CVD-STORM: CROSS-VIEW VIDEO DIFFUSION WITH SPATIAL-TEMPORAL RECONSTRUCTION MODEL FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models have been widely applied to world modeling for environment simulation and future state prediction. With advancements in autonomous driving, there is a growing demand not only for high-fidelity video generation under various controls, but also for producing diverse and meaningful information such as depth estimation. To address this, we propose CVD-STORM, a cross-view video diffusion model utilizing a spatial-temporal reconstruction Variational Autoencoder (VAE) that generates long-term, multi-view videos with 4D reconstruction capabilities under various control inputs. Our approach first fine-tunes the VAE with an auxiliary 4D reconstruction task, enhancing its ability to encode 3D structures and temporal dynamics. Subsequently, we integrate this VAE into the video diffusion process to significantly improve generation quality. Experimental results demonstrate that our model achieves substantial improvements in both FID and FVD metrics. Additionally, the jointly-trained Gaussian Splatting Decoder effectively reconstructs dynamic scenes, providing valuable geometric information for comprehensive scene understanding.

(a) Ground Truth

(b) w/o STORM-VAE (1200 step)

(c) w/ STORM-VAE (1200 step)

Figure 1: **Early-Stage Generation Visualization.** (a) shows the ground-truth sequence. (b) depicts the model's output at training step 1,200 when using a standard VAE. (c) presents the corresponding output generated with our STORM-VAE at the same step. Notably, (c) exhibits significantly improved convergence and visual fidelity compared to (b), demonstrating the effectiveness of our approach even at early stage in training.

1 INTRODUCTION

Autonomous vehicles have emerged as a prominent research domain within artificial intelligence applications. The development of reliable self-driving systems necessitates both extensive data collection for training decision-making algorithms and sophisticated closed-loop simulations to verify planning outputs. These requirements present significant challenges, particularly the need for driving world models that accurately represent the environment and enables precise prediction of future scenarios. Concurrently, diffusion models have become the state-of-the-art approach for video

generation, offering a promising solution for realistic simulation. Recent advances in this field have demonstrated the remarkable capability of these models to generate photorealistic videos Hong et al. (2022); Zheng et al. (2024); Peng et al. (2025), with successful applications extending to complex driving scenarios Kim et al. (2021); Zhao et al. (2025).

To serve as comprehensive driving world models, diffusion-based approaches must be capable of generating long-term, multi-view, and controllable videos. Early attempts such as Gao et al. (2023); Kim et al. (2021) struggled with generating extended sequences and following complex conditional inputs. Recent advancements, however, have significantly addressed these limitations by adopting architectures and methodologies from high-performing diffusion models. For example, Chen et al. (2024); Gao et al. (2024b); Ren et al. (2025) have all implemented spatial-temporal diffusion transformer (DiT) architectures and employed multi-stage training strategies, progressively enhancing generative fidelity and temporal consistency. Nevertheless, despite incorporating cross-view generation, these approaches lack explicit 3D information, which constrains their applicability as world models. To overcome this limitation, Gao et al. (2024a) directly applies an enhanced 3D Gaussian Splatting (3DGS) technique to diffusion outputs, though internal inconsistencies in the generated images remain inadequately resolved. UniScene Li et al. (2025) incorporates semantic occupancy as conditional guidance for LiDAR generation, but requires additional annotation during the training process. Other approaches Hassan et al. (2025); Liang et al. (2025) produce depth maps supervised by Depth Anything V2 Yang et al. (2024c), but these relative depth estimates cannot accurately represent real-world geometry. While Wu et al. (2024b) attempts to generate LiDAR data and video simultaneously, the LiDAR is not well aligned with the images.

To address these challenges, we propose CVD-STORM, a framework that generates long sequential multi-view driving videos while simultaneously decoding reconstructed scenes represented by dynamic 3D Gaussian Splatting (3DGS) Kerbl et al. (2023). First, we finetune an image VAE with an affiliated Gaussian decoder as described in STORM Yang et al. (2025), enabling the decoding of VAE latents into 3D Gaussians. This finetuned model, dubbed as STORM-VAE, serves as the latent encoder for training a cross-view video diffusion model with the same architecture as Chen et al. (2024). Recent research Yu et al. (2025); Leng et al. (2025); Fuest et al. (2024) has established that representation learning is crucial to diffusion model performance. Aligned with these findings, our experiments demonstrate that the latents encoded by STORM-VAE, which fuse information from LiDAR and across frames, significantly improve the generative quality and convergence rate. Figure 1 illustrates the impressive denoising ability of CVD-STORMat an early step, compared with the one without STORM-VAE. During inference, CVD-STORM can generate long-term six-view videos conditioned on text, bounding boxes (BBox), and high-definition maps (HDMap), while the Gaussian Splatting (GS) Decoder can directly reconstruct 4D scenes from the generated latents.

In summary, our main contributions are:

- We introduce STORM-VAE, an extended VAE model incorporating a Gaussian Splatting decoder for 4D scene reconstruction. This auxiliary network integrates spatial and temporal information into the latent representation, moving beyond RGB-only encoding. Meanwhile, it can also achieve 4D reconstruction in the driving scenarios.
- We propose CVD-STORM, a novel pipeline for driving world modeling that simultaneously generates multi-view videos and reconstructs 4D scenes. We separate the training of these complex tasks into two stages, by training the
- Our experiments demonstrate that CVD-STORM not only significantly improve the generative quality of the current world model by enhancing representation learning, but also addresses the challenges of 4D absolute depth estimation.

2 Related Work

2.1 Video Diffusion and Driving World Model

The diffusion approach has become the mainstream for generative tasks. With the advancements in 2D image diffusion models Rombach et al. (2022); Zhang et al. (2023); Labs et al. (2025); Li et al. (2024), this technique has rapidly extended to video generation Hong et al. (2022); Yang et al. (2024d); Gao et al. (2025); Zheng et al. (2024); Peng et al. (2025); Kong et al. (2024), yielding

impressive visualizations and enabling precise control. In addition, related studies highlight its significant potential as a real-world simulator.

In the field of autonomous driving, research started to focus on constructing driving world models based on video generation to simulate realistic driving scenarios. For instance, GenAD Yang et al. (2024a) leverages large-scale web video datasets to enhance long-duration video generation capabilities, while Vista Gao et al. (2024c) incorporates action inputs to control vehicle trajectories. However, these approaches are limited to single-view generation and do not include other conditions to simulate road conditions. There still exists a significant gap between their capabilities and real-world driving requirements.

Therefore, generating multi-view video with precise control and long-term consistency has attracted significant research attention. Early approaches such as Gao et al. (2023); Zhao et al. (2025); Xie et al. (2025) achieved promising results for short-term videos but struggled to extend sequence length effectively. The emergenece of DiT Peebles & Xie (2023) substantially improved diffusion model scalability, prompting numerous researchers to incorporate transformer architectures into driving world models. UniMLVG Chen et al. (2024) enhancs Stable Diffusion 3.5 Esser et al. (2024) with temporal and multi-view modules, successfully unifying multiple datasets with heterogeneous structures during training. Similarly, MagicDriveV2 Gao et al. (2024b) also employs this design but encodes videos through 3D VAE to achieve greater data compression. This architecture has demonstrated exceptional performance when applied to larger-scale datasets Ren et al. (2025); Russell et al. (2025). Additionally, researchers also have successfully incorporated action control mechanisms to enable the generation of precisely controllable multi-view videos Ni et al. (2025b). Despite these advancements, current generative methods still fail to adequately capture important structural information, particularly depth data.

