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

Jiachen Lu[1], Ze Huang[1], Zeyu Yang[1], Jiahui Zhang[1], and Li Zhang[1⋆]

School of Data Science, Fudan University

https://github.com/fudan-zvg/WoVoGen

Fig. 1: Our WoVoGen is crafted to generate future world volumes (*i.e.*, HD maps and occupancy) and high-quality multi-camera street-view images, with the input of past world-volumes. The bottom row shows the weather-based control of WoVoGen. Specifically, with the predicted world volume at time t_4 and a weather description, the multi-camera images transit from rainy to sunny conditions, while maintaining the street layout.

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

---

⋆ Li Zhang (lizhangfd@fudan.edu.cn) is the corresponding author.

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.

Keywords: Diffusion · Autonomous driving · World volume

1   Introduction

The burgeoning field of vision-based autonomous driving perception [20, 23, 25] underscores the need for high-quality multi-camera street-view datasets [2]. 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 driving scene content synthesis techniques can be categorized into two main groups: rendering-based [3,10,43,44,46] and diffusion-based [9,18,34,40,45] 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 the noted advantages, diffusion-based approaches [9, 18, 34, 40, 45] that employ bounding boxes or HD maps as control signals offer a sparse representation of the driving scene. Consequently, the generated videos often lack critical information about the scene, leading to an incomplete understanding. This limitation underscores the necessity for a diffusion-based model that incorporates a dense representation of the world as a condition, ensuring a more comprehensive and accurate depiction of the driving environment.

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. 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 [38] 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 [47]. Alongside image features, we employ textual prompts as scene guidance akin to those used in Stable Diffusion [27] 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. **(iii)** Our system not only excels in producing street-view videos that exhibit consistency within the world and across various cameras, driven by vehicle control inputs, but also facilitates scene editing tasks.

2 Related work

Diffusion model for visual content generation Diffusion models, as demonstrated in [5, 12, 27, 32], exhibit the capability to generate a diverse array of images through a learned denoising process. Building upon the foundation of a well-trained image generation model, several works have emerged to augment its controllable generation capabilities, such as text-guided approaches [6,26,28,29], as well as layout-guided approaches [15,19,47,48], etc. Further advancements in image generation extend to video generation, emphasizing temporal consistency in recent works [1,31,42,49]. Additionally, there has been a surge in interest in novel view generation techniques [21,22,24,30], focused on generating diverse
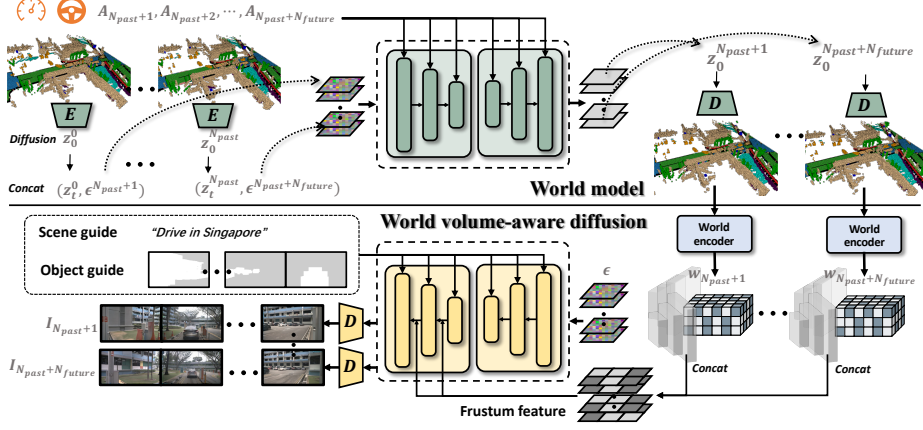
Fig. 2: Overall framework of WoVoGen. ***Top: world model branch.*** We finetune the AutoencoderKL and train the 4D diffusion model from scratch to generate future world volumes based on past world volumes and the actions of the ego car. ***Bottom: world volume-aware synthesis branch.*** Leveraging the 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.

perspectives from existing images. These advancements necessitate models with enhanced capabilities to ensure spatial consistency in generated outputs.

We consolidate various functionalities from preceding diffusion models and introduce WoVoGen, designed specifically for generating driving scene content. WoVoGen encompasses several innovative features: **(i)** Text-guided and world volume-aware controllable image generation, **(ii)** Spatially consistent multi-view image generation and novel view image generation, and **(iii)** Temporally consistent video generation.

