LSD-3D: Large-Scale 3D Driving Scene Generation with Geometry Grounding

Julian Ost[1*], **Andrea Ramazzina**[2*], **Amogh Joshi**[1*], **Maximilian Bömer**[3], **Mario Bijelic**[1,3], **Felix Heide**[1,3]

[1]Princeton University [2]Mercedes-Benz [3]Torc Robotics

Abstract

Large-scale scene data is essential for training and testing in robot learning. Neural reconstruction methods have promised the capability of reconstructing large physically-grounded outdoor scenes from captured sensor data. However, these methods have baked-in static environments and only allow for limited scene control – they are functionally constrained in scene and trajectory diversity by the captures from which they are reconstructed. In contrast, generating driving data with recent image or video diffusion models offers control, however, at the cost of geometry grounding and causality. In this work, we aim to bridge this gap and present a method that directly generates large-scale 3D driving scenes with accurate geometry, allowing for causal novel view synthesis with object permanence and explicit 3D geometry estimation. The proposed method combines the generation of a proxy geometry and environment representation with score distillation from learned 2D image priors. We find that this approach allows for high controllability, enabling the prompt-guided geometry and high-fidelity texture and structure that can be conditioned on map layouts – producing realistic and geometrically consistent 3D generations of complex driving scenes. Interactive demonstrations and results are available on our project webpage: https://light.princeton.edu/LSD-3D.

1 Introduction

Large-scale public datasets have driven significant advancements in robot learning over the last decade. For autonomous driving, large volumes of data [26, 10, 82, 51, 18, 6] have unlocked new capabilities for perception and planning. Initially limited to a few sparsely labeled scenes [26], existing datasets now offer thousands of multi-modal, fully annotated scenes from cities around the world [6, 82] – unlocking broader capabilities for robot learning. However, achieving generalized robot autonomy demands even greater scales of both *diverse* and *complete* data, to capture the long tail of driving scene distributions – a challenge given the high costs of capturing and annotating real data. Research in various layers of the robotics stack has demonstrated that training models on large quantities of data and in simulated environments [16, 29, 4, 72, 75, 83] can result in robust and more generalized autonomy.

Recent work in neural scene reconstruction promises to bridge this gap by reconstructing previously simulated driving environments directly from sensor data [62, 88, 107, 98]. Such scenes can be replayed in real time, synthesize novel views, and allow for unseen actor variations in closed-loop testing [88, 54, 11]. However, neural reconstruction methods are fundamentally limited in that they cannot produce novel content beyond recorded scenes – they do not offer data-driven simulation with great scene diversity.

Video diffusion models have been proposed to increase data volumes and diversity. Pretrained on internet-scale datasets and subsequently fine-tuned on autonomous driving data they generate videos which mimic driving datasets and multi-view camera setups [93, 55, 25, 69]. While these methods can generate a large corpus of novel driving data, outside of their training data, and provide feature control, they also come with inherent limitations. High computational costs, on the order of seconds per generated multi-view frame, restrict real-time replay and scalability for closed-loop simulation tasks. Furthermore, they lack explicit spatial modeling, which prevents causality, object permanence, and 3D consistency. The latter also prohibits them from replaying novel trajectories within a pregenerated environment [22]. As such, video diffusion models struggle as data-driven simulators, especially for safe robot learning.

Explicit 3D scene models *guarantee causality* and 3D consistency. Directly generating explicit 3D scenes, however, poses a challenge: both *geometry* and *texture* have to be generated consistently and with high quality. Some approaches use LiDAR point clouds to produce pure geometry without detailed texture [46, 112, 67, 119] – which cannot be used to train image-based autonomous driving models – while more recent approaches adopt coarse 3D geometry as a conditioning mechanism for video synthesis [55, 56]. As such, these methods inherit video diffusion limitations, such as a lack of causality. Distillation methods [64] instead address the 3D data gap by transferring priors from 2D image models into 3D representations via inverse rendering techniques [59, 42, 109]. However, existing techniques are limited to object-centric generation [64, 101, 99] and lack realism [47] — and so, they are not suitable for the complexity [115] or spatial scale [76] of large-scale driving scenes. Only recent work has explored distillation approaches for scene extrapolation from sparse real data cap-

[*]Indicates equal contribution.

tures [95, 97, 53, 20], achieving improved reconstruction quality from few images and hinting at the potential for complete outdoor scene generation.

We propose a novel approach that overcomes the aforementioned limitations of existing generative methods for large-scale driving scenes. Our method generates explicit 3D models of entirely novel environments – with both geometry and high-fidelity texture – by fusing the diversity of image and video generation with the efficiency of explicit 3D representations. We first generate a coarse geometry of street scenes, optionally controlled by a road map layout. This proxy is then used to guide the generation of fine structural details and high-fidelity textures via image space distillation with a high-quality image generation model. Specifically, we introduce geometry-grounded distillation sampling (GGDS), an image space sampling approach, that incorporates explicit geometry control and exact noise sampling by DDIM inversion in a single method. We find that the combination of geometry guidance and consistent noise sampling through inversion can deliver successful and 3D consistent scene generation via distillation. The method produces diverse, realistic, and large-scale 3D scene models for autonomous driving. Generated scenes guarantee causality and can produce unlimited novel trajectories in real-time – enabling *scalability* – while maintaining 3D consistency and appearance fidelity. Furthermore, precise prompt control over weather, season, time-of-day, and location, in the form of explicit environmental lighting, enables further fine-grained customization of these virtual scenes.

We validate our method on the Waymo Open Dataset [82], generating novel scenes which not only inherit the data prior distribution, but leverage the implicit prior of 2D diffusion models to provide enhanced scene *diversity*. Our approach, generating complete and coherent large-scale 3D scenes, outperforms state-of-the-art existing generative methods in image synthesis of unseen camera angles by **18% in FVD** and maintains prompt adherence on the level of pure video-based approaches.

