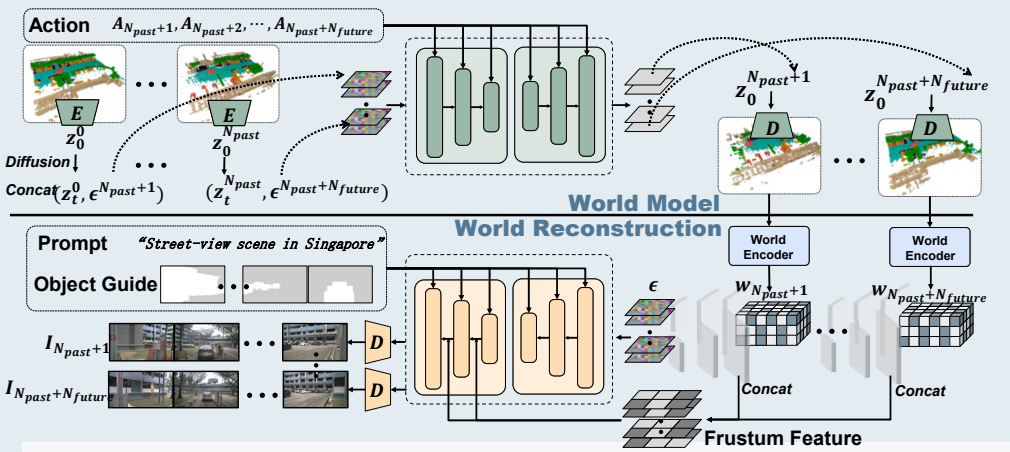
WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation

Jiachen Lu[1], Ze Huang[1], Zeyu Yang[1], Jiahui Zhang[1], Li Zhang[1]
[1]Fudan University

Multi-camera driving scene generation

Our model operates in two distinct phases:

1. Envisioning the future 4D temporal world volume based on vehicle control sequences
2. Generating multi-camera videos, informed by this envisioned 4Dtemporal world volume and sensor interconnectivity



Bottom: word volume-aware synthesis branch. Subsequent sampling yields $\mathcal{F}_{img}$, which are then aggregated. The process is finalized by applying panoptic diffusion to produce future videos.

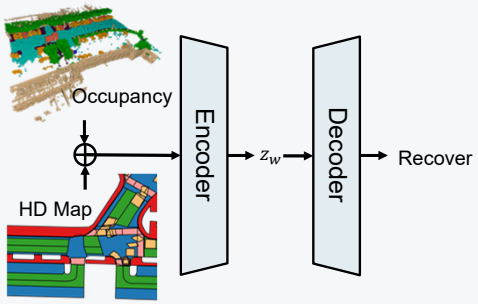
Overall framework

WoVoGen comprises 2 branches: **the world model branch** and **the world volume-aware generation branch**. The world model branch is responsible for generating future world volumes, while the world volume-aware generation branch focuses on the generation of multi-camera video.

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. $\mathcal{F}_w$ are derived through the world encoder.

World model 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.



$$z_w = \text{rearrange}(z_w, \langle b\ n\rangle\ h\ w\ c \rightarrow \langle b\ n\rangle\ \langle h\ w\rangle\ c)$$
$$z_w = \text{MHSA}(\text{Norm}(z_w)) + z_w$$
$$z_w = \text{rearrange}(z_w, \langle b\ n\rangle\ \langle h\ w\rangle\ c \rightarrow \langle b\ h\ w\rangle\ n\ c)$$
$$z_w = \text{MHSA}(\text{Norm}(z_w)) + z_w$$
$$z_w = \text{rearrange}(z_w, \langle b\ h\ w\rangle\ n\ c \rightarrow \langle b\ n\rangle\ \langle h\ w\rangle\ c)$$
$$z_w = \text{MHCA}(\text{Norm}(z_w, A)) + z_w$$
$$z_w = \text{FFN}(\text{Norm}(z_w)) + z_w$$

World volume-aware 2D feature

1. **World volume encoding:** we employ a featurization process utilizing CLIP.
$$\mathcal{F}_w = \text{SPConv}(\text{PCA}(\text{CLIP}(\mathcal{W})))$$
2. **Camera volume sampling:** to incorporate the world volume into image generation, we sample from it using dense rays emitted from the camera.
$$\mathcal{F}_{cam} = \text{interpolate}(p_c, \mathcal{F}_w)$$
3. **Squeeze-and-excitation operation:** We apply a squeeze-and-excitation operation on the depth channel and sum along the depth to obtain the world volume-aware 2D image feature:
$$\mathcal{F}_{img} = \sum_{i=1}^{D_c} SE(\mathcal{F}_{cam})[:,:,:,i,:,:]$$

World volume-aware diffusion generation

- **Panoptic diffusion:** we aggregate the world volume-aware 2D image feature from different view into a single panoptic feature
$$\mathcal{F}_{pano} = \begin{bmatrix} \mathcal{F}_{img}^{front\ left} & \mathcal{F}_{img}^{front} & \mathcal{F}_{img}^{front\ right} \\ \mathcal{F}_{img}^{back\ right} & \mathcal{F}_{img}^{back} & \mathcal{F}_{img}^{back\ left} \end{bmatrix}$$
- **Scene guidance:** text prompt-based scene guidance
- **Object guidance:**
$$z_{pano} = \text{MHCA}(z_{pano}(m_{\text{class}} = 1), CLIP(\text{class})) + z_{pano}$$

Multi-camera image editing