Rendering-based driving scene synthesis The rendering-based approaches [10, 43, 44, 46] initially reconstruct comprehensive world structures, followed by image or video rendering based on this reconstructed data. While these methods yield high-quality synthetic results, they often lack robust control capabilities. Efforts like NeuralField-LDM [16] have attempted to introduce diffusion models into volume rendering to address diversity issues and improve control in rendering-based synthesis. However, they still rely heavily on numerous real-world samples from various perspectives or additional supervision during model training, raising concerns about their efficacy.

Diffusion-based driving scene synthesis The diffusion-based approaches [9, 18, 34, 40, 45] leverage pre-trained diffusion models and incorporate sparse conditions, i.e., HD maps and bounding boxes, enabling precise control over driving scene content generation. Part of the works [9, 34, 45] primarily focus on single-frame image generation. As a unified pipeline, they encode ground truth condition information and inject conditional features into the diffusion model to gen-

erate images corresponding to the specified condition. Other works [18, 40] have ventured into driving scene video generation. Building upon trained single-frame generation models, they finetune the temporal module to ensure the temporal coherence of generated elements across consecutive frames in the video.

The proposed WoVoGen excels in generating high-quality driving scene images and videos. Distinguishing itself from other driving scene video generation works [18, 40], WoVoGen possesses unique traits: **(i)** It employs a self-organized dense world volume as a condition, facilitating the natural generation of geometrically consistent images from multiple perspectives. **(ii)** WoVoGen integrates a world model capable of generating future world volumes, eliminating the necessity for ground truth conditions for each video frame generation. **(iii)** With the incorporation of explicit 4D information, *i.e.*, temporal world volumes, WoVoGen offers a comprehensive and accurate representation of the driving environment, resulting in a strong correlation with critical scene details in the generated content.
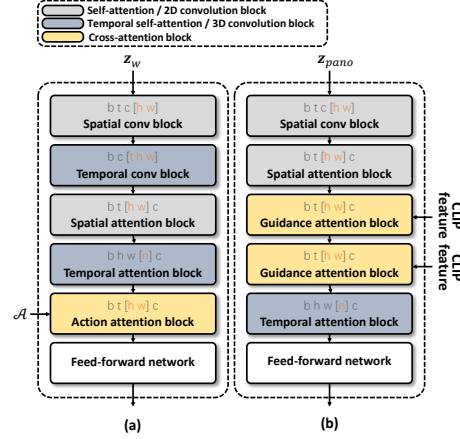

Fig. 3: (a): an action attention block enhances the model by incorporating action information. **(b)**: a guidance attention block integrates the CLIP feature of a specific object into the latent representation.

3 Method

3.1 Preliminary

The latent diffusion model (LDM) [27] 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) [12] 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:

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

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

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:

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

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.xi

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 predefined 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 BEVDet-occ [14] 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 [7], aimed at minimizing the variance within the latent space.

4D 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 [12]. 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 [39], 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.

As shown in Figure 3, we build upon the basic spatial Transformer block of Stable Diffusion [27], 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:

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

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 $\mathcal{W} = \{\mathcal{W}_t\}_{t=1,2,\ldots}$ 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 [25]:

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

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) [4] 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 $\mathcal{F}_{cam} \in \mathbb{R}^{B \times N_{\text{future}} \times C \times D_c \times H_c \times W_c}$ represents the camera frustum. We then apply a squeeze-and-excitation operation [13] on the depth channel and sum along the depth dimension to obtain the world volume-aware 2D image feature:

