—Supplementary Material—
DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation

1. Metric Depth Curation

The proposed metric depth curation pipeline relies on LiDAR point clouds and visual reconstruction results to generate fine-grained depth maps. Due to the high sparsity of LiDAR point clouds in the nuScenes [2] dataset, we aggregate multi-frame LiDAR data using a window of three frames. For the MVS reconstruction network, we set the maximum depth to 100 meters, and the final aggregated point cloud undergoes voxel downsampling with a resolution of 0.1 meters to reduce redundancy.

Since MVS static point clouds contain significant noise, we filter out ground-level noise by retaining only points located above the ego vehicle's LiDAR. When merging LiDAR and MVS points, we prioritize LiDAR data and utilize MVS points only in regions where LiDAR coverage is unavailable. We apply nearest-neighbor interpolation to obtain an initial dense metric depth prompt. Finally, we utilize a generative depth completion network to obtain a dense and accurate metric depth map, while a semantic segmentation network is applied to identify the sky region and assign it a depth of 100 meters, ensuring consistency in depth representation.

The effectiveness of the proposed metric depth curation pipeline is illustrated in Fig.1. When relying solely on LiDAR points, depth estimation errors for distant objects tend to be significant. Incorporating the static scene point cloud reconstructed via MVS substantially alleviates this issue. Moreover, since both LiDAR and MVS-reconstructed point clouds maintain cross-camera consistency, our depth completion pipeline not only enhances fine-grained details within individual frames but also ensures high temporal and multi-view consistency across the entire scene.

We further evaluate our pseudo metric depth ground truth on the nuScenes validation set [2], using multi-frame LiDAR depth as the reference. As shown in Tab.1, our pseudo depth GT achieves higher accuracy compared to the estimated depth results from[11, 16].

More results about our metric depth are in Fig.14

Method	Abs. Rel. ↓	RMSE ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑
Multiple Frame LiDAR GT				
SD [11]	0.27 / 0.28	6.50 / 6.59	0.67 / 0.63	0.87 / 0.85
M^{2}Depth [16]	0.25 / 0.26	6.02 / 6.16	0.72 / 0.72	0.89 / 0.88
DiST-T (Ours-D)	0.20 / 0.25	5.33 / 5.58	0.79 / 0.75	0.92 / 0.91
DiST-T (Ours)	0.24 / 0.35	6.24 / 6.89	0.71 / 0.61	0.88 / 0.83
Pseudo GT	0.13 / 0.20	3.46 / 3.31	0.84 / 0.76	0.95 / 0.93

Table 1. Quantitative evaluation of pseudo depth GT and generated depth on nuScense dataset [2].

2. DiST-T

Model Setup. We use the pre-trained 3D VAE from CogVideoX [14] and train the diffusion model from scratch. First, we train DiST-T for RGB video generation with a resolution of 224×400 for 7 days. Then we train the DiST-T for RGB-D video generation with the resolution of 224×400 for 3 days, followed by training at 424×800 for 3 days. All training phases are conducted on $8\times$ NVIDIA H20 GPUs. The backbone of STDIT has the same layer $N = 28$ and hidden size $d = 1152$ following the previous work.

Loss Function We use simulation-free rectified flow [7] and v-prediction loss [5]:

$$\mathbf{z}_t = (1-t)x_0 + t\epsilon \tag{1}$$

$$\mathcal{L} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\|\mathcal{G}_\theta(\mathbf{z}_t, t) - (\epsilon - x_0)\|_2^2, \tag{2}$$

where $t \sim \text{lognorm}(0, 1)$ is timestep and $\mathcal{G}_\theta$ indicates DiST-T network with parameters θ.

More experiment results We chose 17 frames for a fair evaluation with prior works [6], and our model can support 65 frames when trained on a single H20 GPU, as shown in Fig. 2. In order to further demonstrate the capability of DiST-T in OOD cases, We conduct zero-shot experiments on the Waymo dataset, as shown in Fig. 3.

Besides, DiST-T can support generation with fewer control signals. During training we use random dropout on various conditions, enabling the model to infer using only the BEV map, as shown in Fig. 4. More visualization results

