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: 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 [27, 11, 87, 55, 19, 7] have unlocked new capabilities for perception and planning. Initially limited to a few sparsely labeled scenes [27], existing datasets now offer thousands of multi-modal, fully annotated scenes from cities around the world [7, 87] – 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 [17, 30, 5, 76, 80, 88] 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 [66, 93, 112, 103]. Such scenes can be replayed in real time, synthesize novel views, and allow for unseen actor variations in closed-loop testing [93, 58, 12]. 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 [98, 59, 26, 72]. 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 [50, 117, 71, 124] – 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 [59, 60]. As such, these methods inherit video diffusion limitations, such as a lack of causality. Distillation methods [68] instead address the 3D data gap by transferring priors from 2D image models into 3D representations via inverse rendering techniques [63, 46, 114]. However, existing techniques are limited to object-centric generation [68, 106, 104] and lack realism [51] — and so, they are not suitable for the complexity [120] or spatial scale [81] of large-scale driving scenes. Only recent work has explored distillation approaches for scene extrapolation from sparse real data cap-

[*]Indicates equal contribution.

tures [100, 102, 57, 21], 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 [87], 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) [28, 6, 42, 78], likelihood-based methods [49, 70, 94], to more prominently, diffusion models [15, 84, 16, 18, 34, 75, 95] – have been recently extended to video generation [4, 121, 83, 75, 31, 33, 35, 23, 113, 36, 90]. Early methods offer control via text prompt [83], image [121], or camera position [77, 110]. Recent methods like GEN3C [74, 89] or GeoDrive [8] for driving scenes incorporate 3D conditioning and point projection to improve geometric consistency. Specialized models, such as Vista [26], MagicDrive [24, 23], or DriveDreamer [98], have been developed on top of foundational video models [4] 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 [85, 61], video generation models remain computationally intensive [113, 90, 96] 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 [98]	✗	✗	✗	✗	✗	✓	✓	✗
WonderJourney [116]	✗	✗	✗	✗	✗	✓	✓	✗
Streetscapes [14]	✗	✗	✗	✗	✗	✓	✓	✓
Vista [26]	✗	✗	✗	✗	✗	✓	✓	✗
MagicDriveDiT [23]	✗	✗	✗	✗	✗	✓	✓	✗
WoVoGen [59]	✗	✗	✗	✗	✗	✓	✓	✓
WonderWorld [115]	✗	✗	✗	✗	✗	✓	✓	✗
NF-LDM [48]	✗	✗	✗	✓	✗	✓	✓	✓
InfiniCity [56]	✗	✗	✓	✓	✓	✗	✗	✓
MagicDrive3D [26]	✗	✓	(✓)	✓	✓	✓	✓	✓
CityDreamer [105]	✗	✗	✓	✓	✓	✗	✗	✓
InfiniCube [60]	✗	✓	(✓)	✓	✓	✓	✓	✓
GEN3C [74]	✗	✗	(✓)	✗	✓	✓	✓	✗
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 [104, 53, 101, 25, 41, 77, 91, 106] 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 [13, 10], 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 [48, 111, 50, 1, 97, 2, 56], various 3D representations have been proposed: triplanes [82, 50, 1], semantic occupancy grids, bounding boxes, and 3D maps [111, 107]. 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 [56, 105]. More recently, hierarchical voxel diffusion methods [71,

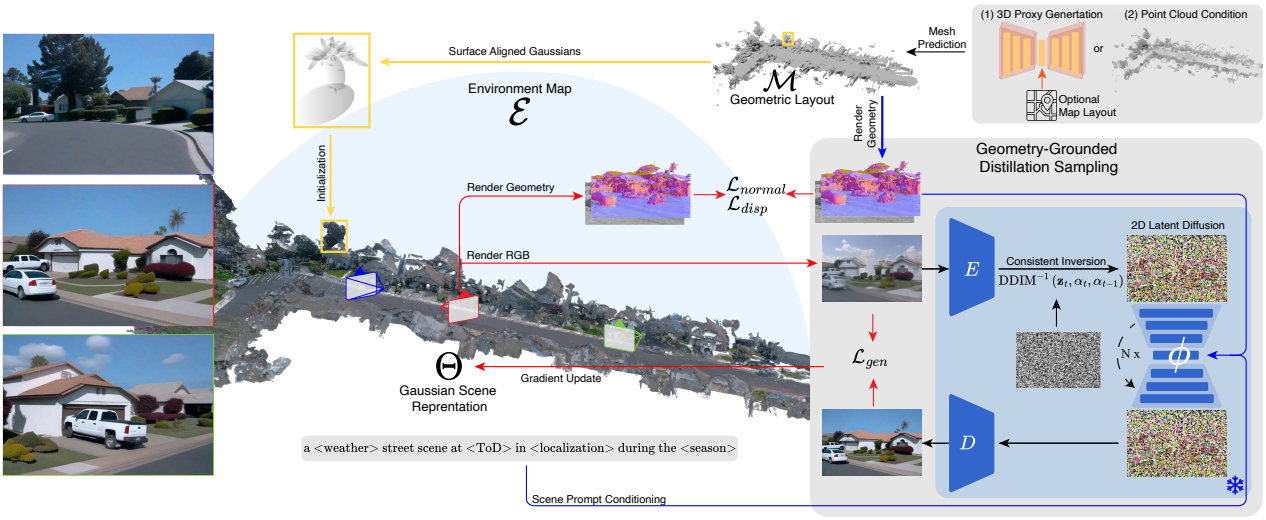

Figure 1: **Geometry-Grounded Large-Scale 3D Scene Generation.** We generate a large-scale scene as a combination of a coarse geometric layout, an environment map, and a set of Gaussians for texture details, discussed in Sec. 3.1. The geometric layout is either generated, conditioned on a map, or predicted 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.

73], have used accumulated LiDAR point clouds to supervise the generation of 3D driving scene meshes or directly generate realistic LiDAR point clouds [124]. 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 [63] on the other side has shown promising results for the production of high-quality, photorealistic visual data. 3D Gaussians [46] have emerged as prime scene representations, able to explicitly model geometry while also allowing for real-time rendering, enabling scalability. Methods such as OmniRe [9], SplatAD [32], SCube [73] or STORM [109] 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 [59] and InfiniCube [60], 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 [68, 1, 54, 62] and sparse reconstruction [100, 102, 57, 21, 25] 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 [29]. 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

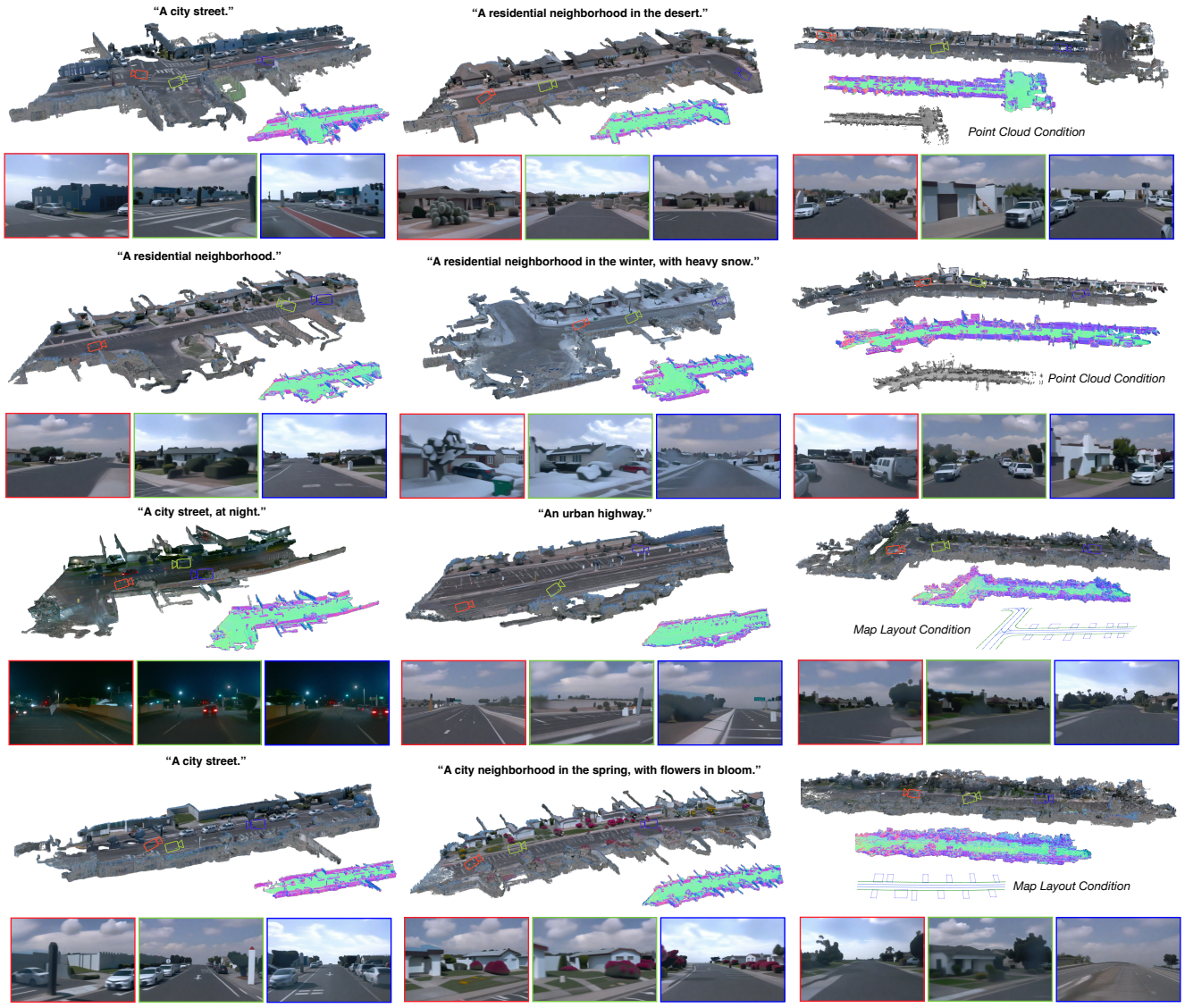

Figure 2: **Geometry-Grounded LSD-3D Generations**. We visualize 3D scenes generated via our method, alongside the corresponding map of surface normals and selection of novel viewpoints at street level for each of them. In the first two columns, we provide samples of scenes with diversity in time-of-day, 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.* [2024]. 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 [75], which makes use of a dense, low-resolution and a sparse, high-resolution 3D UNet [123, 71] as its respective backbones (see the Supplementary Mate-

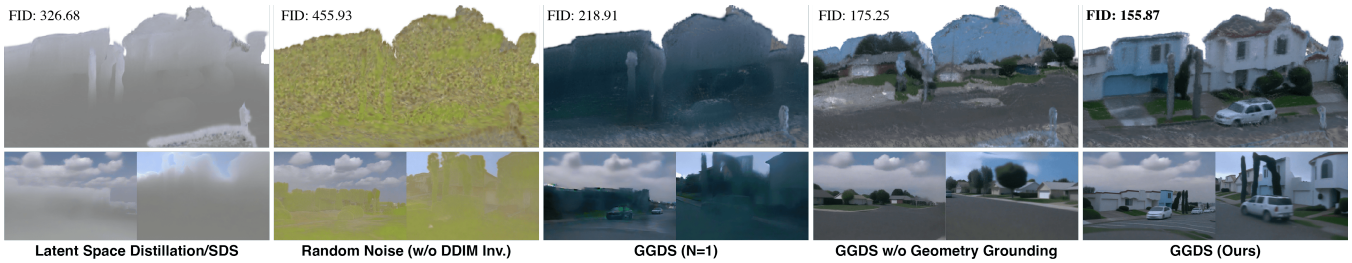

Figure 3: **Ablation Experiments**. We qualitatively validate the core components of our optimization method. With vanilla SDS, scenes completely fail to converge, necessitating our Gaussian Optimization approach. The proposed texture regularization and initialization approach ensures that scenes converge 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 [87]. From the generated voxel grid, we then predict the enclosing coarse surface mesh geometry $\mathcal{M}$ with neural kernel surface reconstruction [38, 99].

