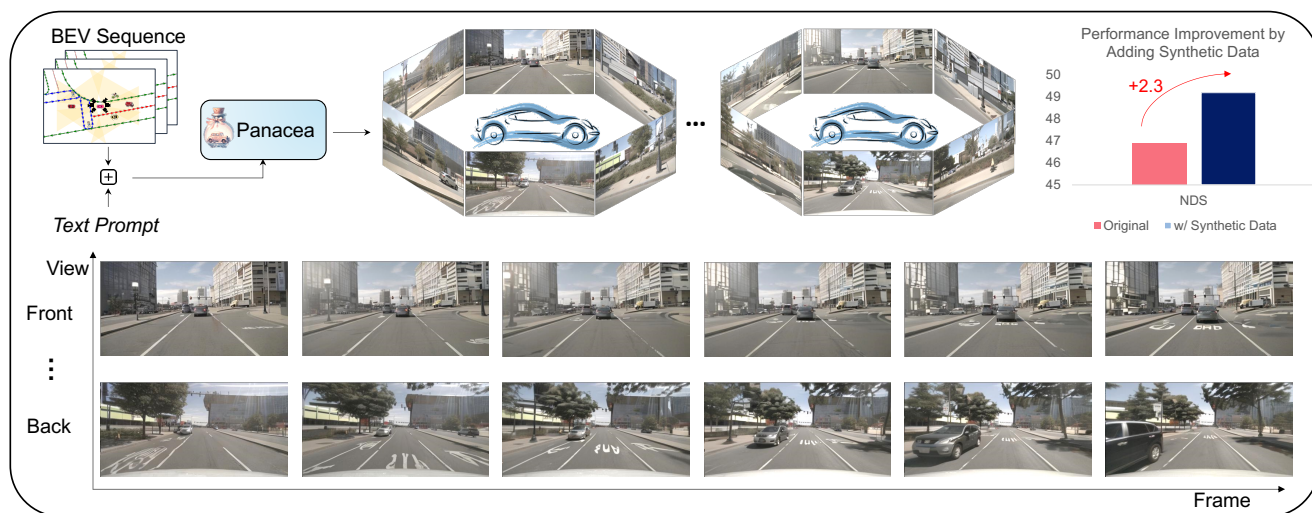# Panacea: Panoramic and Controllable Video Generation for Autonomous Driving

Yuqing Wen[1][*][†] Yucheng Zhao[2][*] Yingfei Liu[2][*] Fan Jia[2], Yanhui Wang[1], Chong Luo[1]
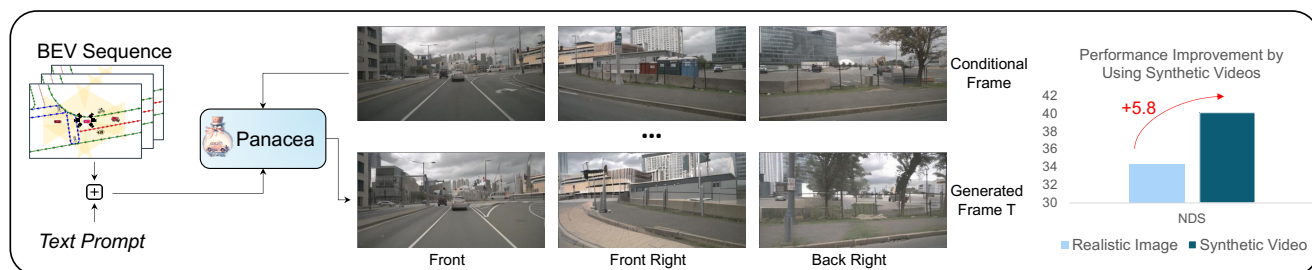Chi Zhang[3], Tiancai Wang[2][‡] Xiaoyan Sun[1][‡] Xiangyu Zhang[2]
[1]University of Science and Technology of China   [2]MEGVII Technology   [3]Mach Drive
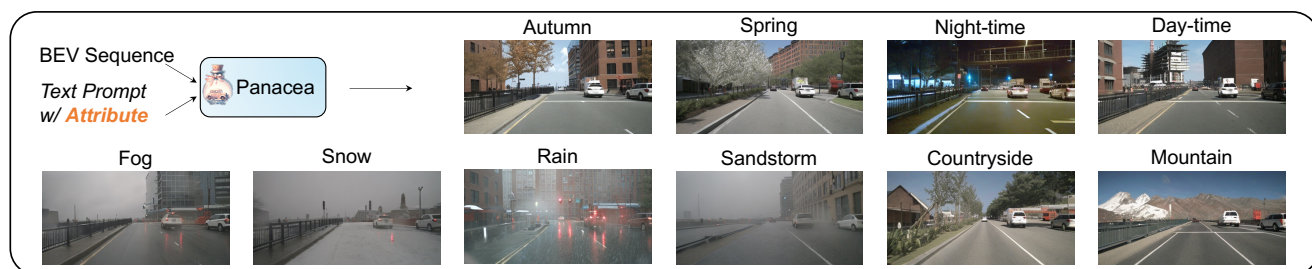*Project Page*: https://panacea-ad.github.io/

Figure 1. **Visualizations of the Panacea's capability.** (a). Panoramic video generation based on BEV (Bird's-Eye-View) layout sequence facilitates the establishment of a synthetic video dataset, which enhances perceptual tasks. (b). Producing panoramic videos with conditional images and BEV layouts can effectively elevate image-only datasets to video datasets, thus enabling the advancement of video-based perception techniques. (c). Video generation with variable attribute controls, such as weather, time, and scene.