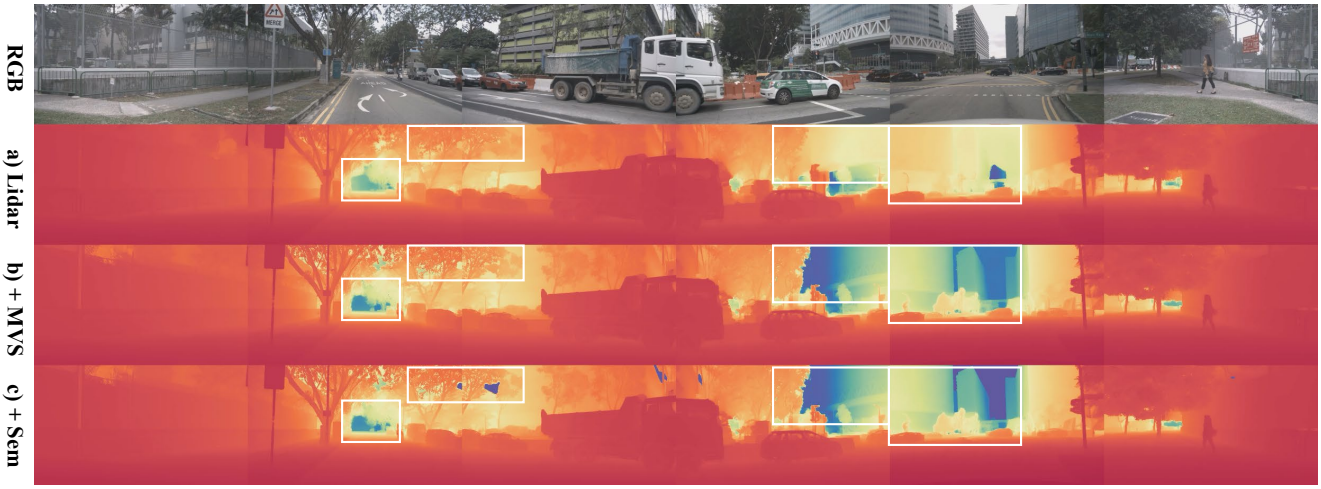

Figure 1. The comparison of the dense metric depth completion performance on a) only multi-frame LiDAR point cloud, b) add static scene point cloud with MVS, c) add sky semantic mask

about RGB-D video generation of DiST-T are provided in Fig. 8 ∼ Fig. 11.

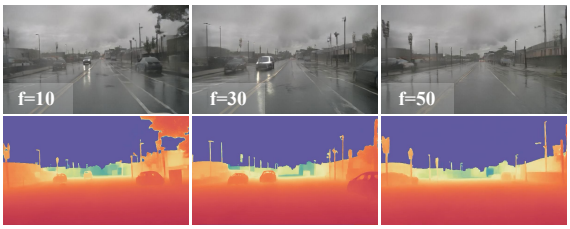

Figure 2. Generating more frames with DiST-T

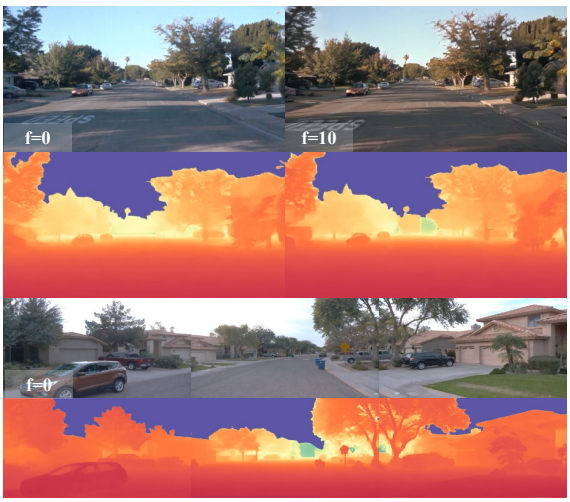

Figure 3. Zero-shot performance on Waymo

Figure 4. Inference with only BEV map condition

3. DiST-S

Model Setup. The resolution of generated results is set to 448×768 with the video length $T = 6$. Initializing DiST-S from SVD [1], we train DiST-S for 1 day with $8\times$ NIVIDA H20 GPUs. SCC strategy is employed to full trainset for generating novel trajectories. DiST-S with the SCC strategy will require an additional day of training.

More experiment results We qualitatively compare DiST-S with another Point2Video diffusion model, as illustrated in Fig.6. For FreeVS* [9], we retrain the model on the nuScenes[2] dataset using the official code and aggregated LiDAR point clouds from more frames ($n = 10$). However, this method relies on accurate and dense LiDAR point clouds. Even after multi-frame aggregation, the density of LiDAR point clouds in nuScenes still falls short of that in the Waymo Open Dataset [8]. Besides, LiDAR-based methods struggle to handle distant buildings. From the visual comparisons, it can be observed that FreeVS* exhibits significantly poorer

performance in distant objects and the sky.

For ViewCrafter*[15], DUSt3R[10] is used for relative depth estimation in the official model, making control-specific locations unfeasible. Therefore, we utilized our processed pseudo-image as the conditional input and the image in the recorded trajectory as a reference image. However, as this method is specifically designed for static scenes, its effectiveness in driving scenarios is limited, and it lacks the capability to accurately synthesize novel views based on given conditions.