2.2 4D Reconstruction in Driving Scenarios

Capturing 3D information is crucially important in driving scenarios and numerous studies has explored how to predict the depth or reconstruct the 4D scene in the front-view driving videos. Some research incorporates the structure prediction in the generative procedure. For instance, UniFuture Liang et al. (2025) directly unified the depth prediction into the video generation to attain highly aligned RGB-Depth correspondence. However, this work needs Depth Anything V2 Yang et al. (2024c) to generate pseudo supervision for depth. Additionally, this approach can only produce relative depth, which is insufficient for the autonomous driving application. Another unified framework GEM Hassan et al. (2025) mitigates problems with consistencies in long-range video generation, yet still preserves similar problem in depth estimation as UniFuture.

On the other hand, a considerable body of research has focused on incorporating reconstruction tasks into driving scenarios. MagicDrive3D Gao et al. (2024a) employs a two-stage pipeline that integrates Gaussian splatting for 3D reconstruction. Although presented as a unified framework, the second-stage reconstruction process exerts minimal influence on the generative model in the first stage, limiting true end-to-end interaction. More approaches concentrate primarily on pure reconstruction objectives. For instance, ReconDreamer Ni et al. (2025a) introduces a network trained to correct artifacts in novel views reconstructed from a pretrained 3D Gaussian representation. Similarly, OmniScene Wei et al. (2025) leverages forward Gaussian mapping to obtain a 3D scene representation in bird's-eye view (BEV) format. Building upon this, STORM Yang et al. (2025) advances the paradigm by employing forward 4D Gaussian splatting to capture spatiotemporal dynamics through sequential scene modeling.

While both generative modeling and 3D reconstruction have been extensively studied in autonomous driving contexts, few works have explored the integration of these two tasks in a synergistic manner. The potential of jointly optimizing generation and reconstruction remains largely underexplored.

2.3 Representation Learning in Diffusion

Recent research has devoted considerable effort to exploring better latent representations for improving diffusion model performance Fuest et al. (2024). Yang et al. (2022); Tian et al. (2023); Deja et al. (2023) involves incorporating additional tasks during generation training, such as classification

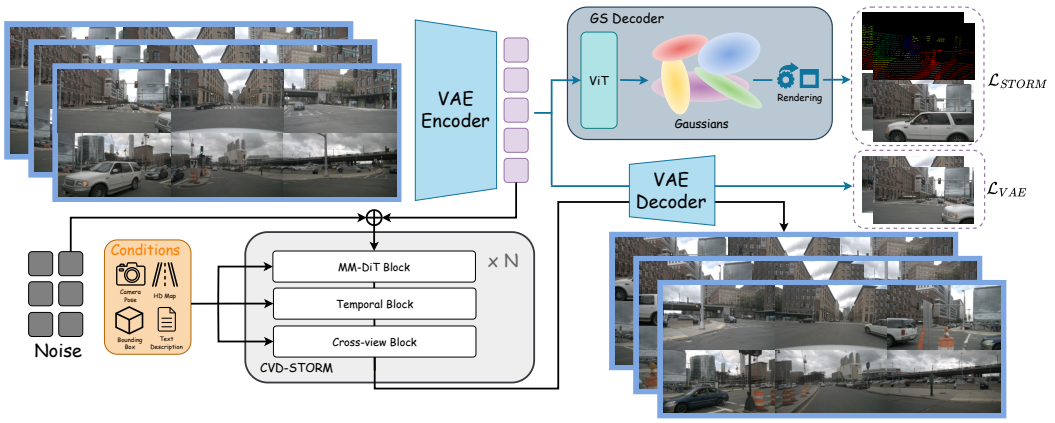

Figure 2: **Overall framework of the model.** Our pipeline contains two models. The upper section illustrates STORM-VAE training, with the forward process indicated by blue arrows. STORM-VAE takes multi-view images from context timesteps and processes the image latents through two decoders: the VAE Decoder performs image reconstruction (updated by $\mathcal{L}_{\text{VAE}}$), while the GS Decoder performs scene reconstruction (updated by $\mathcal{L}_{\text{STORM}}$). The lower section illustrates the inference pipeline of CVD-STORM, with the forward process shown by solid block arrows. The diffusion part can either use STORM-VAE latents as reference frames for prediction or generate from noise, while incorporating various conditioning inputs for guidance.

and segmentation. Other works focuses on aligning the latent space with that of foundation models. For example, Pernias et al. (2023) divides diffusion training into two stages, with the first stage dedicated to training an additional encoder that extracts image semantic features. REPA Yu et al. (2025) takes intermediate features from the diffusion model and projects them to align with features from pretrained models, while VA-VAE Yao et al. (2025) performs this alignment during variational auto-encoder (VAE) training. Building upon REPA, REPA-E Leng et al. (2025) finetunes the entire model end-to-end, allowing the alignment loss to update VAE parameters and thereby accelerating generation performance.

Inspired by these advances, we extend this representation learning approach to video diffusion models by introducing a reconstruction task during training and simultaneously tuning the VAE. This approach aims to enhance generation performance while achieving a significant additional capability — 4D reconstruction.

3 Method

Figure 2 illustrates the overall pipeline of our proposed method. Our framework generates multi-view driving videos conditioned on various inputs, including text prompts, bounding boxes, HD maps—with or reference frames, while simultaneously producing scene reconstructions represented as dynamic 3DGS. Our approach extends UniMLVG Chen et al. (2024) by enhancing its variational autoencoder (VAE) architecture and refining the training procedure. Specifically, we first finetune the pretrained image VAE to create STORM-VAE, which incorporates an additional reconstruction task adapted from STORM Yang et al. (2025). This modification introduces a Gaussian Decoder capable of reconstructing 3D Gaussians and their associated velocities. We then leverage STORM-VAE to train a DiT-based diffusion model that employs three distinct transformer blocks operating along different data dimensions, which improve both spatial coherence and temporal consistency in the generated outputs.

3.1 Preliminary: STORM

Given a set of images $\{\boldsymbol{I}_t^v \in \mathbb{R}^{H\times W\times 3}\}$ with corresponding camera poses from timestamps $t \in T_c$ and viewpoints $v \in V$, STORM fuses image features through a Vision Transformer (ViT) and generate pixel-level Gaussians $\{\boldsymbol{G}_t^v \in \mathbb{R}^{H\times W\times 12}\}$. Each Gaussian is characterized by its center $\boldsymbol{\mu} \in \mathbb{R}^3$, orientation $\mathcal{R} \in \mathbb{SO}(3)$, scale $s \in \mathbb{R}$, opacity $o \in \mathbb{R}$, and color $\boldsymbol{c} \in \mathbb{R}^3$. The center $\boldsymbol{\mu}$ is positioned along the ray cast from the camera center, allowing the Gaussian decoder to only

output the depth value. Additionally, the model predicts the velocity of each Gaussian to model dynamic scene elements. To render target viewpoints at timestamp t', the 3D Gaussian Splatting (3DGS) elements $\boldsymbol{G}_t^v$ are transformed according to their predicted velocities into Gaussians at time t', denoted as $\boldsymbol{G}_{t\to t'}^v$. The target images are then rendered based on the union of all $\boldsymbol{G}_{t\to t'}^v$. To enhance image quality, STORM incorporates auxiliary tokens to compose sky colors and adopts view-based exposure variations.

The training process is supervised by target views randomly sampled within a predefined sampling range. The image rendering loss $\mathcal{L}_{\text{rgb}}$ is formulated as:

$$\mathcal{L}_{\text{rgb}} = \sum_{t'\in T_t, v\in V} \|D(F(\{\boldsymbol{I}_t^v\}), t', v) - \boldsymbol{I}_{t'}^v\|_2^2, \tag{1}$$

where F represents the ViT encoder, D denotes the decoder and rendering, including all image post-processing operations, and T_t is the set of target timesteps. Additionally, the Gaussian rendering can also produce depth so we use the depth map obtained by projecting LiDAR on camera views to supervise the training. We define the overall loss as $\mathcal{L}_{\text{STORM}}$ and omit discussion of additional loss terms not directly relevant to this paper. For more detailed description of the methodology, please refer to Yang et al. (2025).