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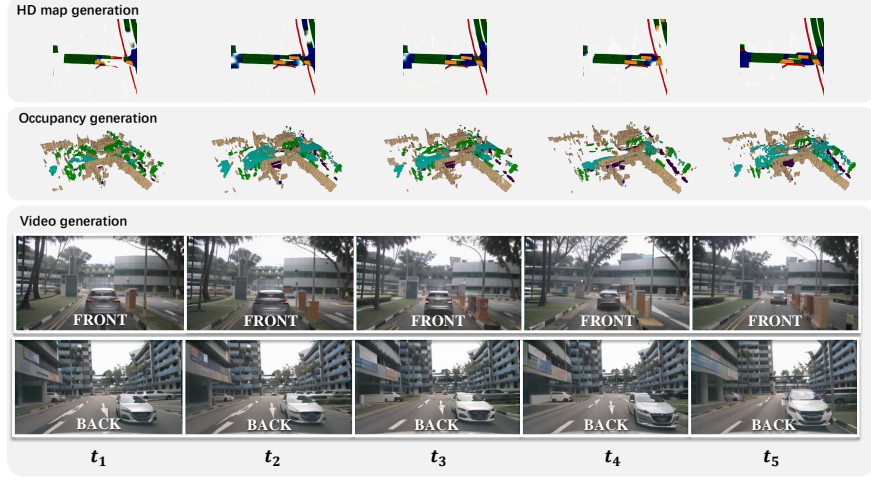$$

Fig. 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.

3.6 World volume-aware diffusion generation

Based on the ControlNet framework [47], 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 [27].

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 add temporal attention block and only train the temporal attention block as shown in the right side of Figure 3 and freeze the other blocks.
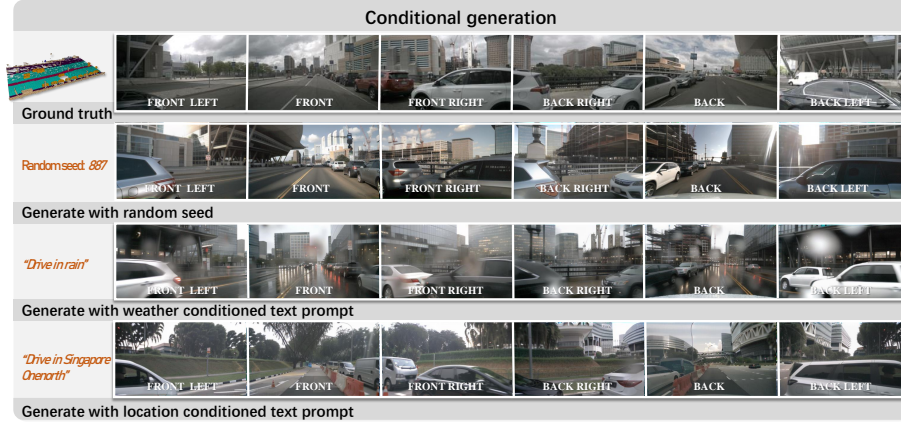

Fig. 5: Examples of conditional generation on nuScenes [2] validation dataset. WoVo-Gen empowers diverse and controllable scene generation. Altering the random seed allows for the generation of various scenarios. Additionally, adjustment to weather (such as sunny, rainy, night, etc.) and location (Singapore, Boston, etc.) within the prompt enables the modification of weather conditions and city styles within the generated scene.

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

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

Fig. 6: Examples of controlled editing on nuScenes [2] 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.

4   Experiments

4.1   Experimental setup

Dataset We conducted experiments using the nuScenes [2] dataset which comprises 700 training videos and 150 validation videos. Our study utilized images from the six camera views.

For ground truth occupancy, we use the data from CVPR 2023 3D Occupancy Prediction Challenge [35,36], aligned with semantic labels consistent with the nuScenes-lidarseg dataset [8]. 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 256×448, 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 [11] (FID) and Fréchet Video Distance [37] (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

4D 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.

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. 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. When prompted with "*Drive in Singapore Onenorth,*" the model skillfully generates a scene embodying the garden city essence typical of that region.

The quantitative results in Table 1 underscore that WoVoGen showcases markedly lower FID (27.6) compared to DriveGan [17] and DriveDreamer [40], 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.

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.

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.

Method	Multi-view	Multi-frame	FID↓	FVD↓
DriveGan [17]	√		73.4	502.3
DriveDreamer [40]		√	52.6	452.0
Ours(single-frame)	√		27.6	-
Ours(video)	√	√	-	417.7

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

4.3   Ablation studies

Effectiveness of world volume and object guidance In this study, we examine the significance of world volume and object guidance, as detailed in Table 2. We analyze three scenarios: object guidance alone, world volume alone, and the combination of world volume with object guidance. When solely relying on object guidance, interpreted as object mask information, the generated results are notably inferior. In contrast, incorporating world volume, regardless of the presence of object guidance, significantly enhances the generation outcomes. This underscores the critical importance of world volume in the generative process. Additionally, the incorporation of object guidance leads to more precise object localization, especially when evaluating the fidelity of the generated images, as indicated by the highest FID scores.