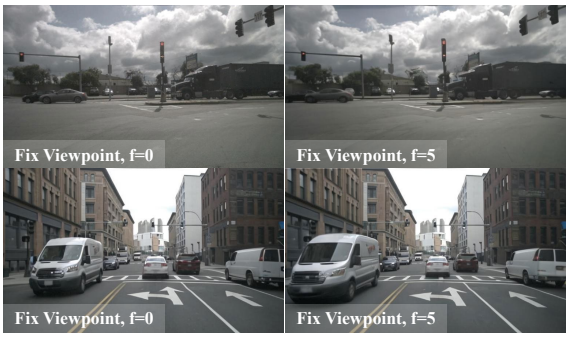

Figure 5. Fixing viewpoint whilt varying timestamp.

In addition to horizontal shifts, DiST-S is capable of simulating various translations and rotations, as shown in Fig. 7. The model can support various other tasks, such as fixing viewpoint while varying timestamp (Fig. 5).

Our model maintains robust inter-frame consistency in dynamic cars and background, though degradation may occur with distant objects, dynamic pedestrians, and areas with ambiguous conditions, especially in distant views. More visualization results about spatial novel view synthesis of DiST-S are provided in Fig. 12 and Fig. 13.

Implementation Details of Reconstruction Methods To evaluate the effectiveness of our method in novel view synthesis (NVS), we compare it against NeRF-based (EmerNeRF [13]) and 3DGS-based (StreetGaussian [12], PVG [3], OmniRe [4]) approaches on 30 scenes from the nuScenes validation dataset. The results are presented in Table 5 and Figure 7 in the main text.

For these reconstruction methods, we use all frames in each scene for training. Specifically, **EmerNeRF**: We use the official implementation. Since EmerNeRF is configured to train with 100 frames by default on the nuScenes dataset, we split each scene into two subsets, each containing approximately 100 frames. **StreetGaussian, PVG, and OmniRe**: We utilize the official code and training configuration provided by OmniRe.

The 30 selected scenes for validation (nuScenes-devkit [2] order) are: *11-scene, 12-scene, 13-scene, 14-scene, 36-scene, 75-scene, 79-scene, 83-scene, 84-scene, 87-scene, 88-scene, 90-scene, 91-scene, 92-scene, 214-scene, 257-scene, 259-scene, 261-scene, 262-scene, 410-scene, 412-scene, 414-scene, 436-scene, 439-scene, 442-scene, 443-scene, 444-scene, 445-scene, 446-scene, 447-scene*.

4. Notations

The notations used in paper are listed in Tab. 2.

References

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

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2, 3

[3] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv preprint arXiv:2311.18561*, 2023. 3

[4] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024. 3

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1

[6] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *ICLR*, 2024. 1

[7] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[8] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[9] Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. *arXiv preprint arXiv:2410.18079*, 2024. 2, 4

[10] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 3

[11] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on robot learning*, pages 539–549. PMLR, 2023. 1

Figure 6. Comparison of FreeVS* [9] and ViewCrafter* [15]. We present novel view synthesis results under shifted viewpoints (shift left by 2 meters), where our method produces higher-quality images. While ViewCrafter* uses the same sparse conditions as ours, it struggles to adhere accurately to conditions, resulting in mismatched novel view outputs.

[12] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024. 3

[13] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. 3

[14] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[15] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3, 4

[16] Yingshuang Zou, Yikang Ding, Xi Qiu, Haoqian Wang, and Haotian Zhang. M2depth: Self-supervised two-frame m ulti-camera m etric depth estimation. In *European Conference on Computer Vision*, pages 269–285. Springer, 2024. 1

