WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation

Jiachen Lu
Fudan University

Ze Huang
Fudan University

Jiahui Zhang
Fudan University

Li Zhang
Fudan University

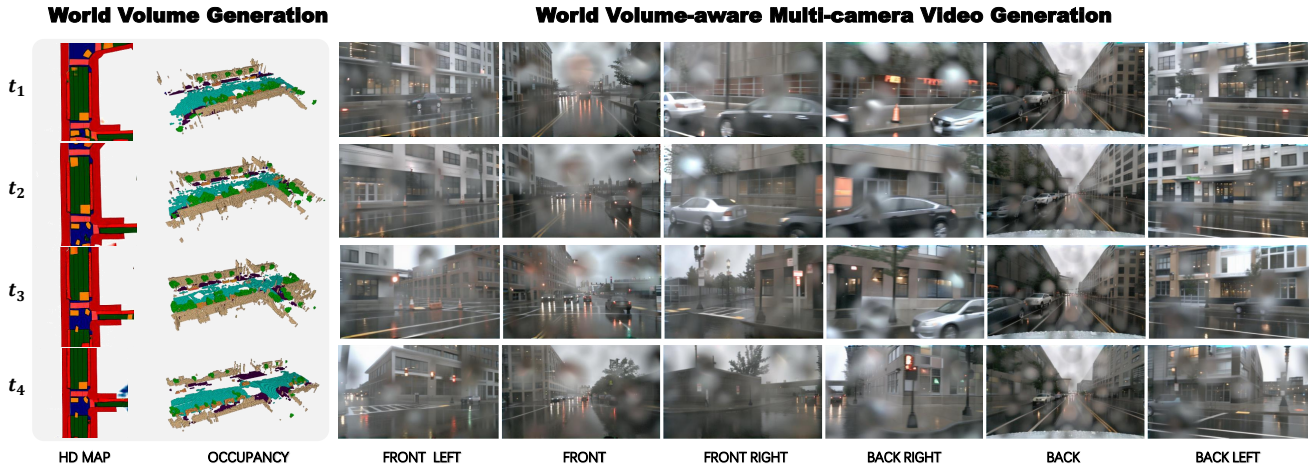

Figure 1. WoVoGen aims to generate real-world future world volumes (left) given vehicle actions and world volume-aware multi-camera street-view videos (right). WoVoGen is trained in two phases by latent diffusion models [20, 33]: first envisioning the future 4D world volume, followed by generating multi-camera videos based on this envisioned world volume.

Abstract

Generating multi-camera street-view videos is critical for augmenting autonomous driving datasets, addressing the urgent demand for extensive and varied data. Due to the limitations in diversity and challenges in handling lighting conditions, traditional rendering-based methods are increasingly being supplanted by diffusion-based methods. However, a significant challenge in diffusion-based methods is ensuring that the generated sensor data preserve both intra-world consistency and inter-sensor coherence. To address these challenges, we combine an additional explicit world volume and propose World Volume-aware Multi-camera Driving Scene Generator (WoVoGen). This system is specifically designed to leverage 4D world volume as a foundational element for video generation. Our model operates in two distinct phases: (i) envisioning the future 4D temporal world volume based on vehicle control sequences, and (ii) generating multi-camera videos, informed by this envisioned 4D temporal world volume and sensor interconnectivity. The incorporation of the 4D world volume empowers WoVoGen not only to generate high-quality street-view videos in response to vehicle control inputs but also to facilitate scene editing tasks.

1. Introduction

The burgeoning field of vision-based autonomous driving perception [15, 18, 19] underscores the need for high-quality multi-camera street-view datasets [1]. With the notorious costs of labeling in autonomous driving datasets, there is a significant demand for generating high-quality multi-camera street-view videos that accurately mirror real-world 3D data distributions and maintain consistency across multiple sensors.

Recent generation techniques can be categorized into two main groups: Rendering-based [7, 29, 30, 32] and diffusion-based [6, 12–14, 28, 31] methods. Rendering-based methods benefit from an explicit 3D or 4D world structure, ensuring stringent 3D consistency. Yet, this approach tends to offer limited diversity and requires considerable effort to produce multi-camera videos that comply with real-world lighting, weather conditions, *etc*. On the other hand, methods based on finetuned stable diffusion models boast high diversity and can easily generate images

that follow real-world distributions. Despite these advantages, diffusion-based approaches often lack a clear 3D or 4D (time) world volume, leading to lower 3D consistency and a shortfall in temporal and multi-sensor coherence.

To overcome the limitations of diffusion-based methods, we introduce World Volume-aware Multi-camera Driving Scene Generator (WoVoGen), a framework designed to endow diffusion-based generative models with an explicit 4D world volume. This enhancement imposes strict 3D consistency and temporal coherence on the diffusion process. Our approach operates in two stages: initially, we envision a 4D world volume using a reference scene combined with a future vehicle control sequence. Subsequently, this volume guides the generation of multi-camera footage.

Concretely, our 4D world volume manifests as a dense voxel volume spanning four dimensions: time, height, length, and width, corresponding to the scope of the bird's-eye view (BEV) domain. This representation encapsulates the scene's comprehensive data, comprising object occupancy, high-definition maps, background details, and road attributes. Each voxel within this 4D construct is annotated with a class label, providing a rich, multi-faceted understanding of the environment.

In the preliminary stage, our method is to train an autoencoder model [26] that encodes a single-frame 3D world volume into a 2D latent representation. Subsequently, we stack these 2D latents along the temporal axis to form a 2D temporal latent series. We further refine the UNet by introducing temporal versions of its residual and cross-attention blocks, which can effectively process the time-varying information, guided by the vehicle control sequence as the conditional context. Upon generating the future 2D temporal latent, we proceed to decode it back into the 4D world voxel volume using the autoencoder's decoder.

Having generated the future 4D world voxel volume, we employ a combination of CLIP and 3D sparse CNN to convert it into a 4D world feature. This feature is then transformed geometrically to sample 3D image volumes for each camera, corresponding to each time step. Subsequently, these 3D image volumes are condensed into 2D image features, which serve as conditional inputs for ControlNet [33]. Alongside image features, we employ textual prompts as scene guidance akin to those used in Stable Diffusion [20] to govern the overall scene conditions such as weather, lighting, location, and the scene at large. For more precise object location control within the scene, textual prompts are utilized with greater specificity. We map the 4D world volume labels onto each 2D pixel, utilizing the label names as textual conditions to objectively guide over each pixel's characteristics.

For achieving inter-sensor consistency, we concatenate surround-view images into a meta-image and leverage the diffusion model to learn the distribution of real-world multi-view image sequences. To ensure temporal coherence, we utilize the same temporal transformer blocks previously described.

In summary, we make the following contributions: (i) We propose WoVoGen, a framework that leverages an explicit world volume to guide the diffusion model generation, ensuring intra-world and inter-sensor consistency. (ii) We introduce a two-phase strategy: initially envisioning the future 4D world volume, followed by generating multi-camera videos based on this envisioned world volume.

2. Related works

Diffusion model for visual content generation Diffusion models, as demonstrated in [3, 9, 22], exhibit the capability to generate a diverse array of images and videos through a learned denoising process. Latent diffusion models [20], by sampling from a learned posterior distribution over latent variables, adeptly capture intricate dependencies and correlations in the data, thereby facilitating the generation of high-quality images. In the realm of controllable image generation, ControlNet [33] adopts a unique approach by duplicating the neural network weights. Specifically, the duplicated networks are exclusively trained to comprehend input conditions, with the use of zero initialization to safeguard against parameter degradation. To ensure image consistency across multiple views, some studies have integrated cross-view attention mechanisms to establish connections between different perspectives [16, 17, 21]. Notably, none of the aforementioned works incorporates 3D information. In contrast, our work distinguishes itself by extracting 3D information from future world volumes and injecting it into the model, thereby significantly enhancing its overall performance.

Autonomous driving image synthesis In the landscape of autonomous driving image and video synthesis, two primary categories have emerged. Render-based approaches have been explored [7, 29, 32]. These approaches involve the reconstruction of comprehensive explicit 3D or 4D world structures and the generation of high-quality images by manipulating components within these structures and rendering them. Diffusion-based approaches have gained traction [6, 12, 14, 28, 31]. These approaches employ pre-trained diffusion models and incorporate 2D conditions, such as HD maps and bounding boxes, for precise control over the diffusion process. Different from the above methods, our approach can generate future world volumes through actions and past world volumes, and then produce multi-view consistent videos based on a 4D world volume.