We summarize our specific contributions as follows:

- To our knowledge, our method is the first to utilize a distillation approach to directly *generate* and optimize explicit 3D driving scenes with high-quality geometry and texture – guaranteeing causal generation.
- We introduce Geometry-Grounded Distillation Sampling (GGDS), a method combining controlled proxy mesh generation with a conditional diffusion prior to produce novel, view-consistent Gaussian splatting scenes, with real-time rendering and composability with 3D assets.
- We generate diverse large-scale scenes which can be rendered into physically-grounded videos controlled by trajectories through the scene, enabling the creation of unlimited, completely unseen environments, controlled by scene descriptions, traffic map layouts, or text prompts.

2 Related Work

Image and Video Synthesis. Recent advances in image generation have enabled the synthesis of high-resolution, photorealistic imagery. These approaches – from generative adversarial networks (GANs) [27, 5, 40, 74], likelihood-based methods [45, 66, 89], to more prominently, diffusion models [14, 79, 15, 17, 32, 71, 90] – have been recently extended to video generation [3, 116, 78, 71, 30, 33, 34, 23, 108, 35, 85]. Early methods offer control via text prompt [78], image [116], or camera position [73, 105]. Recent methods like GEN3C [70, 84] or GeoDrive [7] for driving scenes incorporate 3D conditioning and point projection to improve geometric consistency. Specialized models, such as Vista [25], MagicDrive [21, 23], or DriveDreamer [93], have been developed on top of foundational video models [3] for autonomous driving applications – employing bounding box, HD-map, and dense voxel guidance for feature control. These methods extend pre-trained video diffusion models to generate videos which mimic driving sensor setups. However, despite efficiency improvements [80, 57], video generation models remain computationally intensive [108, 85, 91] and preclude the scalability necessary for real-time simulation. Furthermore, they struggle with coherent novel-view rendering over long driving trajectories and suffer from a lack of causality. In the top section of Tab. 1, we provide a comparison of relevant video synthesis methods for large-scale driving scenes.

Method	Unlimited Viewpoints	Compos-ability	Causal 3D Geometry	Real-Time Rendering	View Extrapolation	Control		
						Weather	Time	Map
DriveDreamer [93]	✗	✗	✗	✗	✗	✓	✓	✗
WonderJourney [111]	✗	✗	✗	✗	✗	✓	✓	✗
Streetscapes [13]	✗	✗	✗	✗	✗	✓	✓	✓
Vista [25]	✗	✗	✗	✗	✗	✓	✓	✗
MagicDriveDiT [23]	✗	✗	✗	✗	✗	✓	✓	✗
WoVoGen [55]	✗	✗	✗	✗	✗	✓	✓	✓
WonderWorld [110]	✗	✗	✗	✗	✗	✓	✓	✗
NF-LDM [44]	✗	✗	✗	✓	✗	✓	✓	✓
InfiniCity [52]	✗	✗	✓	✓	✓	✗	✗	✓
MagicDrive3D [25]	✗	✓	(✓)	✓	✓	✓	✓	✓
CityDreamer [100]	✗	✗	✓	✓	✓	✗	✗	✓
InfiniCube [56]	✗	✓	(✓)	✓	✓	✓	✓	✓
GEN3C [70]	✗	✗	(✓)	✗	✓	✓	✓	✗
LSD-3D (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Video and 3D Generation for Driving Scenes.** We review the recent body of video (top) and 3D generation work (bottom). *Please zoom in digital document for details.*

3D Generation. To guarantee consistency and causality, recent approaches separate from 2D image and video generation and focus on the explicit generation of individual, object-centric 3D structures. These methods [99, 49, 96, 24, 39, 73, 86, 101] generate consistent 3D structure by leveraging well-defined features and latent spaces, and multi-view observations – enabling them to directly perform explicit diffusion of 3D objects. However, to train these 3D generative models, they rely on high-quality synthetic 3D data [12, 9], and most works are thus constrained to the generation of individual objects – generation of large-scale 3D outdoor scenes remains an open challenge.

Scene Reconstruction and Generation. For large-scale scene generation [44, 106, 46, 1, 92, 2, 52], various 3D representations have been proposed: triplanes [77, 46, 1], semantic occupancy grids, bounding boxes, and 3D maps [106, 102]. However, reliance on generation of explicit priors requires expensive annotation data, and these early methods' generations lack significantly in photorealism and scale. In contrast, satellite imagery has been used for the city-scale 3D generation of urban environments [52, 100]. More recently, hierarchical voxel diffusion methods [67,