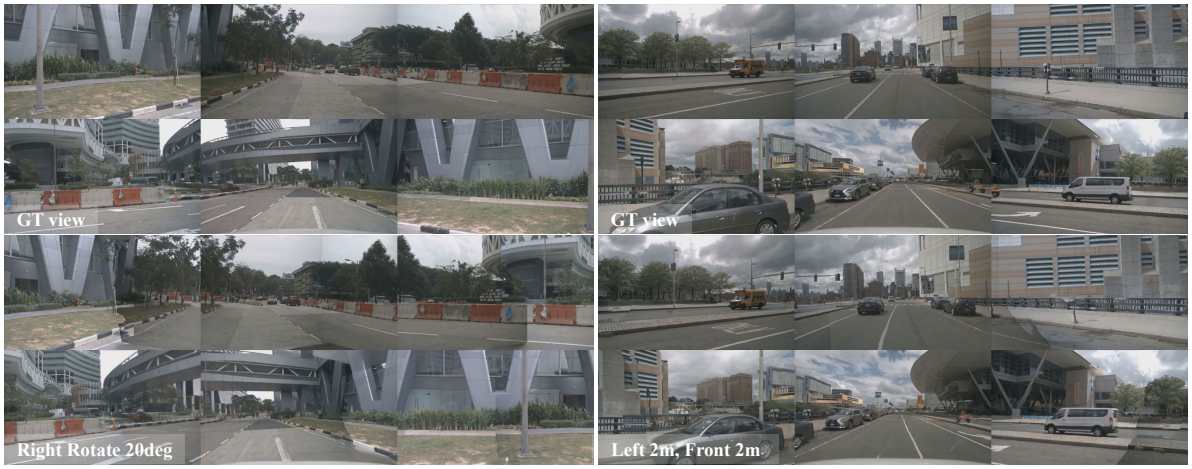

Figure 7. More results of translations and rotations in NVS.

Notation	Description
E	VAE encoder
D	VAE decoder
T	number of frames
C	number of cameras
$\mathbf{I}_{\text{ref},c}$	reference image of c-th camera
$\mathbf{I}_{\text{ref}}$	reference images of the all cameras
$\mathbf{Z}^I$, $\mathbf{Z}^D$	image, depth latent feature
$\mathbf{Z}^I_{\text{cond}}$, $\mathbf{Z}^D_{\text{cond}}$	latent feature of image, depth condition
N	number of blocks in the DiST-T
$\mathbf{P}_t$,	camera pose at t-th frame
$\mathbf{B}_t$	3D bounding boxes at t-th frame including class information and corner points of boxes
$\mathbf{A}_t$	ego trajectory information at t-th frame
$\mathbf{M}_t$	map information
$\mathbf{A}_{\text{ori}}$	the original trajectory
$\mathbf{A}_{\text{novel}}$	the novel trajectory
V_t, V_{t+n}	viewpoint of the t-th ,and $t+n$-th frame in the original trajectory $\mathbf{A}_{\text{ori}}$
$V_{t'}$	viewpoint of the t-th frame in the novel trajectory $\mathbf{A}_{\text{novel}}$
$T^{t\to t+n}$	transform matrix from the t-th frame to the $t+n$-th frame in $\mathbf{A}_{\text{ori}}$
$T^{t\to t'}$	transform matrix from the t-th frame in $\mathbf{A}_{\text{ori}}$ to the t-th frame in $\mathbf{A}_{\text{novel}}$
τ	laterally shift distance from $\mathbf{A}_{\text{ori}}$ to $\mathbf{A}_{\text{novel}}$
$\mathbf{I}^p_{t+n}$	image condition projected at the $t+n$-th frame
$\mathbf{D}^p_{t+n}$	depth condition projected at the $t+n$-th frame, including depth and valid mask
$\mathbf{C}_{t+n}$	projected condition at the $t+n$-th frame concatenated from image, depth and valid mask