3.2 STORM-VAE

We introduce STORM-VAE, a novel variational autoencoder that incorporates STORM as an auxiliary network within the VAE framework. The upper part of Figure 2 illustrates the architecture of our proposed model. STORM-VAE builds upon a general VAE structure, specifically utilizing the pretrained VAE from Stable Diffusion 3.5 (SD3.5) Esser et al. (2024) in our setting. In the STORM-VAE pipeline, the VAE encoder E first encodes input images into latent representations, which are subsequently processed through two parallel branches. In the first branch, the latents are processed by the VAE decoder D_{VAE} to ensure high-fidelity image reconstruction, supervised by the loss function $\mathcal{L}_{\text{VAE}}$. In the second branch, sampled context latents are fed into the Gaussian Splatting decoder (D_{GS}), which shares architectural similarities with STORM. The key distinction is that STORM processes RGB images directly while the D_{GS} operates on the VAE latent representations. Consequently, the new RGB rendering loss is formulated as:

$$\mathcal{L}_{\text{rgb}} = \sum_{t'\in T_t, v\in V} D_{\text{GS}}(E(\boldsymbol{I}t^v), t', v) - \boldsymbol{I}t'v|_2 2, \tag{2}$$

where D_{GS} is equivalent to $D \cdot F$ described in Section 3.1. The comprehensive training objective combines the VAE and STORM components as follows:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda \mathcal{L}_{\text{STORM}}. \tag{3}$$

Here, $\mathcal{L}_{\text{VAE}}$ comprises three components: the reconstruction loss $\mathcal{L}_{\text{MSE}}$, the perceptual loss $\mathcal{L}_{\text{LPIPS}}$, and the KL divergence loss $\mathcal{L}_{\text{KL}}$. We deliberately excluded the GAN loss from our implementation as our experiments indicated it compromised training stability. In our experiments, we set λ to 0.5.

3.3 CVD-STORM

The lower part of Figure 2 illustrates the architecture of CVD-STORM. Following UniMLVG Chen et al. (2024), we adopt SD3.5 as initialization and append a temporal block and a cross-view block after each Multi-Modality DiT (MM-DiT) block of SD3.5. The input latent of CVD-STORM is $z_t \in \mathbb{R}^{T\times V\times C\times H\times W}$, where T is the number of frames, V is the number of viewpoints, C is the latent dimension, and H, W are the latent spatial dimensions of a single image. The MM-DiT block performs attention only at the image level (i.e., across $H \times W$ dimensions), which requires reshaping the input to $HW \times TV \times C$ before processing. Similarly, the temporal block operates on the sequence length dimension and the cross-view block operates on the view dimension, requiring to reshape the input to $T \times VHW \times C$ and $V \times THW \times C$, respectively. We also incorporate the multiple conditioning approaches and multi-task framework from UniMLVG in our training. For details regarding these components, please refer to their paper. The training loss utilizes rectified flow Liu et al. (2022), formulated as:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\|\epsilon_\theta(z_t, t, c) - (z_0 - \epsilon)\|^2\right], \tag{4}$$

Method	Duration	FID↓	FVD↓	mAP$_{obj}$↑	mIoU$_r$↑	mIoU$_v$↑
DreamForge Mei et al. (2024)	20s	16	224.8	13.80	-	-
UniScene Li et al. (2025)	-	6.1	70.5	-	-	-
Glad Xie et al. (2025)	-	11.2	118.0	-	-	-
DriveScape Wu et al. (2024a)	-	8.3	76.4	-	64.43	28.86
MagicDrive2 Gao et al. (2024b)	5s	19.1	218.1	12.30	61.05	27.01
DriveSphere Yan et al. (2025)	-	-	103.4	21.45	-	-
DiVE Jiang et al. (2024)	20s	-	94.6	<u>24.55</u>	-	-
UniMLVG Chen et al. (2024)	20s	5.8	36.1	22.50	**70.81**	<u>29.12</u>
CVD-STORM	20s	**3.8**	**14.0**	**25.21**	<u>66.11</u>	**29.84**

Table 1: **Comparison of the generation quality and condition-following metrics on nuScenes validation set.** The best results are in **bold**, while the second best results are in <u>underlined</u>. Since most of methods do not release their checkpoints, we list the results reported in their paper. − represents the values not mentioned in the corresponding papers. mIoU$_r$ and mIoU$_v$ are the short of the mean IoU of road and vehicle.

where ϵ_θ denotes the model, z represents the STORM-VAE latent, z_t is the noisy latent, ϵ is the noise, t is the timestep, and c is the conditioning information.

Different from UniMLVG, we replace the SD3.5 VAE with our STORM-VAE, which provides enhanced latent representations and the capability to estimate absolute depth. Furthermore, rather than employing a multi-stage training process to progressively develop temporal and spatial generation capabilities, we jointly train the temporal blocks, spatial blocks, and MM-DiT blocks in a single stage. This integrated approach significantly simplifies the training procedure and reduces computational costs.

4 Experiments

4.1 Experiment Details

4.1.1 Datasets

We adopt both single-view and multi-view datasets in our training: OpenDV-Youtube Yang et al. (2024b) for single-view data, and nuScenes Caesar et al. (2020), Waymo Sun et al. (2020), and Argoverse2 Wilson et al. (2023) for multi-view data. We set the sequence length of a single simple as 19. To enhance the extensibility and diversity of our model, we incorporate three different image resolutions: 144×256, 176×304, and 256×448, with sampling ratios of 0.1, 0.3, and 0.6, respectively. All the models are trained on H100 with batch size 32. For diffusion training, we leverage available dataset annotations, including 3D bounding boxes, HD maps, and camera parameters. For nuScenes specifically, we utilize 12 Hz interpolated annotations. Text descriptions for all frames and views are generated at 2 Hz (key frames for evaluation).

4.1.2 Evaluation Metrics

To assess the effectiveness of our method in terms of realism, continuity, ad precise control, we selected four key metrics to compare against existing multi-view image and video generation methods. We use the widely recognized Fréchet Inception Distance(FID) Heusel et al. (2017) for realism evalution and Fréchet Video Distance (FVD) Unterthiner et al. (2018) for temporal coherence estimation. To evaluate controllablity, we evaluate two perception tasks: 3D object detection and BEV segmentation of road maps. These tasks serve as proxies for measuring the spatial accuracy and consistent geometry of our generated content. We adopt BEVFormer Li et al. (2022) and cross-view transformers Zhou & Krähenbühl (2022) to evaluate the performance on these two tasks respectively.