Abstract

The field of autonomous driving increasingly demands high-quality annotated training data. In this paper, we propose Panacea, an innovative approach to generate panoramic and controllable videos in driving scenarios, capable of yielding an unlimited numbers of diverse, annotated samples pivotal for autonomous driving advancements. Panacea addresses two critical challenges: 'Consistency' and 'Controllability.' Consistency ensures temporal and cross-view coherence, while Controllability ensures the alignment of generated content with corresponding annotations. Our approach integrates a novel 4D attention and a two-stage generation pipeline to maintain coherence, supplemented by the ControlNet framework for meticulous control by the Bird's-Eye-View (BEV) layouts. Extensive qualitative and quantitative evaluations of Panacea on the nuScenes dataset prove its effectiveness in generating high-quality multi-view driving-scene videos. This work notably propels the field of autonomous driving by effectively augmenting the training dataset used for advanced BEV perception techniques.

1. Introduction

In the domain of autonomous driving, there has been a surge of interest in Bird's-Eye-View (BEV) perception methods, which have demonstrated significant potential across key perception tasks including 3D detection [16, 23, 45], map segmentation [20, 26], and 3D lane detection [6, 17]. Cutting-edge BEV perception approaches, exemplified by StreamPETR [45], are trained on multi-view videos. As a result, the crux of building robust autonomous driving system lies in high-quality, large-scale annotated video datasets. Yet, the acquisition and annotation of such data present formidable challenges. Assembling diverse video datasets encompassing a spectrum of weather, environmental, and lighting conditions not only poses challenges but can occasionally entail risks. Moreover, the annotation of video data necessitates significant resources in both effort and cost.

Inspired by the success of leveraging synthetic street images to improve the performance of perception tasks [7, 41, 52, 57], our proposal focuses on generating synthetic multi-view driving video data to bolster the training of cutting-edge video-based perception methods. To mitigate high annotation costs, we aim to utilize BEV layout sequences, which encompass 3D bounding boxes and road maps, for the generation of corresponding videos. Such BEV sequences can be acquired from annotated video

---

*Equal contribution.

†This work was done during the internship at MEGVII Technology.

‡Corresponding author.

datasets [4, 40, 51] or synthesized using advanced simulators [11, 50, 53]. This initiative can therefore be formulated as the diverse multi-view driving video generation conditioned on BEV sequences. The effectiveness of our generation model rests on two critical criteria: **controllability** and **consistency**. Empowering users to govern the generated videos via input BEV sequences and descriptive text prompts define controllability, while consistency underscores temporal coherence within individual single-view videos and coherence across multiple views.

Thanks to recent advancements in diffusion-based generative models and their extensions [13, 34, 38, 54], what was once a formidable challenge has become more tractable. In particular, Stable Diffusion [34] pioneers the adoption of diffusion models within latent spaces, amplifying computational efficiency with minimal compromise in generation quality. Extending this innovation, Video Latent Diffusion Model [3] expands the paradigm to high-resolution video generation by integrating temporal dimensions into established image frameworks. Furthermore, ControlNet [54] introduces an innovative neural architecture adept at modulating pretrained diffusion models, significantly enhancing their controllability and unlocking pathways for advanced applications. Nonetheless, seamlessly amalgamating these technologies to achieve panoramic and controllable video generation remains a huge challenge.

In this paper, we present Panacea, an innovative video generation approach tailored specifically for panoramic and controllable driving scene synthesis. Panacea operates as a two-stage system: the initial stage crafts realistic multi-view driving scene images, while the subsequent stage expands these images along the temporal axis to create video sequences. For panoramic video generation, Panacea introduces decomposed 4D attention, enhancing both multi-view and temporal coherence. Moreover, we employ ControlNet[54] to allow for the injection of BEV sequences. Beyond these core designs, our model retains the versatility to manipulate global scene attributes via textual descriptions, such as weather, time, and scene, offering a user-friendly interface for generating specific samples.