Table 2. Table of notations and descriptions

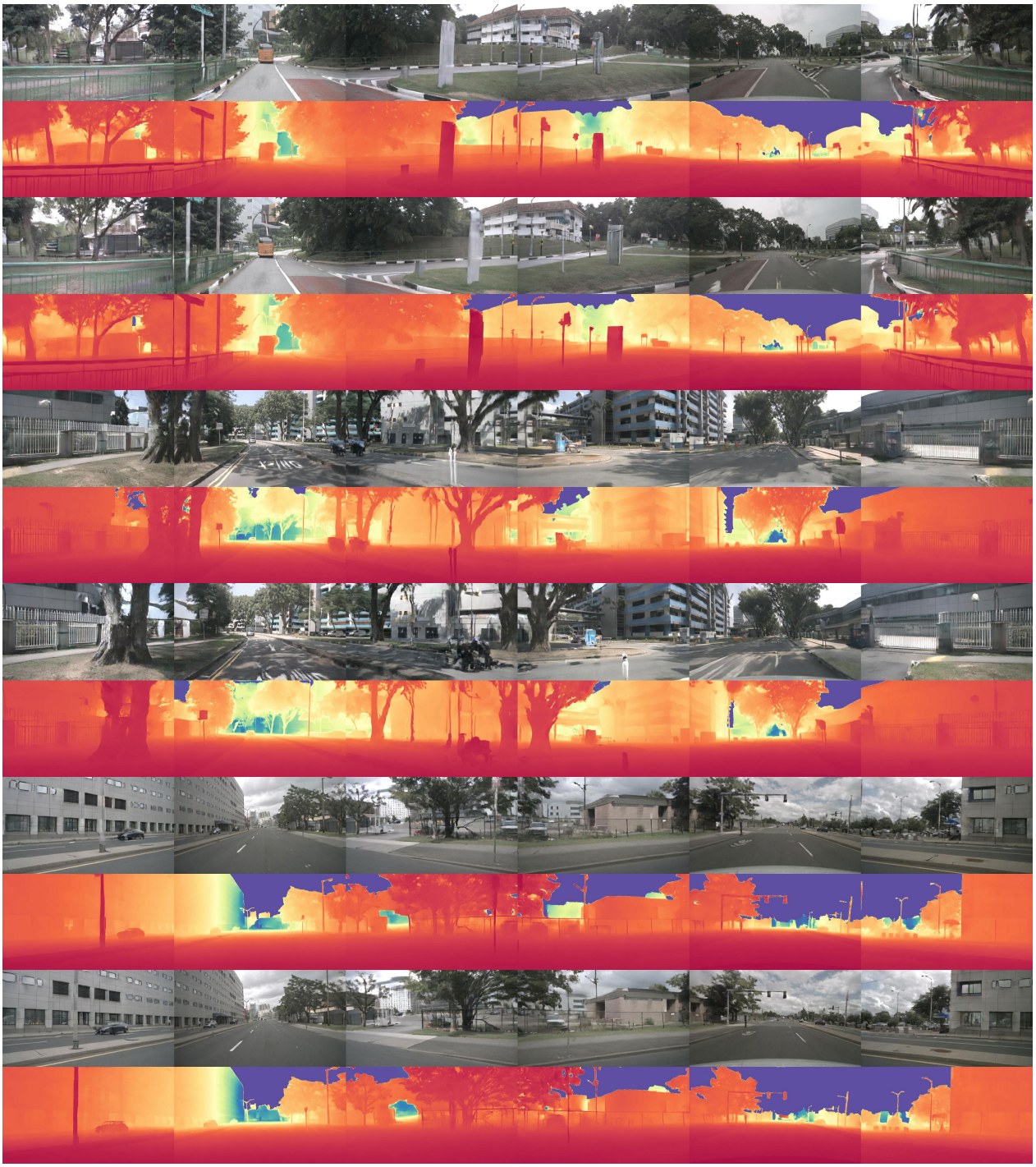

Figure 8. Additional visualizations of video generation using DiST-T. Our model can produce high-quality RGB videos along with corresponding metric depth sequences.