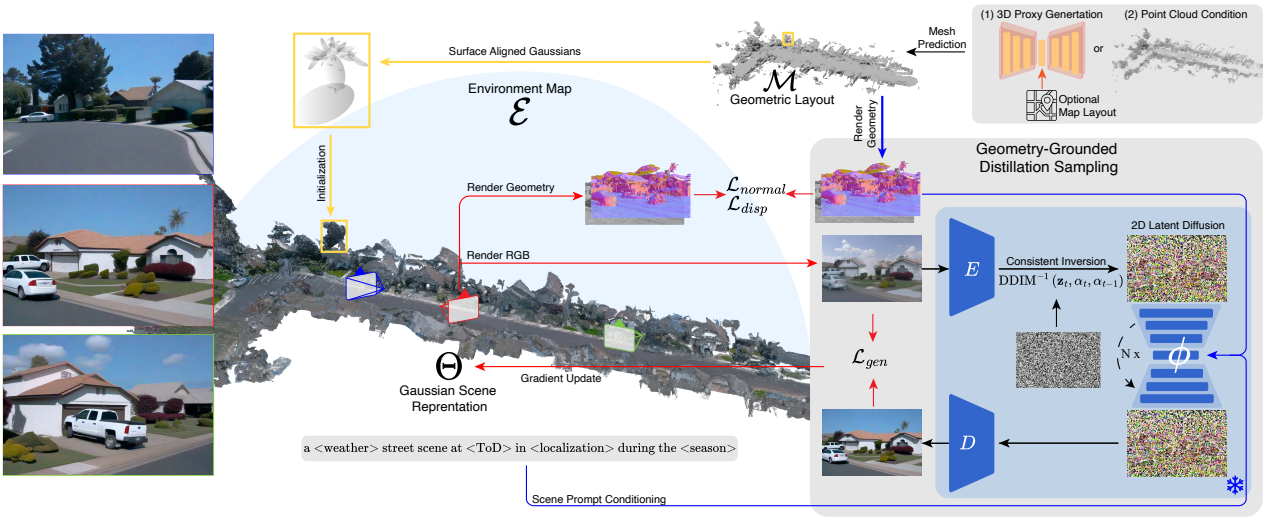

Figure 1: **Geometry-Grounded Large-Scale 3D Scene Generation.** We generate a large-scale scene as a combination of a coarse geometric layout, an environment map, and a set of Gaussians for texture details, discussed in Sec. 3.1. The geometric layout is either generated, conditioned on a map, or predicted from point-cloud data and guides the overall scene structure. We can further control the setting with a scene prompt, describing time-of-day, season, and weather. Through Geometry-Grounded Distillation Sampling (GGDS), we then further optimize the Gaussian-based scene representation by leveraging 2D priors from the conditional latent diffusion model through consistent diffusion sampling and image-space optimization – together with a set of geometry-grounding regularizers – and generate a causal large-scale scene representation as described in Sec. 3.3.

68], have used accumulated LiDAR point clouds to supervise the generation of 3D driving scene meshes or directly generate realistic LiDAR point clouds [119]. However, these methods exclusively generate geometry, without the texture or appearance needed for training perception models in simulation – see the bottom section of Tab. 1.

Neural scene reconstruction [59] on the other side has shown promising results for the production of high-quality, photorealistic visual data. 3D Gaussians [42] have emerged as prime scene representations, able to explicitly model geometry while also allowing for real-time rendering, enabling scalability. Methods such as OmniRe [8], SplatAD [31], SCube [68] or STORM [104] are capable of reconstructing 3D texture and geometry as 3D Gaussians from real-world driving videos, allowing for the exploration of novel trajectories. Nevertheless, pure reconstruction methods are still fundamentally limited by the availability of real data to reconstruct from. A natural extension of these works has been undertaken by works like WoVoGen [55] and InfiniCube [56], which replace the real data required for scene reconstruction with generated videos conditioned on scene geometry – the latter then fits these videos onto a set of deformable Gaussian Splats, a first approach in 3D grounding. However, this approach is still inherently limited to the generated video trajectory. Furthermore, despite the explicit 3D representation, the lack of causality in a single original video results in visual artifacts and inconsistencies being baked into the scene representation. In contrast, distillation is a paradigm which has recently emerged in 3D object generation [64, 1, 50, 58] and sparse reconstruction [95, 97, 53, 20, 24] that is centered around the learning of a neural 3D representation guided by pretrained text-to-image models. This approach makes it possible to synthesize novel 2D views from arbitrary camera positions, while ensuring 3D consistency and visual fidelity throughout the optimization, thereby bypassing the problem of limited 3D data availability. However, generation by distillation works for object-centric views but does not naturally scale to complex textures and large-scale scenes, as is necessary for driving scene generation. Our work investigates how distillation can be extended for large-scale driving scene generation, allowing for the *distillation* of image priors as part of the 3D generation – as opposed to explicit *supervision*.

3 Geometry-Grounded 3D Generation

Our method generates large-scale driving scenes with 3D-consistent geometry and texture. We provide an overview of the generative process in Fig. 1. In the following, we first describe our large-scale scene representation before introducing the proposed conditional generation process.

3.1 Scene Representation

Geometric Layout and Background Environment. The coarse geometric layout of the scene encodes road, rough vegetation, static vehicles, and building facades. The layout is represented by a mesh $\mathcal{M} = \{\mathbf{F_1}, ..., \mathbf{F_N}\}$ where each triangular face is defined by three vertices $\mathbf{F} = [\mathbf{V}_a, \mathbf{V}_b, \mathbf{V}_c], \mathbf{V} \in \mathbb{R}^3$.

We model the background texture at infinity with an environment map [28]. For a given time of day, weather, and seasonal setting, we introduce into our scenes a corresponding background environment in the form of an equirectangular