4.1.3 Implementation details

For STORM-VAE training, we designate the 1st, 7th, 13th and 19th frames as the context frames while 3 timesteps are randomly sampled as targets. Since the Opendv-Youtube is a single-view

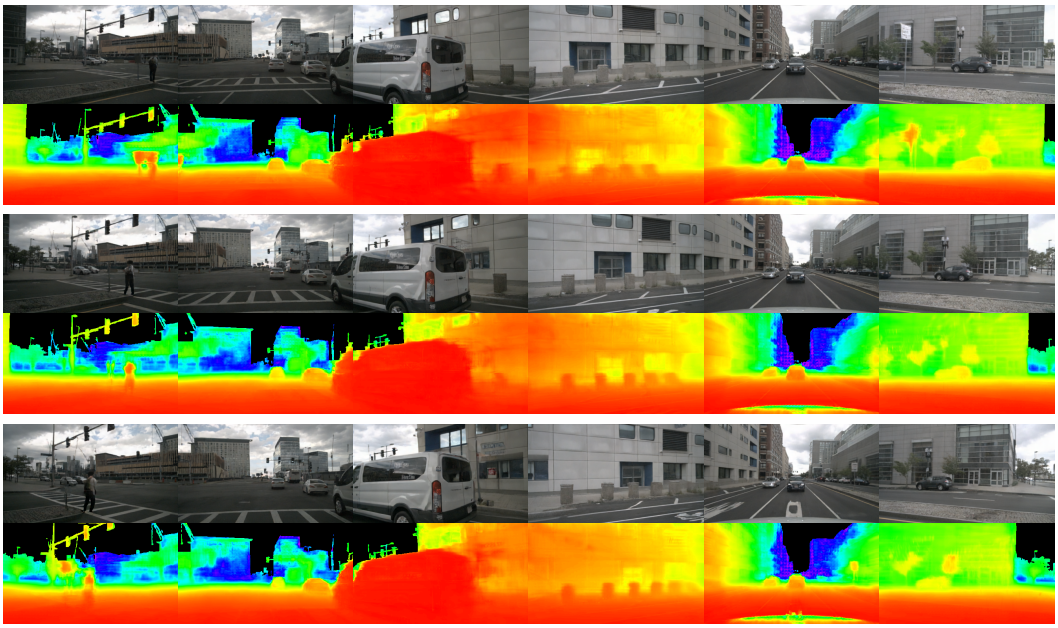

Figure 3: **Qualitative results of Depth Estimation.** This figure illustrates the depth of the videos generated by CVD-STORM at frame 0, 5, 10. Our GS decoder can successfully extract the depth information of dynamic and static objects.

Table 2: **Ablation Study**

# Ref. frames	FID	FVD
0	8.7	39.0
1	**3.6**	17.2
3	3.8	**14.0**

(a) **Ablation study of the number of reference frames**. The best results are in **bold**. The FID is about the same with reference frame, while FVD strictly decreases with larger reference frame count.

VAE used	FID	FVD
w/o STORM-VAE	9.36	52.85
w/ STORM-VAE	**7.92**	**34.37**

(b) **Ablation study of the use of VAE**. The best results are in **bold**. w/o STORM-VAE means using default vae of SD3.5. Both models are trained for 40k steps with Opendv, nuScenes, Waymo, Argoverse2. No pretarined weight is loaded for diffusion for fair comparison.

dataset without LiDAR data, it is used exclusively to train the VAE image reconstruction branch. The other three datasets are utilized for both VAE and STORM training. To address viewpoint inconsistency across datasets, we standardize inputs to 6 views for all datasets and implement attention masking to avoid redundant data fusion.

For diffusion training, we freeze the encoder of STORM-VAE. As discussed in Section 3.3, we implement the single-stage training so we have to deal with the invariance across datasets. For OpenDV-Youtube, the cross-view block is omitted due to its single-view nature. For multi-view datasets, we randomly drop temporal and cross-view blocks to enhance the generative capability of each individual block, thereby improving the overall model stability and robustness. During inference, we use 3 frames as reference for autoregressive prediction. A cosine scheduler is used with initial learning rate of 6×10^{-5}. The minimum learning rate is set to 1×10^{-7}. The optimize is widely used AdamW. The inference steps are set to 50. All Experiments are conducted on H100 GPUs.

4.2 Experiment Results

4.2.1 Comparison

Generation Tasks. Following the common evaluation protocols, we report quantitative metrics on the nuScenes validation set, shown in Table 1. Our model demonstrates exceptional perfor-

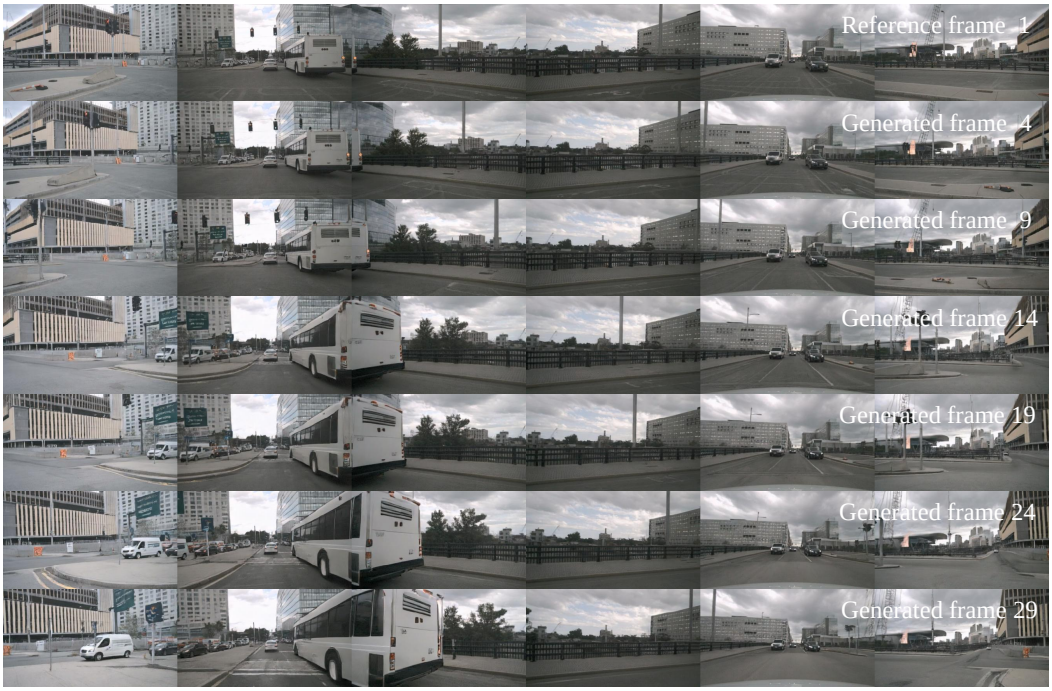

Figure 4: **Qualitative Results of Video Prediction.** We produce this example using three reference frames. The first line is the first reference frame and the following lines are the predicted frames. Our method demonstrates strong temporal consistency in the video prediction task.

mance compared to previous SOTA methods DiVE Jiang et al. (2024) and UniMLVG Chen et al. (2024). Specifically, our approach achieves significant improvements of 34.48% in Fréchet Inception Distance (FID) and 61.21% in Fréchet Video Distance (FVD) relative to the second-best method. Additionally, our model can generate high-quality videos with durations up to 20 seconds. Regarding condition consistency, our approach outperforms competing methods on mAP of object detection (mAP_{obj}) and IoU of road (IoU_r) of. It ranks second in IoU of vehicle (IoU_v), performing marginally below UniMLVG in this particular metric.

STORM-VAE Results. We provide the visualization of the depth maps of the generative images in Figure 3. We put the more detailed evaluation and discussion in the Appendix.

4.3 Ablation Study

Number of Reference Frames. The number of reference frames represents different types of tasks in the generative model. Without reference frames, the model conducts pure video generation, producing content based solely on conditional inputs. On the contrary, the model perform video prediction when the reference frames are given. We present qualitative results in Figures 4 and 5, with quantitative evaluations in Table 2a. As shown in the table, the FVD score is steadily improved as the number of reference frames increases, indicating that additional reference frames provide richer temporal information from the ground truth, thereby improving temporal consistency. Conversely, when reference frames are provided, the model performs video prediction. For more results, please refer to the Appendix.

Effect of STORM-VAE. Table 2b demonstrates that our STORM-VAE significantly improves generation quality over the standard VAE baseline. Specifically, STORM-VAE yields a 15.38% reduction in Fréchet Inception Distance (FID) and a 34.97% decrease in Fréchet Video Distance (FVD), indicating substantial enhancements in both image and video generation quality. Furthermore, Figure 1 illustrates that our model accelerates convergence compared to the baseline. To ensure fair evaluation in this ablation study, we compare models trained for the same number of steps.

