

UniMLVG: Unified Framework for Multi-view Long Video Generation with Comprehensive Control Capabilities for Autonomous Driving

Supplementary Material

6. Details on Explicit Perspective Modeling

Explicit perspective modeling aims to inject spatial information into UniMLVG to enhance the coherence of generated videos. Specifically, we utilize the camera’s intrinsic parameters $K \in \mathbb{R}^{B \times T \times V \times 3 \times 3}$ and the extrinsic transformations $E \in \mathbb{R}^{B \times T \times V \times 3 \times 4}$ to obtain ray maps that match the size of the images. Importantly, we establish a unified coordinate system with the optical center of the forward-facing camera in the first frame as the origin. In this way, we can obtain the camera’s origin coordinates $Ray_o = E_4$, where $E_4 \in \mathbb{R}^{B \times T \times V \times 3}$ represent the latest columns of the camera extrinsic matrices. We extend the camera origins to $\mathbb{R}^{B \times T \times V \times 3 \times H \times W}$ to match the number of pixels. We then define a three-dimensional pixel index plane (in homogeneous coordinates) $P \in \mathbb{R}^{B \times T \times V \times 3 \times H \times W}$ with the same dimensions as the image latent space. Using the scaled camera intrinsic parameters and transformations, the ray directions from the origin to the plane can be computed as follows: $Ray_d = E_{3,:3} \times K^{-1} \times P$, $E_{3,:3}$ refers to the upper-left 3×3 rotation matrix of the extrinsic matrix E . After transforming the encoded Ray_o and Ray_d through an MLP, the resulting features are added to the image latent before feeding it into the cross-view and temporal modules.

7. Details on fusion of Cross-view and Temporal Information

The fusion of cross-view and temporal information is essential to this task. We believe that the original text-to-image generation model [10] already possesses strong image generation capabilities, requiring only minor modifications to the image latent to achieve cross-view consistency and temporal coherence. Therefore, we simply use a learnable parameter to perform a weighted summation of the outputs from the cross-view or temporal models with the backbone’s image latent. Specifically, both the cross-view and temporal modules are GPT-2-style self-attention mechanisms [36]. The fusion process can be formulated as:

$$z_l = \text{Sigmoid}(\alpha) \cdot z'_l + (1 - \text{Sigmoid}(\alpha)) \cdot \mathcal{F}_l(z'_l), \quad (3)$$

where α is a learnable parameter, initially set to 2 to facilitate gradual model training, z'_l denotes the output image latent from the backbone at the l -th layer, and \mathcal{F} represents either the cross-view module or the temporal module. In addition, we found that it is not necessary to add these two modules after every layer of the backbone; adding them at intervals does not affect performance.

8. Details on Image-level Description

In previous works [12, 23, 52, 64], scene descriptions were exclusively used as textual conditions for multi-view video generation. However, such descriptions often lack the fine-grained details necessary for high-quality and consistent generation. To address this limitation, we incorporate image-level descriptions, enabling more precise control over the generation process and improving the consistency and realism of multi-view videos. Specifically, we leverage the multimodal model Drivemlm [50]. For each view, we input the image along with two question prompts: “Describe the time, weather, environment, objects, and each value should be a single string and less than 20 words.” and “Describe objects in this image within about 30 words.” to generate detailed annotations of the time, weather, environment, and objects present in the viewpoint. Figure 8 presents the statistical information on time, weather, and environment annotations across the four datasets. We can observe that daytime scenes dominate across all four datasets, accounting for more than 80% of the data. In terms of weather, the two most frequently occurring descriptors are *overcast* and *clear sky*. It is worth noting that snowy scenes in Argoverse2 constitute less than 2.6%. However, UniMLVG can still modify weather text conditions to transform scenes under other weather conditions into snowy scenes, demonstrating its robust generalization and semantic understanding capabilities. In terms of environmental descriptions, we can observe the primary focus of data collection for each dataset. For instance, nuScenes, Waymo, and Argoverse2 primarily capture urban street environments, whereas OpenDV-Youtube exhibits a broader range of scenes, including highways and deserts.

9. More Qualitative Results

We provide additional examples to further demonstrate the capabilities of UniMLVG in generating long-duration, multi-view consistent videos. Figures 9, 10 and 11 showcase multi-view long video samples under various weather conditions and times, utilizing reference frames. Conversely, Figures 12, 13 and 14 illustrate multi-view long video samples under similar conditions but without reference frames. Additionally, Figures 15, 16 and 17 demonstrate the transformation of scenes with different times and weather conditions into snowy scenes through text editing. Finally, Figures 18, 19, and 20 showcase the model’s generalization capability, enabling the generation of realistic scenes based on conditions from virtual simulations.