Consistency To assess the quantitative outcomes of multi-camera and temporal consistency, we adopt a quantitative methodology utilizing Key Points Matching (KPM) [41]. This approach involves a pre-trained matching model [33] to calculate the average number of matching keypoints. Table 2 reveals that there is an impressive match rate of up to 80.2% for keypoints across successive frames, underscoring a significant accomplishment in maintaining temporal consistency. Regarding multi-view consistency, we observe that 90.0% of keypoints match between adjacent cameras, further highlighting our methodology's efficacy in ensuring robust multi-camera consistency.

Table 2: ***Left:*** Quantitative comparison of image generation quality on the nuScenes validation set with and without object guidance. ***Right:*** Quantitative results of multi-view consistency and tmerpoal consistency. KPM is used for evaluation

World Volume	Object Guidance	**FID**↓		Consistency	**KPM**(%)↑
×	✓	28.6		Multi-view	90.0
✓	×	27.8		Temporal	80.2
✓	✓	**27.6**			

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 technologies.

Acknowledgments

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint (2023) 3
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020) 2, 10, 11, 12
3. Chen, Y., Gu, C., Jiang, J., Zhu, X., Zhang, L.: Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. arXiv preprint (2023) 2
4. Contributors, S.: Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv (2022) 8
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021) 3
6. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers (2021) 3
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021) 7
8. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking (2022) 12
9. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint (2023) 2, 4
10. Guo, J., Deng, N., Li, X., Bai, Y., Shi, B., Wang, C., Ding, C., Wang, D., Li, Y.: Streetsurf: Extending multi-view implicit surface reconstruction to street views. arXiv preprint (2023) 2, 4
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2017) 12
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020) 3, 5, 7
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018) 8
14. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint (2021) 7
15. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint (2023) 3
16. Kim, S.W., Brown, B., Yin, K., Kreis, K., Schwarz, K., Li, D., Rombach, R., Torralba, A., Fidler, S.: Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In: CVPR (2023) 4
17. Kim, S.W., Philion, J., Torralba, A., Fidler, S.: Drivegan: Towards a controllable high-quality neural simulation. In: CVPR (2021) 13

18. Li, X., Zhang, Y., Ye, X.: Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. arXiv preprint (2023) 2, 4, 5

19. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023) 3

20. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022) 2

21. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (2023) 3

22. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint (2023) 3

23. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: ECCV (2020) 2

24. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint (2023) 3

25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 2, 8

26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint (2022) 3

27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 3, 5, 7, 9

28. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023) 3

29. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding (2022) 3

30. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint (2023) 3

31. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint (2022) 3

32. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2023) 3

33. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021) 14

34. Swerdlow, A., Xu, R., Zhou, B.: Street-view image generation from a bird's-eye view layout. IEEE Robotics and Automation Letters (2024) 2, 4

35. Tian, X., Jiang, T., Yun, L., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. arXiv preprint (2023) 12

36. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: ICCV (2023) 12

37. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint (2018) 12

38. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning (2017) 3
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need (2017) 7
40. Wang, X., Zhu, Z., Huang, G., Chen, X., Lu, J.: Drivedreamer: Towards real-world-driven world models for autonomous driving. arXiv preprint (2023) 2, 4, 5, 13
41. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. arXiv preprint (2023) 14
42. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV (2023) 3
43. Wu, Z., Liu, T., Luo, L., Zhong, Z., Chen, J., Xiao, H., Hou, C., Lou, H., Chen, Y., Yang, R., et al.: Mars: An instance-aware, modular and realistic simulator for autonomous driving. arXiv preprint (2023) 2, 4
44. Xie, Z., Zhang, J., Li, W., Zhang, F., Zhang, L.: S-nerf: Neural radiance fields for street views. In: ICLR (2023) 2, 4
45. Yang, K., Ma, E., Peng, J., Guo, Q., Lin, D., Yu, K.: Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. arXiv preprint (2023) 2, 4
46. Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R.: Unisim: A neural closed-loop sensor simulator. In: CVPR (2023) 2, 4
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 3, 9
48. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: CVPR (2023) 3
49. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint (2022) 3