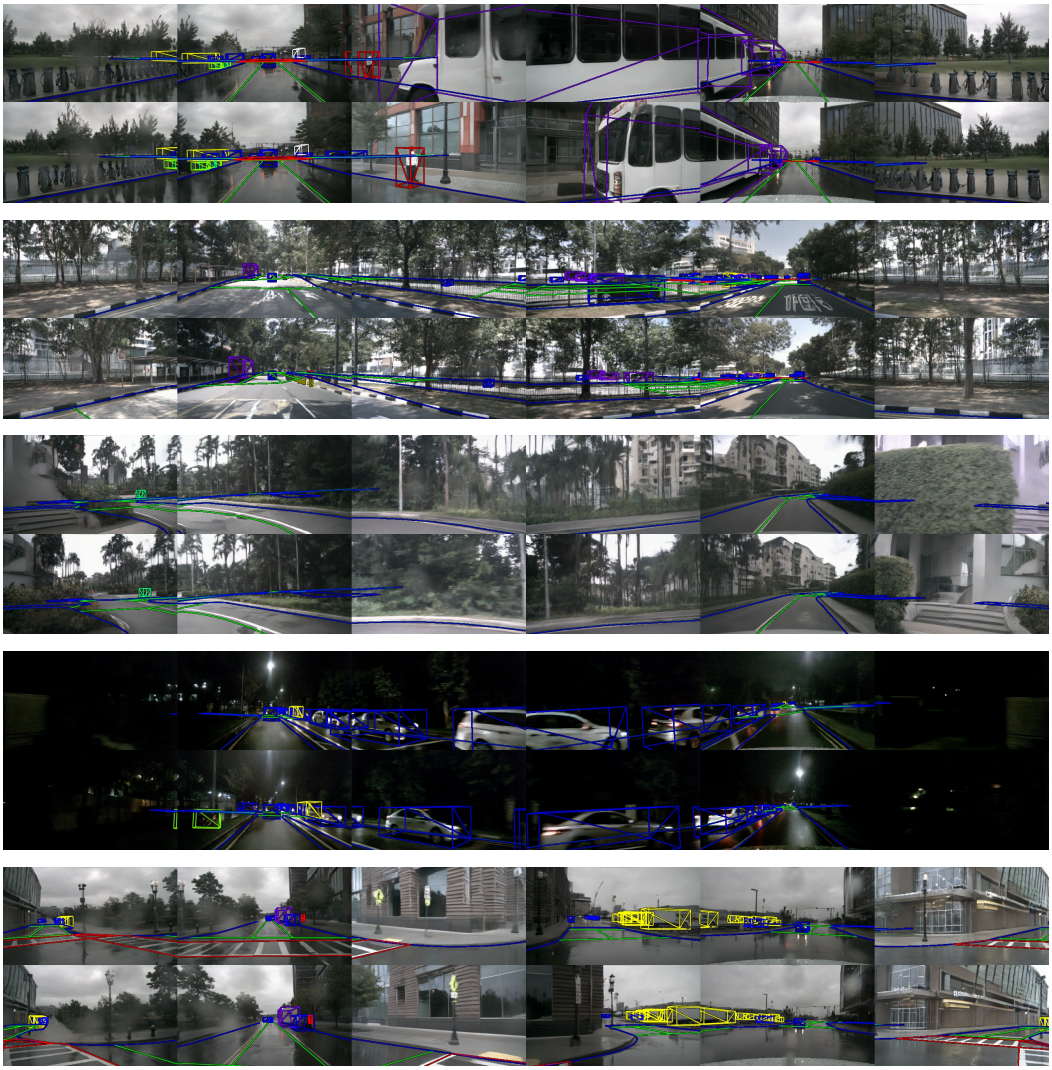

Figure 5: **Qualitative Results of Video Generation.** We provide the examples generated with the conditons only, without any reference frame. For each scene, we list the 1st frame in the first line and the 10th frame in the second line. The bounding boxes and road maps are overlapping over the generative images. The object in the bounding boxes with the same color are should be of the same class. For example, cars should be generated in the blue 3D bounding boxes.

5 Conclusion

We introduce CVD-STORM, a novel framework that unifies long-sequence, multi-view video generation with dynamic 4D scene reconstruction. Our approach extends the traditional VAE architecture by incorporating a Gaussian Splatting Decoder, namely STORM-VAE. This design not only enables high-quality 4D scene reconstruction but also substantially enhances representation learning, thereby improving the generative capabilities of our downstream diffusion model. Leveraging the pre-trained STORM-VAE, we train CVD-STORM using multiple datasets and support various conditioning types across diverse generative tasks. Experimental results demonstrate that CVD-STORM surpasses SOTA methods, particularly in image quality and temporal coherence. Furthermore, the Gaussian Splatting Decoder directly estimates absolute depth through neural rendering, providing richer 3D structural information than previous approaches.

REFERENCES

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024.

Kamil Deja, Tomasz Trzciński, and Jakub M Tomczak. Learning data representations with joint diffusion models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 543–559. Springer, 2023.

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

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Bjorn Ommer. Diffusion models and representation learning: A survey. *arXiv preprint arXiv:2407.00783*, 2024.

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 arXiv:2310.02601*, 2023.

Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024a.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024b.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024c.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22404–22415, 2025.

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. *Advances in neural information processing systems*, 30, 2017.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Junpeng Jiang, Gangyi Hong, Lijun Zhou, Enhui Ma, Hengtong Hu, Xia Zhou, Jie Xiang, Fan Liu, Kaicheng Yu, Haiyang Sun, et al. Dive: Dit-based video generation with enhanced control. *arXiv preprint arXiv:2409.01595*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5820–5829, 2021.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and Others. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11971–11981, 2025.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

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. *arXiv preprint arXiv:2203.17270*, 2022.

Dingkang Liang, Dingyuan Zhang, Xin Zhou, Sifan Tu, Tianrui Feng, Xiaofan Li, Yumeng Zhang, Mingyang Du, Xiao Tan, and Xiang Bai. Seeing the future, perceiving the future: A unified driving world model for future generation and perception. *arXiv preprint arXiv:2503.13587*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Jianbiao Mei, Tao Hu, Xuemeng Yang, Licheng Wen, Yu Yang, Tiantian Wei, Yukai Ma, Min Dou, Botian Shi, and Yong Liu. Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*, 2024.

Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1559–1569, 2025a.

Jingcheng Ni, Yuxin Guo, Yichen Liu, Rui Chen, Lewei Lu, and Zehuan Wu. Maskgwm: A generalizable driving world model with video mask reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22381–22391, 2025b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in 200k. *arXiv preprint arXiv:2503.09642*, 2025.

Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

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

Changyao Tian, Chenxin Tao, Jifeng Dai, Hao Li, Ziheng Li, Lewei Lu, Xiaogang Wang, Hongsheng Li, Gao Huang, and Xizhou Zhu. Addp: Learning general representations for image recognition and generation with alternating denoising diffusion process. *arXiv preprint arXiv:2306.05423*, 2023.

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 arXiv:1812.01717*, 2018.

Dongxu Wei, Zhiqi Li, and Peidong Liu. Omni-scene: Omni-gaussian representation for ego-centric sparse-view scene reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22317–22327, 2025.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Wei Wu, Xi Guo, Weixuan Tang, Tingxuan Huang, Chiyu Wang, Dongyue Chen, and Chenjing Ding. Drivescape: Towards high-resolution controllable multi-view driving video generation. *arXiv preprint arXiv:2409.05463*, 2024a.

Zehuan Wu, Jingcheng Ni, Xiaodong Wang, Yuxin Guo, Rui Chen, Lewei Lu, Jifeng Dai, and Yuwen Xiong. Holodrive: Holistic 2d-3d multi-modal street scene generation for autonomous driving. *arXiv preprint arXiv:2412.01407*, 2024b.