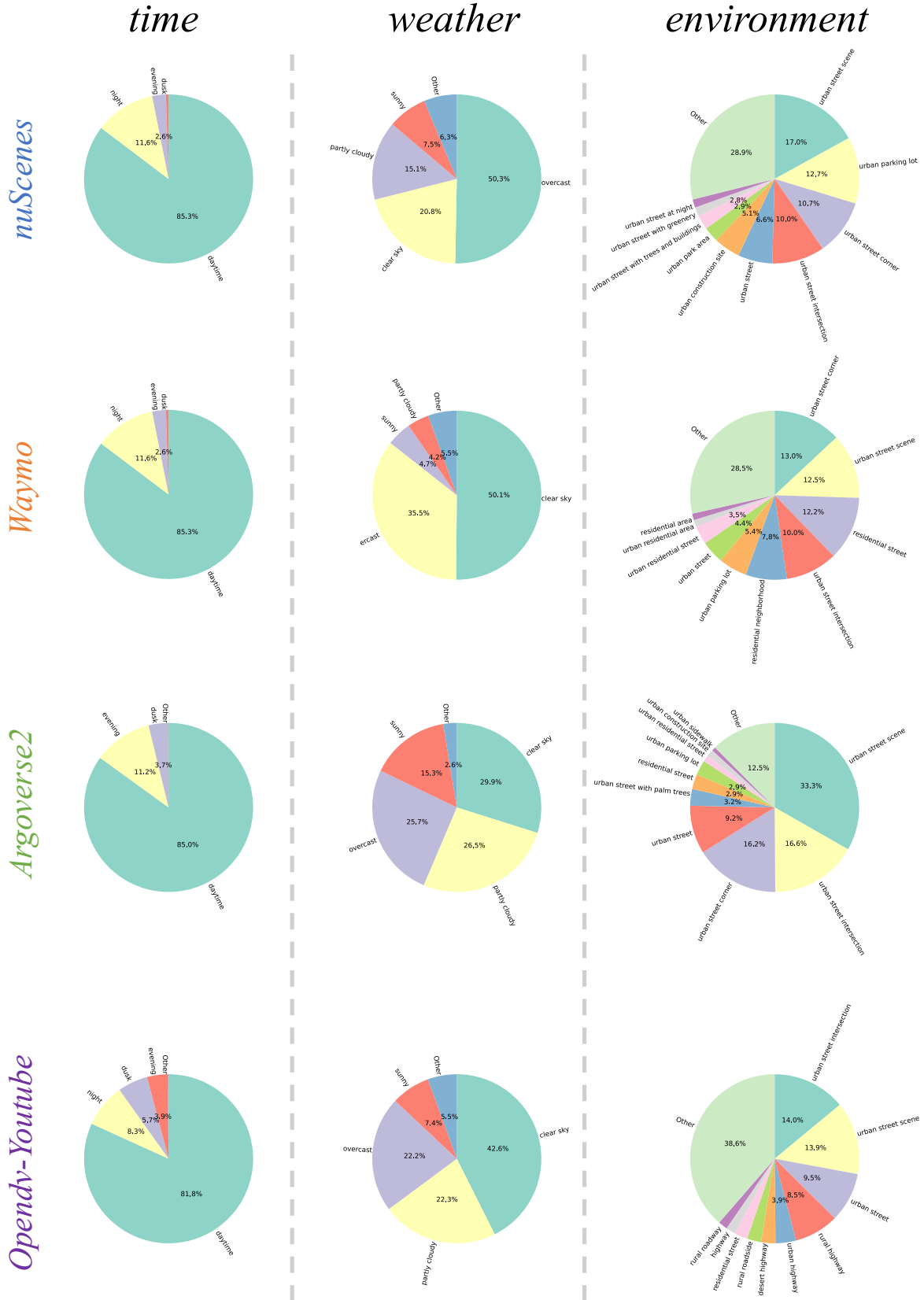


Figure 8. Statistical Analysis of time, weather, and environment in text descriptions on four datasets.

Ground Truth



$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Generation



$t=0s$



$t=2s$



$t=5s$



$t=10s$



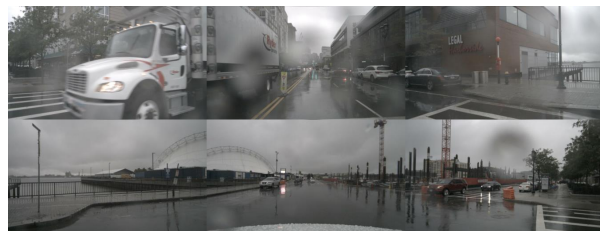
$t=15s$



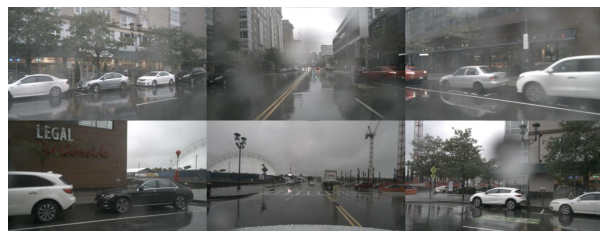
$t=20s$

Figure 9. Sample of 20s multi-view video in a sunny scene with reference frames.