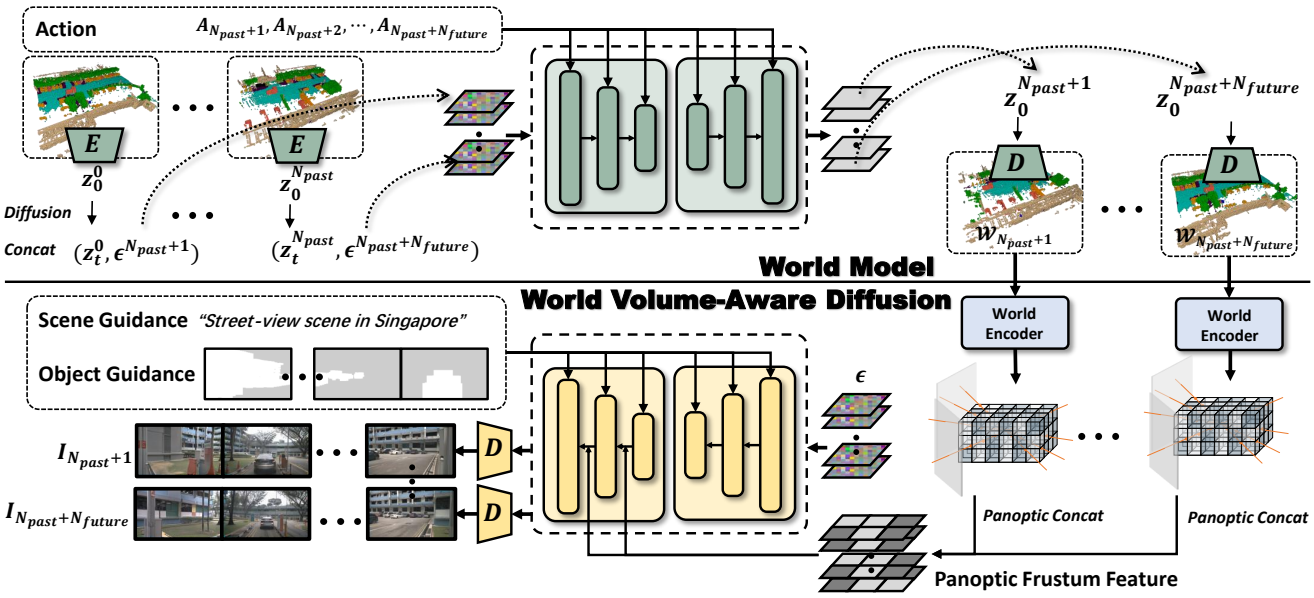

Figure 2. Overall framework of WoVoSyn. ***Top: World Model Branch.*** We finetune the AutoencoderKL and train the 4D diffusion model from scratch to generate future world volumes based on past world volumes and the actions of the ego car. ***Bottom: World Volume-Aware Synthesis Branch.*** Leveraging the generated future volumes as input, $\mathcal{F}_w$ are derived through the world encoder. Subsequent sampling yields $\mathcal{F}_{img}$, which are then aggregated. The process is finalized by applying panoptic diffusion to produce future videos.

3. Method

3.1. Preliminary

The latent diffusion model (LDM) [20] is a generative model adept at generating high-resolution images in two stages. Initially, an autoencoder compresses data into a latent space (perceptual compression), where $z = E(x)$ and $\hat{x} = D(z)$ represent encoding and decoding processes, respectively. Subsequently, a denoising diffusion probabilistic model (DDPM) [9] models image distributions in this latent space.

The generative process reverses the diffusion sequence $\mathbf{z}_1, \ldots, \mathbf{z}_T$ using a learned Gaussian transition, formalized as:

$$q\left(\mathbf{z}_\tau|\mathbf{z}_{\tau-1}\right) = \mathcal{N}\left(\mathbf{z}_\tau; \sqrt{1-\beta_\tau}\mathbf{z}_{\tau-1}, \beta_\tau \mathbf{I}\right), \quad (1)$$

$$p_\theta\left(\mathbf{z}_{\tau-1}|\mathbf{z}_\tau\right) = \mathcal{N}\left(\mathbf{z}_{\tau-1}; \mu_\theta(\mathbf{z}_\tau, \tau), \tilde{\beta}_\tau\right). \quad (2)$$

Here, $\mu_\theta(\mathbf{z}_\tau, \tau)$ is given by a trainable noise predictor $\epsilon_\theta(\mathbf{z}_\tau, \tau)$.

DDPMs, viewed as a series of weight-sharing denoising autoencoders, train to predict the initial noise from a noisy input. The training goal is to maximize the variational lower bound of the negative log-likelihood:

$$\mathbb{E}_{z,\epsilon,\tau}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_\tau, \tau)\|_2^2\right]. \quad (3)$$

Finally, sampling from the latent distribution and then using the latent decoder generates novel images from Gaussian noise.

3.2. Overall architecture

The proposed architecture, referred to as WoVoGen, comprises two distinct operational branches: the ***World Model Branch*** and the ***World Volume-Aware Generation Branch***. The World Model Branch is responsible for generating future world volumes, incorporating action inputs and several initial frames of the world volume to inform its predictions. Meanwhile, the World Volume-Aware Generation Branch focuses on the Generation of multi-camera video outputs based on temporal world volumes. An overview of our comprehensive pipeline is visually represented in Figure 2.

3.3. World volume

To effectively harness the potential of rapidly evolving generative architectures, it is imperative to transform the environmental context into a standardized format. Our primary focus herein lies on the extraction of high-level, abstract structural data pertinent to autonomous driving scenarios, encompassing road layouts, semantic occupancies, and other related elements.

Recognizing the critical role of three-dimensional data in rendering processes and the benefits derived from fine-grained constraints in view generation, we propose the encoding of scene information within a three-dimensional world volume, denoted as $\mathcal{W} \in \mathbb{R}^{Z \times H \times W \times C}$. Specifically, at any given time instance t, we amalgamate various facets of driving-related information to encapsulate the environmental context around the ego vehicle within a pre-

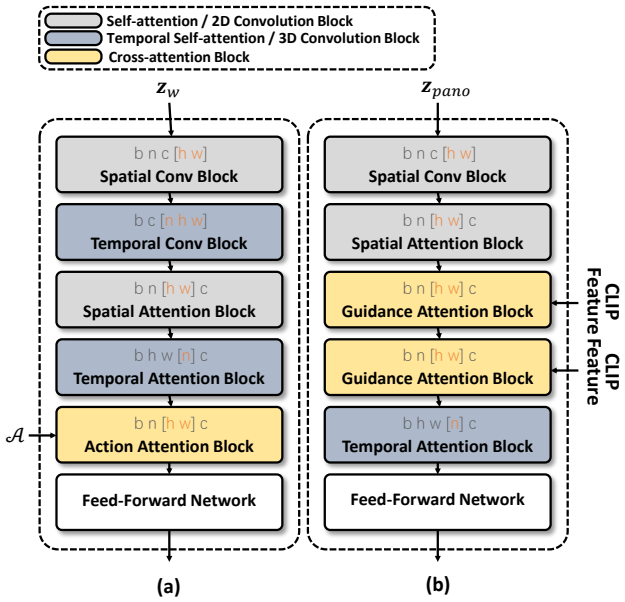

Figure 3. ***(a)***: An action attention block enhances the model by incorporating essential action information. ***(b)***: A guidance attention block integrates the CLIP feature of a specific object into the latent representation, further refining the model's capabilities.

defined range. This integration is achieved through the concatenation process, *i.e.*, $\mathcal{W} = \text{concat}(\mathcal{O}, \mathcal{M})$, where $\mathcal{O} \in \mathbb{R}^{Z \times H \times W \times C_{\text{occ}}}$ represents the three-dimensional semantic occupancy grid, with C_{occ} indicating the number of semantic classes. Concurrently, $\mathcal{M} \in \mathbb{R}^{1 \times H \times W \times C_{\text{map}}}$ captures the road map information, constrained to the voxel plane at zero height.

To maintain uniformity across the height dimension, we apply zero-padding to $\mathcal{M}$ along the Z axis. For a more streamlined representation, both road elements and semantic classes are encoded into the RGB spectrum, *i.e.*, setting $C_{\text{map}} = 3$.

3.4. World model

Given a clip of multi-camera videos captured during driving, denoted as $\{\mathcal{I}_0, \cdots, \mathcal{I}_{N_{\text{past}}}\}$, we initially employ off-the-shelf scene understanding models to infer the environment. This allows us to obtain the 3D world volume at these moments, represented as $\{\mathcal{W}_t\}_{t=1,\cdots,N_{\text{past}}}$. Subsequently, the world volume for the time instance $N_{\text{past}} + 1$ is inferred, utilizing the past world volumes and the corresponding driving actions.