Bin Xie, Yingfei Liu, Tiancai Wang, Jiale Cao, and Xiangyu Zhang. Glad: A streaming scene generator for autonomous driving. *arXiv preprint arXiv:2503.00045*, 2025.

Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27531–27541, 2025.

Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Maximilian Igl, Apoorva Sharma, Peter Karkus, Danfei Xu, Boris Ivanovic, Yue Wang, and Marco Pavone. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. *arXiv preprint arXiv:2501.00602*, 2025.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14662–14672, June 2024a.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14662–14672, 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024c.

Xiulong Yang, Sheng-Min Shih, Yinlin Fu, Xiaoting Zhao, and Shihao Ji. Your vit is secretly a hybrid discriminative-generative diffusion model. *arXiv preprint arXiv:2208.07791*, 2022.

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

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

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*, pp. 3836–3847, 2023.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, 2022.

A APPENDIX

A.1 THE USE OF LARGE LANGUAGE MODELS (LLMS)

LLMs are only used to help polish paper writing. The retrieval of references and ideation of research are performed solely by human authors.

Table 3: **Comparison of STORM and STORM-VAE**

Method	PSNR ↑	D-RMSE ↓
STORM	20.89	5.52
STORM-VAE	21.18	4.55

(a) **Comparison of Reconstruction.** We extend the original STORM to a 6-view rendering model and evaluate the performance on NuScene. Our STORM-VAE also slightly outperforms the STORM in the reconstruction task.

Method	AbsRel ↓	$\boldsymbol{\delta_1}$ ↑
UniMLVG + STORM	30.825	49.7
CVD-STORM	16.05	49.7

(b) **Comparison of Zero-shot Depth Estimate.** We evaluate the performance of our models on depth estimation in the generation results. We use the pesudo-groundtruth produced by Depth Anything V2.

A.2 ADDITIONAL EXPERIMENTS

A.2.1 VIDEO RESULTS

We provide a video in the supplementary material for better visualization.

A.2.2 COMPARISON OF STORM AND STORM-VAE.

We evaluate the performance of STORM-VAE in comparison to STORM, as illustrated in Table 3. Specifically, Table 3a demonstrates STORM-VAE's reconstruction capabilities relative to STORM. For quantitative assessment, we evaluate the reconstructed images and depth maps of STORM-VAE on the nuScenes dataset using two metrics: Peak Signal-to-Noise Ratio (PSNR) for image quality and Depth Root Mean Square Error (D-RMSE) for depth accuracy. Our experimental results demonstrate that STORM-VAE even slightly exceeds its performance.

In generation task, we compare the performance of CVD-STORM against UniMLVG + STORM, which first employs UniMLVG to generate videos and subsequently applies STORM to reconstruct the 4D scene. During inference, we set the context timesteps equal to the target timesteps, which are the four adjacent frames spanning the interval $[t, t+3]$. The GS Decoder processes frames [t+3, t+6] as context in next iterations and continues this progressive reconstruction strategy until reaching the end of the sequence. To assess its zero-shot depth estimation, we employ two metrics: Absolute Relative Error (AbsRel) and δ_1, where δ_1 represents the percentage of pixels satisfying $\max(\frac{d}{\hat{d}}, \frac{\hat{d}}{d}) < 1.25$, shown in Table 3b. Since ground truth depth is unavailable for generated results, we utilize Depth Anything V2 Yang et al. (2024c) to produce pseudo ground-truth depth maps. While these metrics provide valuable comparative insights, we acknowledge their limitation in assessing absolute depth accuracy, which remains an open challenge in generative depth evaluation. We provide more qualitative results in Figure 6,7.

A.2.3 MORE QUALITATIVE RESULTS

We provide more qualitative results in Figure 8, 9, 10, 11, 12.