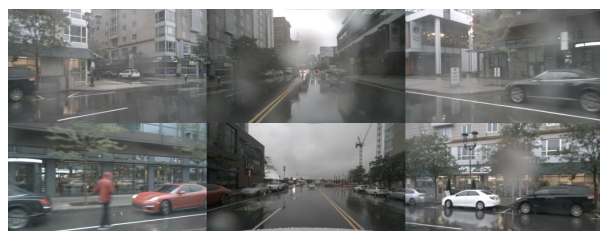
Ground Truth



$t=0s$



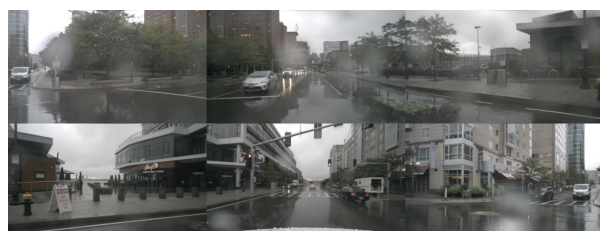
$t=2s$



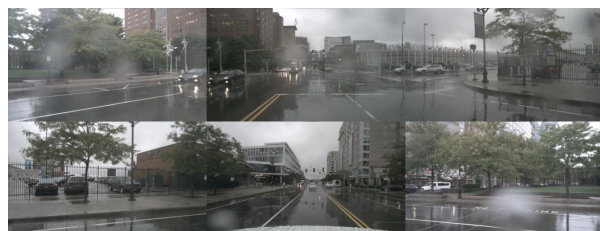
$t=5s$



$t=10s$

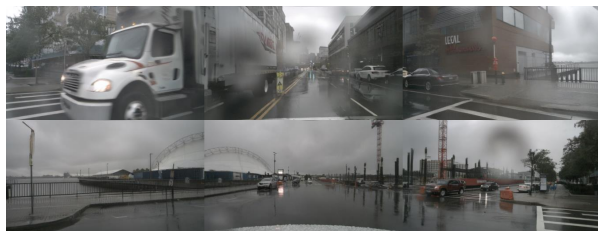


$t=15s$

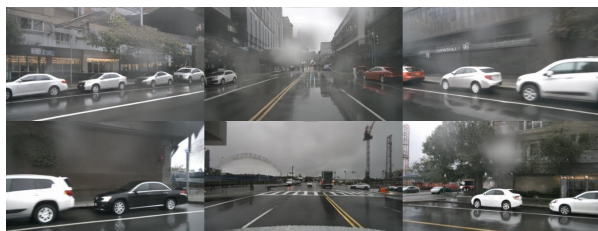


$t=20s$

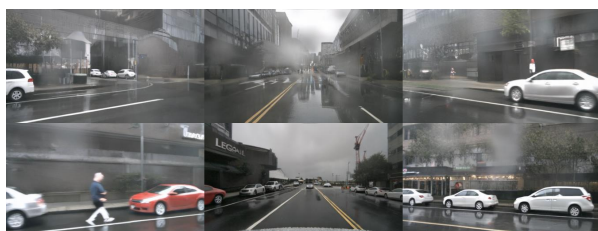
Generation



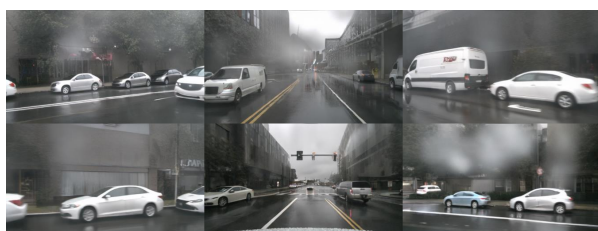
$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Figure 10. Sample of 20s multi-view video in a rainy scene with reference frames.