Latent world volume autoencoder Given the world volume $\mathcal{W} \in \mathbb{R}^{Z \times H \times W \times C}$, the encoder is designed to compress $\mathcal{W}$ into its latent representation, $z_w = E_{\mathcal{W}}(\mathcal{W}) \in \mathbb{R}^{\frac{Z}{s} \times \frac{H}{s} \times \frac{W}{s} \times C_z}$. Here, s represents the downsampling factor, and is set equal to Z to achieve a 2D latent representation. The decoder, $D_{\mathcal{W}}$, is then employed to reconstruct the original world volume from the latent representation z_w. The autoencoder is trained using a reconstruction loss and vq-regularization [4], aimed at minimizing the variance within the latent space.

Latent world diffusion After the training of the latent autoencoder, we commence the training of the world model, with both the encoder and the decoder remaining frozen. As illustrated in Figure 2, our world model is a diffusion model that captures the conditional distribution of the latent world volume in the future, conditioned on the past world volumes and the driving actions of the ego vehicle. The noise predictor ϵ_ϕ in this stage is implemented as a time-conditioned UNet [9]. The world volumes of preceding frames are first encoded by the same latent encoder to dimensions commensurate with the noised latent z_i, and then channel-wise concatenate with z_i to serve as the input for the noise predictor.

For the driving actions, we initially tokenize the velocity v_i and the steering angle a_i of the ego vehicle through Fourier embedding [27], resulting in a sequence of action tokens $\mathcal{A} = \{v_1, a_1, \cdots, v_{N_{\text{past}}}, a_{N_{\text{past}}}\}$. These action tokens are then refined by a shallow transformer. Finally, the updated action tokens are injected into the noise predictor via cross-attention layers.

Jointly modeling of the consecutive frames To enhance the temporal consistency of predicted future world volumes and achieve more realistic scene dynamics, we transition to generating consecutive world volumes by considering a joint distribution of a world volume sequence, $p(\mathcal{W}_{N_{\text{past}}+1}, \cdots, \mathcal{W}_{N_{\text{past}}+N_{\text{future}}} | \mathcal{W}_1, \cdots, \mathcal{W}_{N_{\text{past}}}; \mathcal{A})$.

As shown in Figure 3, we build upon the basic spatial transformer block of Stable Diffusion [20], which comprises spatial convolution, spatial self-attention, and spatial cross-attention. Our temporal variant augments this framework by integrating temporal attention with residual connections into the spatial transformers.

Specifically, let us denote $z_w \in \mathbb{R}^{B \times N_{\text{future}} \times C \times H \times W}$ as the sequence of latent world volumes, adding the batch dimension for clarity. The computation within the modified spatial-temporal transformer is formalized as follows:

$$
\begin{aligned}
\mathbf{z}_w &= \text{rearrange}\left(\mathbf{z}_w, (b\ n)\ h\ w\ c \rightarrow (b\ n)\ (h\ w)\ c\right), \\
\mathbf{z}_w &= \text{MHSA}\left(\text{Norm}\left(\mathbf{z}_w\right)\right) + \mathbf{z}_w, \quad \text{(Spatial)} \\
\mathbf{z}_w &= \text{rearrange}\left(\mathbf{z}_w, (b\ n)\ (h\ w)\ c \rightarrow (b\ h\ w)\ n\ c\right), \\
\mathbf{z}_w &= \text{MHSA}\left(\text{Norm}\left(\mathbf{z}_w\right)\right) + \mathbf{z}_w, \quad \text{(Temporal)} \\
\mathbf{z}_w &= \text{rearrange}\left(\mathbf{z}_w, (b\ h\ w)\ n\ c \rightarrow (b\ n)\ (h\ w)\ c\right), \\
\mathbf{z}_w &= \text{MHCA}\left(\text{Norm}\left(\mathbf{z}_w, \mathcal{A}\right)\right) + \mathbf{z}_w, \quad \text{(Action)} \\
\mathbf{z}_w &= \text{FFN}\left(\text{Norm}\left(\mathbf{z}_w\right)\right) + \mathbf{z}_w,
\end{aligned}
$$

where `Norm` represents the group normalization, `MHSA` stands for multi-head self-attention, `MHCA` denotes multi-head cross-attention, and `FFN` refers to the feed-forward network. Both `MHSA` and `MHCA` perform attention operations on the second dimension l of the $(b\ l\ c)$ configuration.

3.5. World volume-aware 2D feature

With the temporally continuous world volumes $\mathcal{W} = \{\mathcal{W}_t\}_{t=1,2,\cdots}$ generated by our diffusion-based world model, we further decode them into relational camera images for autonomous driving video generating.

World volume encoding The world volumes inherently possess semantic information, initially represented in the form of simplistic labels. To enhance their informativeness, we employ a featurization process utilizing CLIP [19]:

$$\mathcal{F}_w = \text{SPConv}\left(\text{PCA}\left(\text{CLIP}\left(\mathcal{W}\right)\right)\right), \tag{4}$$

where `CLIP` encodes the class name into the class feature. `PCA` is used to decrease the dimension of the CLIP feature, thereby reducing the computational cost. `SPConv` (Sparse Convolution) [2] then processes the reduced-dimensional world volume feature.

Camera volume sampling To incorporate the world volume into image generation, we sample from it using dense rays emitted from the camera. We construct a 3D grid for each camera's frustum, denoted as p_c, with dimensions $D_c \times H_c \times W_c$. This grid is based on the camera's intrinsic and extrinsic properties. The process is formalized as follows:

$$\mathcal{F}_{cam} = \text{interpolate}\left(p_c, \mathcal{F}_w\right), \tag{5}$$

where $\mathcal{F}_c \in \mathbb{R}^{B\times N_{\text{future}}\times C\times D_c\times H_c\times W_c}$ represents the camera frustum. We then apply a squeeze-and-excitation operation [10] on the depth channel and sum along the depth dimension to obtain the world volume-aware 2D image feature:

$$\mathcal{F}_{img} = \sum_{i=1}^{D_c} \text{SE}(\mathcal{F}_{cam})_{[:,\ :,\ :,\ i,\ :,\ :]}. \tag{6}$$

3.6. World volume-aware diffusion generation

Based on the ControlNet framework [33], we implement a controller that injects the aforementioned world volume-aware 2D image feature $\mathcal{F}_{\text{img}}$ into the pretrained latent diffusion model. The transformer architecture of the UNet is illustrated in Figure 3. However, we initially omit the Temporal Attention Block when training the single-frame Generation model.

Panoptic diffusion We aggregate the world volume-aware 2D image feature from different view into a single panoptic feature $\mathcal{F}_{pano}$ input to the diffusion model:

$$\mathcal{F}_{pano} = \begin{bmatrix} \mathcal{F}_{img}^{\text{front left}} & \mathcal{F}_{img}^{\text{front}} & \mathcal{F}_{img}^{\text{front right}} \\ \mathcal{F}_{img}^{\text{back right}} & \mathcal{F}_{img}^{\text{back}} & \mathcal{F}_{img}^{\text{back left}} \end{bmatrix}. \tag{7}$$

Naturally, the decoding target is transformed into the corresponding panoptic image. This operation shifts the focus from ensuring consistency between latent codes for different views to maintaining inherent consistency within a single latent code $z_{pano} \in \mathbb{R}^{C\times 2H_c\times 3W_c}$, resulting in a unified and coherent appearance in multi-camera image generation.

Scene guidance Except for the conditional feature, we also introduce text prompt-based scene guidance into the latent representations. This involves extracting the CLIP feature from a text-based description of an image and mapping this text feature to the intermediate layers of the latent diffusion model, following a similar approach as LDM [20].

Object guidance When it comes to objects that require specific placement within the generated image, we emphasize their pixel-level locations in the latent space using a cross-attention calculation.

Specifically, we employ the voxel-based projection of the initial world volume onto each camera plane, resulting in the creation of preliminary masks m_{class} distinguished by the world volumes' semantic categories. Then, cross-attention is calculated by:

$$\mathbf{z}_{pano} = \text{MHCA}(\mathbf{z}_{pano}(m_{\text{class}} = 1), \text{CLIP}(\text{class})) + \mathbf{z}_{pano},$$

where CLIP(class) represents the category-specific CLIP feature, with semantic categories `class` being defined as bus, car, pedestrian, truck, construction, and vegetation.

3.7. Video Generation

Before we generate video, we first train the single-frame multi-camera generation model, and then we only train the Temporal Attention Block as shown in the right side of Figure 3 and freeze the other block.

