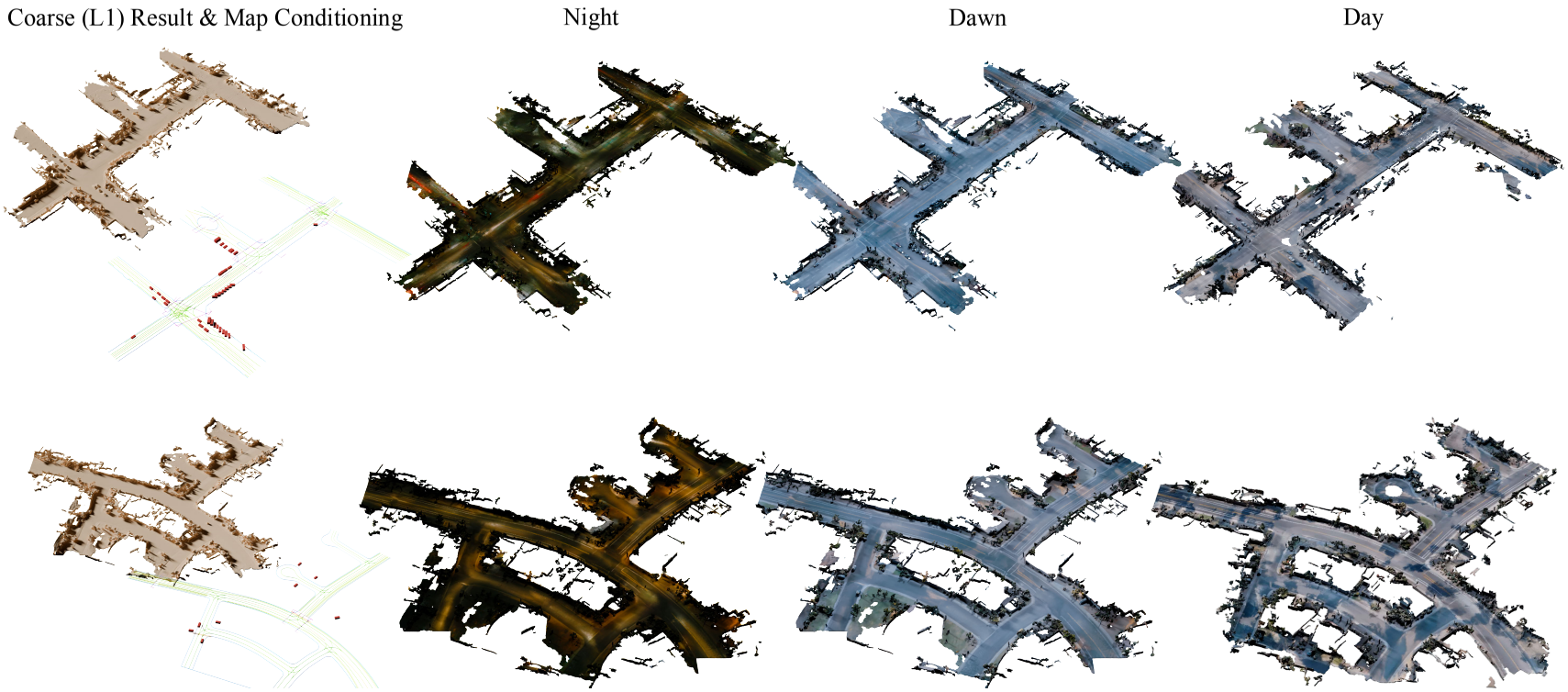


Figure 6. **Large-Scale Outdoor Scene Generations with Map Layout and Scene Attribute Controllability (2)**. We showcase outdoor driving scene examples generated with WorldFlow3D, showing the underlying road map layout control used in the first column, and three diverse texture variations for different `Time-Of-Day`s showcasing our approach's ability to strictly follow the provided structure and texture control, and produce high-quality, causal, and coherent large-scale 3D worlds.