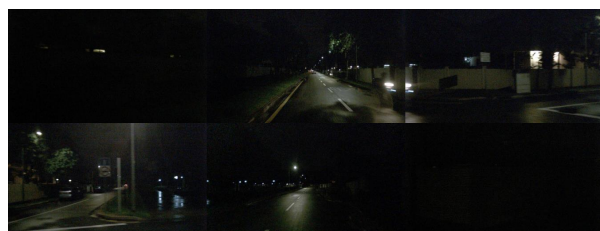
Ground Truth



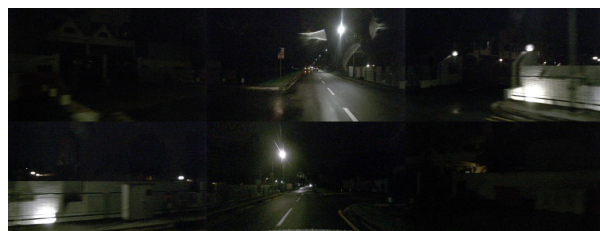
$t=0s$



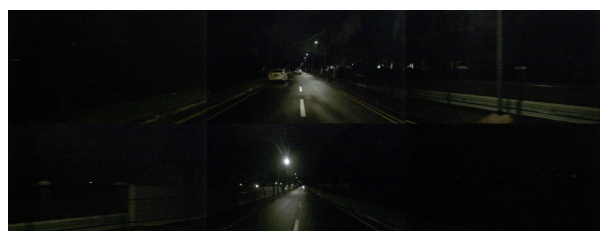
$t=2s$



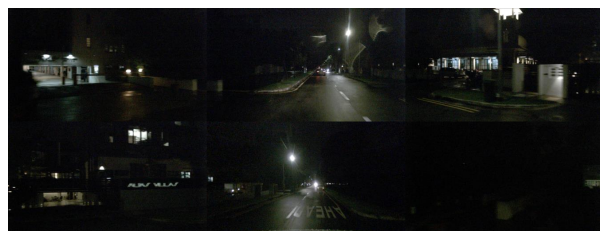
$t=5s$



$t=10s$



$t=15s$



$t=20s$

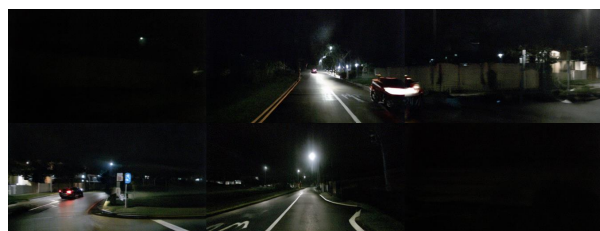
Generation



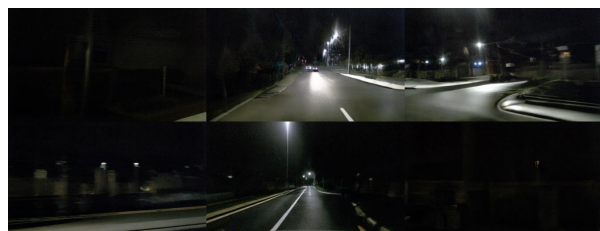
$t=0s$



$t=2s$



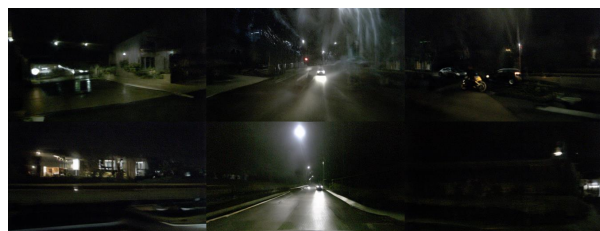
$t=5s$



$t=10s$



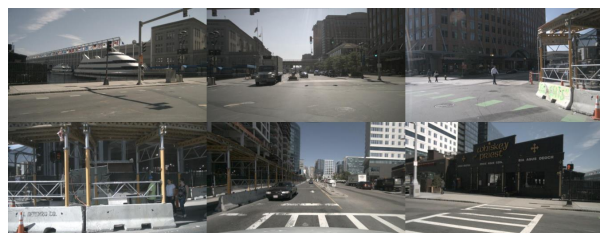
$t=15s$



$t=20s$

Figure 11. Sample of 20s multi-view video at night with reference frames.

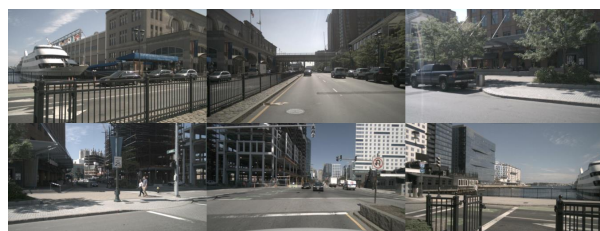
Ground Truth



$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Generation



$t=0s$



$t=2s$



$t=5s$



$t=10s$



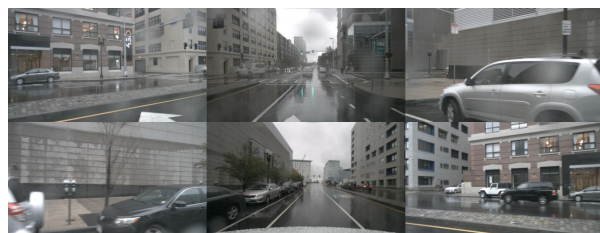
$t=15s$



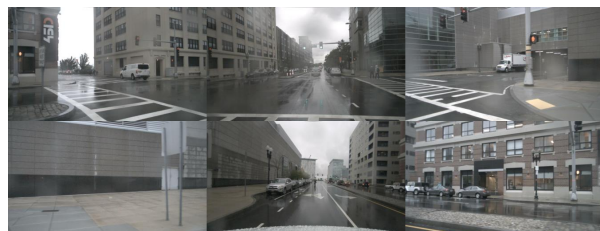
$t=20s$

Figure 12. Sample of 20s multi-view video in a sunny scene without reference frames.