Temporal consistency Temporal Attention Block is added to ensure the generation of multi-frame images consistent, which calculates the attention by:

$$\begin{aligned} \mathbf{z}_{pano} &= \text{rearrange}\left(\mathbf{z}_{pano}, (b\ n)\ (h\ w)\ c \rightarrow (b\ h\ w)\ n\ c\right), \\ \mathbf{z}_{pano} &= \text{MHSA}\left(\text{Norm}\left(\mathbf{z}_{pano}\right)\right) + \mathbf{z}_{pano}, \quad \text{(Temporal)} \end{aligned}$$

3.8. Training target

With all the conditions serving as inputs, the training objective can be formulated as:

$$\min_{\theta} \mathcal{L} = \mathbb{E}_{\mathbf{z},\epsilon\sim\mathcal{N}(0,1),\tau}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_\tau, \tau, c)\|_2^2\right], \tag{8}$$

where θ is the learnable parameters in our network, c denotes the conditions.

4. Experiments

4.1. Experimental setup

Dataset We conducted experiments using the NuScenes [1] dataset which comprises 700 training videos and 150 validation videos. Our study utilized images generated by the six camera views trained on the NuScenes dataset.

For ground-truth occupancy data, we utilized CVPR 2023 3D Occupancy Prediction Challenge [23, 24], aligned with semantic labels consistent with the nuScenes-lidarseg

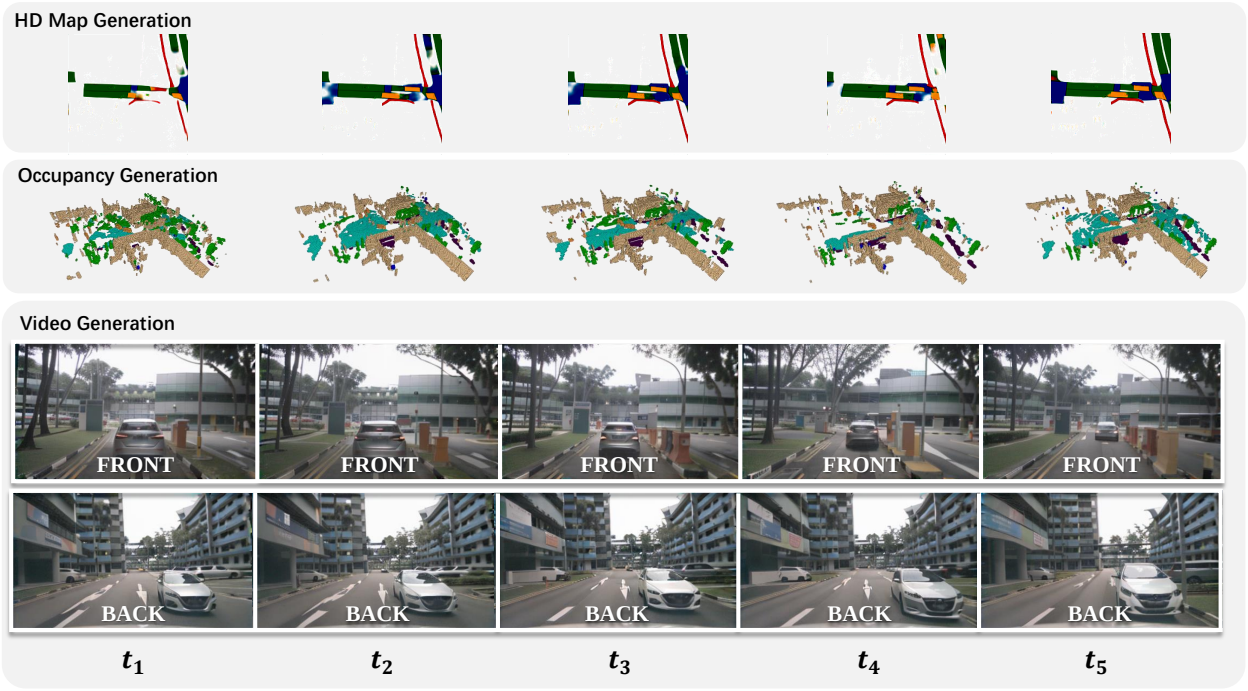

Figure 4. WoVoGen excels in producing future world volumes (top two rows) with temporal consistency. Subsequently, it utilizes the world volume-aware 2D image features derived from the world model's outputs to synthesize a driving video (bottom two rows) with both multi-camera consistency and temporal consistency.

dataset [5]. The ego car-centric local map is derived from the global HD map using the vehicle's ego pose and precisely aligned with the occupancy data.

Data preprocessing Images are resized to a resolution of 256 $\times$ 448, and six consecutive frames are selected for video training. Text prompts are constructed using a structured template: "*Drive in* {*weather description*} *in* {*location*}. *The driving scene is in* {*environment description*}*, captured by multi-view camera.*" The descriptive details enclosed in brackets are filled using scene descriptions from NuScenes.

Training The training process utilized eight NVIDIA A6000 GPUs, encompassing 30,000 iterations for the world volume auto-encoder, 50,000 iterations for volume diffusion, 50,000 iterations for single-frame generation training, and 3,500 iterations for refining video generation.

Inferring We begin with three initial world volumes organized from ground-truth data and proceed to generate the subsequent three world volumes by incorporating actions as input. This iterative process continues, producing a quasi-long world volume video sequence. Ultimately, the world volume video sequence is decoded to form a multi-view camera video.

Evaluation We conducted a comparative analysis of our method's image quality against other approximate works in the following text. To quantitatively assess our work, we employed established metrics such as Fréchet Inception Distance [8] (FID) and Fréchet Video Distance [25] (FVD) for evaluating the quality of the generated video. Additionally, we provided qualitative demonstrations that underscore the advantages of our proposed method.

4.2. Results

4.2.1 World volume generation

The distinctive feature of WoVoGen lies in its ability to predict high-level driving environments in the first stage, including occupancy and HD maps, as vividly illustrated in the left two columns of Figure 1 and the top two rows of Figure 4. Given vehicle actions from the dataset, WoVoGen successfully foresees high-quality future driving environments. Notably, our model retains its effectiveness in predicting single frames while ensuring a high degree of consistency across the generated future frames. This dual capability of WoVoGen — excelling in both short-term prediction and long-term scene evolution — highlights its potential as a powerful tool for advanced driving environment simulation.

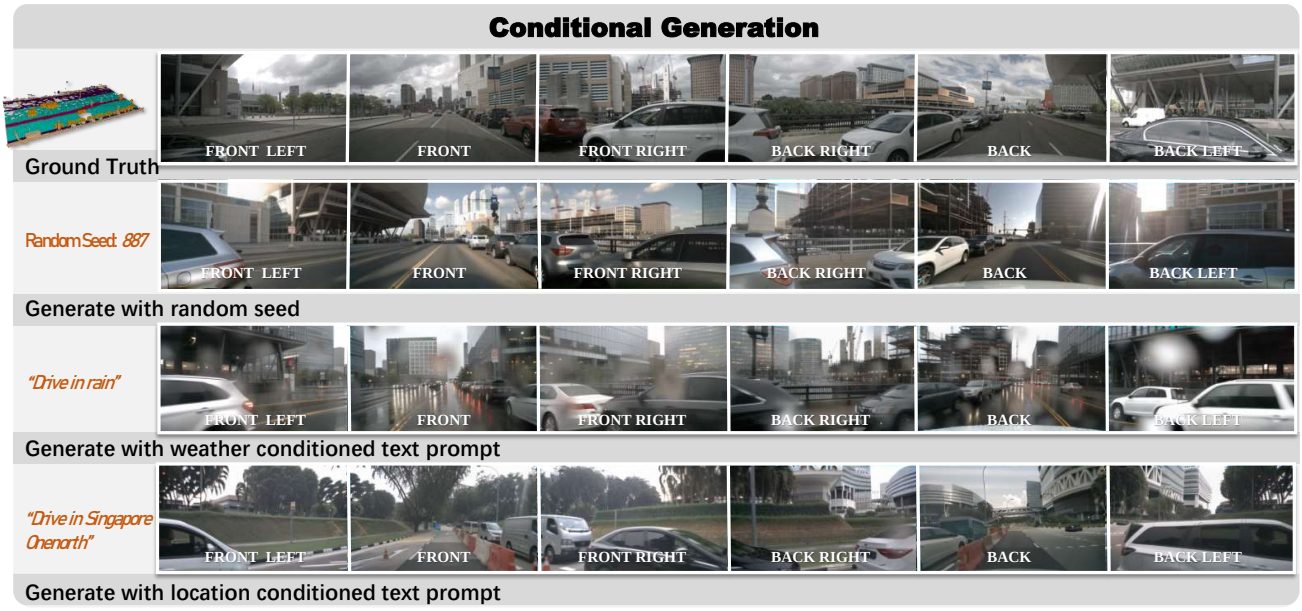

Figure 5. Examples of conditional generation on nuScenes [1] validation dataset. WoVoGen empowers diverse and controllable scene generation. Altering the random seed allows for the generation of various scenarios. Additionally, adjustment to weather (such as sunny, rainy, night, etc.) and location (Singapore, Boston, etc.) within the prompt enables the modification of weather conditions and city styles within the generated scene.