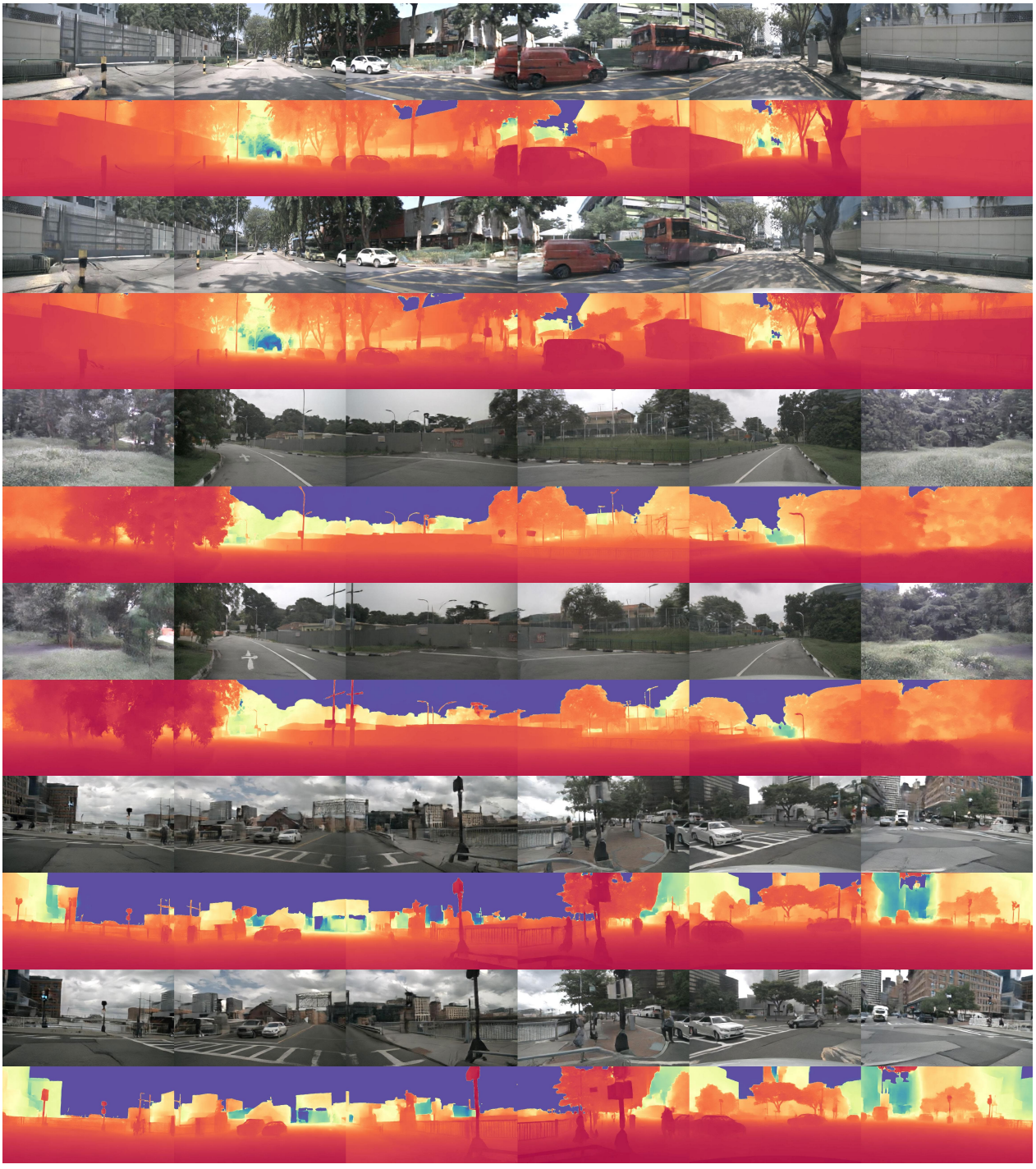

Figure 9. Additional visualizations of video generation using DiST-T. Our model can produce high-quality RGB videos along with corresponding metric depth sequences.