Ground Truth



$t=0s$



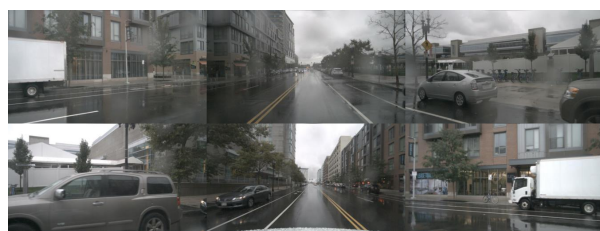
$t=2s$



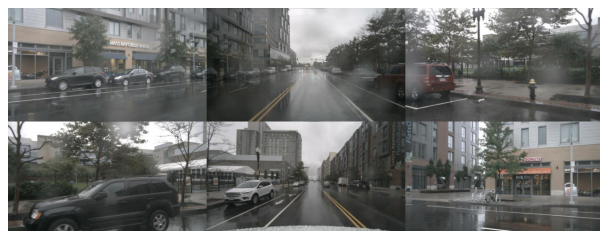
$t=5s$



$t=10s$

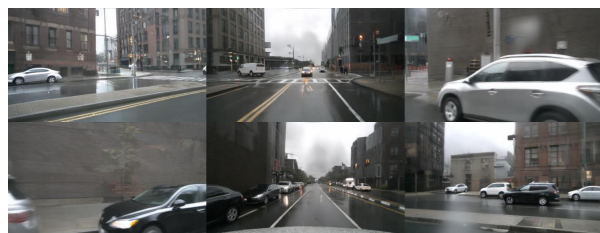


$t=15s$

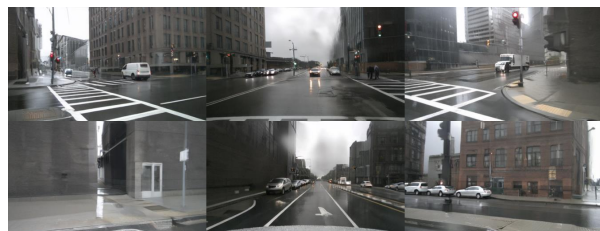


$t=20s$

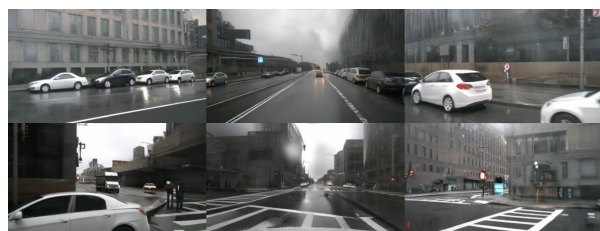
Generation



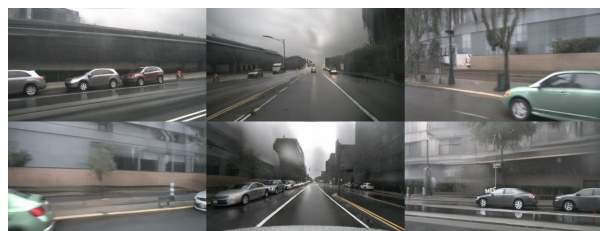
$t=0s$



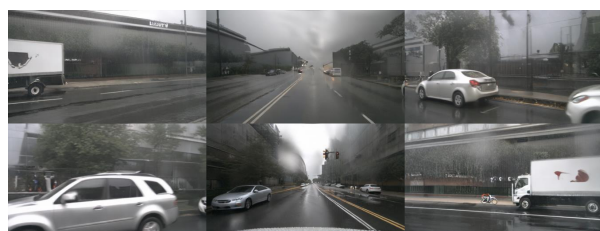
$t=2s$



$t=5s$



$t=10s$



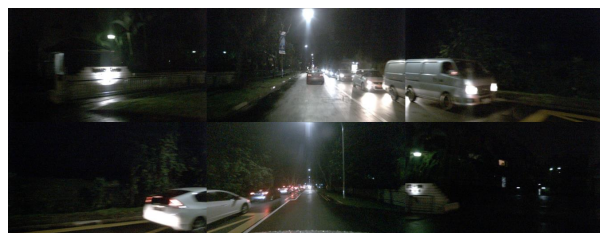
$t=15s$



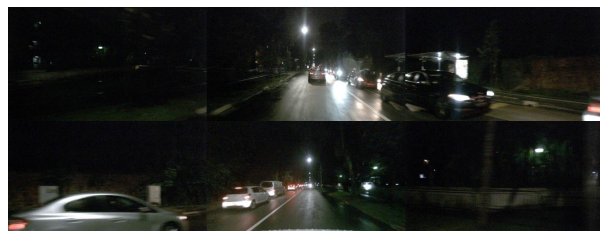
$t=20s$

Figure 13. Sample of 20s multi-view video in a rainy scene without reference frames.