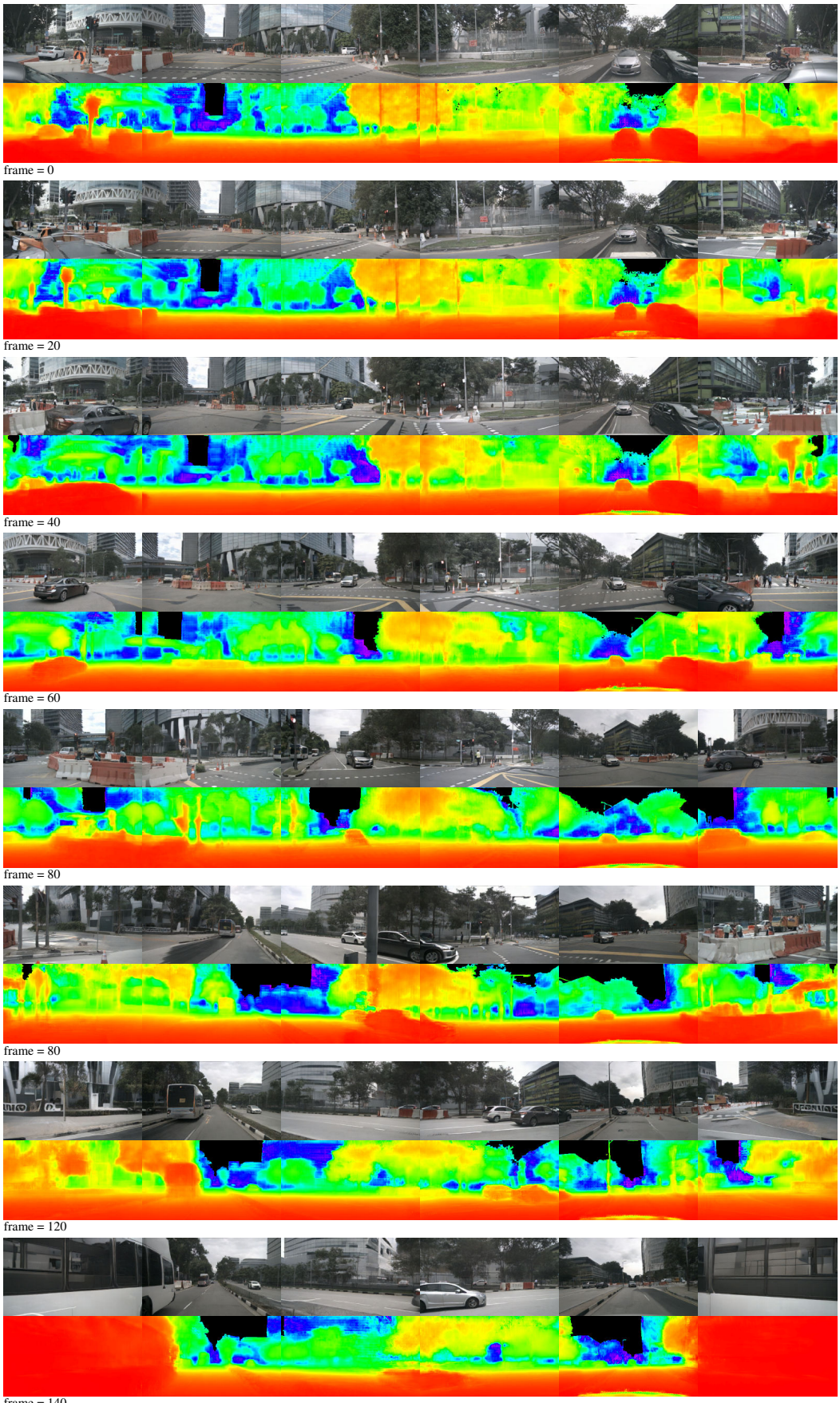

Figure 6: **Qualitative results of Depth Estimation.**

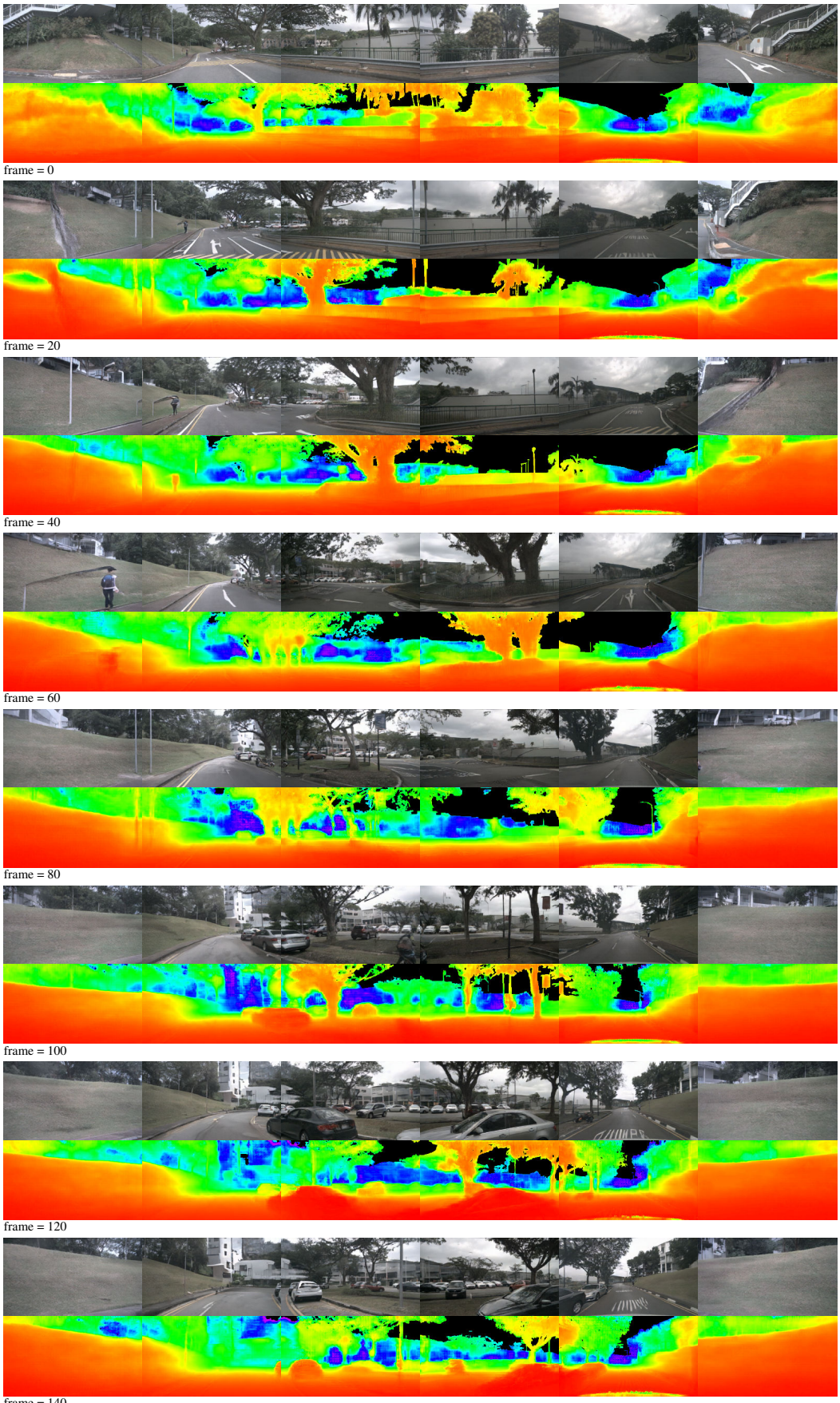

Figure 7: **Qualitative results of Depth Estimation.**

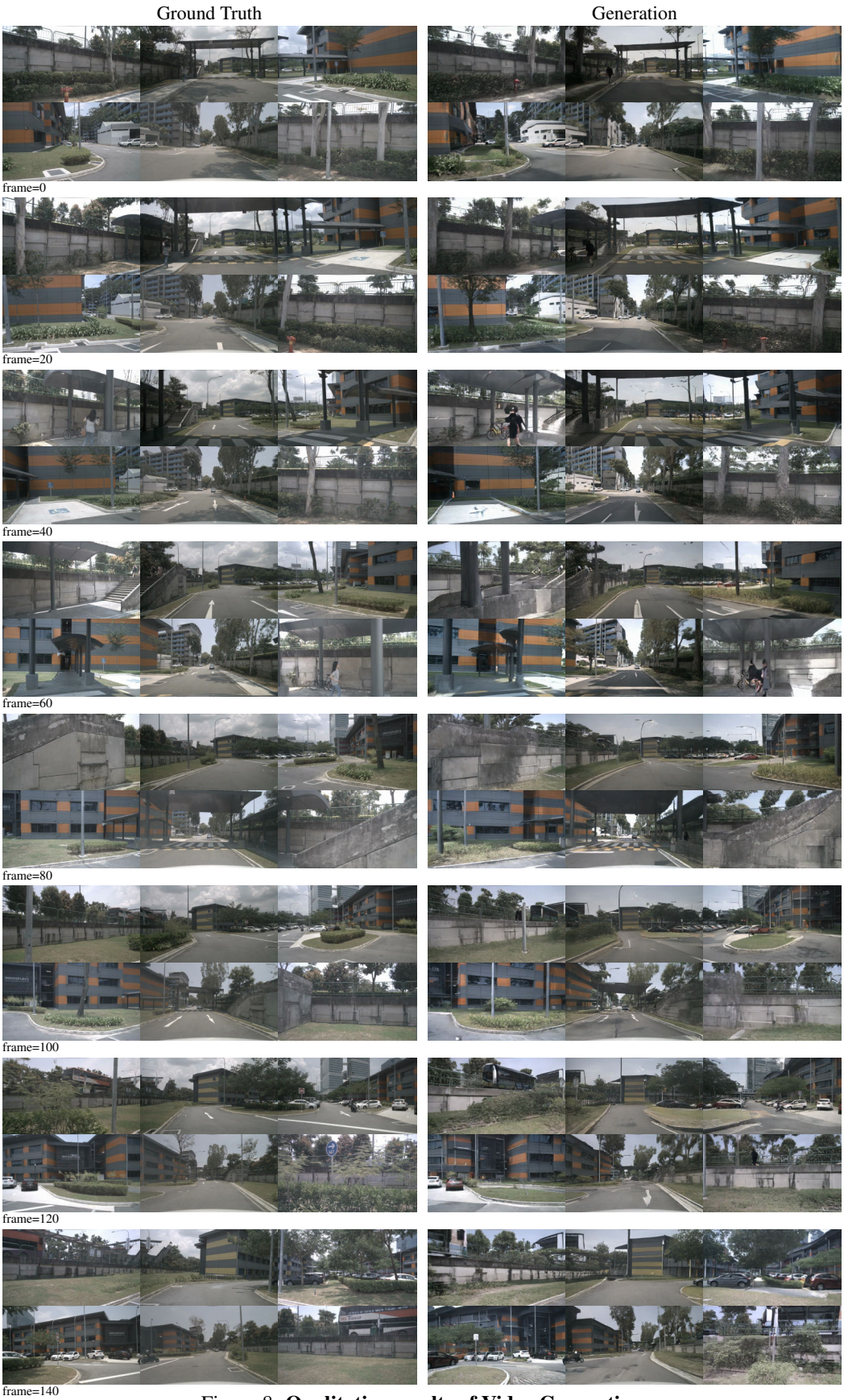

Figure 8: **Qualitative results of Video Generation**

Figure 9: **Qualitative results of Video Generation**