4.2.2 Multi-camera single-frame image generation

Figure 5 shows the qualitative results of the multi-camera single-frame image generation under diverse conditions. We can observe that the generated images have highly consistency, which clearly shows that our model indeed learns to understand the geometric and photometric relationships between different views of the same object or scene. Furthermore, from the first row of Figure 5, we can see that the generated images accurately correspond to the occupancy guide. Beyond that, we can also control the style of the generated images through natural language. In the last two rows of the Figure 5, our WoVoGen demonstrates a profound comprehension of the specified textual conditions, skillfully creating highly realistic simulations of rainy day scenarios. Remarkably, this model also exhibits the capability to interpret and respond to location-specific instructions through natural language processing. When prompted with "*Drive in Singapore Onenorth,*" the model skillfully generates a scene embodying the garden city essence typical of that region. This notable feature not only illustrates the model's advanced understanding of geographical nuances but also highlights its potential to generate a wide array of diverse and contextually accurate data.

The quantitative results in Table 1 underscore that WoVoGen showcases markedly lower FID (27.6) compared to DriveGan [11] and DriveDreamer [28], while these methods are limited to single-view camera image generation. This result solidifies the improved capacity of our approach in generating more realistic autonomous driving images.

4.2.3 Multi-camera single-frame image editing

Given the world volume-aware character of WoVoGen, we can realize editing based on the edited world volume.
Rearranging objects: We can add, delete, or transform the objects within the occupancy. As shown in Figure 6, we can add trees, vehicles, or rearrange the location of objects.
Camera extrinsic editing: More interestingly, we can rotate the camera stereo when projecting the world volume to the camera volume and generate the corresponding scene under a new camera setting.

Method	Multi-view	Multi-frame	FID↓	FVD↓
DriveGan [11]	√		73.4	502.3
DriveDreamer [28]		√	52.6	452.0
Ours(single-frame)	√		27.6	-
Ours(video)	√	√	-	417.7

Table 1. Quantitative comparison of image/video generation quality on Nuscenes validation set. WoVoGen achieve both multi-view and multi-frame generation, demonstrating the lowest FID and FVD scores among all methods.

4.2.4 Multi-camera video generation

Figure 4 showcases the qualitative results of the multi-camera video generation. Thanks to the high-quality world

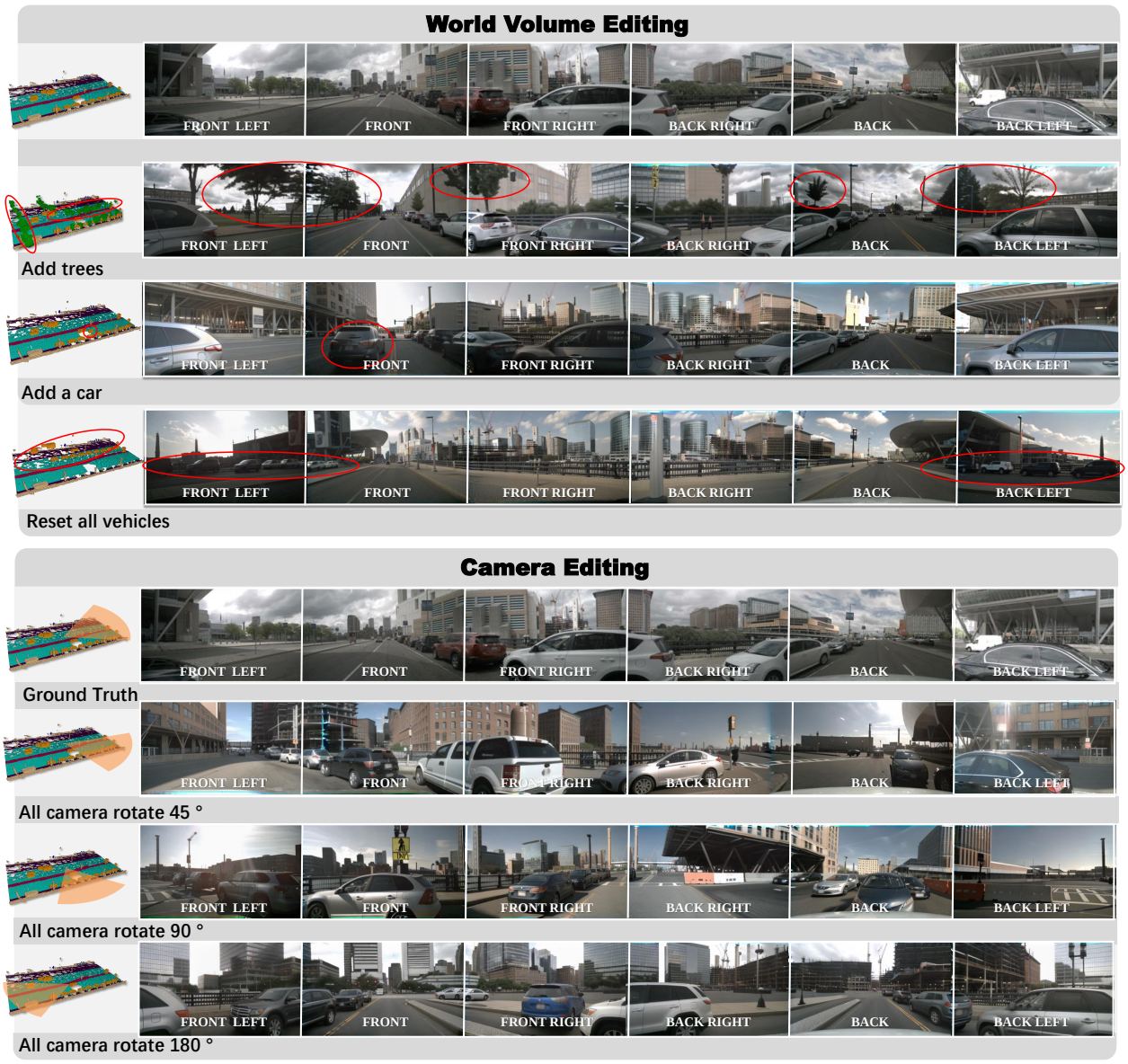

Figure 6. Examples of controlled editing on nuScenes [1] validation dataset. ***Top:*** The ability to selectively add or remove specific objects (such as trees, buildings, cars, etc., highlighted by red circles in the figure) within the world volume empowers the precise and coherent generation of diverse driving scenarios across multiple cameras. ***Bottom:*** Due to WoVoGen's advanced 3D understanding capabilities, the rotation of perspectives across multiple cameras can be achieved by modifying the camera's extrinsic parameters. This functionality enables the generation of driving scenes from any desired angle.

volume generation and finetuning for temporal consistency, the generated video demonstrates a remarkable level of temporal coherence. The appearance of objects and background within the video maintains consistency throughout the sequence. Furthermore, the quantitative results of video quality, indicated by the FVD (417.7) in Table 1, underscores the superiority of our method compared to the other works.

5. Conclusion

In this paper, we propose WoVoGen, which marks a significant advancement in generating multi-camera driving scene videos. Leveraging a novel 4D world volume, it adeptly integrates time with spatial data, addressing the complexity of creating content from multi-sensor data while ensuring consistency. This two-phase system not only produces high-quality videos based on vehicle controls but also enables complex scene editing, showcasing its potential as a comprehensive tool for advancing autonomous driving tech-

nologies.

WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation

Supplementary Material

6. Masks for object guidance

We have devised an efficient approach to compute masks for object guidance.

Specifically, for the original occupancy voxel grids belongs to a certain class $P_{class} \in \mathbb{R}^{N\times 3}$, our initial step involves projecting them onto the camera imaging plane using the camera's intrinsic properties K and extrinsic properties T, performed as follows:

$$P_{proj} = K(RP_{class} + t), \tag{9}$$

where $R = T^{-1}[:3, :3]$ and $t = T^{-1}[:3, 3]$ represent the rotation matrix and translation vector of the inverse of T, respectively. Following this, we initialize an all-zero image m_{class}, matching the dimensions of the camera image. For each normalized projected point P^i_{proj}, we efficiently simulate voxel projections by marking a square-shaped region in m_{class} as occupied:

$$m_{class}[P^{i_x}_{proj} - \delta : P^{i_x}_{proj} + \delta, P^{i_y}_{proj} - \delta : P^{i_y}_{proj} + \delta] = 1, \tag{10}$$

where $P^{i_x}_{proj}$ and $P^{i_y}_{proj}$ represent the normalized x and y components of P^i_{proj}, while δ is inversely proportional to the depth of the projection points $P^{i_z}_{proj}$:

$$\delta = d/P^{i_z}_{proj}, \tag{11}$$

where d is a hyperparameter set to 375. Several generated mask samples are illustrated in Figure 7.