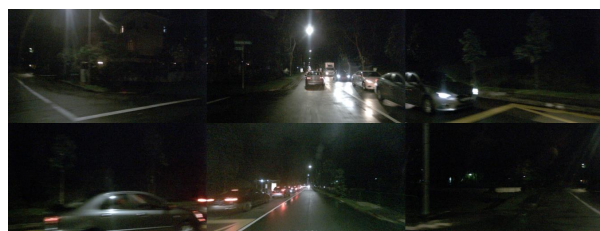
Ground Truth



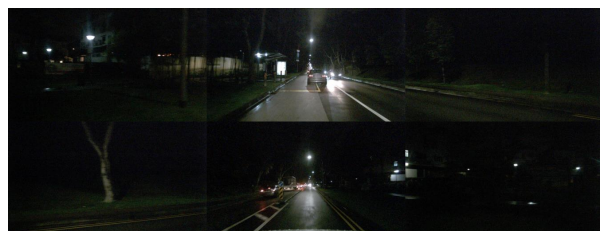
$t=0s$



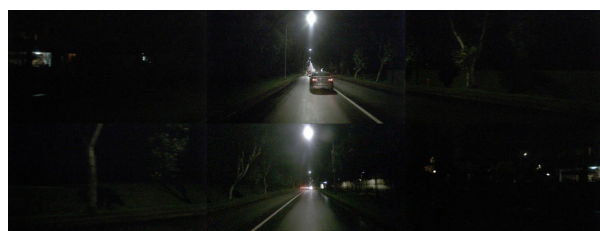
$t=2s$



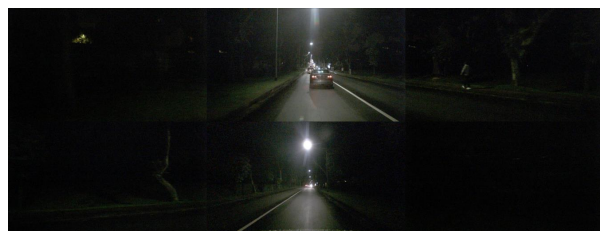
$t=5s$



$t=10s$



$t=15s$

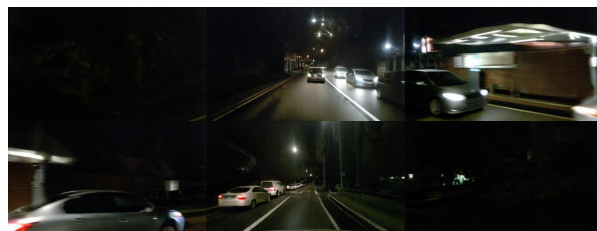


$t=20s$

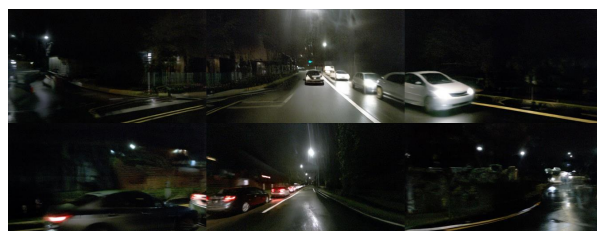
Generation



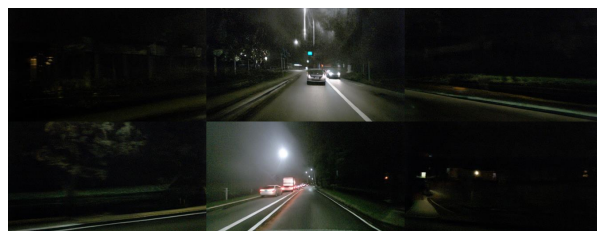
$t=0s$



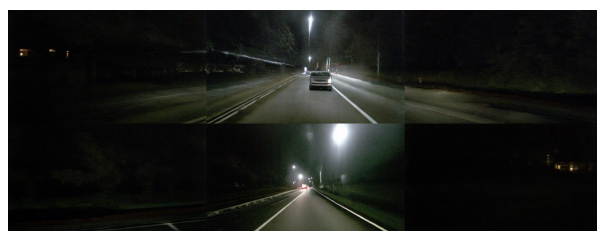
$t=2s$



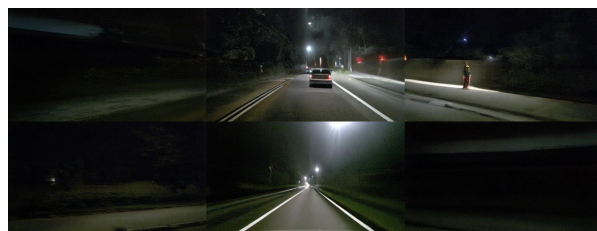
$t=5s$



$t=10s$



$t=15s$



$t=20s$

Figure 14. Sample of 20s multi-view video at night without reference frames..