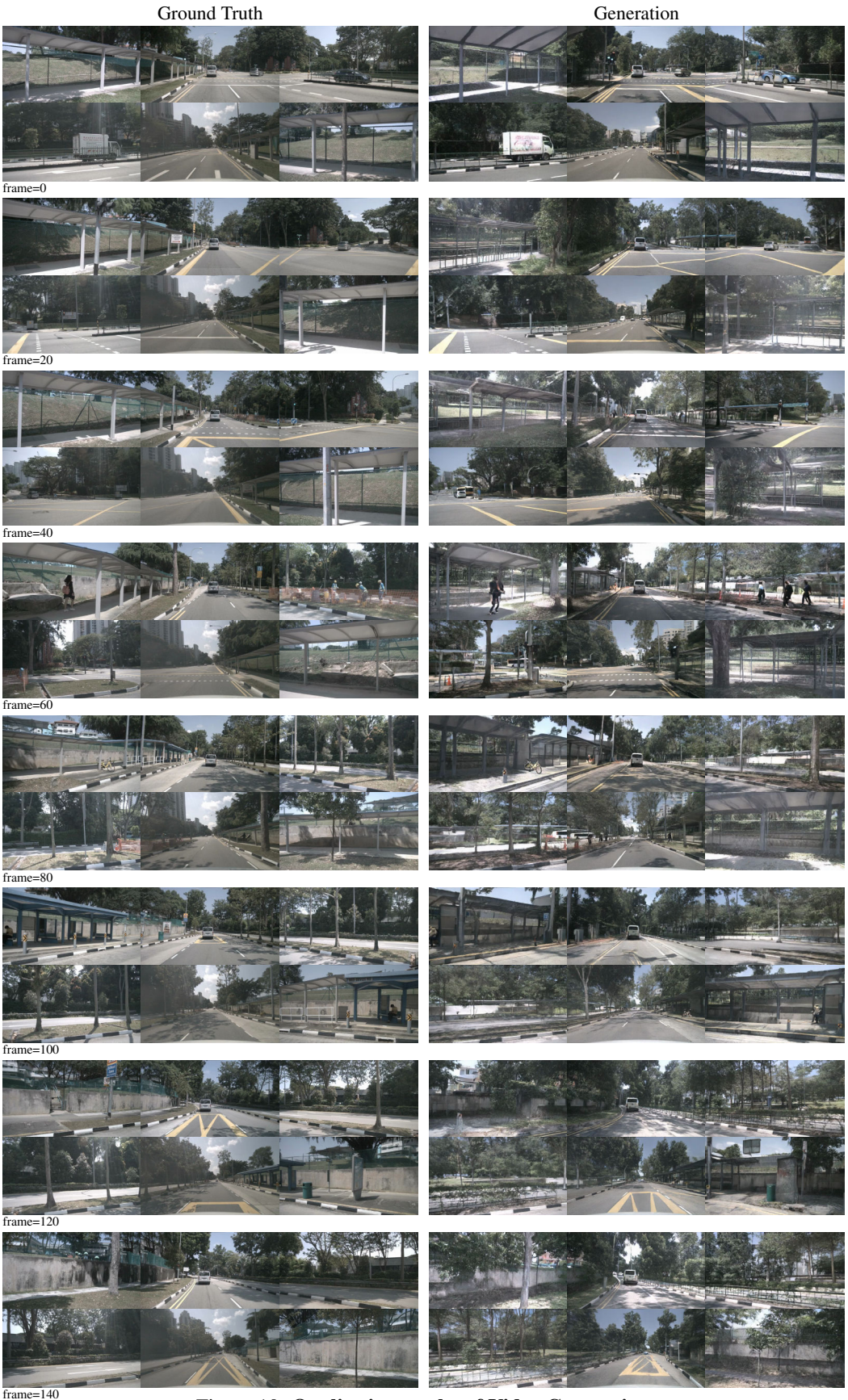

Figure 10: **Qualitative results of Video Generation**

Figure 11: **Qualitative results of Video Generation from 3 reference frames.**

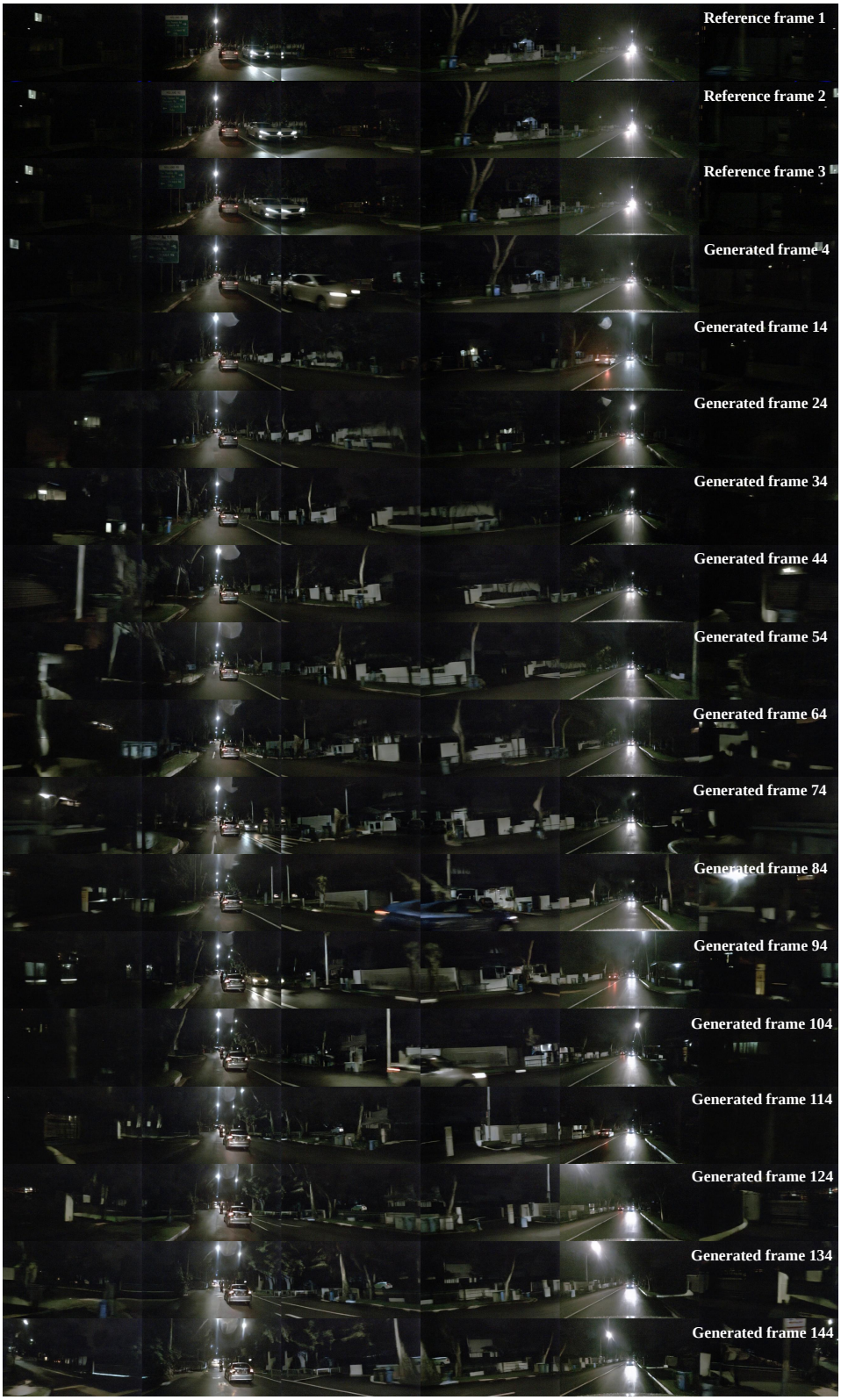

Figure 12: **Qualitative results of Video Generation from 3 reference frames at night.** Our model imitated the blur of fast motion.