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 [68, 54, 39] 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 $\mathbf{x}_i = g\left(\Theta, \psi_i\right)$:

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

where $\omega\left(t\right)$ is the noise-level dependent weight and $\mathcal{L}_{LPIPS}$ is the perceptual similarity [119]. 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 [84] 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 [47] with

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

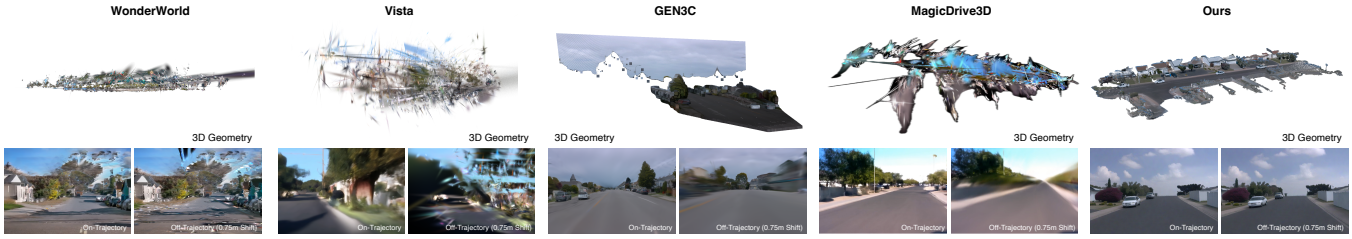

Figure 4: **Qualitative Comparisons to Video and Scene Generation Methods.** Our approach generates an accurate and 3D-consistent scene representation, enabling high-quality novel view synthesis and the generation of unlimited off-trajectory viewpoints. In contrast, existing baselines WonderWorld [115], Vista [26] combined with Gaussian Splatting [37, 46], GEN3C [74] 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 ↓	[2021] ↑
WonderWorld [2024] + 2DGS [2024]	WOD [2020]	130.17	1333.88	1315.53	<u>220.61</u>	<u>1489.51</u>	1424.56	28.88
Vista [2024] + 2DGS	OpenDV [2024]/WOD	<u>111.93</u>	1510.25	1023.4	242.03	1805.96	<u>1190.76</u>	26.51
MagicDrive3D [2024] [1]	NuScenes [2020]/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 [2025] with Cosmos [2025]	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** with our proposed method and 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 [67] for all models, and the same 2DGS [37] pipeline for optimization, to provide fair comparisons. Significantly lower FVD and FID score 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 [118] 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 [2024]. 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. [2019]. 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 [26], a full driving video model trained on internet-scale driving videos, and (b) WonderWorld [115], in combination with our fine-tuned image model [67] on the target dataset [87] for fair comparison. For both methods, we fit 2D Gaussian Splats [37] 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 [74], which fuses geometry prediction from a single image with the Cosmos-Predict [64] video diffusion model. Due to the unavailability of public code and models of any candidate [22, 60] 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 [54] 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 [65] based Fréchet Distance $\text{FD}_{\text{DINOv2}}$ [86]. For temporal quality, we use the Fréchet Video Distance (FVD), with a subset of the respective training dataset [87, 7] 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 [69] with the implementation from [122] 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 [116].

Figure 5: **Composability with Dynamic Actors.** We simulate driving trajectories in a residential street scenes for a Waymo-representative [87] 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

The composability of generated scenes with dynamic actor assets and autonomous agents 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 the use of standard traffic

[1] No code publicly available (or from authors). Reimplemented with MagicDriveDiT [2024] and 2DGS [37].

[2] We compare scores on the Waymo [87] distributions and released video generations. NuScenes [7] results are reported first for completeness of the evaluation.

generation [43, 17, 52, 20] and planning modules, providing realistic asset placement and integration in driving scenes, as seen with cars from image-to-3D pipelines [104] 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] Anciukevicius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N. J.; and Guerrero, P. 2022. RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation. *arXiv*.

[2] Bahmani, S.; Skorokhodov, I.; Rong, V.; Wetzstein, G.; Guibas, L.; Wonka, P.; Tulyakov, S.; Park, J. J.; Tagliasacchi, A.; and Lindell, D. B. 2023. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*.

[3] Bao, Y.; Liu, H.; Gao, X.; Fu, H.; and Kang, G. 2025. FreeInv: Free Lunch for Improving DDIM Inversion. *arXiv preprint arXiv:2503.23035*.

[4] Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

[5] Borkman, S.; Crespi, A.; Dhakad, S.; Ganguly, S.; Hogins, J.; Jhang, Y.-C.; Kamalzadeh, M.; Li, B.; Leal, S.; Parisi, P.; et al. 2021. Unity perception: Generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*.

[6] Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

[7] Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*.

[8] Chen, A.; Zheng, W.; Wang, Y.; Zhang, X.; Zhan, K.; Jia, P.; Keutzer, K.; and Zhang, S. 2025. GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control. arXiv:2505.22421.

[9] Chen, Z.; Yang, J.; Huang, J.; de Lutio, R.; Esturo, J. M.; Ivanovic, B.; Litany, O.; Gojcic, Z.; Fidler, S.; Pavone, M.; Song, L.; and Wang, Y. 2025. OmniRe: Omni Urban Scene Reconstruction. In *The Thirteenth International Conference on Learning Representations*.

[10] Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Yago Vicente, T. F.; Dideriksen, T.; Arora, H.; Guillaumin, M.; and Malik, J. 2022. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. *CVPR*.

[11] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

[12] Dauner, D.; Hallgarten, M.; Li, T.; Weng, X.; Huang, Z.; Yang, Z.; Li, H.; Gilitschenski, I.; Ivanovic, B.; Pavone, M.; Geiger, A.; and Chitta, K. 2024. NAVSIM: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[13] Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; VanderBilt, E.; Kembhavi, A.; Vondrick, C.; Gkioxari, G.; Ehsani, K.; Schmidt, L.; and Farhadi, A. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663*.

[14] Deng, B.; Tucker, R.; Li, Z.; Guibas, L.; Snavely, N.; and Wetzstein, G. 2024. Streetscapes: Large-scale Consistent Street View Generation Using Autoregressive Video Diffusion. In *SIGGRAPH 2024 Conference Papers*.

[15] Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

[16] Dockhorn, T.; Vahdat, A.; and Kreis, K. 2021. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*.

[17] Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

[18] Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

[19] Ettinger, S.; Cheng, S.; Caine, B.; Liu, C.; Zhao, H.; Pradhan, S.; Chai, Y.; Sapp, B.; Qi, C. R.; Zhou, Y.; Yang, Z.; Chouard, A.; Sun, P.; Ngiam, J.; Vasudevan, V.; McCauley, A.; Shlens, J.; and Anguelov, D. 2021. 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)*, 9710–9719.

[20] Feng, L.; Li, Q.; Peng, Z.; Tan, S.; and Zhou, B. 2023. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 3567–3575. IEEE.

[21] Fischer, T.; Bulò, S. R.; Yang, Y.-H.; Keetha, N. V.; Porzi, L.; Müller, N.; Schwarz, K.; Luiten, J.; Pollefeys, M.; and Kontschieder, P. 2025. FlowR: Flowing from Sparse to Dense 3D Reconstructions. *arXiv preprint arXiv:2504.01647*.

[22] Gao, R.; Chen, K.; Li, Z.; Hong, L.; Li, Z.; and Xu, Q. 2024. MagicDrive3D: Controllable 3D Generation for Any-View Rendering in Street Scenes. *arXiv preprint arXiv:2405.14475*.

[23] Gao, R.; Chen, K.; Xiao, B.; Hong, L.; Li, Z.; and Xu, Q. 2024. MagicDriveDiT: High-Resolution Long Video Generation for Autonomous Driving with Adaptive Control. arXiv:2411.13807.

[24] Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2023. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*.

[25] Gao*, R.; Holynski*, A.; Henzler, P.; Brussee, A.; Martin-Brualla, R.; Srinivasan, P. P.; Barron, J. T.; and Poole*, B. 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *Advances in Neural Information Processing Systems*.

[26] Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[27] Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

[28] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

[30] Gulino, C.; Fu, J.; Luo, W.; Tucker, G.; Bronstein, E.; Lu, Y.; Harb, J.; Pan, X.; Wang, Y.; Chen, X.; Co-Reyes, J. D.; Agarwal, R.; Roelofs, R.; Lu, Y.; Montali, N.; Mougin, P.; Yang, Z.; White, B.; Faust, A.; McAllister, R.; Anguelov, D.; and Sapp, B. 2023. 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*.

[31] Harvey, W.; Naderiparizi, S.; Masrani, V.; Weilbach, C.; and Wood, F. 2022. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35: 27953–27965.

[32] Hess, G.; Lindström, C.; Fatemi, M.; Petersson, C.; and Svensson, L. 2025. Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11982–11992.

[33] Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

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

[35] Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

[36] Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

[37] Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery.

[38] Huang, J.; Gojcic, Z.; Atzmon, M.; Litany, O.; Fidler, S.; and Williams, F. 2023. Neural Kernel Surface Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4369–4379.

[39] Hwang, S.; Kim, M.-J.; Kang, T.; Kang, J.; and Choo, J. 2024. Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. In *European Conference on Computer Vision*, 1–18. Springer.

[40] Jain, J.; Li, J.; Chiu, M.; Hassani, A.; Orlov, N.; and Shi, H. 2023. OneFormer: One Transformer to Rule Universal Image Segmentation.

[41] Jun, H.; and Nichol, A. 2023. Shap-E: Generating Conditional 3D Implicit Functions. arXiv:2305.02463.

[42] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

[43] Kazemkhani, S.; Pandya, A.; Cornelisse, D.; Shacklett, B.; and Vinitsky, E. 2025. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[44] Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] Ke, B.; Qu, K.; Wang, T.; Metzger, N.; Huang, S.; Li, B.; Obukhov, A.; and Schindler, K. 2025. Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis. arXiv:2505.09358.

[46] Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

[47] Kheradmand, S.; Rebain, D.; Sharma, G.; Sun, W.; Tseng, J.; Isack, H.; Kar, A.; Tagliasacchi, A.; and Yi, K. M. 2024. 3D Gaussian Splatting as Markov Chain Monte Carlo. arXiv:2404.09591.

[48] Kim, S. W.; Brown, B.; Yin, K.; Kreis, K.; Schwarz, K.; Li, D.; Rombach, R.; Torralba, A.; and Fidler, S. 2023. NeuralField-LDM: Scene Generation With Hierarchical Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8496–8506.

[49] Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[50] Lee, J.; Lee, S.; Jo, C.; Im, W.; Seon, J.; and Yoon, S.-E. 2024. SemCity: Semantic Scene Generation with Triplane Diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[51] Li, H.; Shi, H.; Zhang, W.; Wu, W.; Liao, Y.; Wang, L.; Lee, L.-h.; and Zhou, P. 2024. DreamScene: 3D Gaussian-based Text-to-3D Scene Generation via Formation Pattern Sampling. *arXiv preprint arXiv:2404.03575*.

[52] Li, Q.; Peng, Z.; Feng, L.; Zhang, Q.; Xue, Z.; and Zhou, B. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[53] Li, Y.; Zou, Z.-X.; Liu, Z.; Wang, D.; Liang, Y.; Yu, Z.; Liu, X.; Guo, Y.-C.; Liang, D.; Ouyang, W.; et al. 2025. TripoSG: High-Fidelity 3D Shape Synthesis using Large-Scale Rectified Flow Models. *arXiv preprint arXiv:2502.06608*.

[54] Liang, Y.; Yang, X.; Lin, J.; Li, H.; Xu, X.; and Chen, Y. 2024. 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*, 6517–6526.

[55] Liao, Y.; Xie, J.; and Geiger, A. 2022. 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.

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

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

[58] Ljungbergh, W.; Tonderski, A.; Johnander, J.; Caesar, H.; Åström, K.; Felsberg, M.; and Petersson, C. 2024. NeuroNCAP: Photorealistic Closed-loop Safety Testing for Autonomous Driving. *arXiv preprint arXiv:2404.07762*.

[59] Lu, J.; Huang, Z.; Zhang, J.; Yang, Z.; and Zhang, L. 2024. WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation. In *European Conference on Computer Vision (ECCV)*.

[60] Lu, Y.; Ren, X.; Yang, J.; Shen, T.; Wu, Z.; Gao, J.; Wang, Y.; Chen, S.; Chen, M.; Fidler, S.; and Huang, J. 2024. InfiniCube: Unbounded and Controllable Dynamic 3D Driving Scene Generation with World-Guided Video Models. arXiv:2412.03934.

[61] Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

[62] Miao, X.; Duan, H.; Ojha, V.; Song, J.; Shah, T.; Long, Y.; and Ranjan, R. 2024. Dreamer XL: Towards High-Resolution Text-to-3D Generation via Trajectory Score Matching. *arXiv preprint arXiv:2405.11252*.

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

[64] NVIDIA; Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; Dworakowski, D.; Fan, J.; Fenzi, M.; Ferroni, F.; Fidler, S.; Fox, D.; Ge, S.; Ge, Y.; Gu, J.; Gururani, S.; He, E.; Huang, J.; Huffman, J.; Jannaty, P.; Jin, J.; Kim, S. W.; Klár, G.; Lam, G.; Lan, S.; Leal-Taixe, L.; Li, A.; Li, Z.; Lin, C.-H.; Lin, T.-Y.; Ling, H.; Liu, M.-Y.; Liu, X.; Luo, A.; Ma, Q.; Mao, H.; Mo, K.; Mousavian, A.; Nah, S.; Niverty, S.; Page, D.; Paschalidou, D.; Patel, Z.; Pavao, L.; Ramezanali, M.; Reda, F.; Ren, X.; Sabavat, V. R. N.; Schmerling, E.; Shi, S.; Stefaniak, B.; Tang, S.; Tchapmi, L.; Tredak, P.; Tseng, W.-C.; Varghese, J.; Wang, H.;

Wang, H.; Wang, H.; Wang, T.-C.; Wei, F.; Wei, X.; Wu, J. Z.; Xu, J.; Yang, W.; Yen-Chen, L.; Zeng, X.; Zeng, Y.; Zhang, J.; Zhang, Q.; Zhang, Y.; Zhao, Q.; and Zolkowski, A. 2025. Cosmos World Foundation Model Platform for Physical AI.