7. Additional results

7.1. Downstream tasks

To enhance performance through more comprehensive data training, we integrate our generation results with the NuScenes [1] training set as a supplementary data source. We test our results on 3D object detection, with BEVDet [**?**] as baseline. We evaluate our outcomes in 3D object detection, using BEVDet as our baseline model. Our approach involves generating a complete training dataset based on the ground truth. Subsequently, we train the 3D object detection model using BEVDet on both the original NuScenes dataset and our newly generated NuScenes dataset. Since our world volume class encompasses only cars, trucks, and buses, our testing focuses exclusively on vehicles within these categories.

As indicated in Table 2, we observed a significant improvement in the mean Average Precision (mAP). Upon analyzing the error contributions, it's evident that the most substantial improvement arises from a reduction in orientation error, which decreased from 17.9 to 15.7. This demonstrates that our generated data indeed enhances the training of downstream tasks, such as 3D object detection.

Method	mAP$_v$ ↑	mATE$_v$ ↓	mASE$_v$ ↓	mAOE$_v$ ↓	mAVE$_v$ ↓	mAAE$_v$ ↓
original	34.9	69.2	20.4	17.9	124.6	28.6
+ generated	**36.2**	**68.6**	**20.1**	**15.7**	**123.4**	**28.1**

Table 2. The enhancement brought about by our generated data in 3D object detection is tested using the BEVDet model [**?**]. The validation is on the vehicle classes of cars, trucks, and buses.

7.2. Ablation studies

Effectiveness of object guidance We implement object guidance to emphasize the location of the generated target. The qualitative comparison in Figure 8 illustrates the ablation study, while the quantitative comparison, using the metric that more accurately depict the alignment among image instances, i.e., CLIP FID [**?**], is presented in Table 3.

Method	FID_{clip} ↓
w/o object guidance	20.3
w/ object guidance	12.8

Table 3. Quantitative comparison of image generation quality on the NuScenes validation set with and without object guidance

Clearly, the integration of object guidance results in a more precise alignment of the object within the generated image with the target image. This refinement corresponds to a notable reduction of 7.5 in the metric, FID_{clip}.

Effectiveness of temporal finetuning To verify the effectiveness of the video consistency module, we illustrate the simulation outcomes of the front and back cameras in Figure 9 after removing the temporal finetuning.

Obviously, the finetuned model notably maintains consistent object and background appearances across frames. In contrast, the single-frame model struggles to achieve this level of consistency.

7.3. Additional single-frame image generation results

We provide additional single-frame image generation samples showcasing more diverse editing and controlling capabilities in Figure 10.

Utilizing the Lego-style edit-friendly features of the world volume offers boundless possibilities for editing and controlling, resulting in numerous potential generation outcomes, some of which may even be rare in the real world.

7.4. Additional video generation results

More generated videos along with the corresponding generated world volume are shown in Figure 11, Figure 12, Figure 13 and Figure 14.

The world model accurately predicts future world evolution and generates a comprehensive world volume. Leveraging world volume-aware diffusion, it produces high-quality videos that maintain cross-frame consistency and coherence across multiple cameras.

References

[1] 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, 5, 7, 8

[2] Spconv Contributors. Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv, 2022. 5

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 4

[5] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. 6

[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. *arXiv preprint*, 2023. 1, 2

[7] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint*, 2023. 1, 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 6

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3, 4

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 5

[11] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, 2021. 7

[12] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *CVPR*, 2023. 1, 2

[13] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint*, 2023.

[14] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 1, 2

[15] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1

[16] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 2

[17] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint*, 2023. 2

[18] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 1

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 1, 5

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4, 5

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

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2023. 2

[23] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint*, 2023. 5

[24] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *ICCV*, 2023. 5

[25] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint*, 2018. 6

[26] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[28] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint*, 2023. 1, 2, 7

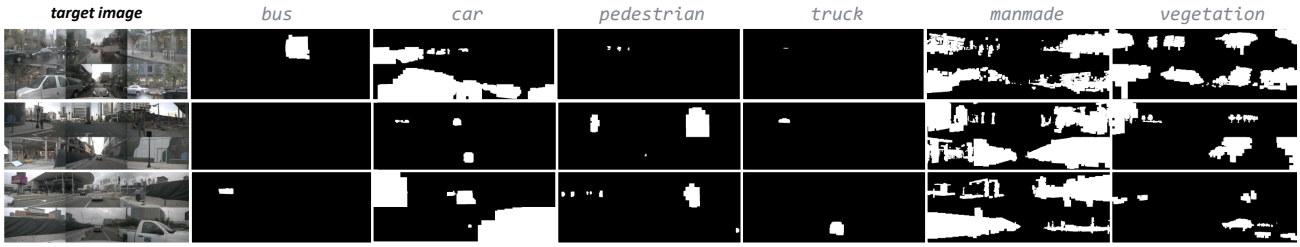

Figure 7. **Object masks calculated by voxel projection.** Origin images and per-class masks are organized as: ***Top:*** front left, front, front right; ***Bottom:*** back right, back, back left.

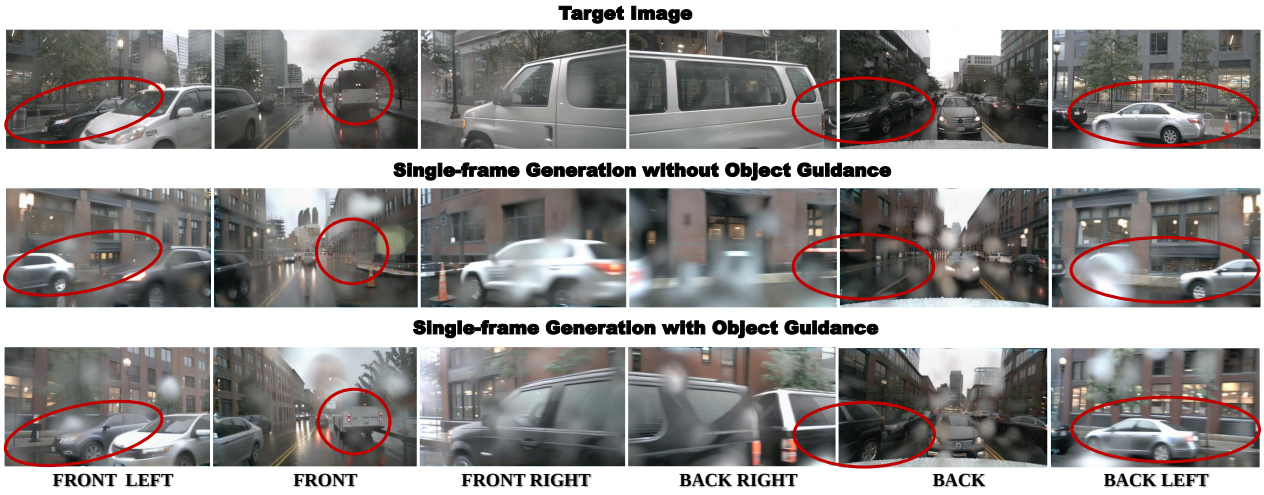

Figure 8. **Qualitative ablation study on object guidance.** The incorporation of object guidance significantly improves the alignment of the generated image's object with the target image.

[29] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. *arXiv preprint*, 2023. 1, 2