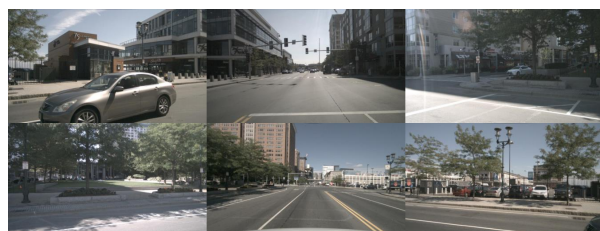
Ground Truth



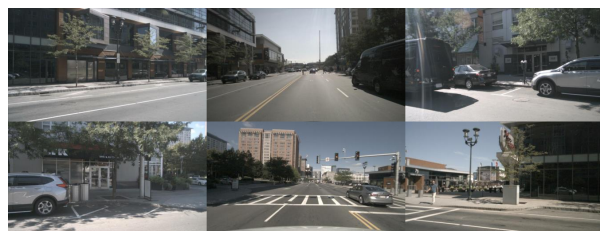
$t=0s$



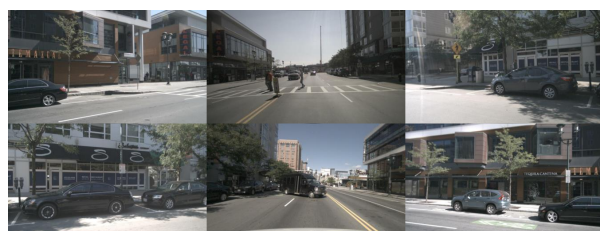
$t=2s$



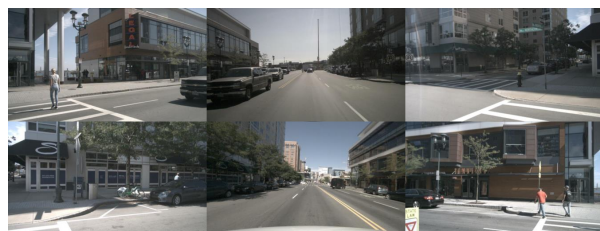
$t=5s$



$t=10s$

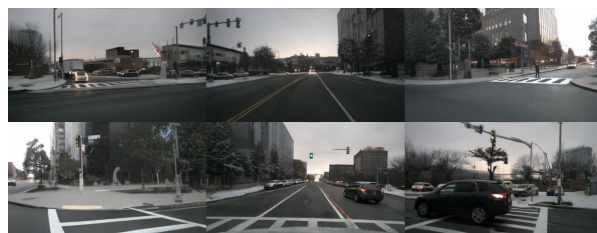


$t=15s$

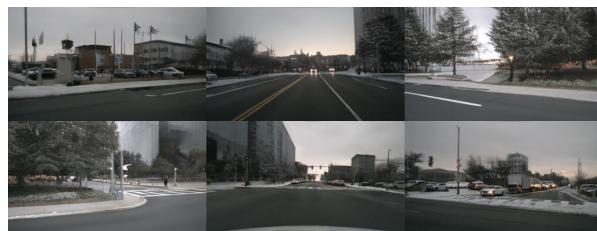


$t=20s$

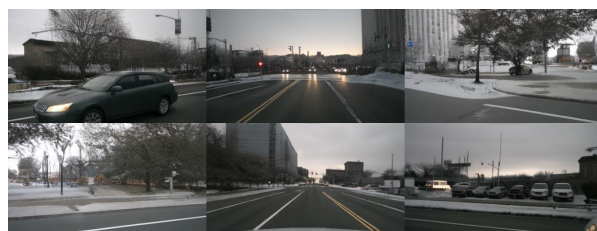
Generation



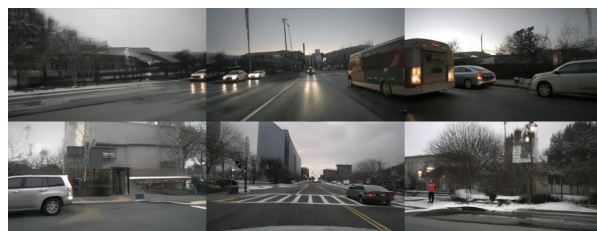
$t=0s$



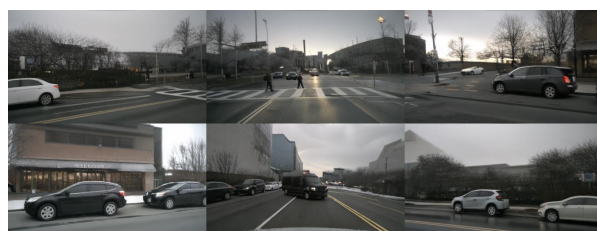
$t=2s$



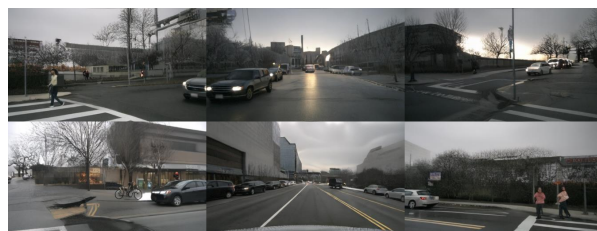
$t=5s$



$t=10s$



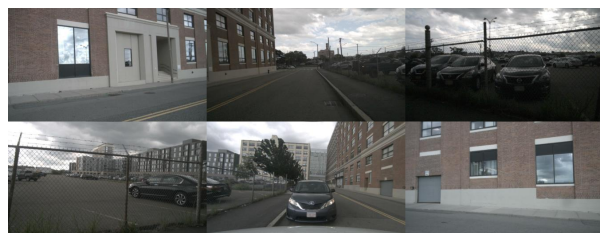
$t=15s$



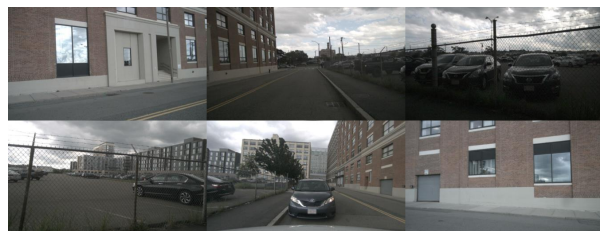
$t=20s$

Figure 15. Sample of a 20s multi-view video transformed from a sunny to a snowy scene through text editing.

Ground Truth



$t=0s$



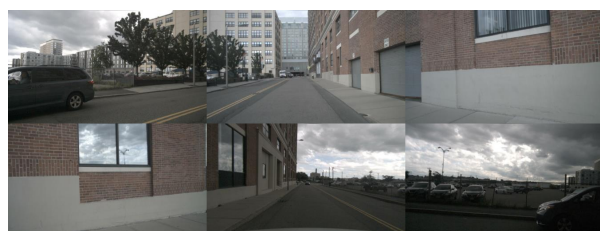
$t=2s$



$t=5s$



$t=10s$

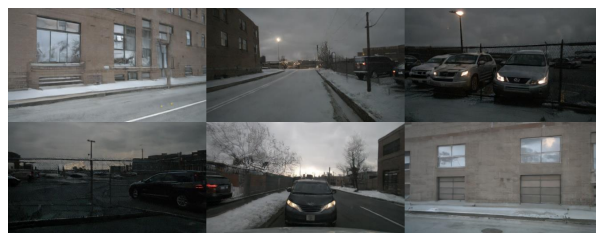


$t=15s$



$t=20s$

Generation



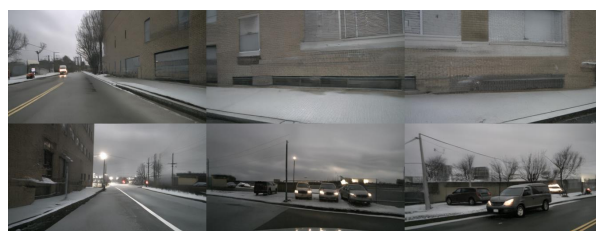
$t=0s$



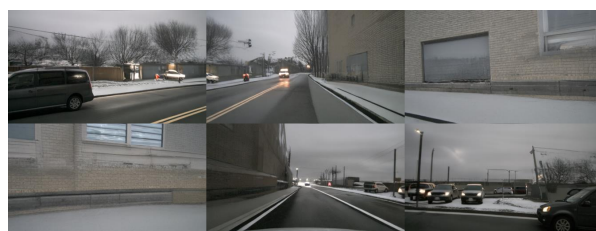
$t=2s$



$t=5s$



$t=10s$



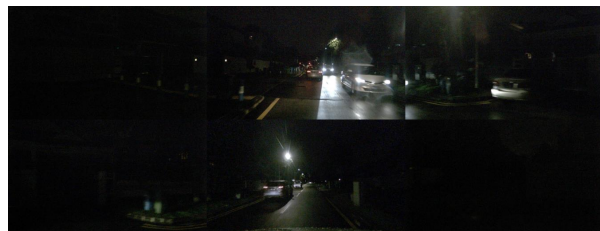
$t=15s$



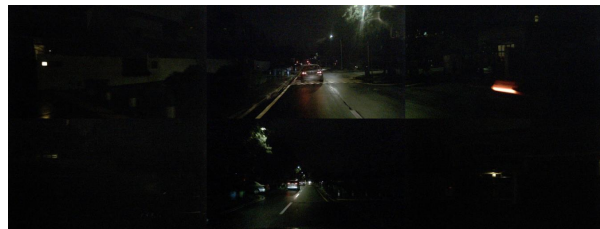
$t=20s$

Figure 16. Sample of a 20s multi-view video transformed from a cloudy to a snowy scene through text editing.