[65] Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[66] Ost, J.; Mannan, F.; Thuerey, N.; Knodt, J.; and Heide, F. 2021. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2856–2865.

[67] Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

[68] Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv*.

[69] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

[70] Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

[71] Ren, X.; Huang, J.; Zeng, X.; Museth, K.; Fidler, S.; and Williams, F. 2024. XCube: Large-Scale 3D Generative Modeling using Sparse Voxel Hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[72] Ren, X.; Lu, Y.; Cao, T.; Gao, R.; Huang, S.; Sabour, A.; Shen, T.; Pfaff, T.; Wu, J. Z.; Chen, R.; Kim, S. W.; Gao, J.; Leal-Taixe, L.; Chen, M.; Fidler, S.; and Ling, H. 2025. Cosmos-Drive-Dreams: Scalable Synthetic Driving Data Generation with World Foundation Models.

[73] Ren, X.; Lu, Y.; Liang, H.; Wu, J. Z.; Ling, H.; Chen, M.; Fidler, F., Sanja annd Williams; and Huang, J. 2024. SCube: Instant Large-Scale Scene Reconstruction using VoxSplats. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[74] Ren, X.; Shen, T.; Huang, J.; Ling, H.; Lu, Y.; Nimier-David, M.; Müller, T.; Keller, A.; Fidler, S.; and Gao, J. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6121–6132.

[75] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

[76] Rong, G.; Shin, B. H.; Tabatabaee, H.; Lu, Q.; Lemke, S.; Možeiko, M.; Boise, E.; Uhm, G.; Gerow, M.; Mehta, S.; et al. 2020. Lgsvl simulator: A high fidelity simulator for autonomous driving. In *2020 IEEE 23rd International conference on intelligent transportation systems (ITSC)*, 1–6. IEEE.

[77] Sargent, K.; Li, Z.; Shah, T.; Herrmann, C.; Yu, H.-X.; Zhang, Y.; Chan, E. R.; Lagun, D.; Fei-Fei, L.; Sun, D.; et al. 2023. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*.

[78] Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

[79] Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

[80] Shah, S.; Dey, D.; Lovett, C.; and Kapoor, A. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, 621–635. Springer.

[81] Shriram, J.; Trevithick, A.; Liu, L.; and Ramamoorthi, R. 2024. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*.

[82] Shue, J. R.; Chan, E. R.; Po, R.; Ankner, Z.; Wu, J.; and Wetzstein, G. 2023. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20875–20886.

[83] Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

[84] Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.

[85] Song, Y.; Sun, Z.; and Yin, X. 2024. SDXS: Real-Time One-Step Latent Diffusion Models with Image Conditions. *arXiv preprint arXiv:2403.16627*.

[86] Stein, G.; Cresswell, J.; Hosseinzadeh, R.; Sui, Y.; Ross, B.; Villecroze, V.; Liu, Z.; Caterini, A. L.; Taylor, E.; and Loaiza-Ganem, G. 2023. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36: 3732–3784.

[87] 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.; and Anguelov, D. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[88] Talwar, D.; Guruswamy, S.; Ravipati, N.; and Eirinaki, M. 2020. Evaluating validity of synthetic data in perception tasks for autonomous vehicles. In *2020 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 73–80. IEEE.

[89] Team, A.; Zhu, H.; Wang, Y.; Zhou, J.; Chang, W.; Zhou, Y.; Li, Z.; Chen, J.; Shen, C.; Pang, J.; et al. 2025. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*.

[90] Team, G. 2024. Mochi 1. https://github.com/genmoai/models.

[91] Team, T. H. 2025. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. arXiv:2501.12202.

[92] Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4): 1–12.

[93] Tonderski, A.; Lindström, C.; Hess, G.; Ljungbergh, W.; Svensson, L.; and Petersson, C. 2023. NeuRAD: Neural rendering for autonomous driving. *arXiv preprint arXiv:2311.15260*.

[94] Vahdat, A.; and Kautz, J. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33: 19667–19679.

[95] Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34: 11287–11302.

[96] Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

[97] Wang, P.; Xu, D.; Fan, Z.; Wang, D.; Mohan, S.; Iandola, F.; Ranjan, R.; Li, Y.; Liu, Q.; Wang, Z.; et al. 2023. Taming Mode Collapse in Score Distillation for Text-to-3D Generation. *arXiv preprint arXiv:2401.00909*.

[98] Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; and Lu, J. 2023. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*.

[99] Williams, F.; Gojcic, Z.; Khamis, S.; Zorin, D.; Bruna, J.; Fidler, S.; and Litany, O. 2021. Neural Fields as Learnable Kernels for 3D Reconstruction. arXiv:2111.13674.

[100] Wu, J. Z.; Zhang, Y.; Turki, H.; Ren, X.; Gao, J.; Shou, M. Z.; Fidler, S.; Gojcic, Z.; and Ling, H. 2025. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26024–26035.

[101] Wu, R.; Gao, R.; Poole, B.; Trevithick, A.; Zheng, C.; Barron, J. T.; and Holynski, A. 2024. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. *arXiv:2411.18613*.

[102] Wu, R.; Mildenhall, B.; Henzler, P.; Park, K.; Gao, R.; Watson, D.; Srinivasan, P. P.; Verbin, D.; Barron, J. T.; Poole, B.; et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21551–21561.

[103] Wu, Z.; Liu, T.; Luo, L.; Zhong, Z.; Chen, J.; Xiao, H.; Hou, C.; Lou, H.; Chen, Y.; Yang, R.; et al. 2023. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, 3–15. Springer.

[104] Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2024. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*.

[105] Xie, H.; Chen, Z.; Hong, F.; and Liu, Z. 2024. Citydreamer: Compositional generative model of unbounded 3d cities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9666–9675.

[106] Xie, K.; Lorraine, J.; Cao, T.; Gao, J.; Lucas, J.; Torralba, A.; Fidler, S.; and Zeng, X. 2024. LATTE3D: Large-scale Amortized Text-To-Enhanced3D Synthesis. *arXiv preprint arXiv:2403.15385*.

[107] Xu, Y.; Chai, M.; Shi, Z.; Peng, S.; Skorokhodov, I.; Siarohin, A.; Yang, C.; Shen, Y.; Lee, H.-Y.; Zhou, B.; et al. 2023. 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*, 4402–4412.

[108] Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; Li, T.; Dai, B.; Chitta, K.; Wu, P.; Zeng, J.; Luo, P.; Zhang, J.; Geiger, A.; Qiao, Y.; and Li, H. 2024. Generalized Predictive Model for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[109] Yang, J.; Huang, J.; Chen, Y.; Wang, Y.; Li, B.; You, Y.; Igl, M.; Sharma, A.; Karkus, P.; Xu, D.; Ivanovic, B.; Wang, Y.; and Pavone, M. 2025. STORM: Spatio-Temporal Reconstruction Model for Large-scale Outdoor Scenes. *arXiv preprint arXiv:2501.00602*.

[110] Yang, S.; Hou, L.; Huang, H.; Ma, C.; Wan, P.; Zhang, D.; Chen, X.; and Liao, J. 2024. Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion. *arXiv preprint arXiv:2402.03162*.

[111] Yang, Y.; Yang, Y.; Guo, H.; Xiong, R.; Wang, Y.; and Liao, Y. 2023. Urbangiraffe: Representing urban scenes as compositional generative neural feature

fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9199–9210.

[112] Yang, Z.; Chen, Y.; Wang, J.; Manivasagam, S.; Ma, W.-C.; Yang, A. J.; and Urtasun, R. 2023. UniSim: A Neural Closed-Loop Sensor Simulator. In *CVPR*.

[113] Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

[114] Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024. GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. In *CVPR*.

[115] Yu, H.-X.; Duan, H.; Herrmann, C.; Freeman, W. T.; and Wu, J. 2024. WonderWorld: Interactive 3D Scene Generation from a Single Image. *arXiv preprint arXiv:2406.09394*.

[116] Yu, H.-X.; Duan, H.; Hur, J.; Sargent, K.; Rubinstein, M.; Freeman, W. T.; Cole, F.; Sun, D.; Snavely, N.; Wu, J.; et al. 2024. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6658–6667.

[117] Zhang, J.; Zhang, Q.; Zhang, L.; Kompella, R. R.; Liu, G.; and Zhou, B. 2024. Urban Scene Diffusion through Semantic Occupancy Map. *arXiv preprint arXiv:2403.11697*.

[118] Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

[119] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[120] Zhang, S.; Zhang, Y.; Zheng, Q.; Ma, R.; Hua, W.; Bao, H.; Xu, W.; and Zou, C. 2024. 3D-SceneDreamer: Text-Driven 3D-Consistent Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10170–10180.

[121] Zhang, Z.; Long, F.; Pan, Y.; Qiu, Z.; Yao, T.; Cao, Y.; and Mei, T. 2024. TRIP: Temporal Residual Learning with Image Noise Prior for Image-to-Video Diffusion Models. *arXiv preprint arXiv:2403.17005*.

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

[123] Zhou, L.; Du, Y.; and Wu, J. 2021. 3D Shape Generation and Completion Through Point-Voxel Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5826–5835.

[124] Zyrianov, V.; Che, H.; Liu, Z.; and Wang, S. 2024. LidarDM: Generative LiDAR Simulation in a Generated World. *arXiv preprint arXiv:2404.02903*.

LSD-3D: Large-Scale 3D Driving Scene Generation with Geometry Grounding (Supplemental Document)

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

This supplementary document provides additional information and experiments in support of the findings of our main manuscript. In Section 1, we provide additional implementation details for GGDS, including a thorough derivation of the distillation, expansion on all regularizations and the proxy geometry generation procedure.

In Section 2, we provide additional elaboration on our experimental procedure for evaluating our method and existing baseline methods. In Section 3, we include additional details on the composability of our generated scenes with dynamic actor assets. Finally, in Sections 4, we provide a significant amount of additional results generated using our method, and in Section 5, we provide additional visual comparisons to existing methods.

Table of Contents

*Indicates equal contribution.

1 Implementation Details

In this section, we describe additional implementation details of the loss and regularization term that we use in order to achieve stable and consistent geometry grounding. In the main paper, we provide a derivation of all of the primary losses and their implementation, in GGDS as the basis for our geometry-grounded distillation sampling (GGDS) objective, which incorporates such as multi-step diffusion, multi-step DDIM-Inversion, and LDM fine-tuning. Here, we provide further elaboration on our optimization objective.

1.1 Objectives

We detail all loss components and the optimization objective of GGDS and its integration with large-scale 3D scene generation with 2D Gaussian Splatting [37]. The final combined optimization objective consists of a distillation component $\mathcal{L}_{gen}$, geometrical regularization terms $\mathcal{L}_{geo}$ and common regularization terms $\mathcal{L}_{gauss}$ on the surfel primitives, that is

$$\mathcal{L} = \lambda_{gen}\mathcal{L}_{gen} + \mathcal{L}_{geo} + \mathcal{L}_{gauss}. \tag{5}$$

Multi-Step Image Space Diffusion. As described in paragraph **GGDS** of Sec. 3.3 and ablated in Sec. 4.1 of the main paper, stable optimization of non-overlapping camera views in large-scale driving scenes is achieved via DDIM inversion [84] in our work. The DDIM denoising process from $t = T$ to $t = 0$ follows

$$\begin{aligned} &\mathbf{z}_{t-1,i} = \sqrt{\alpha_{t-1}}\hat{\mathbf{z}}^t_{0,i} + \Delta\mathbf{z}_{t,i} \\ &\text{with } \hat{\mathbf{z}}^t_0 = (\frac{z_t - \sqrt{1-\alpha_t}\epsilon_\Phi(\mathbf{z}_t)}{\sqrt{\alpha_t}} \text{ and } \Delta\mathbf{z}_t = \sqrt{1-\alpha_t-\sigma_t^2}\cdot\epsilon_\Phi(\mathbf{x}_t) + \sigma_t\epsilon_t, \end{aligned} \tag{6}$$

where z_0 is the encoded latent variable, ϵ_Φ is the noise predicted by the latent diffusion model parametrized with Φ, and σ_t defines the randomness of the process. For $\sigma_t = 0$ and as chosen in this work, the noise sampling process is deterministic, and the respective inversion process can then be formulated as

$$\begin{aligned} \mathbf{z}_{t,i} &= \text{DDIM}^{-1}\left(\mathbf{z}_{t-1,i}, \alpha_t, \alpha_{t-1}\right) \\ &= \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_t - 1}}\left(z_{t-1} - \sqrt{1-\alpha_{t-1}}\epsilon_\Phi(\mathbf{z}_{t-1})\right) + \sqrt{1-\alpha_t}\epsilon_\Phi(\mathbf{z}_{t-1}). \end{aligned} \tag{7}$$

The consistency of the deterministic inversion still depends on the step-size and starting point z_0, which we mitigate with multiple inversion steps. This results in lower dependence of the initial direction computed from $\epsilon_\Phi(z_0)$. Various additions proposed recently and consolidated by Bao et al. [3] might show further improvements, and we would like to explore their usage in future work.