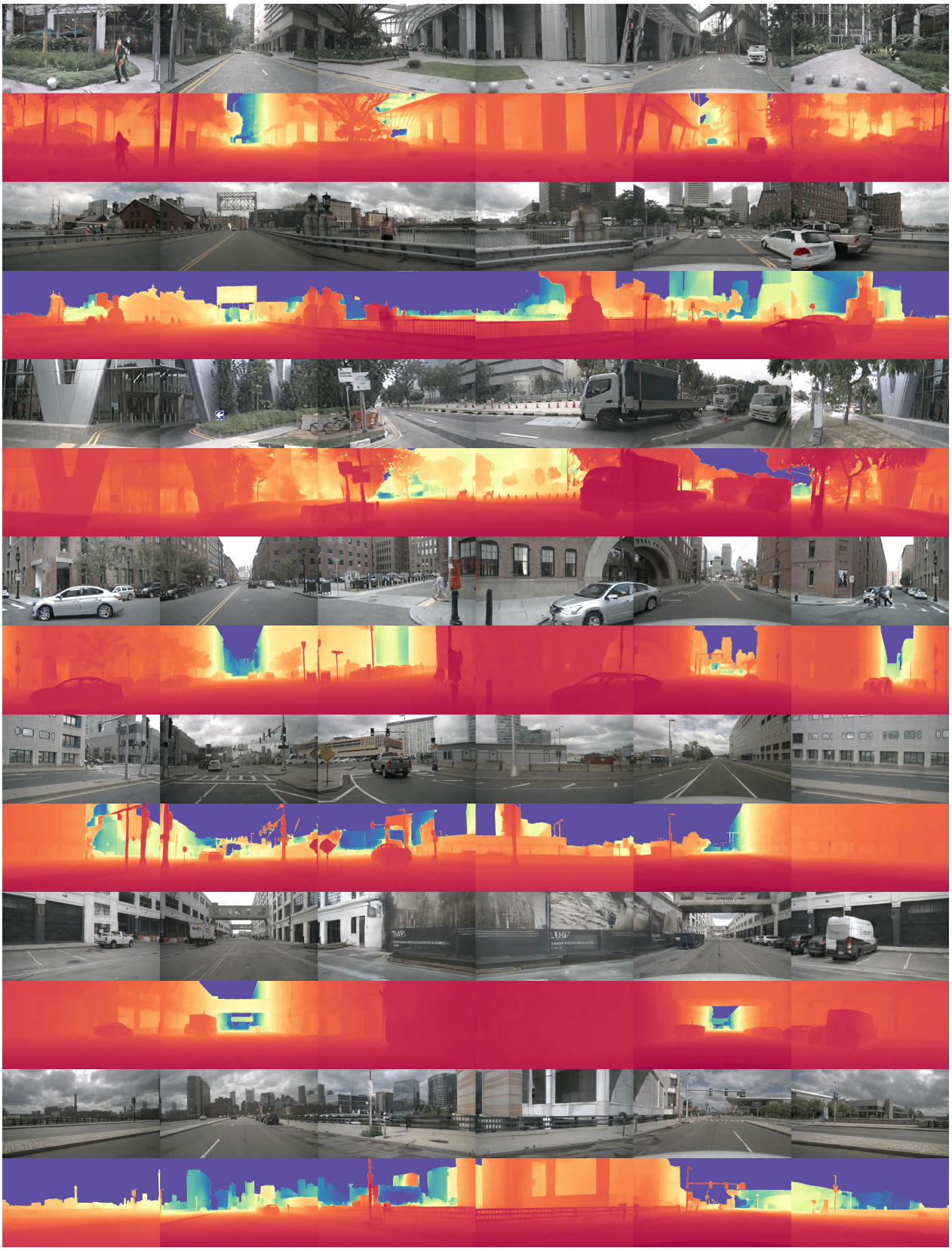

Figure 10. Additional visualizations of video generation using DiST-T. Our model can produce high-quality RGB videos along with corresponding metric depth sequences.

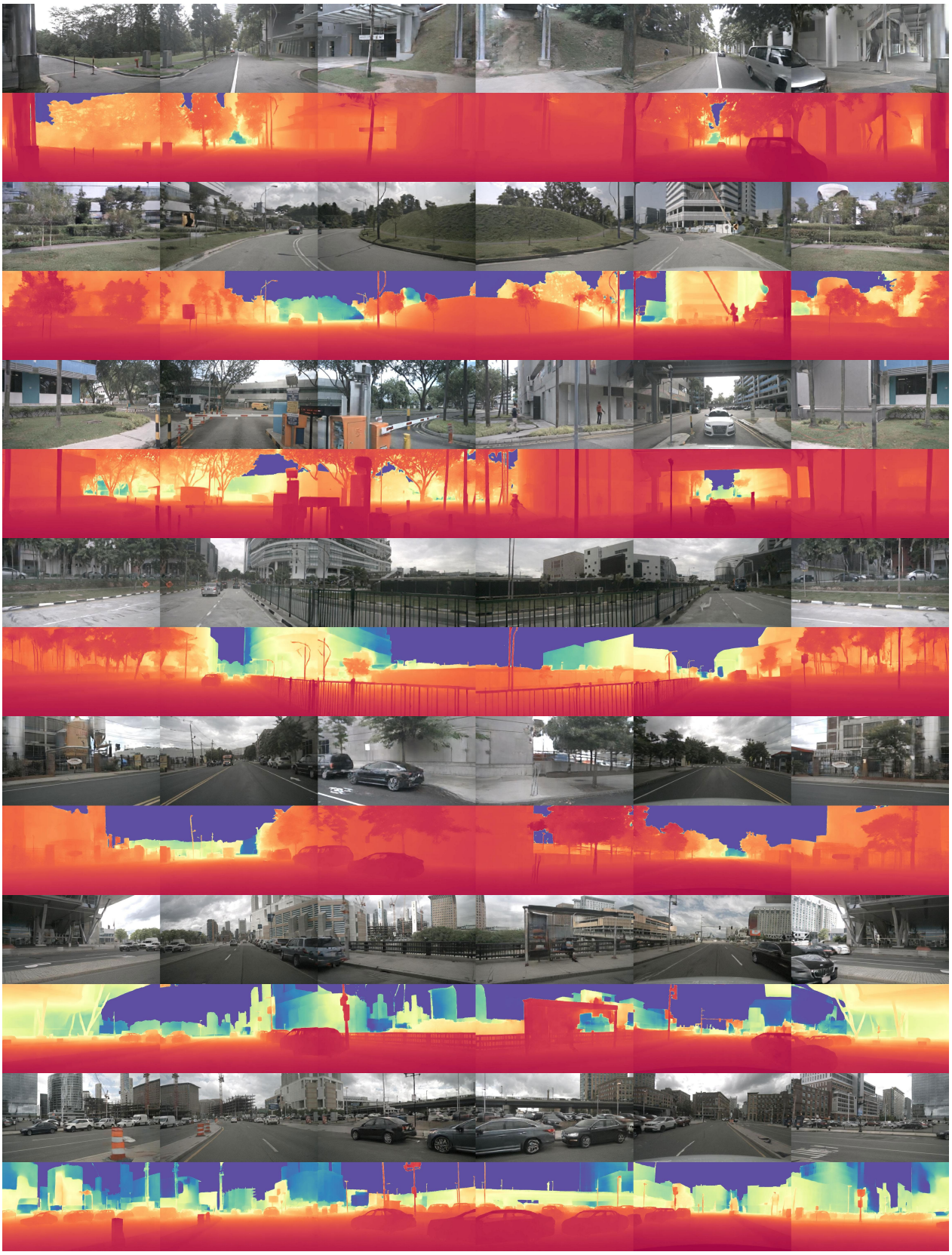

Figure 11. Additional visualizations of video generation using DiST-T. Our model can produce high-quality RGB videos along with corresponding metric depth sequences.