[30] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. In *ICLR*, 2023. 1

[31] Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint*, 2023. 1, 2

[32] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *CVPR*, 2023. 1, 2

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 5

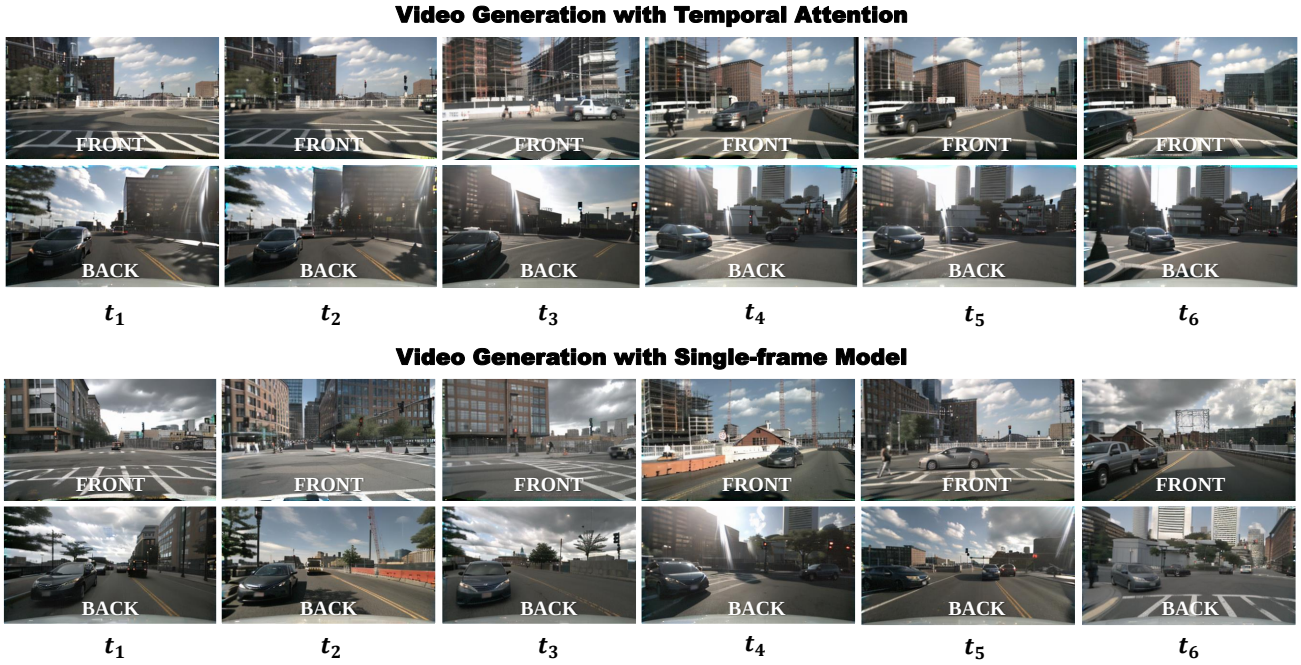

Figure 9. **Qualitative ablation study on temporal finetuning.** Consistency in object and background preservation across frames is evident when conduct temporal finetuning.