Given the noisy latent $\mathbf{z}_{t,i}$, we then use the forward process in Eq. 6 to predict $\hat{\mathbf{z}}_0$ conditioned on the respective text prompt and mesh depth D. The denoised latent is then decoded into the image $\hat{\mathbf{x}} = \mathcal{D}\left(\hat{\mathbf{z}}_0\right)$ and used to guide the scene generation through a reconstruction loss

$$\mathcal{L}_{gen}\left(\Theta\right) = \mathbb{E}_{\psi_i,t}[\omega\left(t\right)\left(\|g\left(\Theta,\psi_i\right) - \hat{\mathbf{x}}\| + \mathcal{L}_{LPIPS}(g\left(\Theta,\psi_i\right),\hat{\mathbf{x}})\right)], \tag{8}$$

that consists of a reconstruction and perceptual similarity [119] component.

Geometry Regularization. We prevent the Gaussians to diverge from the proxy geometry by enforcing the geometrical consistency priors through

$$\mathcal{L}_{geo} = \lambda_{norm}\mathcal{L}_{norm} + \lambda_{disp}L_{disp}. \tag{9}$$

Specifically, we penalize large deviations of the rasterized normals $\mathcal{D}_{\Theta,\psi_i}$ and disparity maps $\mathcal{N}_{\Theta,\psi_i}$ (generated from a viewpoint ψ_i) from the proxy geometry

$$\mathcal{L}_{norm} = \|\mathcal{N}_{\Theta,\psi_i} - \mathcal{N}_{\mathcal{M},\psi_i}\|, \quad \mathcal{L}_{disp} = \|\mathcal{D}_{\Theta,\psi_i} - \mathcal{D}_{\mathcal{M},\psi_i}\|. \tag{10}$$

The weight of each loss, λ, for all terms in the above expression, and the terms in each loss sub-expression as defined in our main paper, are set as reported in Tab. 1.

Gaussian Regularization. Following 2D Gaussian Splatting [37] and Markov Chain Monte Carlo (MCMC) [47] optimization, we add common regularization terms. The appearance and position of the Gaussian disks is kept optimal through

$$\mathcal{L}_{gauss} = \lambda_{dist}\mathcal{L}_{dist} + \lambda_{nc}\mathcal{L}_{nc} + \lambda_o\mathcal{L}_{\text{opa}} + \lambda_k\mathcal{L}_{\text{scale}} + \lambda_{tv}\mathcal{L}_{tv}. \tag{11}$$

Here, $\mathcal{L}_{nc}$ is the normal consistency loss and $\mathcal{L}_{dist}$ is the depth distortion loss [37]. $\mathcal{L}_{\text{opa}}$ and $\mathcal{L}_{\text{scale}}$ are regularization on Gaussian opacity and scale [47], to prohibit large and Gaussians in non-visible areas across the scene. $\mathcal{L}_{tv}$ is the total-variation loss on the rasterized image, with the objective of reducing high-frequency artifacts in rendered images.

We specifically promote the local alignment between the estimated-surface normal and the splat normal via

$$\mathcal{L}_{nc} = \sum_{a} \left(1 - \mathbf{n}_a \mathcal{N}_{\Theta,\psi_i}\right), \tag{12}$$

where $\mathbf{n}_i$ is the normal orientated towards the camera of each 2D Gaussian disk intersected by the ray that corresponds to each pixel $s = (u, v)$ and

$$\mathcal{N}_{\Theta,\psi_i}(u, v) = \frac{\nabla_u \mathbf{p}_s \times \nabla_v \mathbf{p}_s}{|\nabla_u \mathbf{p}_s \times \nabla_v \mathbf{p}_s|}. \tag{13}$$

Moreover, we encourage the Gaussians to lie close to the estimated surface through

$$\mathcal{L}_{dist} = \sum_{i,j} \omega_i \omega_j \left|z_i - z_j\right|, \tag{14}$$

where z is the intersection depth and ω the blending weight. This constraint helps to avoid erroneous weights spread along the rays.

Finally, to enforce spatial coherence and reduce the visual noise in the rendered images, we introduce a Total Variation (TV) loss on the splatted images as

$$\mathcal{L}_{tv}(x) = \sum_{i,j} \sqrt{(x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2}, \tag{15}$$

where i, j here denote the pixel indexes of the image. By encouraging smooth color transitions while preserving sharp edges, this loss helps suppress noise during optimization and leads to more visually consistent renderings from the Gaussian representation. We provide the selected λ values for each optimization objective in Table 1.

Weight	Regularized Attribute/Information	Weight Value
λ_{gen}	Multi-Step RGB Recon. & Percept.	0.001
λ_{disp}	Gaussian & Mesh Disparity	1.0
λ_{norm}	Gaussian & Mesh Normals	1.0
λ_{dist}	Gaussian Centers	0.05
λ_{nc}	Gaussian Normals & rendered depth	0.05
λ_{opa}	Gaussian Opacity	0.01
λ_{scale}	Gaussian Scale	1.0
λ_{tv}	Rendered Image	10.0

Table 1: Loss weighting terms and components affected by each regularization term.

1.2 Deferred Rendering.

Deferred Image Rendering. For deferred rendering, we follow the deterministic noise sampling process in Eq. 7 and denoise the noisy latent with the same latent diffusion model ϵ_Φ. We choose a very low, fixed timestep of $t = 0.22$.

1.3 Geometry Layout Generation

We train the coarse and fine VAE for 3 days and both diffusion models for 12 days on 4 H100 GPUs each. As a basis for our hierarchical latent voxel diffusion model, we employ XCube [71] with a single coarse and fine layer. All models are supervised by point-cloud data[2] that is obtained from the Waymo Open Dataset [87]. We additionally extract road and object voxel occupancies from the map and 3D bounding box data for all actors and train the coarse 3D diffusion model by concatenating the latent with respective occupancy. The fine layer is solely conditioned on the generated coarse voxel grid, serving as a refinement step. At inference, we control the generation through the same conditioning, but keep bounding boxes of actors on the road optional, allowing us to generate scenes which are composable with dynamic actors.

For unconditional generation, we generate meshes $\mathcal{M}$ for which we utilize the checkpoints provided by the authors in [87], which were trained on semantically annotated and aggregated point clouds from the Waymo Open Dataset. Point cloud conditioning is achieved through direct voxelization of the aggregated point clouds of the dataset. In all three cases, the final mesh $\mathcal{M}$ is predicted through Neural Kernel Surface Reconstruction [38] with varying metric voxel sizes of 0.1 for point clouds and 0.18 for conditional and unconditional generation. This coarse mesh provides the initial proxy geometry of our scene generation. To this end, we initialize a set of 2D Gaussian disks [37] on top, as follows.

[2]We thank the authors of [73] for the aggregated data.

Gaussian Initialization Given the coarse proxy mesh $\mathcal{M} = \{\mathbf{F_1}, ..., \mathbf{F_N}\}$, we initialize a Gaussian disk θ_k at the center of each triangle face $\mathbf{F}$ at the location

$$\mathbf{p}_k = (\mathbf{V}_i + \mathbf{V}_j + \mathbf{V}_l)/3, \tag{16}$$

where each face is denoted by its three vertices $\mathbf{F} = [\mathbf{V}_i, \mathbf{V}_j, \mathbf{V}_k]$. Further, the normal $\mathbf{n}_F$, and the area a_F of each triangle mesh is used to initialize the orientation, scale, and tangential axis $\mathbf{t}_u \times \mathbf{t}$ such that

$$\mathbf{n}_F = \mathbf{t}_u \times \mathbf{t}_u, l = l_F \text{ and } \|\mathbf{s}\| = a_F. \tag{17}$$

1.4 Camera Pose Sampling

Large-scale street scenes are characterized by a higher degree of complexity, topological variation, and size. As such, we cannot assume a standard spherical viewpoint distribution around a centered object (as seen in prior score distillation and object generation works, as discussed in the Related Work). Furthermore, naïvely sampling cameras through the scene may conflict with the prior image distribution and geometry conditioning (as described in the main paper). Instead, we develop a method that aims to replicate the prior image distribution, while fully covering all relevant visible portions of the scene.

Trajectory Sampling. To produce such a viewpoint distribution, we aim to replicate natural driving trajectories through our generated scenes. For map-layout and point-cloud conditioned geometry, we leverage map data to produce natural driving trajectories with 8 cameras with 50°FoV sampled at multiple points along the trajectory. In the case of unconditional generation, this is accomplished by using the initial occupancy and semantics in the proxy mesh generation stage. We identify road points using the semantic labels, and produce a dense point grid from these. This point grid is then projected into the XY plane, creating a discretized representation of the drivable area with high resolution. From this occupancy, we apply a standard Euclidean distance transform to identify ridge points corresponding to the centerline or medial axis of driving areas - in essence, the center of the road. The Euclidean distance transform computes for each grid cell the Euclidean distance to the nearest background pixel, that is

$$\mathcal{D}(i,j) = \min_{(i',j') \notin \mathcal{O}} \|(i,j) - (i',j')\|_2, \tag{18}$$

where $\mathcal{O}$ represents the set of occupied road cells.

To order these ridge points into a continuous trajectory, we select the point closest to the grid boundary and iteratively connect to the closest neighboring point, which ensures path connectivity with a high degree of confidence. Finally, we fit a B-spline curve through these ordered ridge points and resample $N = 150$ points along this spline to represent the positions for our camera rig. At each of those points, we spawn eight cameras with 50°FoV and align them with a 45°horizontal pitch to achieve 360°coverage. Lifting the camera rig from the ground at each point we elevate the cameras by their typical mounting height $h_w = 2.1m$, which corresponds to the height of the camera in the prior dataset.

2 Experimental Details

Baseline Method Evaluation. For a quantitative and qualitative evaluation of our approach, we compare LSD-3D with four competing generative approaches, namely WonderJourney [116], Vista [26], MagicDrive3D [22], and SVD [4] conditioned on depth (inspired by [59]). We detail the experiments below.

- For WonderWorld [115], we use the official code base, and start from a frame generated using our finetuned Stable Diffusion XL model. Following their paper, we use the Stable Diffusion 2 Inpaint model [75] for outpainting, Marigold[44, 45] for geometry estimation, and OneFormer [40] for sky and foreground segmentation. As the rendering trajectory, we adopt a straight-line path to simulate a forward-driving car motion, with camera positions sampled at uniform spatial intervals along this path to ensure consistent and realistic viewpoint transitions throughout the sequence.
- For Vista [26], we use the publicly available weights and, similar to WonderWorld use as a starting frame the generated image from our finetuned Stable Diffusion XL. As in the original implementation, the video is then generated autoregressively by performing six temporal sampling steps, where each step produces a new chunk of frames conditioned on the previously generated ones.
- For MagicDrive3D [22], as the code is not publicly available, we use the open-sourced MagicDriveDit [23] model. In particular, we use the layout and prompt conditioning, removing dynamic actors for a fairer comparison.
- For GEN3C [74], we use the publicly available code and the COSMOS [64] checkpoints. For our rendering trajectory, we use a straight line trajectory starting from an initial frame condition, or, alternatively, load trajectories generated from our LSD-3D pipeline and scale them to the GEN3C bounding scale. As a starting frame, we once again use the generated image from our finetuned Stable Diffusion XL.

For each of the methods, we use ten different prompts, that we include in Table 2. All these approaches are used to generate photorealistic scene videos, which we then use as part of the 3D scene reconstruction via 2D Gaussian splatting [37], enabling a fair comparison with our approach. We use the official 2D Gaussian Splatting implementation and optimize the scene for 30'000 steps.

Prompts
A photography of a neighborhood street scene
A photography of a neighborhood street scene at sunset
A photography of a neighborhood street scene at night
A photography of a neighborhood street scene on a cloudy day
A photography of a neighborhood street scene on a rainy day
A photography of a neighborhood street scene in the fall, with leaves on the ground
A photography of a neighborhood street scene in the winter, with snow on the ground
A photography of a neighborhood street scene in the spring, with flowers in bloom
A photography of a desert neighborhood street, with cacti and sand
A photography of a San Francisco neighborhood street, with hills and fog

Table 2: List of prompts used for neighborhood street scene variations

3 Novel Scene Composability Implementation

Leveraging the map layout and real-world trajectories from the Waymo open dataset [87], we employ TrafficGen [20] to generate realistic traffic scenarios. To ensure scene diversity, we begin by sampling multiple initial vehicle placements, which serve as inputs for autoregressive trajectory generation. Ego trajectories are then determined either by selecting from the generated set, or utilizing a policy-driven autonomous agent. For this, we integrate with the MetaDrive simulator [52] and apply a pre-trained PPO policy [79], which takes ego states, navigation information, surrounding agent data and a 2D-LiDAR-like point cloud as input. In future work, we aim to scale the proposed approach and incorporate end-to-end agents that directly process realistically rendered sensor data, rather than relying on a combination of approximated sensor inputs and ground truth environment information.