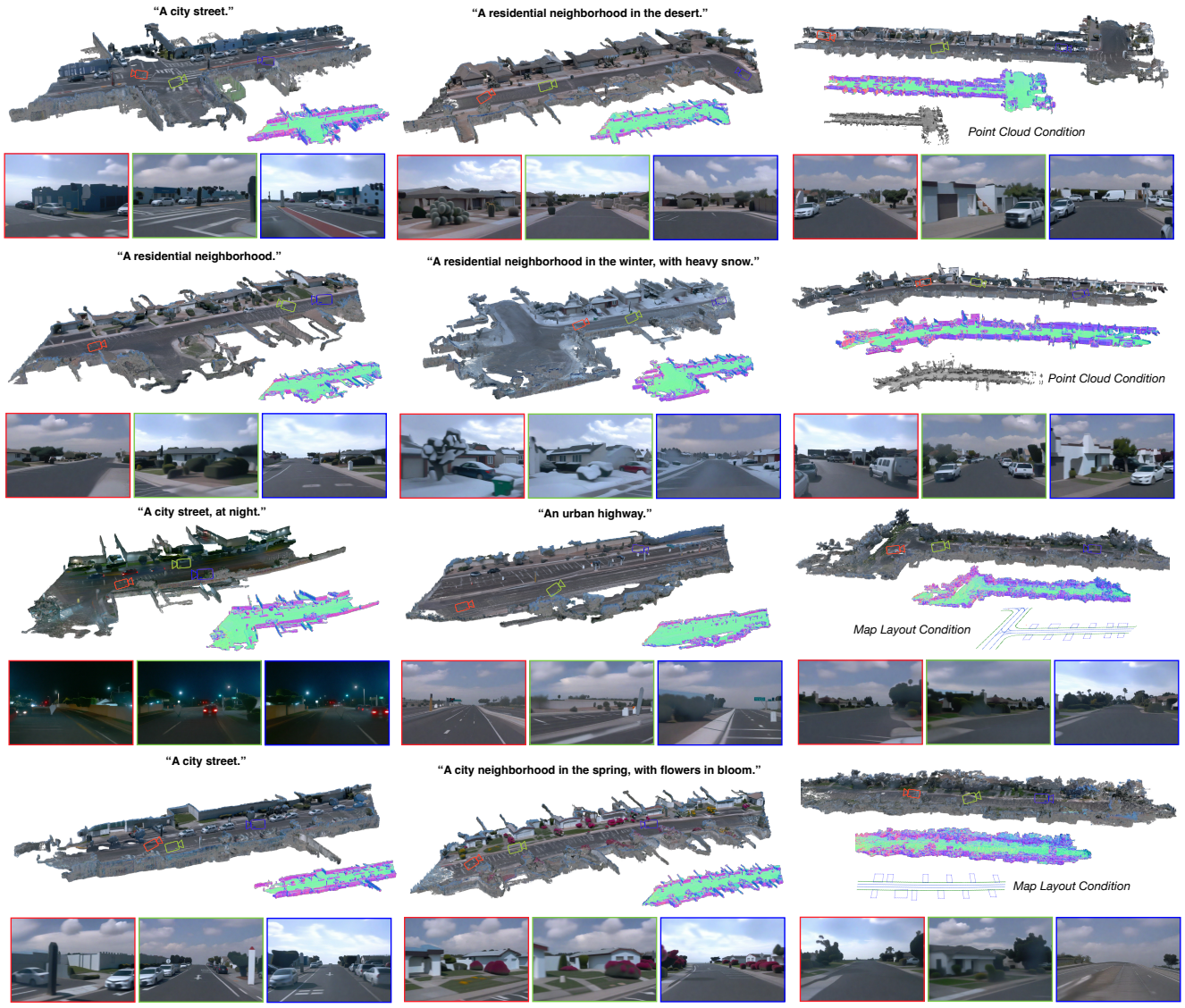

Figure 2: **Geometry-Grounded LSD-3D Generations**. We visualize 3D scenes generated via our method, alongside the corresponding map of surface normals and selection of novel viewpoints at street level for each of them. In the first two columns, we provide samples of scenes with diversity in time-of-day, season, location, and scene type. In the third column, we provide examples of generated scenes with a point cloud condition and map layout condition. We confirm the method generates diverse, explicit, and causal 3D scenes.

map $\mathcal{E}$ of size $H_{env} \times W_{env} \times 3$, that offers explicit environment lighting control. Queried as a spherical environment map $f_{env}(\mathbf{d}, \mathcal{E})$, it returns a color $\mathbf{c}$ for any given viewing directions $\mathbf{d} = (\varphi, \eta) \in \mathbb{R}, (0, 2\pi]$.

Gaussian Structure and Texture. On a finer level, we represent detailed foreground geometry and texture as a set of 2D oriented planar Splats introduced by Huang *et al.* [36]. Each splat θ_k is parametrized by its central point $\mathbf{p}_k \in \mathbb{R}^3$, two principal tangential vectors $\mathbf{t} = (\mathbf{t_u}, \mathbf{t_v})$ that define their orientation, and a variance controlling its scale per axis $\mathbf{s} = (s_u, s_v)$. Our complete scene representation Θ is defined as the set of all K individual Gaussian $\Theta = [\theta_0, ..., \theta_K]$. Complex textures are further modeled by the Gaussian appearance $\mathbf{c}$ (stored as a set of spherical harmonics) and opacity $\mathbf{o}$. The rendered foreground is alpha-composited with the environment map rendered at infinity for each pixel's viewing direction.

3.2 Geometric Layout Generation

To generate the foreground mesh geometry, we first generate a voxel occupancy V from a hierarchical latent voxel diffusion model [71], which makes use of a dense, low-resolution and a sparse, high-resolution 3D UNet [118, 67] as its respective backbones (see the Supplementary Mate-

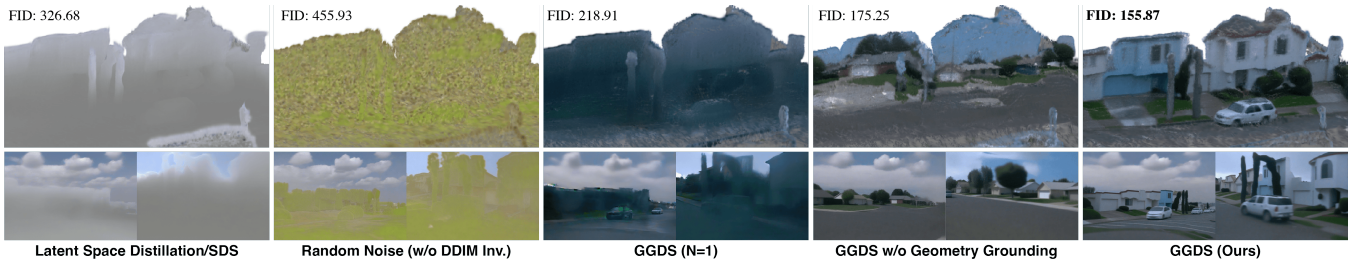

Figure 3: **Ablation Experiments**. We qualitatively validate the core components of our optimization method. With vanilla SDS, scenes completely fail to converge, necessitating our Gaussian Optimization approach. The proposed texture regularization and initialization approach ensures that scenes converge to a reasonable color distribution, while scenes without them fail. The bottom table reports FID scores without the same components, quantitatively confirming that optimization, geometry-grounding, and texture regularization prove critical for producing high-quality 3D driving scenes.

rial for details). To enable explicit control of the layout of V, we condition the diffusion model on a map layout M to model the conditional probability $p\left(V|M\right)$. As single-scene generations are limited to $100m \times 100m$, we expand the generation size by introducing chunk-wise outpainting. The initial chunk is exclusively conditioned on the locally aligned map M_e, but each subsequent chunk e is also conditioned on an overlapping zone with the previous chunk $e-1$ with $p\left(V_e|M_e, V_{e-1}\right)$. This diffusion model is trained from scratch on the aggregated point-cloud data and maps of the target street scene dataset [82]. From the generated voxel grid, we then predict the enclosing coarse surface mesh geometry $\mathcal{M}$ with neural kernel surface reconstruction [37, 94].

3.3 Geometry-Grounded Scene Generation

Given the generated coarse mesh $\mathcal{M}$ and a monochromatically initialized environment map $\mathcal{E}$, we next generate a textured scene with causal consistency.