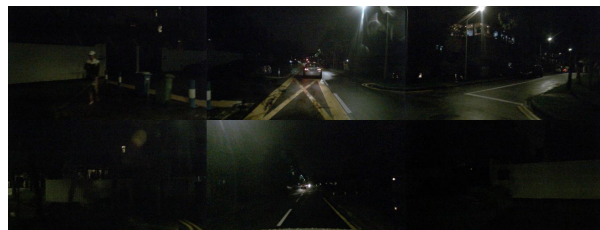
Ground Truth



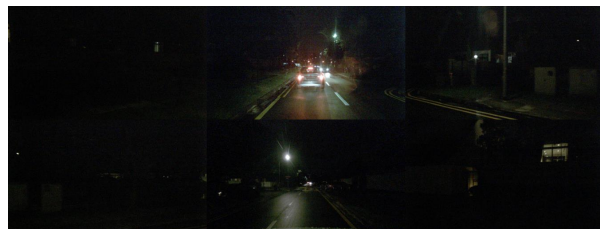
$t=0s$



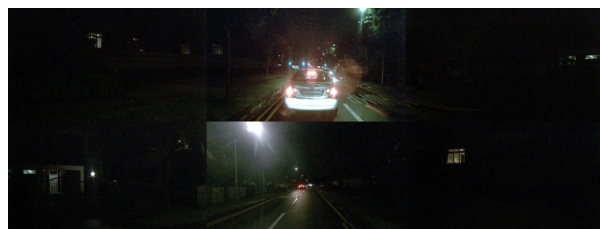
$t=2s$



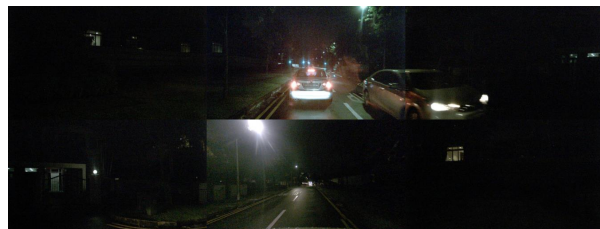
$t=5s$



$t=10s$

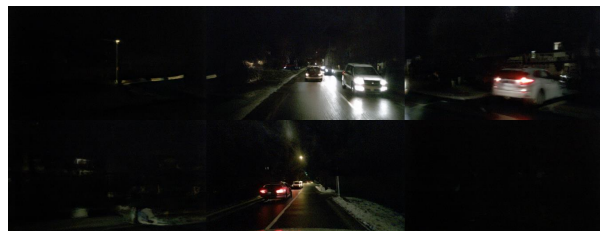


$t=15s$

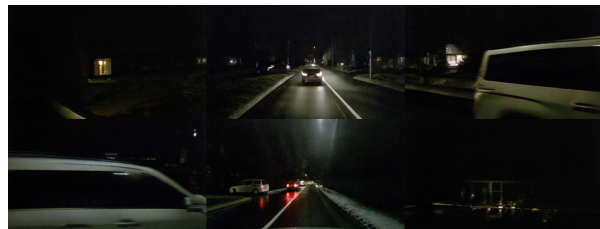


$t=20s$

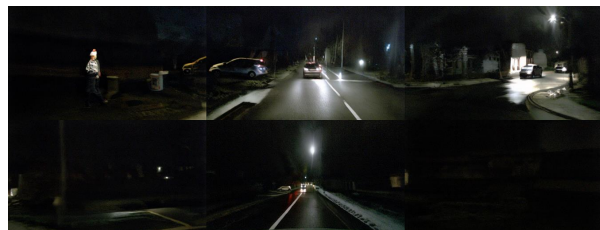
Generation



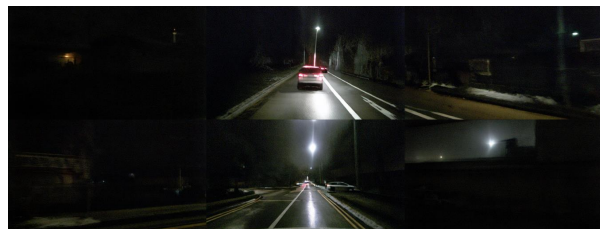
$t=0s$



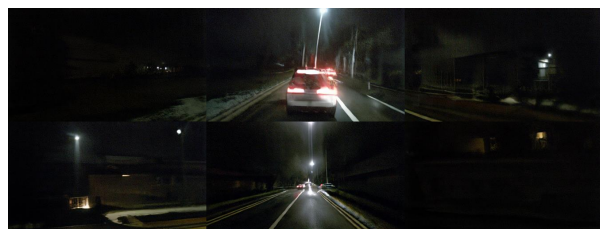
$t=2s$



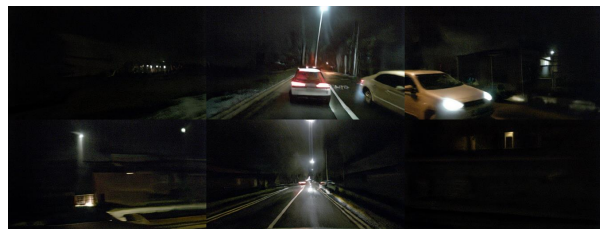
$t=5s$



$t=10s$



$t=15s$



$t=20s$

Figure 17. Sample of a 20s multi-view video generated as a snowy night scene through text editing.

Ground Truth



$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Generation



$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Figure 18. Sample of a realistic scene generation based on CARLA conditions.

Ground Truth



$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Generation



$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Figure 19. Sample of a realistic scene generation based on CARLA conditions.

Ground Truth



$t=0s$



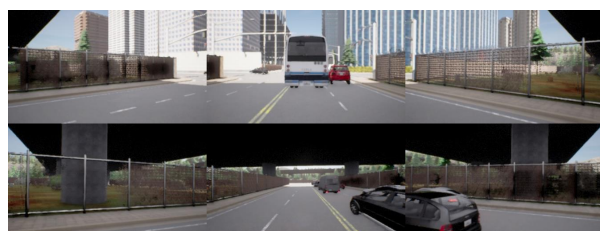
$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Generation



$t=0s$



$t=2s$



$t=5s$



$t=10s$



$t=15s$



$t=20s$

Figure 20. Sample of a realistic scene generation based on CARLA conditions.