4 Additional Results

We report additional results in support of the findings from our main manuscript:

- In Figures 1 and 2, we provide additional examples of generated scenes from our method, along with the corresponding novel views at from a driving perspective.
- In Figure 3, we provide additional generated scenes using point clouds as conditioning.
- In Figures 4 and 5, we provide a total of four example scenes with six prompt variations each, demonstrating the capability of our method to produce stylistically diverse scenes independent of geometry (supporting our prompt adherence claims).
- Finally, in Figure 6, we provide examples of more significant off-trajectory shifts in our generated scenes. These are excluded from the comparison figures, since such shifts cause a total collapse for competing methods – however, our scenes retain not only geometry and texture but view-consistency even at such off-trajectory shifts. We validate the capability of LSD-3D to produce high-quality, explicit 3D generated scenes with accurate texture and geometry grounding.

5 Additional Comparisons to Baselines

In this section, we provide additional comparisons to existing generative methods. Specifically, in Figures 7 and 8, we provide examples of generated scenes with WonderWorld, Vista, GEN3C, and LSD-3D (our method), demonstrating that our method is the only approach capable of producing high quality, causal geometry and texture that retains view consistency off the initial seen trajectory.

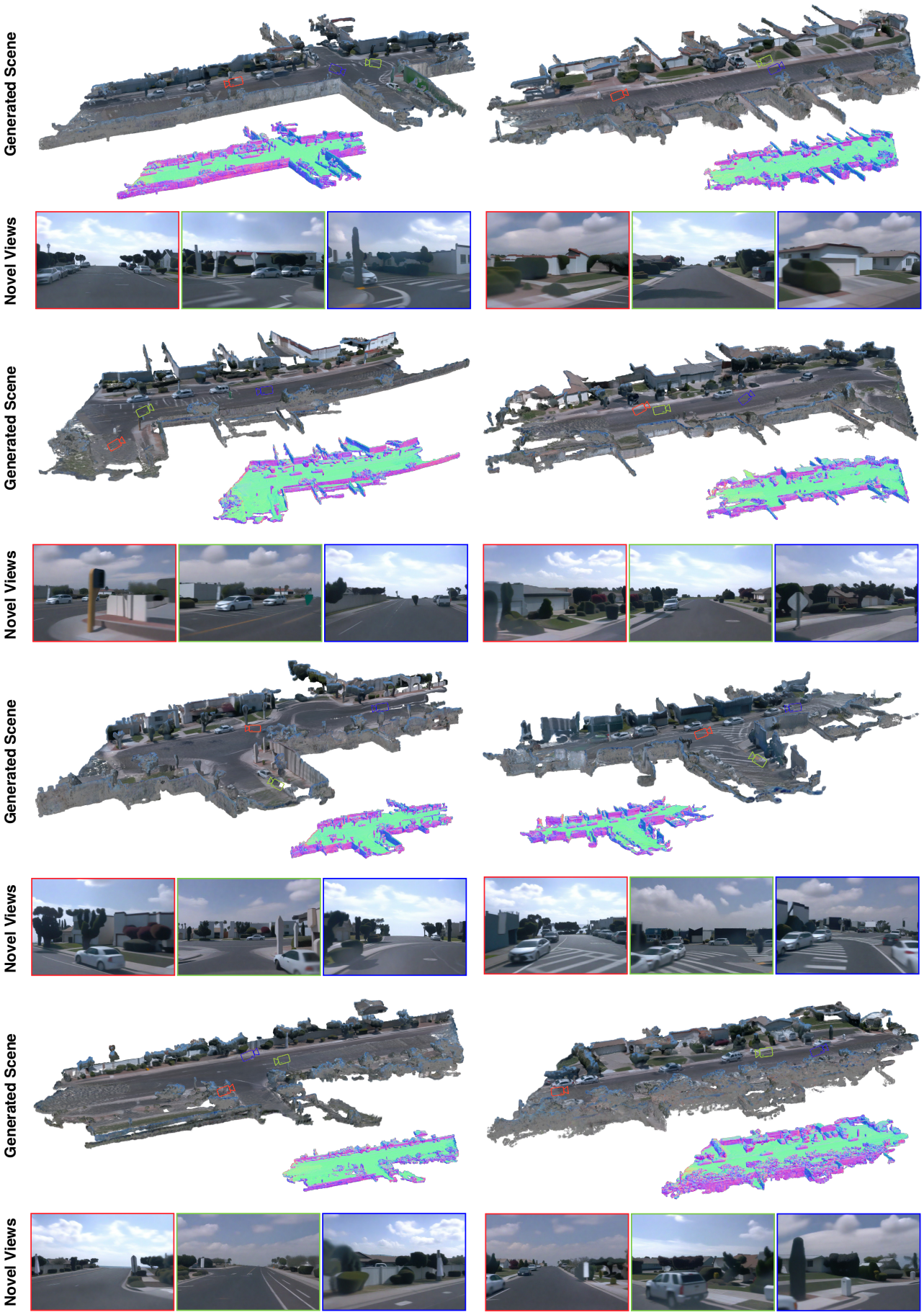

Figure 1: **Additional Results Generated by LSD-3D**. We present a collection of additional scenes which are generated by our method, alongside novel rendered viewpoints.

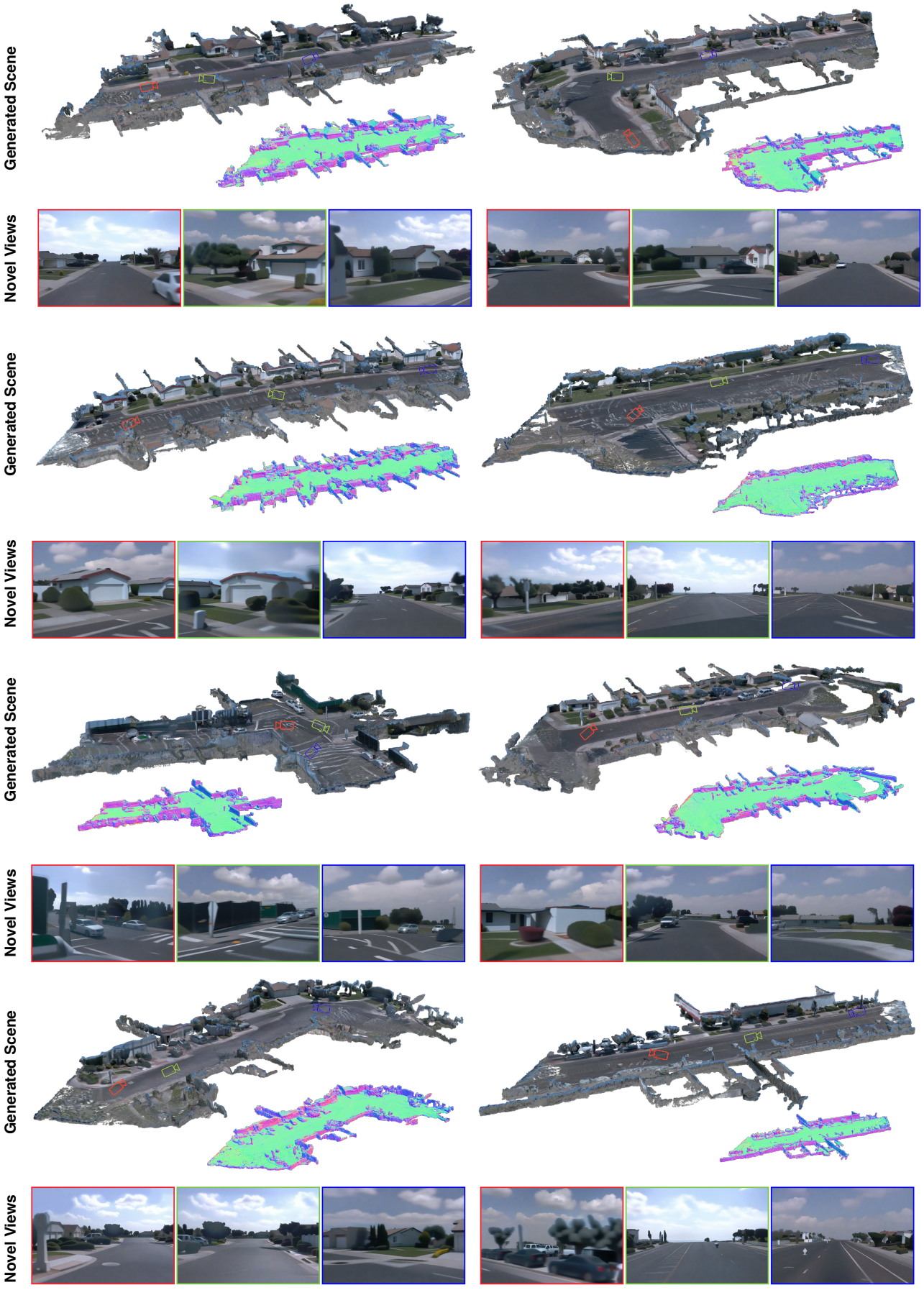

Figure 2: **Additional Results Generated by LSD-3D**. We present another collection of scenes which are generated by our method, alongside novel rendered viewpoints.

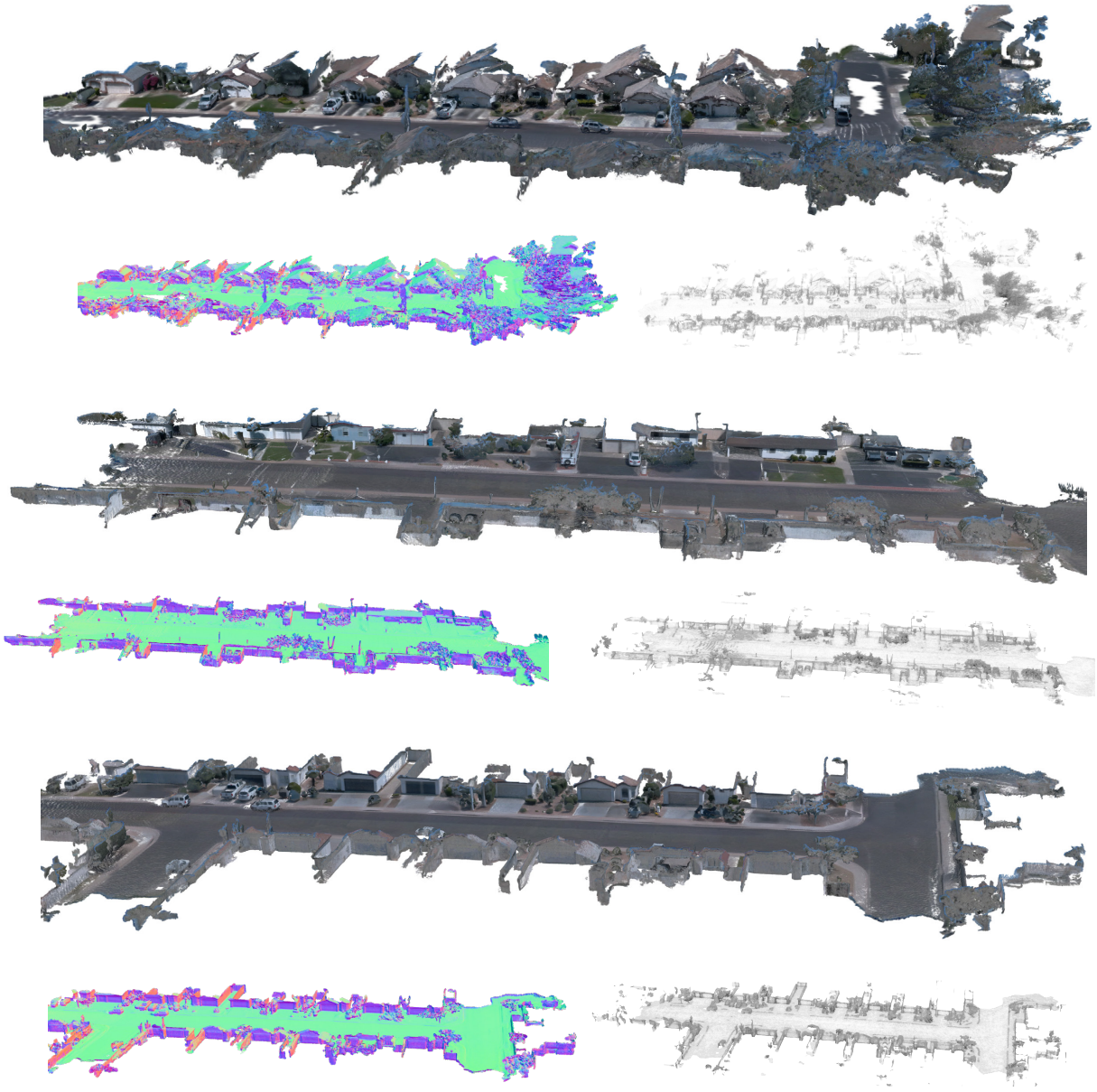

Figure 3: **Point-Cloud Conditioned Generated Scenes with LSD-3D**. We provide additional samples of generated scenes with point cloud conditioning for the proxy geometry generation. The condition is shown on the bottom right of each scene, next to the surface normal geometry.