Figure 10. **Additional controlling and editing samples.**

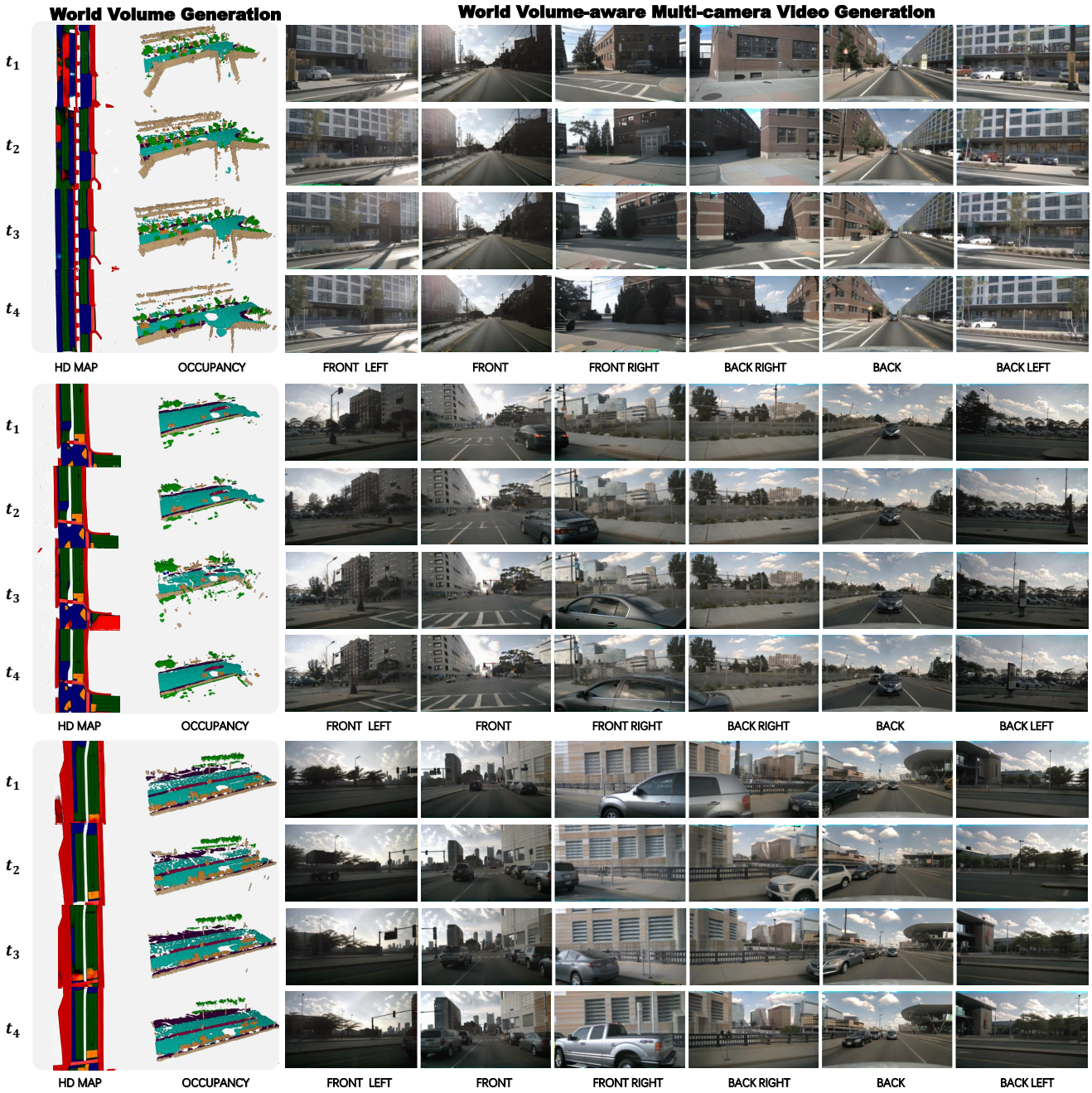

Figure 11. **Additional videos generated from the generated world volumes**

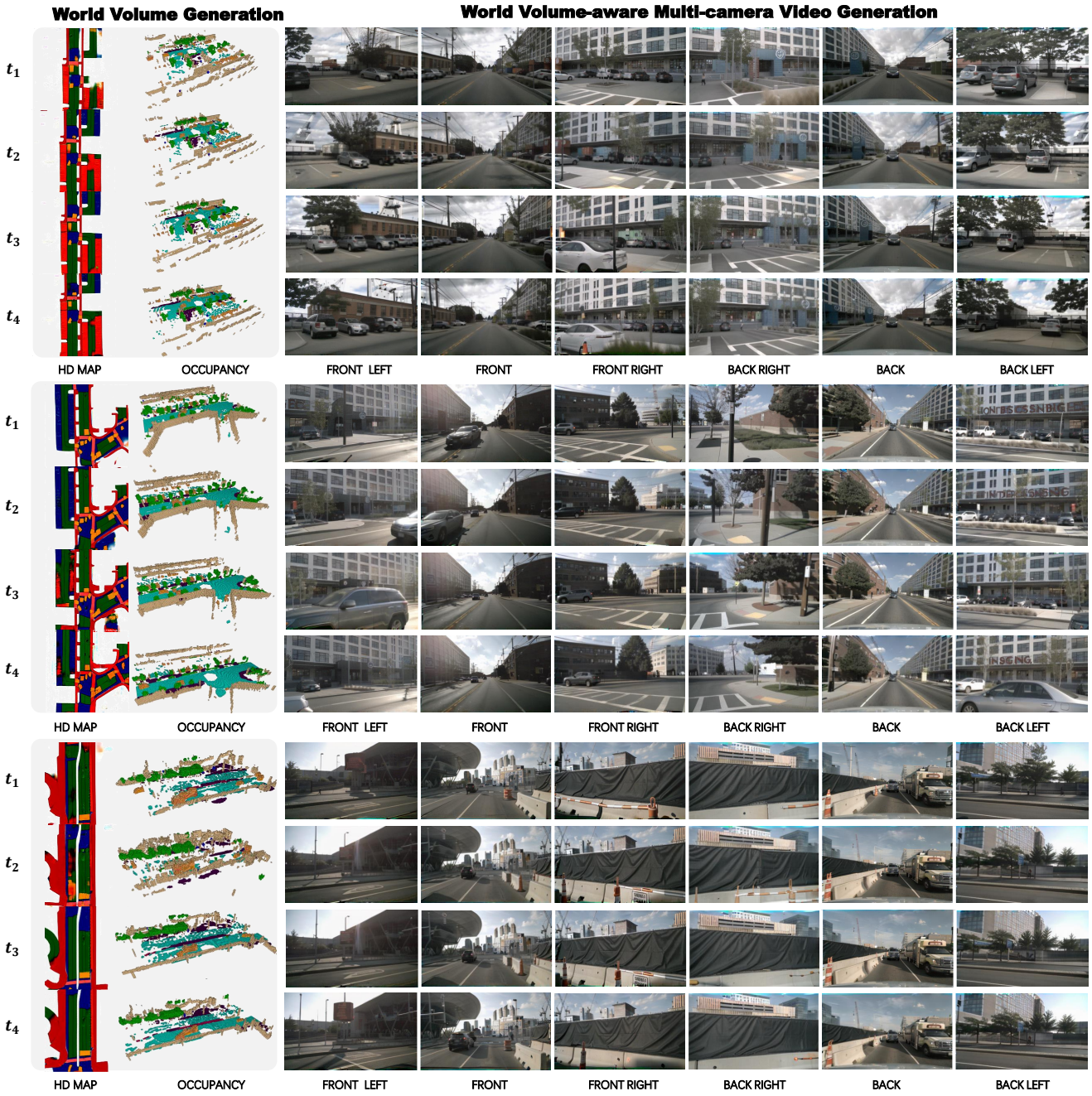

Figure 12. **Additional videos generated from the generated world volumes**

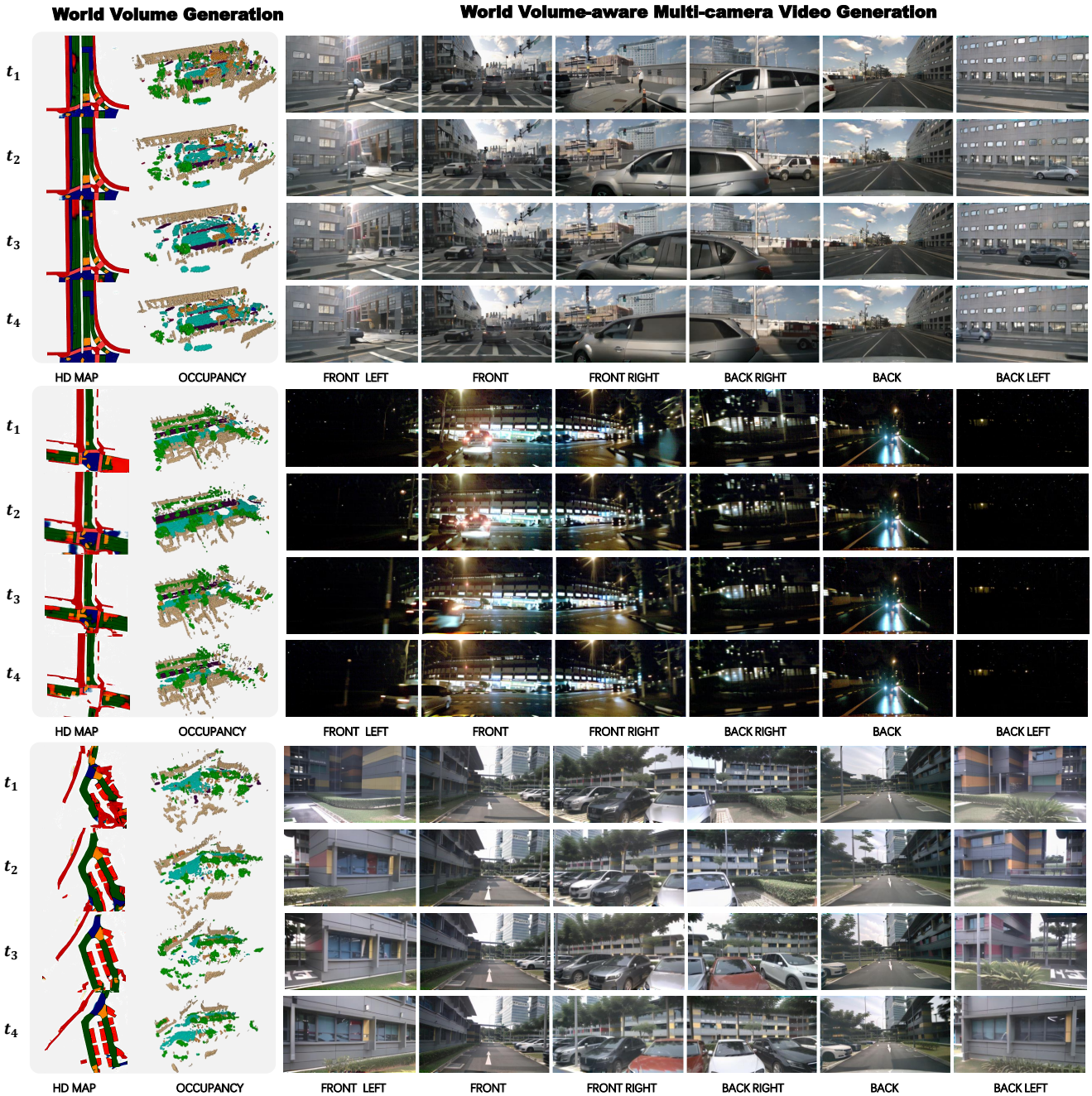

Figure 13. **Additional videos generated from the generated world volumes**

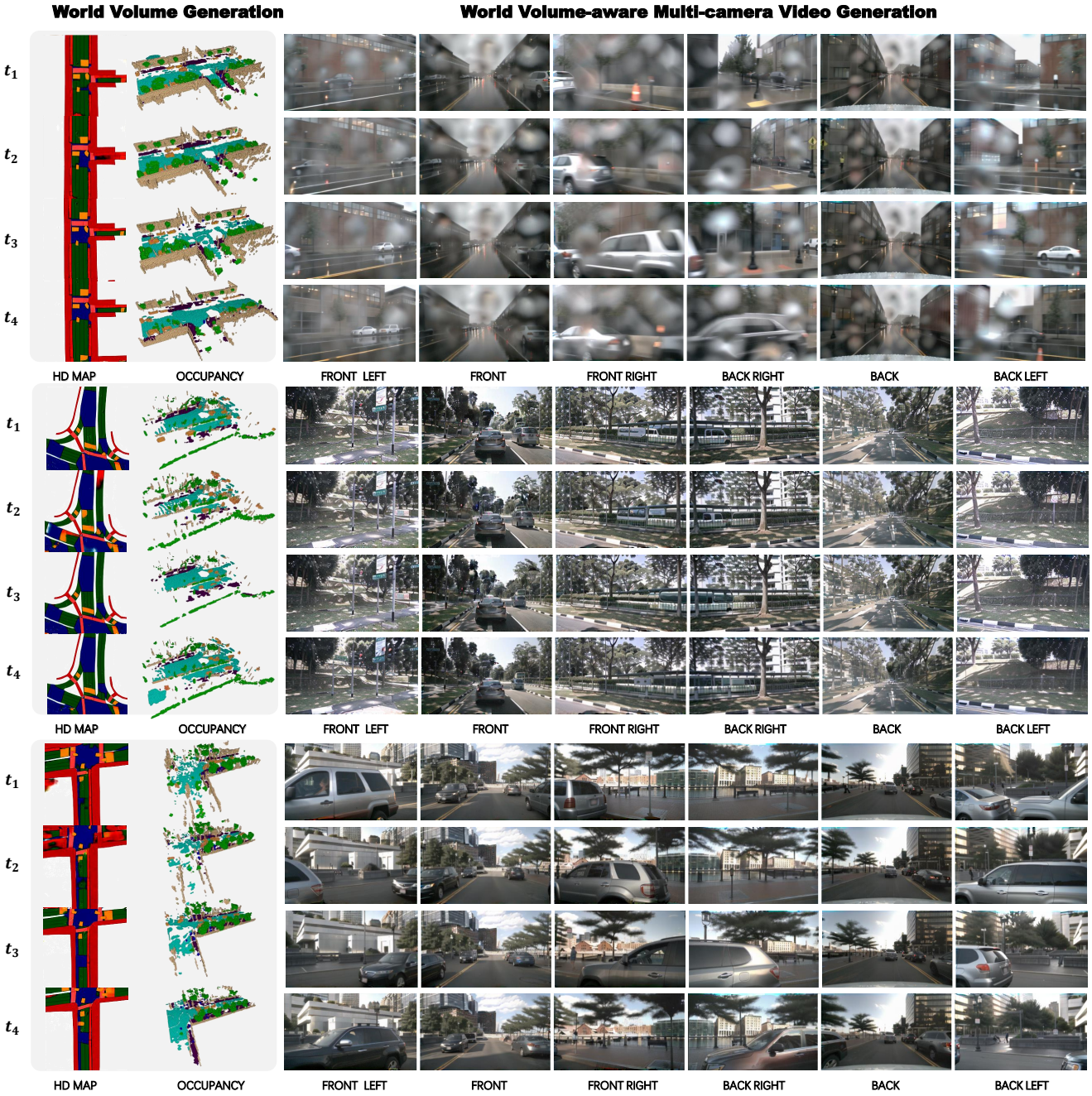

Figure 14. **Additional videos generated from the generated world volumes**