Mesh to Gaussian Representation. We place Gaussians Θ at $\mathbf{p}_k$, to represent mesh faces $\mathbf{F}$ and set orientation, scale, and tangential axes according to the triangle normal $\mathbf{n}_F$, and area a_F, with orientation $\mathbf{n}_F = \mathbf{t}_u \times \mathbf{t}_v$ and scale $|\mathbf{s}| = a_F$.

Geometry-Grounded Distillation Sampling (GGDS). We next distill a latent diffusion model (LDM) $p_{\phi,data}\left(\mathbf{z}\right)$ on the set of Gaussians Θ through a novel iterative optimization method, which we term Geometry-Grounded Distillation Sampling (GGDS). This optimization method is designed to avoid the artifacts that are typically present in existing latent space distillation-based models [64, 50, 38] and in ego-centric scenes (see Fig. 3).

In each distillation step, we first obtain an image $x_i = g\left(\Theta, \psi_i\right)$ for viewpoint ψ_i with the rasterization function g. We encode the latent $z_{0,i} = E\left(x_i\right)$ from this image and add noise ϵ of noise level t to obtain the noisy latent z_t. The noise level t is sampled uniformly between t_{max} and t_{min}. The noisy latent z_t is the denoised for N steps and decoded, generating a ground-truth image $\hat{\mathbf{x}}_i = D\left(\hat{\mathbf{z}}_{0,i}\right)$ for the respective viewpoint, inducing a loss in image-space. We formulate the objective as the image reconstruction loss between the generated image $\hat{\mathbf{x}}_i$ and the rendered image $\mathbf{x}_i = g\left(\Theta, \psi_i\right)$:

$$\begin{aligned}\mathcal{L}_{gen}\left(\Theta\right) =& \mathbb{E}_{\psi_i,t}\left[\omega\left(t\right)\left(\|g\left(\Theta,\psi_i\right) - \hat{\mathbf{x}}\| \right.\right. \\ &\left.\left. + \mathcal{L}_{LPIPS}(g\left(\Theta,\psi_i\right), \hat{\mathbf{x}})\right)\right]\end{aligned} \tag{1}$$

where $\omega\left(t\right)$ is the noise-level dependent weight and $\mathcal{L}_{LPIPS}$ is the perceptual similarity [114]. We choose $N = 5$ independently of t, which we show allows for higher generation quality for lower noise level t in the later stages of the scene optimization. To enforce progressive optimization from coarse to fine, the respective generation strength from the image prior is linearly annealed by dropping the lower sampling bound t_{min}. Directly optimizing in image space has significant advantages over score distillation, as our ablation experiments validate, where latent optimization is unable to converge on non-overlapping viewpoints.

To further mitigate randomness that leads to diverging optimization objectives, we enforce consistency between optimization steps at the same viewpoint through DDIM inversion instead of random noise sampling from noise level t. This ensures a higher level of consistency between the rendered $\mathbf{x}_i$ and generated images $\hat{\mathbf{x}}_i$ even in later steps of the optimization, which is in contrast to random sampling, where high noise levels t can lead to extreme disagreement. We propose a fixed N-step DDIM inversion [79] at any noise-level t and directly predict

$$\begin{aligned}\mathbf{z}_{t,i} =& \text{DDIM}^{-1}\left(\mathbf{z}_{t-1,i}, \alpha_t, \alpha_{t-1}\right) \\ =& \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_t - 1}}\left(\mathbf{z}_{t-1} - \sqrt{1-\alpha_{t-1}}\epsilon_\Phi(\mathbf{z}_{t-1})\right) \\ &+ \sqrt{1-\alpha_t}\epsilon_\Phi(\mathbf{z}_{t-1}).\end{aligned} \tag{2}$$

where $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$ represent the noisy latent, ϵ_Φ the predicted noise, and $\{\alpha_t\}_{t=0}^{T}$ indicate noise level indexing a monotonically increasing time schedule. This allows the model to only introduce changes exactly where needed in each optimization step to satisfy the 2D diffusion prior, also confirmed in Fig. 3. Given this loss objective, we then optimize the Gaussian scene representation through Stochastic Gradient Langevin Dynamics (SGLD) updates [43] with

$$\Theta_{k+1} = \Theta_k + \xi\left(\nabla_\Theta \mathcal{L}_{\text{gen}}\left(\Theta_k\right)\right) + \lambda_{noise}\epsilon, \tag{3}$$

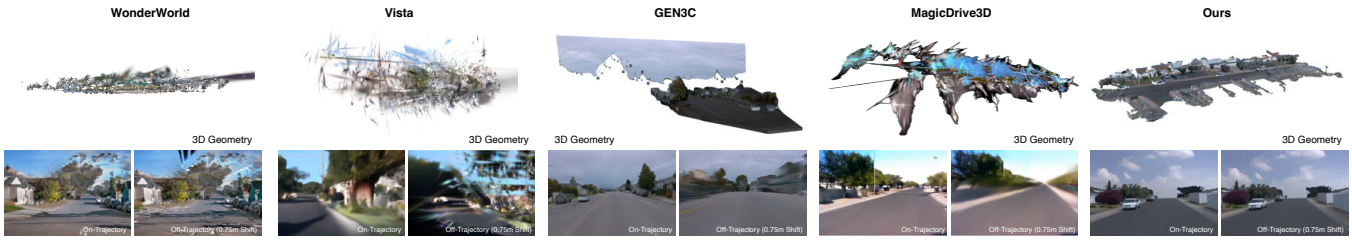

Figure 4: **Qualitative Comparisons to Video and Scene Generation Methods.** Our approach generates an accurate and 3D-consistent scene representation, enabling high-quality novel view synthesis and the generation of unlimited off-trajectory viewpoints. In contrast, existing baselines WonderWorld [110], Vista [25] combined with Gaussian Splatting [36, 42], GEN3C [70] and MagicDrive3D [22], which generate driving videos and thus lack implicit spatial modeling, fail to generate a consistent and 3D-plausible scene, precluding the production of novel driving trajectories (please zoom into PDF version for details).