Figure 12. Additional visualizations of spatial novel view synthesis using DiST-S. We present NVS results of DiST-S from various shifted viewpoints, demonstrating our method's ability to generate photorealistic images with high consistency to the original scene.

Figure 13. Additional visualizations of spatial novel view synthesis using DiST-S. We present NVS results of DiST-S from various shifted viewpoints, demonstrating our method's ability to generate photorealistic images with high consistency to the original scene.

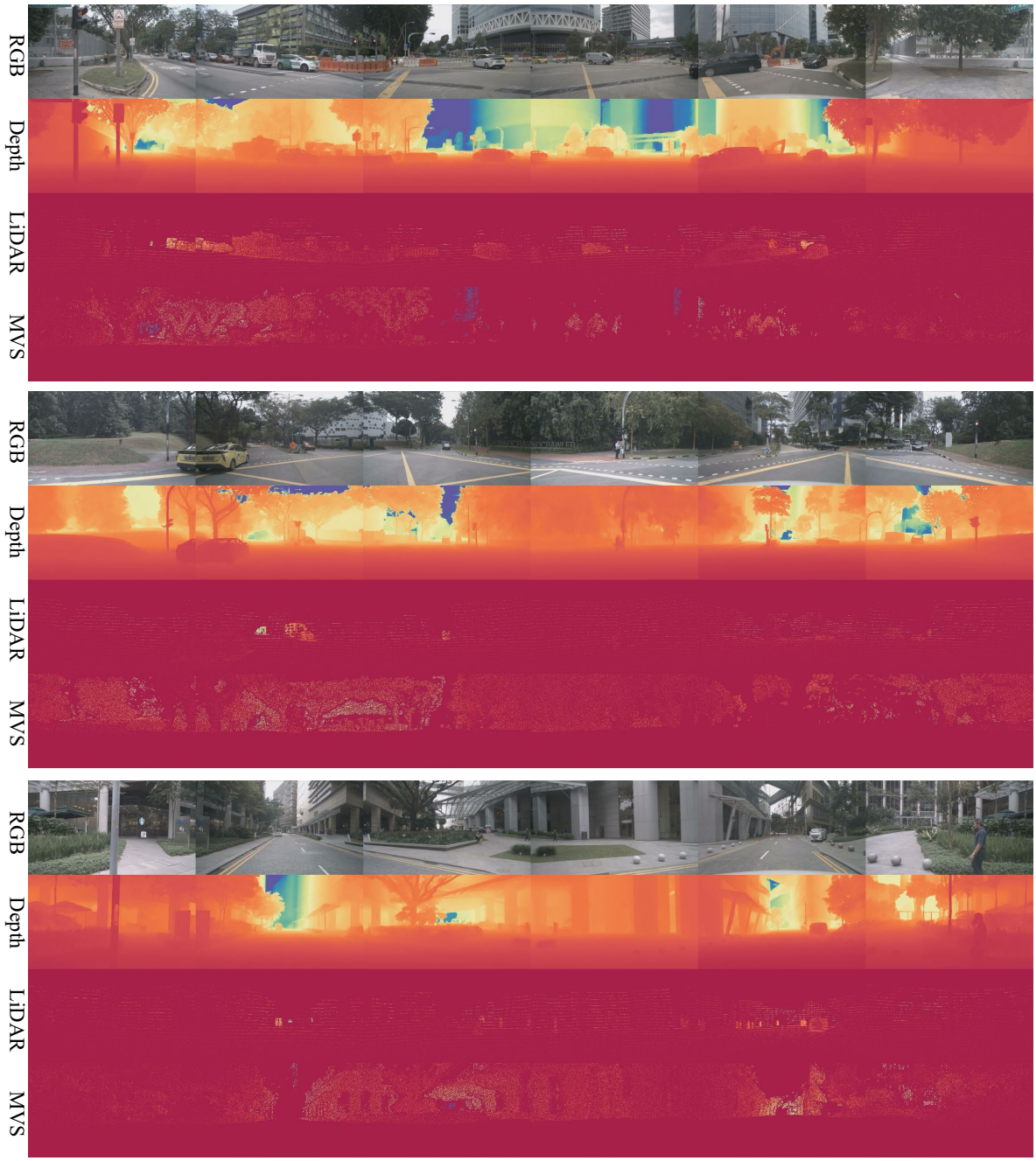

Figure 14. The visualization results of the corresponding LiDAR points and MVS points for the processed metric depth pseudo ground truth.