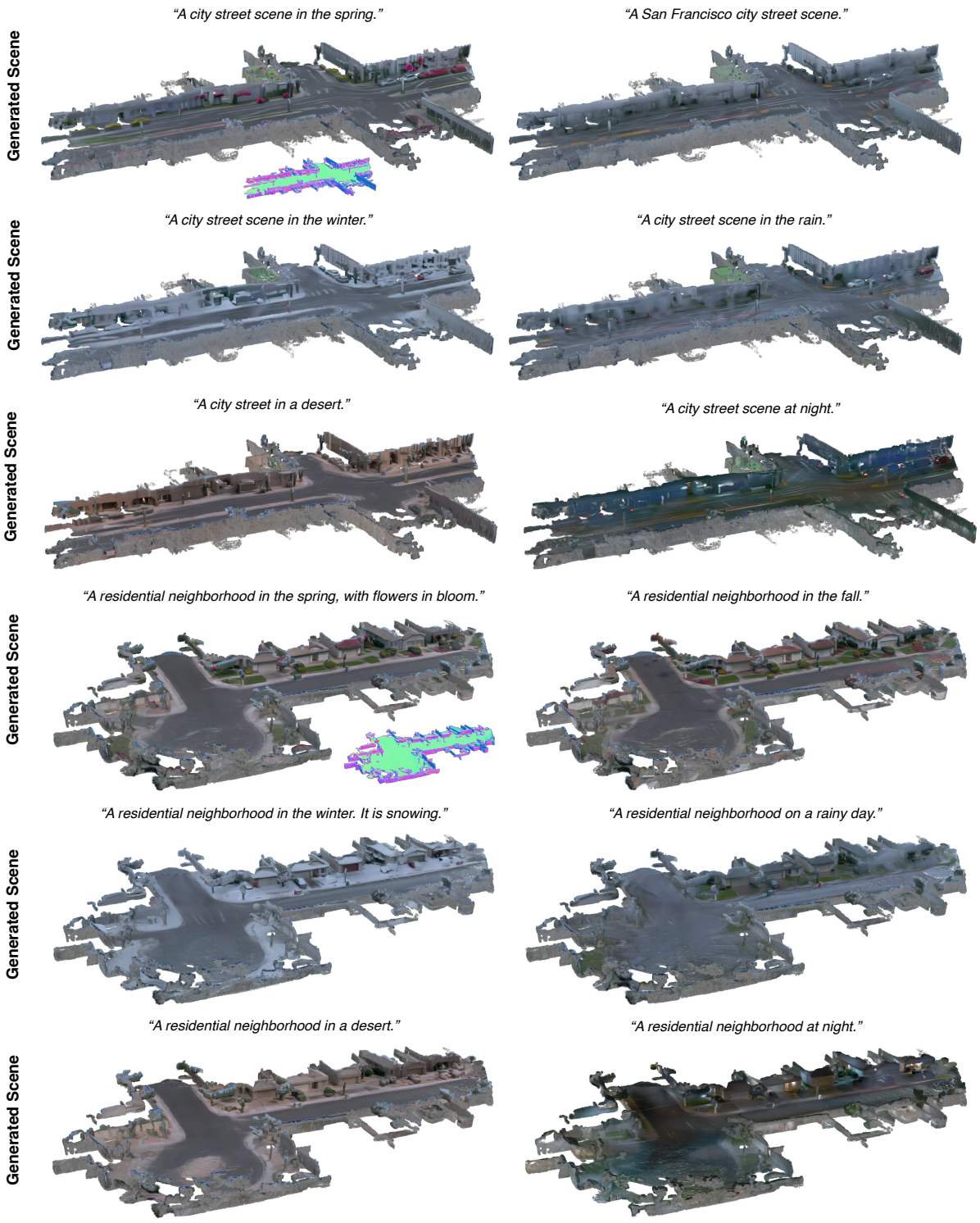

Figure 4: **Prompt Style Variation.** We demonstrate that our method is capable of producing stylistically diverse scenes based on varying text prompts. We display two different scenes, each with six prompt variations, in order to demonstrate the effectiveness of our text conditioning and environment map on producing stylistically diverse results. Our method produces diverse, geometrically-grounded scenes, and retains high-fidelity texture across stylistic variation.

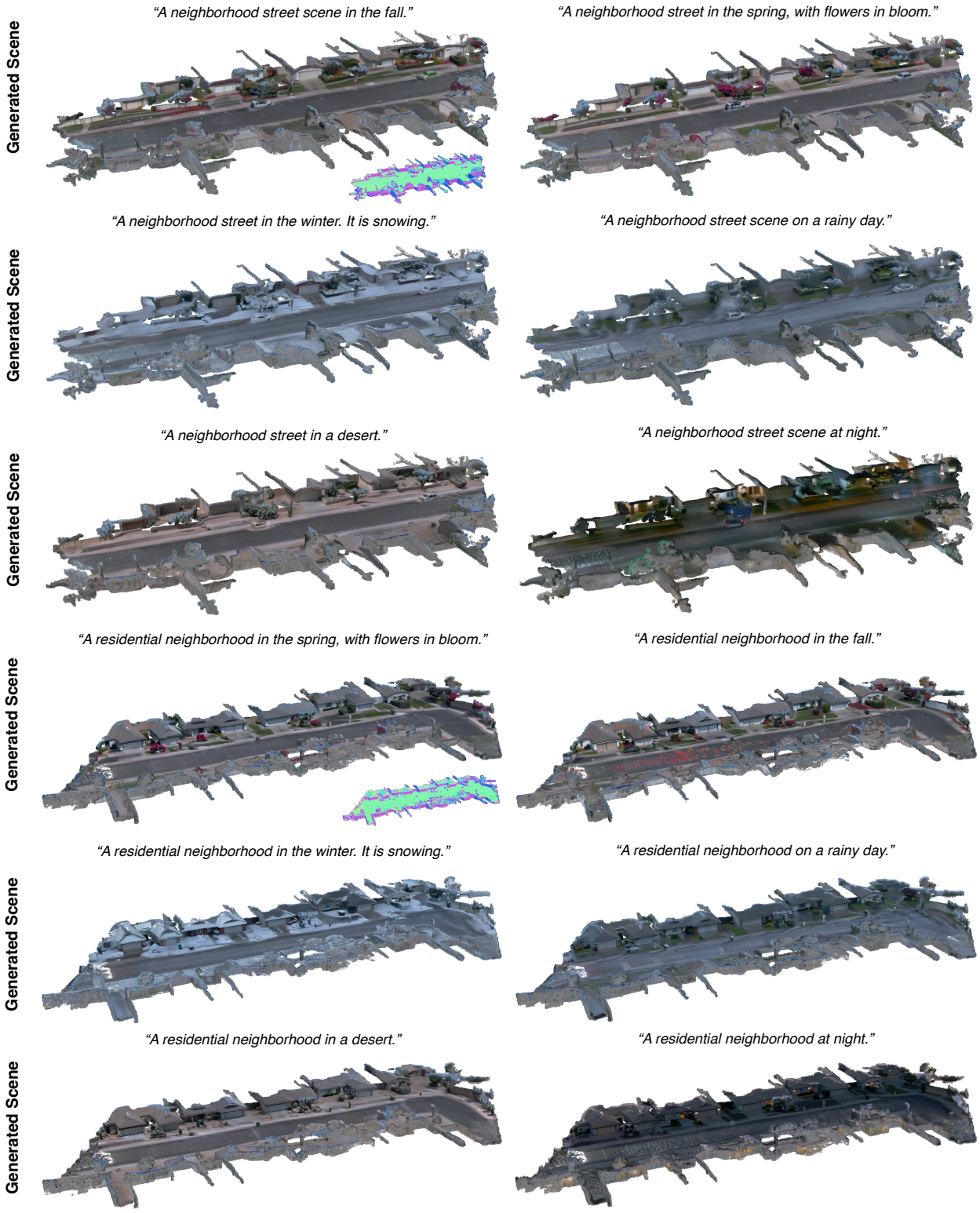

Figure 5: **Additional Prompt Style Variation.** We provide additional results, with six variations of prompt style variations for two different scenes – demonstrating the stylistic diversity capabilities of our method.

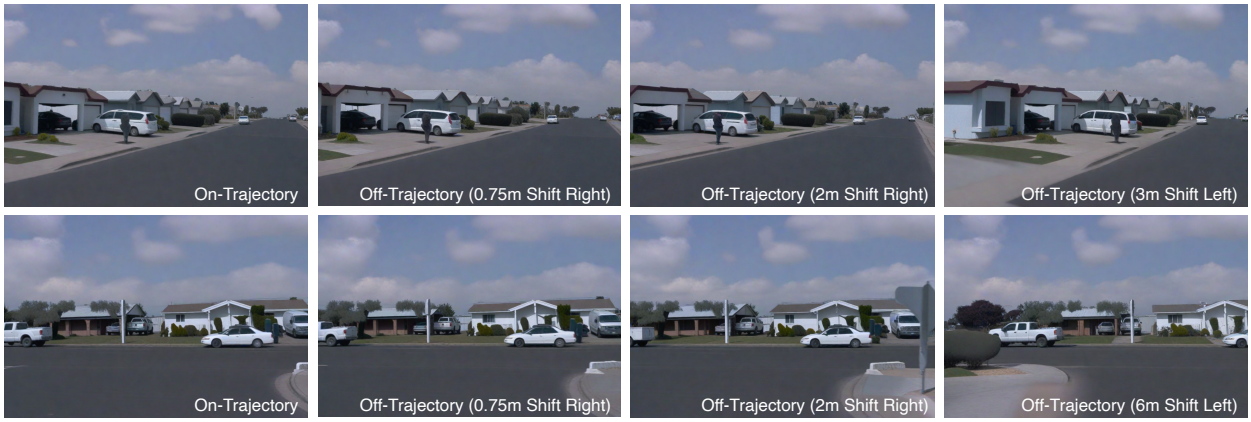

Figure 6: **Large Shift Off-Trajectory Visualizations of LSD-3D Generated Scenes**. We show examples of large off-trajectory shifts for our generated scenes, demonstrating the effectiveness of explicit and consistent 3D modeling. In the left column, we display an image from a seen viewpoint, and in the column next to it we display the 0.75 meter shift as used for baselines. We additionally show a 2 meter shift in the same direction, as well as a 3-6 meter shift (depending on the scene), demonstrating the capability of our method to generate view-consistent and causal results even at such major shifts off of seen trajectories.