Method	Training / Finetuning	Seen			Novel			CLIP
		FID ↓	FD_{DINOv2} ↓	FVD ↓	FID ↓	FD_{DINOv2} ↓	FVD ↓	[65] ↑
WonderWorld [110] + 2DGS [36]	WOD [82]	130.17	1333.88	1315.53	<u>220.61</u>	<u>1489.51</u>	1424.56	28.88
Vista [25] + 2DGS	OpenDV [103]/WOD	<u>111.93</u>	1510.25	1023.4	242.03	1805.96	<u>1190.76</u>	26.51
MagicDrive3D [22][1]	NuScenes [6]/WOD[2]	139.11/178.71[2]	1950.67/1965.31[2]	2285.5/1585.48[2]	163.73/186.36[2]	2004.21	1665.30	21.03
GEN3C [70] with Cosmos [60]	Proprietary/WOD	**97.71**	<u>1331.16</u>	**962.02**	273.27	2237.03	1617.62	26.57
LSD-3D Layout + 2DGS Dist. (**Ours**)	WOD	119.38	**1247.19**	<u>989.94</u>	**119.18**	**1227.62**	**974.36**	26.03

Table 2: **Quantitative Evaluation of Generation Quality for 3D Scene Generation.** We compare our proposed method against existing approaches. Best results in each visual quality category are in **bold** and the second best <u>underlined</u>. We report CLIP scores on the right, confirming that our generations adhere to prompts on par with existing methods. We apply the same pre-trained T2I model [63] for all models, and the same 2DGS [36] pipeline for optimization, to provide fair comparisons. Significantly lower FVD and FID scores on novel views confirm temporal and geometric consistency is achieved by our method through geometrically grounded generations.

where $\lambda_{noise}\epsilon$ is the perturbation of each Gaussian in Θ.

As text conditioning c alone is not a suitable prior to achieve the style of street scenes, we first finetune a LDM $p_{\phi,fine-tune}(\mathbf{x})$ on the desired image distribution. Additionally, to avoid scene geometry drifting from the initial mesh, we incorporate disparity conditioning [113] on disparity maps computed from the rendered mesh depth $\mathcal{D} = g_d(\mathcal{M}, \psi_i)$ to the denoising process $p_{\phi,fine-tune}(x|c_{text}, g_d(\mathcal{M}, \psi_i))$. This step is crucial for consistency across views and the gradient signal from the 2D diffusion prior on the generated proxy geometry.

3D Geometry Loss. Alongside geometry conditioning in the 2D diffusion process, we also regularize all θ_k to retain high-quality, smooth 3D geometry. This is achieved by penalizing splat orientation of normal maps $\mathcal{N}_{\Theta,\psi_i}$ and disparity $\mathcal{D}_{\Theta,\psi_i}$ from viewpoint ψ_i with respect to $\mathcal{N}_{\mathcal{M},\psi_i}$ and $\mathcal{D}_{\mathcal{M},\psi_i}$ rendered from the normals of the proxy geometry as

$$\mathcal{L}_{norm} = \|\mathcal{N}_{\Theta,\psi_i} - \mathcal{N}_{\mathcal{M},\psi_i}\| \text{ and } \mathcal{L}_{disp} = \|\mathcal{D}_{\Theta,\psi_i} - \mathcal{D}_{\mathcal{M},\psi_i}\|. \quad (4)$$

Moreover, we apply Gaussian regularization from Huang et. al [36]. Finally, we employ a total-variation (TV) loss on the rendered images to encourage noise reduction in our scenes. We refer to the Supplemental Material for detailed parameter settings and loss composition.

Deferred Rendering. We propose a deferred rendering process for novel trajectory videos, inspired by Thies et al. [87]. After scene generation, we use the Gaussian rasterizer to produce an initial frame $\mathbf{x}_0$ and encode it into a slightly noisy latent $\mathbf{x}_t$, which is deferred into final output $\mathbf{x}_0^{(1)}$ using a fine-tuned T2I model.

This rendering procedure allows us to generate photorealistic images with both low- and high-frequency textures in image space, for instance the road, or a tree, respectively.

4 Assessment

In this section, we validate our approach both quantitatively and qualitatively. We conduct an ablation study to validate our design choices and investigate the realism of rendered scenes alongside competing baseline methods.

4.1 Ablation Study

We validate the effects of key components of our scene optimization method and visualize the results in Fig 3. Specifically, we analyze the effect of incorporating our improved geometry-grounded distillation sampling (GGDS) approach with SDS, image-space sampling without DDIM inversion, the effectiveness of multi-step denoising and geometry-conditioned diffusion guidance. Without the proposed GGDS, especially in the case of SDS and random noise sampling results, the method experiences catastrophic failure. This highlights the need for two major design choices, which both result in more consistent optimization and distillation of the underlying 3D scene representation. Further multi-step diffusion results in reduced scene generation quality of dark scenes. Without geometry-grounding

diffusion guidance, objects in the scene (such as houses or cars) do not follow the conditioning proxy geometry, resulting in inaccurate and flat 3D geometry and poor novel trajectory results. We confirm the visual trend quantitatively in Fig. 3 bottom, which validates that our method (with all proposed components) produces the best visual quality.

4.2 Experiments

We validate our method by comparing it against four distinct approaches for 3D scene and video generation.

Baselines. We ran two image-to-video generation methods: (a) Vista [25], a full driving video model trained on internet-scale driving videos, and (b) WonderWorld [110], in combination with our fine-tuned image model [63] on the target dataset [82] for fair comparison. For both methods, we fit 2D Gaussian Splats [36] to assess geometric causality and novel view synthesis quality.

We also test two conceptually different 3D generation baselines: (c) MagicDrive3D [22], which proposes a driving scene-specific multi-view video diffusion and Gaussian reconstruction pipeline, and (d) GEN3C [70], which fuses geometry prediction from a single image with the Cosmos-Predict [60] video diffusion model. Due to the unavailability of public code and models of any candidate [22, 56] at the time of submission, we rely on our own implementation (Layout + Geometry Controlled Video Generation) with MagicDrive3D built on top of the latest diffusion transformer version of MagicDriveDiT [23]. See the Supplementary Material for details.