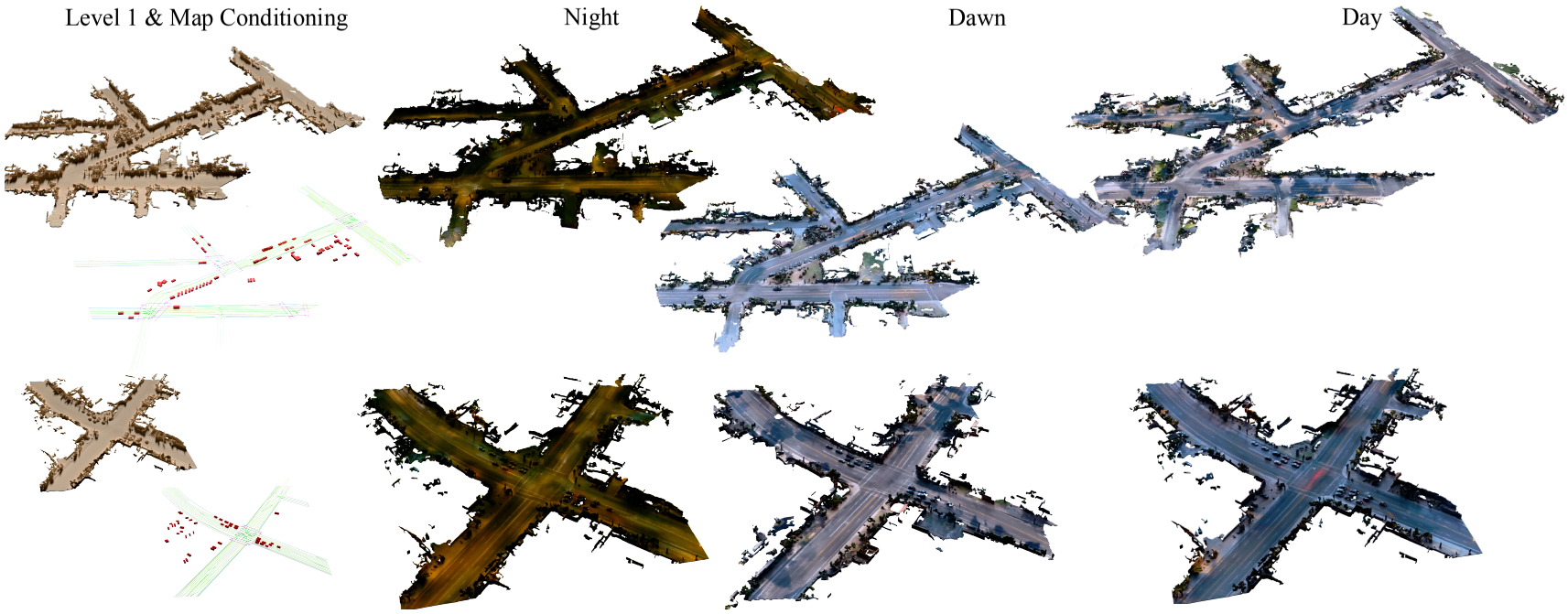


Figure 7. **Large-Scale Outdoor Scene Generations with Map Layout and Scene Attribute Controllability (2)**. We showcase outdoor driving scene examples generated with WorldFlow3D, showing the underlying road map layout control used in the first column, and three diverse texture variations for different `Time-Of-Days` showcasing our approach's ability to strictly follow the provided structure and texture control, and produce high-quality, causal, and coherent large-scale 3D worlds.

References

[1] Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)

[2] Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021)

[3] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[4] Huang, J., Gojcic, Z., Atzmon, M., Litany, O., Fidler, S., Williams, F.: Neural kernel surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4369–4379 (2023)

[5] Li, S., Yang, C., Fang, J., Yi, T., Lu, J., Cen, J., Xie, L., Shen, W., Tian, Q.: Worldgrow: Generating infinite 3d world. arXiv preprint arXiv:2510.21682 (2025)

[6] Meng, Q., Li, L., Nießner, M., Dai, A.: Lt3sd: Latent trees for 3d scene diffusion (2025)

[7] Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)

[8] Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[9] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[10] Wu, Z., Li, Y., Yan, H., Shang, T., Sun, W., Wang, S., Cui, R., Liu, W., Sato, H., Li, H., et al.: Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. ACM Transactions on Graphics (ToG) **43**(4), 1–17 (2024)

[11] Zhang, Y., Wu, X., Lao, Y., Wang, C., Tian, Z., Wang, N., Zhao, H.: Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. In: NeurIPS (2025)

[12] Zyrianov, V., Che, H., Liu, Z., Wang, S.: Lidardm: Generative lidar simulation in a generated world. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6055–6062. IEEE (2025)