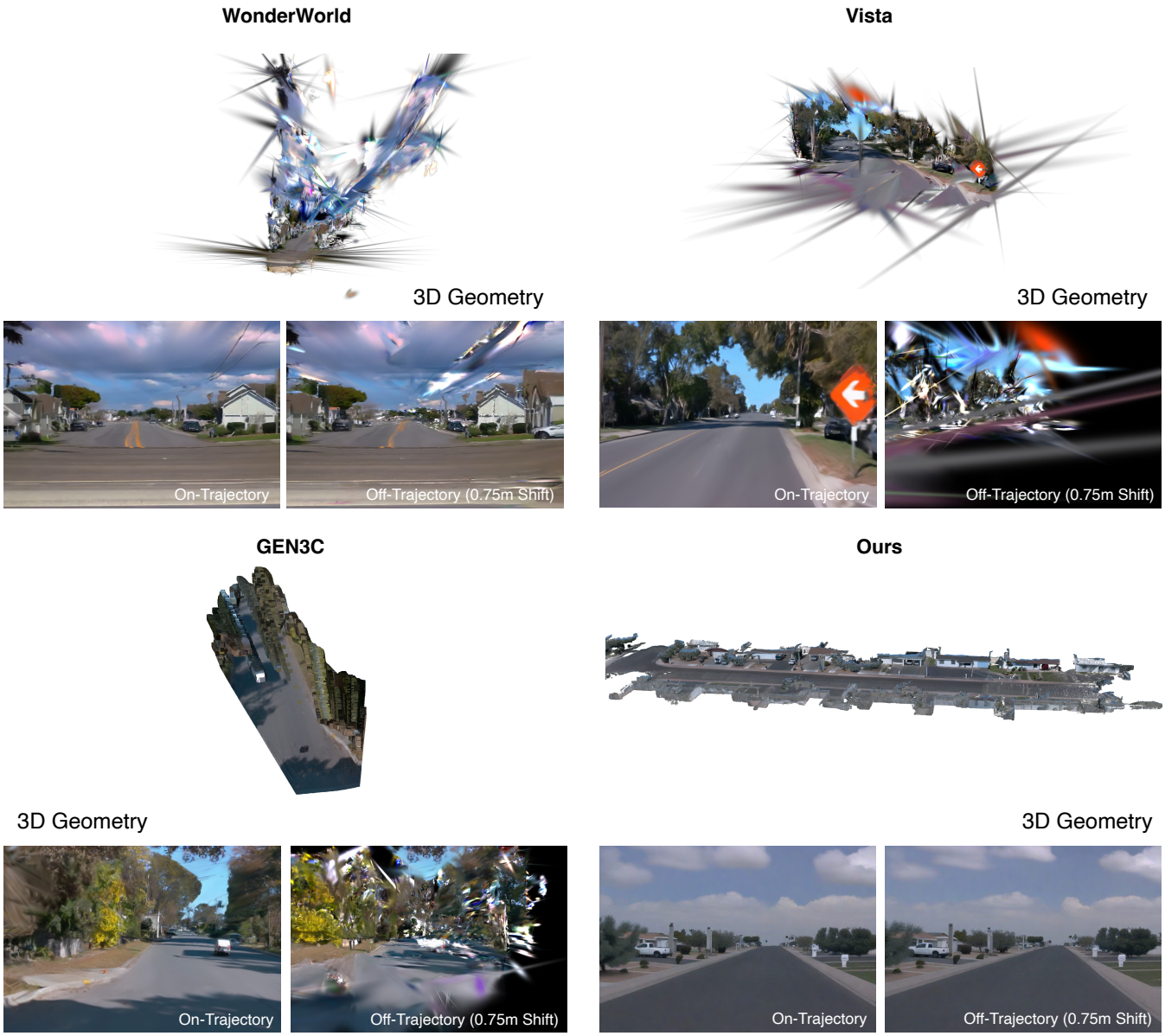

Figure 7: **Qualitative Comparisons to Video and Scene Generation Methods.** Our approach generates an accurate and 3D-consistent scene representation, enabling high-quality novel view synthesis and the generation of virtually unlimited off-trajectory viewpoints. In contrast, existing baselines WonderJourney [116], Vista [26], and GEN3C [74], fail to generate a consistent and 3D-plausible scene even for slighlty offset trajectories, precluding the production of novel driving scenes (please zoom into PDF version for details).

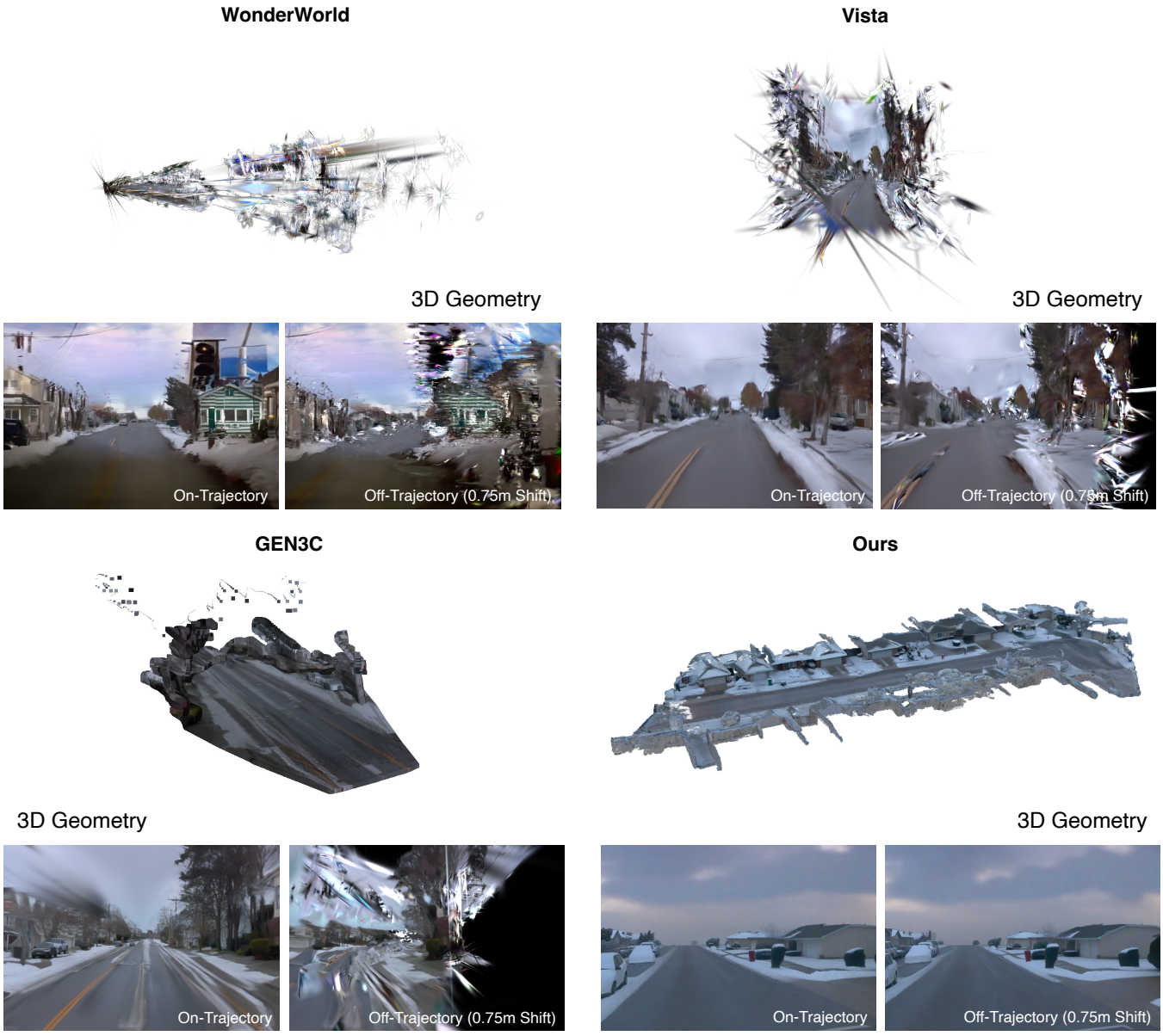

Figure 8: **Additional Qualitative Comparisons to Video and Scene Generation Methods.** We provide additional qualitative comparison for a snowy scene between our method and existing generative baseline methods.

References

[1] Anciukevicius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N. J.; and Guerrero, P. 2022. RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation. *arXiv*.

[2] Bahmani, S.; Skorokhodov, I.; Rong, V.; Wetzstein, G.; Guibas, L.; Wonka, P.; Tulyakov, S.; Park, J. J.; Tagliasacchi, A.; and Lindell, D. B. 2023. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*.

[3] Bao, Y.; Liu, H.; Gao, X.; Fu, H.; and Kang, G. 2025. FreeInv: Free Lunch for Improving DDIM Inversion. *arXiv preprint arXiv:2503.23035*.

[4] Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

[5] Borkman, S.; Crespi, A.; Dhakad, S.; Ganguly, S.; Hogins, J.; Jhang, Y.-C.; Kamalzadeh, M.; Li, B.; Leal, S.; Parisi, P.; et al. 2021. Unity perception: Generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*.

[6] Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

[7] Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*.

[8] Chen, A.; Zheng, W.; Wang, Y.; Zhang, X.; Zhan, K.; Jia, P.; Keutzer, K.; and Zhang, S. 2025. GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control. arXiv:2505.22421.

[9] Chen, Z.; Yang, J.; Huang, J.; de Lutio, R.; Esturo, J. M.; Ivanovic, B.; Litany, O.; Gojcic, Z.; Fidler, S.; Pavone, M.; Song, L.; and Wang, Y. 2025. OmniRe: Omni Urban Scene Reconstruction. In *The Thirteenth International Conference on Learning Representations*.

[10] Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Yago Vicente, T. F.; Dideriksen, T.; Arora, H.; Guillaumin, M.; and Malik, J. 2022. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. *CVPR*.

[11] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

[12] Dauner, D.; Hallgarten, M.; Li, T.; Weng, X.; Huang, Z.; Yang, Z.; Li, H.; Gilitschenski, I.; Ivanovic, B.; Pavone, M.; Geiger, A.; and Chitta, K. 2024. NAVSIM: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[13] Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; VanderBilt, E.; Kembhavi, A.; Vondrick, C.; Gkioxari, G.; Ehsani, K.; Schmidt, L.; and Farhadi, A. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663*.

[14] Deng, B.; Tucker, R.; Li, Z.; Guibas, L.; Snavely, N.; and Wetzstein, G. 2024. Streetscapes: Large-scale Consistent Street View Generation Using Autoregressive Video Diffusion. In *SIGGRAPH 2024 Conference Papers*.

[15] Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

[16] Dockhorn, T.; Vahdat, A.; and Kreis, K. 2021. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*.

[17] Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

[18] Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

[19] Ettinger, S.; Cheng, S.; Caine, B.; Liu, C.; Zhao, H.; Pradhan, S.; Chai, Y.; Sapp, B.; Qi, C. R.; Zhou, Y.; Yang, Z.; Chouard, A.; Sun, P.; Ngiam, J.; Vasudevan, V.; McCauley, A.; Shlens, J.; and Anguelov, D. 2021. 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)*, 9710–9719.

[20] Feng, L.; Li, Q.; Peng, Z.; Tan, S.; and Zhou, B. 2023. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 3567–3575. IEEE.

[21] Fischer, T.; Bulò, S. R.; Yang, Y.-H.; Keetha, N. V.; Porzi, L.; Müller, N.; Schwarz, K.; Luiten, J.; Pollefeys, M.; and Kontschieder, P. 2025. FlowR: Flowing from Sparse to Dense 3D Reconstructions. *arXiv preprint arXiv:2504.01647*.

[22] Gao, R.; Chen, K.; Li, Z.; Hong, L.; Li, Z.; and Xu, Q. 2024. MagicDrive3D: Controllable 3D Generation for Any-View Rendering in Street Scenes. *arXiv preprint arXiv:2405.14475*.

[23] Gao, R.; Chen, K.; Xiao, B.; Hong, L.; Li, Z.; and Xu, Q. 2024. MagicDriveDiT: High-Resolution Long Video Generation for Autonomous Driving with Adaptive Control. arXiv:2411.13807.

[24] Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2023. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*.

[25] Gao*, R.; Holynski*, A.; Henzler, P.; Brussee, A.; Martin-Brualla, R.; Srinivasan, P. P.; Barron, J. T.; and Poole*, B. 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *Advances in Neural Information Processing Systems*.

[26] Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[27] Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

[28] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

[30] Gulino, C.; Fu, J.; Luo, W.; Tucker, G.; Bronstein, E.; Lu, Y.; Harb, J.; Pan, X.; Wang, Y.; Chen, X.; Co-Reyes, J. D.; Agarwal, R.; Roelofs, R.; Lu, Y.; Montali, N.; Mougin, P.; Yang, Z.; White, B.; Faust, A.; McAllister, R.; Anguelov, D.; and Sapp, B. 2023. 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*.

[31] Harvey, W.; Naderiparizi, S.; Masrani, V.; Weilbach, C.; and Wood, F. 2022. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35: 27953–27965.

[32] Hess, G.; Lindström, C.; Fatemi, M.; Petersson, C.; and Svensson, L. 2025. Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11982–11992.

[33] Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

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

[35] Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

[36] Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

[37] Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery.

[38] Huang, J.; Gojcic, Z.; Atzmon, M.; Litany, O.; Fidler, S.; and Williams, F. 2023. Neural Kernel Surface Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4369–4379.

[39] Hwang, S.; Kim, M.-J.; Kang, T.; Kang, J.; and Choo, J. 2024. Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. In *European Conference on Computer Vision*, 1–18. Springer.

[40] Jain, J.; Li, J.; Chiu, M.; Hassani, A.; Orlov, N.; and Shi, H. 2023. OneFormer: One Transformer to Rule Universal Image Segmentation.

[41] Jun, H.; and Nichol, A. 2023. Shap-E: Generating Conditional 3D Implicit Functions. arXiv:2305.02463.

[42] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

[43] Kazemkhani, S.; Pandya, A.; Cornelisse, D.; Shacklett, B.; and Vinitsky, E. 2025. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[44] Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] Ke, B.; Qu, K.; Wang, T.; Metzger, N.; Huang, S.; Li, B.; Obukhov, A.; and Schindler, K. 2025. Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis. arXiv:2505.09358.

[46] Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

[47] Kheradmand, S.; Rebain, D.; Sharma, G.; Sun, W.; Tseng, J.; Isack, H.; Kar, A.; Tagliasacchi, A.; and Yi, K. M. 2024. 3D Gaussian Splatting as Markov Chain Monte Carlo. arXiv:2404.09591.

[48] Kim, S. W.; Brown, B.; Yin, K.; Kreis, K.; Schwarz, K.; Li, D.; Rombach, R.; Torralba, A.; and Fidler, S. 2023. NeuralField-LDM: Scene Generation With Hierarchical Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8496–8506.