Computational Requirements. Each scene is generated with 6000 steps, corresponding to an average time of 2 hours on a single NVIDIA H100 GPU. Comparable methods [50] that generate single objects, i.e., dramatically smaller scenes, require similar runtime at 5000 steps. Our method outputs frames at 960p resolution at rates higher than 60 fps, providing real-time rendering capabilities. Scenes are initialized with 1.8 to 2.2 million Gaussians, and the maximum is set to 4 million Gaussians.

Evaluation Metrics. For all methods, we quantitatively assess the diversity and quality of our results by computing the Fréchet Inception Distance (FID) and the recently established DINOv2 [61] based Fréchet Distance $\text{FD}_{\text{DINOv2}}$ [81]. For temporal quality, we use the Fréchet Video Distance (FVD), with a subset of the respective training dataset [82, 6] as reference distribution. Following [22], we evaluate FID and $\text{FD}_{\text{DINOv2}}$ score on generated results from views seen during the Gaussian Splatting optimization (FID *seen*) as well as from novel views sampled at randomly selected distances from the training ones (FID *novel*). Additionally, we evaluate prompt adherence using the CLIP score [65] with the implementation from [117] on 10 common weather, time-of-day, and localization prompts with 3 samples each.

Quantitative Results. We evaluate all methods on a set of 40 generated scenes and across ten different scene attributes (time of day, season, weather). Exact prompts are provided in the Supplementary material. As reported in Tab. 2, the quality of rendered images from our model is on par with state-of-the-art 2D generative methods and is capable of generating scenes closely matching in style and content with the source distribution However, *competing methods cannot generate a 3D-consistent scene*, resulting in novel views with inferior quality – as seen in the resulting high FID and FVD score for novel views all other methods. We also confirm prompt adherence at par with video models, validating that 3D grounding enables fine-grained text control comparable to generalist models [111].

Figure 5: **Composability with Dynamic Actors.** We simulate driving trajectories in a residential street scenes for a Waymo-representative [82] sensor stack. From bottome to top, we show a third-person view of the ego capture vehicle followed by a set of rendered front cameras. Both columns correspond to a different timestamp of the same traffic configurations + *fully generated* scene & asset models. The second vehicle is hidden behind the corner in the first frame.

Qualitative Results. We also find significant qualitative differences between baselines and our method and baselines in Fig. 4. Baseline image-to-video models produce scene renderings of variable quality. In fact, as views deviate further from the input image, the rendering quality and 3D consistency deteriorate, yielding a Gaussian representation which is inconsistent beyond the original trajectory. This difference is more pronounced when departing from the generated video trajectory - even 2.5D methods fail to produce consistent novel views for unlimited viewpoints. In contrast, our method generates plausible rendering throughout the entire scene, from any realistic viewpoint - without any loss of appearance or geometric quality.

4.3 Composability with Dynamic Actors

Being able to place dynamic actors in the generated scene is crucial for downstream applications, including real-time closed-loop simulation. Our representation directly allows for plug-and-play usage with other components of the simulation stack. We illustrate the integration with actor assets, traffic generation, and sensor stack rendering in Fig. 5. Using either generated 3D objects, reconstructed objects, or even synthetic objects, the environment map can be used to relight added assets. In addition, map conditioning directly supports

[1] No code publicly available (or from authors). Reimplemented with MagicDriveDiT [23] and 2DGS [36].

[2] We compare scores on the Waymo [82] distributions and released video generations. NuScenes [6] results are reported first for completeness of the evaluation.

the use of standard traffic generation [41, 16, 48, 19] and planning modules, providing realistic asset placement and integration in driving scenes, as seen with cars from image-to-3D pipelines [99] placed in our generated scenes. At each timestamp, we sample object poses from the generated trajectories to place agents within the scene. These agents are then relit according to the environment map, and the scene is rendered through the ego vehicle's sensor stack.

5 Conclusion

We introduce, to our knowledge, the first distillation approach to directly generate large-scale explicit 3D driving scenes. To accomplish this, we propose Geometry-Grounded Distillation Sampling (GGDS), which combines controlled proxy mesh generation with a conditional diffusion prior, producing novel and view-consistent Gaussian splatting scenes. Our approach generates completely unseen driving environments controlled by scene descriptions or traffic map layouts. By the design of our method, every scene is generated causally and 3D-consistent, and allows for real-time rendering of physically-grounded videos along novel trajectories. The approach compares favorably against the most successful existing methods – primarily video diffusion approaches – that struggle with view-consistent rendering and causality, and are fundamentally limited to individual trajectories. As a geometry-grounded approach, we hope to integrate the method with driving simulators and extend the domain beyond autonomous driving – ultimately building towards the goal of fully data-driven simulators.

References

[1] Titas Anciukevicius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J. Mitra, and Paul Guerrero. RenderDiffusion: Image diffusion for 3D reconstruction, inpainting and generation. *arXiv*, 2022.

[2] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023.

[3] 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.

[4] Steve Borkman, Adam Crespi, Saurav Dhakad, Sujoy Ganguly, Jonathan Hogins, You-Cyuan Jhang, Mohsen Kamalzadeh, Bowen Li, Steven Leal, Pete Parisi, et al. Unity perception: Generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*, 2021.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[7] Anthony Chen, Wenzhao Zheng, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Kurt Keutzer, and Shanghang Zhang. Geodrive: 3d geometry-informed driving world model with precise action control, 2025.

[8] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[11] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[12] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023.

[13] 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 *SIGGRAPH 2024 Conference Papers*, 2024.

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[15] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.

[16] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[18] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, 2021.

[19] Lan Feng, Quanyi Li, Zhenghao Peng, Shuhan Tan, and Bolei Zhou. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3567–3575. IEEE, 2023.

[20] Tobias Fischer, Samuel Rota Bulò, Yung-Hsu Yang, Nikhil Varma Keetha, Lorenzo Porzi, Norman Müller, Katja Schwarz, Jonathon Luiten, Marc Pollefeys, and Peter Kontschieder. FlowR: Flowing from sparse to dense 3d reconstructions. *arXiv preprint arXiv:2504.01647*, 2025.