[49] Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[50] Lee, J.; Lee, S.; Jo, C.; Im, W.; Seon, J.; and Yoon, S.-E. 2024. SemCity: Semantic Scene Generation with Triplane Diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[51] Li, H.; Shi, H.; Zhang, W.; Wu, W.; Liao, Y.; Wang, L.; Lee, L.-h.; and Zhou, P. 2024. DreamScene: 3D Gaussian-based Text-to-3D Scene Generation via Formation Pattern Sampling. *arXiv preprint arXiv:2404.03575*.

[52] Li, Q.; Peng, Z.; Feng, L.; Zhang, Q.; Xue, Z.; and Zhou, B. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[53] Li, Y.; Zou, Z.-X.; Liu, Z.; Wang, D.; Liang, Y.; Yu, Z.; Liu, X.; Guo, Y.-C.; Liang, D.; Ouyang, W.; et al. 2025. TripoSG: High-Fidelity 3D Shape Synthesis using Large-Scale Rectified Flow Models. *arXiv preprint arXiv:2502.06608*.

[54] Liang, Y.; Yang, X.; Lin, J.; Li, H.; Xu, X.; and Chen, Y. 2024. 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*, 6517–6526.

[55] Liao, Y.; Xie, J.; and Geiger, A. 2022. 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.

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

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

[58] Ljungbergh, W.; Tonderski, A.; Johnander, J.; Caesar, H.; Åström, K.; Felsberg, M.; and Petersson, C. 2024. NeuroNCAP: Photorealistic Closed-loop Safety Testing for Autonomous Driving. *arXiv preprint arXiv:2404.07762*.

[59] Lu, J.; Huang, Z.; Zhang, J.; Yang, Z.; and Zhang, L. 2024. WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation. In *European Conference on Computer Vision (ECCV)*.

[60] Lu, Y.; Ren, X.; Yang, J.; Shen, T.; Wu, Z.; Gao, J.; Wang, Y.; Chen, S.; Chen, M.; Fidler, S.; and Huang, J. 2024. InfiniCube: Unbounded and Controllable Dynamic 3D Driving Scene Generation with World-Guided Video Models. arXiv:2412.03934.

[61] Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

[62] Miao, X.; Duan, H.; Ojha, V.; Song, J.; Shah, T.; Long, Y.; and Ranjan, R. 2024. Dreamer XL: Towards High-Resolution Text-to-3D Generation via Trajectory Score Matching. *arXiv preprint arXiv:2405.11252*.

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

[64] NVIDIA; Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; Dworakowski, D.; Fan, J.; Fenzi, M.; Ferroni, F.; Fidler, S.; Fox, D.; Ge, S.; Ge, Y.; Gu, J.; Gururani, S.; He, E.; Huang, J.; Huffman, J.; Jannaty, P.; Jin, J.; Kim, S. W.; Klár, G.; Lam, G.; Lan, S.; Leal-Taixe, L.; Li, A.; Li, Z.; Lin, C.-H.; Lin, T.-Y.; Ling, H.; Liu, M.-Y.; Liu, X.; Luo, A.; Ma, Q.; Mao, H.; Mo, K.; Mousavian, A.; Nah, S.; Niverty, S.; Page, D.; Paschalidou, D.; Patel, Z.; Pavao, L.; Ramezanali, M.; Reda, F.; Ren, X.; Sabavat, V. R. N.; Schmerling, E.; Shi, S.; Stefaniak, B.; Tang, S.; Tchapmi, L.; Tredak, P.; Tseng, W.-C.; Varghese, J.; Wang, H.; Wang, H.; Wang, H.; Wang, T.-C.; Wei, F.; Wei, X.; Wu, J. Z.; Xu, J.; Yang, W.; Yen-Chen, L.; Zeng, X.; Zeng, Y.; Zhang, J.; Zhang, Q.; Zhang, Y.; Zhao, Q.; and Zolkowski, A. 2025. Cosmos World Foundation Model Platform for Physical AI.

[65] Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[66] Ost, J.; Mannan, F.; Thuerey, N.; Knodt, J.; and Heide, F. 2021. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2856–2865.

[67] Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

[68] Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv*.

[69] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

[70] Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

[71] Ren, X.; Huang, J.; Zeng, X.; Museth, K.; Fidler, S.; and Williams, F. 2024. XCube: Large-Scale 3D Generative Modeling using Sparse Voxel Hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[72] Ren, X.; Lu, Y.; Cao, T.; Gao, R.; Huang, S.; Sabour, A.; Shen, T.; Pfaff, T.; Wu, J. Z.; Chen, R.; Kim, S. W.; Gao, J.; Leal-Taixe, L.; Chen, M.; Fidler, S.; and Ling, H. 2025. Cosmos-Drive-Dreams: Scalable Synthetic Driving Data Generation with World Foundation Models.

[73] Ren, X.; Lu, Y.; Liang, H.; Wu, J. Z.; Ling, H.; Chen, M.; Fidler, F., Sanja annd Williams; and Huang, J. 2024. SCube: Instant Large-Scale Scene Reconstruction using VoxSplats. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[74] Ren, X.; Shen, T.; Huang, J.; Ling, H.; Lu, Y.; Nimier-David, M.; Müller, T.; Keller, A.; Fidler, S.; and Gao, J. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6121–6132.

[75] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

[76] Rong, G.; Shin, B. H.; Tabatabaee, H.; Lu, Q.; Lemke, S.; Možeiko, M.; Boise, E.; Uhm, G.; Gerow, M.; Mehta, S.; et al. 2020. Lgsvl simulator: A high fidelity simulator for autonomous driving. In *2020 IEEE 23rd International conference on intelligent transportation systems (ITSC)*, 1–6. IEEE.

[77] Sargent, K.; Li, Z.; Shah, T.; Herrmann, C.; Yu, H.-X.; Zhang, Y.; Chan, E. R.; Lagun, D.; Fei-Fei, L.; Sun, D.; et al. 2023. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*.

[78] Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

[79] Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

[80] Shah, S.; Dey, D.; Lovett, C.; and Kapoor, A. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, 621–635. Springer.

[81] Shriram, J.; Trevithick, A.; Liu, L.; and Ramamoorthi, R. 2024. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*.

[82] Shue, J. R.; Chan, E. R.; Po, R.; Ankner, Z.; Wu, J.; and Wetzstein, G. 2023. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20875–20886.

[83] Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

[84] Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.

[85] Song, Y.; Sun, Z.; and Yin, X. 2024. SDXS: Real-Time One-Step Latent Diffusion Models with Image Conditions. *arXiv preprint arXiv:2403.16627*.

[86] Stein, G.; Cresswell, J.; Hosseinzadeh, R.; Sui, Y.; Ross, B.; Villecroze, V.; Liu, Z.; Caterini, A. L.; Taylor, E.; and Loaiza-Ganem, G. 2023. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36: 3732–3784.

[87] 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.; and Anguelov, D. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[88] Talwar, D.; Guruswamy, S.; Ravipati, N.; and Eirinaki, M. 2020. Evaluating validity of synthetic data in perception tasks for autonomous vehicles. In *2020 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 73–80. IEEE.

[89] Team, A.; Zhu, H.; Wang, Y.; Zhou, J.; Chang, W.; Zhou, Y.; Li, Z.; Chen, J.; Shen, C.; Pang, J.; et al. 2025. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*.

[90] Team, G. 2024. Mochi 1. https://github.com/genmoai/models.

[91] Team, T. H. 2025. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. arXiv:2501.12202.

[92] Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4): 1–12.

[93] Tonderski, A.; Lindström, C.; Hess, G.; Ljungbergh, W.; Svensson, L.; and Petersson, C. 2023. NeuRAD: Neural rendering for autonomous driving. *arXiv preprint arXiv:2311.15260*.

[94] Vahdat, A.; and Kautz, J. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33: 19667–19679.

[95] Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34: 11287–11302.

[96] Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

[97] Wang, P.; Xu, D.; Fan, Z.; Wang, D.; Mohan, S.; Iandola, F.; Ranjan, R.; Li, Y.; Liu, Q.; Wang, Z.; et al. 2023. Taming Mode Collapse in Score Distillation for Text-to-3D Generation. *arXiv preprint arXiv:2401.00909*.

[98] Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; and Lu, J. 2023. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*.

[99] Williams, F.; Gojcic, Z.; Khamis, S.; Zorin, D.; Bruna, J.; Fidler, S.; and Litany, O. 2021. Neural Fields as Learnable Kernels for 3D Reconstruction. arXiv:2111.13674.

[100] Wu, J. Z.; Zhang, Y.; Turki, H.; Ren, X.; Gao, J.; Shou, M. Z.; Fidler, S.; Gojcic, Z.; and Ling, H. 2025. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26024–26035.

[101] Wu, R.; Gao, R.; Poole, B.; Trevithick, A.; Zheng, C.; Barron, J. T.; and Holynski, A. 2024. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. *arXiv:2411.18613*.

[102] Wu, R.; Mildenhall, B.; Henzler, P.; Park, K.; Gao, R.; Watson, D.; Srinivasan, P. P.; Verbin, D.; Barron, J. T.; Poole, B.; et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21551–21561.

[103] Wu, Z.; Liu, T.; Luo, L.; Zhong, Z.; Chen, J.; Xiao, H.; Hou, C.; Lou, H.; Chen, Y.; Yang, R.; et al. 2023. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, 3–15. Springer.

[104] Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2024. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*.

[105] Xie, H.; Chen, Z.; Hong, F.; and Liu, Z. 2024. Citydreamer: Compositional generative model of unbounded 3d cities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9666–9675.

[106] Xie, K.; Lorraine, J.; Cao, T.; Gao, J.; Lucas, J.; Torralba, A.; Fidler, S.; and Zeng, X. 2024. LATTE3D: Large-scale Amortized Text-To-Enhanced3D Synthesis. *arXiv preprint arXiv:2403.15385*.

[107] Xu, Y.; Chai, M.; Shi, Z.; Peng, S.; Skorokhodov, I.; Siarohin, A.; Yang, C.; Shen, Y.; Lee, H.-Y.; Zhou, B.; et al. 2023. 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*, 4402–4412.

[108] Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; Li, T.; Dai, B.; Chitta, K.; Wu, P.; Zeng, J.; Luo, P.; Zhang, J.; Geiger, A.; Qiao, Y.; and Li, H. 2024. Generalized Predictive Model for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[109] Yang, J.; Huang, J.; Chen, Y.; Wang, Y.; Li, B.; You, Y.; Igl, M.; Sharma, A.; Karkus, P.; Xu, D.; Ivanovic, B.; Wang, Y.; and Pavone, M. 2025. STORM: Spatio-Temporal Reconstruction Model for Large-scale Outdoor Scenes. *arXiv preprint arXiv:2501.00602*.

[110] Yang, S.; Hou, L.; Huang, H.; Ma, C.; Wan, P.; Zhang, D.; Chen, X.; and Liao, J. 2024. Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion. *arXiv preprint arXiv:2402.03162*.

[111] Yang, Y.; Yang, Y.; Guo, H.; Xiong, R.; Wang, Y.; and Liao, Y. 2023. Urbangiraffe: Representing urban scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9199–9210.

[112] Yang, Z.; Chen, Y.; Wang, J.; Manivasagam, S.; Ma, W.-C.; Yang, A. J.; and Urtasun, R. 2023. UniSim: A Neural Closed-Loop Sensor Simulator. In *CVPR*.

[113] Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

[114] Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024. GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. In *CVPR*.

[115] Yu, H.-X.; Duan, H.; Herrmann, C.; Freeman, W. T.; and Wu, J. 2024. WonderWorld: Interactive 3D Scene Generation from a Single Image. *arXiv preprint arXiv:2406.09394*.

[116] Yu, H.-X.; Duan, H.; Hur, J.; Sargent, K.; Rubinstein, M.; Freeman, W. T.; Cole, F.; Sun, D.; Snavely, N.; Wu, J.; et al. 2024. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6658–6667.

[117] Zhang, J.; Zhang, Q.; Zhang, L.; Kompella, R. R.; Liu, G.; and Zhou, B. 2024. Urban Scene Diffusion through Semantic Occupancy Map. *arXiv preprint arXiv:2403.11697*.

[118] Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

[119] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[120] Zhang, S.; Zhang, Y.; Zheng, Q.; Ma, R.; Hua, W.; Bao, H.; Xu, W.; and Zou, C. 2024. 3D-SceneDreamer: Text-Driven 3D-Consistent Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10170–10180.

[121] Zhang, Z.; Long, F.; Pan, Y.; Qiu, Z.; Yao, T.; Cao, Y.; and Mei, T. 2024. TRIP: Temporal Residual Learning with Image Noise Prior for Image-to-Video Diffusion Models. *arXiv preprint arXiv:2403.17005*.

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

[123] Zhou, L.; Du, Y.; and Wu, J. 2021. 3D Shape Generation and Completion Through Point-Voxel Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5826–5835.

[124] Zyrianov, V.; Che, H.; Liu, Z.; and Wang, S. 2024. LidarDM: Generative LiDAR Simulation in a Generated World. *arXiv preprint arXiv:2404.02903*.