[21] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023.

[22] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024.

[23] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDriveDiT: High-resolution long video generation for autonomous driving with adaptive control, 2024.

[24] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024.

[25] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[26] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[28] Ned Greene. Environment mapping and other applications of world projections. *IEEE computer graphics and Applications*, 6(11):21–29, 1986.

[29] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.

[30] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022.

[31] Georg Hess, Carl Lindström, Maryam Fatemi, Christoffer Petersson, and Lennart Svensson. Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11982–11992, 2025.

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

[33] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[34] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[35] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[36] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024.

[37] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023.

[38] Sungwon Hwang, Min-Jung Kim, Taewoong Kang, Jayeon Kang, and Jaegul Choo. Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[39] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023.

[40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[41] Saman Kazemkhani, Aarav Pandya, Daphne Cornelisse, Brennan Shacklett, and Eugene Vinitsky. Gpudrive: Data-driven, multi-agent driving simulation at 1 million fps. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

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

[43] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Jeff Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo, 2024.

[44] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8496–8506, 2023.

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

[46] Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. Semcity: Semantic scene generation with triplane diffusion. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, 2024.

[47] Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-hang Lee, and Pengyuan Zhou. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. *arXiv preprint arXiv:2404.03575*, 2024.

[48] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[49] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025.

[50] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6517–6526, 2024.

[51] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

[52] 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.

[53] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024.

[54] William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. Neuroncap: Photorealistic closed-loop safety testing for autonomous driving. *arXiv preprint arXiv:2404.07762*, 2024.

[55] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision (ECCV)*, 2024.

[56] Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, and Jiahui Huang. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models, 2024.

[57] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[58] Xingyu Miao, Haoran Duan, Varun Ojha, Jun Song, Tejal Shah, Yang Long, and Rajiv Ranjan. Dreamer xl: Towards high-resolution text-to-3d generation via trajectory score matching. *arXiv preprint arXiv:2405.11252*, 2024.

[59] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[60] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025.

[61] 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.

[62] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021.

[63] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[64] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

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

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

[68] Xuanchi Ren, Yifan Lu, Hanxue Liang, Jay Zhangjie Wu, Huan Ling, Mike Chen, Francis Fidler, Sanja annd Williams, and Jiahui Huang. Scube: Instant large-scale scene reconstruction using voxsplats. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[69] Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, Seung Wook

Kim, Jun Gao, Laura Leal-Taixe, Mike Chen, Sanja Fidler, and Huan Ling. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models, 2025.

[70] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6121–6132, 2025.

[71] 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.

[72] Guodong Rong, Byung Hyun Shin, Hadi Tabatabaee, Qiang Lu, Steve Lemke, Mārtiņš Možeiko, Eric Boise, Geehoon Uhm, Mark Gerow, Shalin Mehta, et al. Lgsvl simulator: A high fidelity simulator for autonomous driving. In *2020 IEEE 23rd International conference on intelligent transportation systems (ITSC)*, pages 1–6. IEEE, 2020.

[73] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023.

[74] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[75] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018.

[76] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024.

[77] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023.

[78] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[79] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020.

[80] Yuda Song, Zehao Sun, and Xuanwu Yin. Sdxs: Real-time one-step latent diffusion models with image conditions. *arXiv preprint arXiv:2403.16627*, 2024.

[81] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

[82] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[83] Deepak Talwar, Sachin Guruswamy, Naveen Ravipati, and Magdalini Eirinaki. Evaluating validity of synthetic data in perception tasks for autonomous vehicles. In *2020 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pages 73–80. IEEE, 2020.

[84] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025.

[85] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

[86] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025.

[87] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[88] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. *arXiv preprint arXiv:2311.15260*, 2023.

[89] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[90] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021.

[91] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[92] Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, et al. Taming mode collapse in score distillation for text-to-3d generation. *arXiv preprint arXiv:2401.00909*, 2023.

[93] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

[94] Francis Williams, Zan Gojcic, Sameh Khamis, Denis Zorin, Joan Bruna, Sanja Fidler, and Or Litany. Neural fields as learnable kernels for 3d reconstruction, 2021.

[95] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025.

[96] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv:2411.18613*, 2024.

[97] 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.

[98] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, pages 3–15. Springer, 2023.

[99] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

[100] 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.

[101] Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. Latte3d: Large-scale amortized text-to-enhanced3d synthesis. *arXiv preprint arXiv:2403.15385*, 2024.

[102] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, et al. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4402–4412, 2023.

[103] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[104] Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Maximilian Igl, Apoorva Sharma, Peter Karkus, Danfei Xu, Boris Ivanovic, Yue Wang, and Marco Pavone. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. *arXiv preprint arXiv:2501.00602*, 2025.

[105] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. *arXiv preprint arXiv:2402.03162*, 2024.

[106] Yuanbo Yang, Yifei Yang, Hanlei Guo, Rong Xiong, Yue Wang, and Yiyi Liao. Urbangiraffe: Representing urban scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9199–9210, 2023.

[107] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *CVPR*, 2023.

[108] 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. *arXiv preprint arXiv:2408.06072*, 2024.

[109] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024.

[110] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024.

[111] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024.

[112] Junge Zhang, Qihang Zhang, Li Zhang, Ramana Rao Kompella, Gaowen Liu, and Bolei Zhou. Urban scene diffusion through semantic occupancy map. *arXiv preprint arXiv:2403.11697*, 2024.

[113] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[114] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[115] Songchun Zhang, Yibo Zhang, Quan Zheng, Rui Ma, Wei Hua, Hujun Bao, Weiwei Xu, and Changqing Zou. 3d-scenedreamer: Text-driven 3d-consistent scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10170–10180, 2024.

[116] Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. *arXiv preprint arXiv:2403.17005*, 2024.

[117] SUN Zhengwentai. clip-score: CLIP Score for PyTorch. https://github.com/taited/clip-score, 2023. Version 0.2.1.

[118] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5826–5835, 2021.

[119] Vlas Zyrianov, Henry Che, Zhijian Liu, and Shenlong Wang. Lidardm: Generative lidar simulation in a generated world. *arXiv preprint arXiv:2404.02903*, 2024.