We apply the Panacea approach to the widely used nuScenes dataset [4]. Comprehensive evaluations across diverse real-world application scenarios indicate Panacea's proficiency in generating valuable video training instances. Not only does it enrich existing video datasets with a plethora of synthesized samples, but it also has the potential to elevate image-only datasets to video datasets, enabling the advancement of video-based perception techniques. Furthermore, the exceptional generation fidelity coupled with heightened controllability positions Panacea as a viable candidate for real-world driving simulation. Overall, our key contributions are two-fold:

• We introduce Panacea, an innovative approach to multi-

view video generation for driving scenes. This two-stage framework seamlessly integrates existing visual generation technologies while making important technical advancements to achieve multi-view and temporal consistency, alongside critical controllability. These technical enhancements play an important role in the solution's success.

- Through comprehensive qualitative and quantitative assessments, Panacea demonstrates its proficiency in producing high-quality driving-scene videos. Particularly significant is the quantitative evidence highlighting the substantial enhancement our synthesized video instances provide to state-of-the-art BEV perception models. Our plan to release these synthesized instances as the "GennuScenes" dataset aims to promote further research in the field of video generation for autonomous driving.

2. Related Work

2.1. Diffusion-based Generative Models

Recent advancements in diffusion models (DMs) have achieved remarkable milestones in image generation [8, 28, 31, 33, 36]. In particular, Stable Diffusion (SD) [34] employs DMs within the latent space of autoencoders, striking a balance between computational efficiency and high image quality. Beyond text conditioning, the field is evolving with the introduction of additional control signals [27, 54]. A noteworthy example among them is ControlNet [54], which incorporates a trainable copy of the SD encoder to integrate control signals. Furthermore, some studies concentrate on generating multi-view images. MVDffusion [42], for instance, processes perspective images in parallel with a pretrained diffusion model.

In addition to image generation, the application of diffusion models for video generation [9, 14, 21, 44, 48, 49, 59] is gaining significant attention. MagicVideo [59] employs frame-wise adaptors and causal temporal attention module for text-to-video generation. Video Latent Diffusion Model (VLDM) [3] integrates temporal layers into a 2D diffusion model to produce temporally aligned videos. Make-A-Video [37] stretches a diffusion based text-to-image model without the necessity for text-video pairs. Imagen Video [14] harnesses a chain of video diffusion models for generating videos according to text inputs.

Our method also falls under the category of video generation. However, unlike previous work, we focus on the creation of controllable multi-view videos within driving contexts, representing an innovative yet complex scenario.

2.2. Generation for Autonomous Driving

The development of Bird's-Eye-View (BEV) representation [19, 22, 23, 25, 45] in multi-view perception has become a critical research area in autonomous driving. This ad-

vancement is instrumental in enhancing downstream tasks such as multi-object tracking [29, 55], motion prediction [29], and planning [15, 18]. Recently, Video-based BEV perception methods [22, 23, 30, 45] have become the mainstream. BEVFormer [23] pioneering in integrating temporal modeling mechanism, yielding a substantial enhancement over single-frame methods such as DETR3D [47] and PETR [25]. Then BEVDepth [22], SOLOFusion [30], and StreamPETR [45] further improve the temporal modeling approach, achieving superior performance.

As the BEV perception methods rely heavily on paired data and BEV ground truth layouts, numerous studies are delving into paired data generation to aid training. Previously, generative efforts in autonomous driving primarily employed BEV layouts to augment image data with synthetic single or multi-view images [7, 41, 52, 57], proving beneficial for single-frame perception methods. For example, BEVGen [41] specializes in generating multi-view street images based on BEV layouts, while BEVControl [52] proposes a two-stage generative pipeline for creating image foregrounds and backgrounds from BEV layouts. However, the generation of paired video data, crucial for more advanced video-based BEV perception methods, remains largely unexplored. The Video Latent Diffusion Model [3] attempts to generate driving videos but its scope is limited to single-view and falls short in effectively bolstering video perception models.

In light of this, our work initiates the first exploration of generating multi-view videos paired with BEV layout sequences, marking a significant leap forward in enhancing video-based BEV perception.

3. Method

In this section, we present Panacea, an innovative approach to generate controllable multi-view videos for driving scenes. Sec. 3.1 provides a brief description of the latent diffusion models that form the foundation of our approach. Following this, Sec. 3.2 delves into our novel method that enables the generation of high-quality multi-view videos in a feasible and efficient manner. Finally, Sec. 3.3 elaborates on the controlling modules integral to Panacea, which is the central design feature that renders our model an invaluable asset for the advancement of autonomous driving systems.

3.1. Preliminary: Latent Diffusion Models

Diffusion models (DMs) [13, 38] learn to approximate a data distribution $p(x)$ via iteratively denoising a normally distributed noise ϵ. Specifically, DMs first construct the diffused inputs x_t through a fixed forward diffusion process in Eq. 1. Here α_t and σ_t represent the given noise schedule, and t indicates the diffusion time step. Then, a denoiser model ϵ_θ is trained to estimate the added noise ϵ from the diffused inputs x_t. This is achieved by minimizing
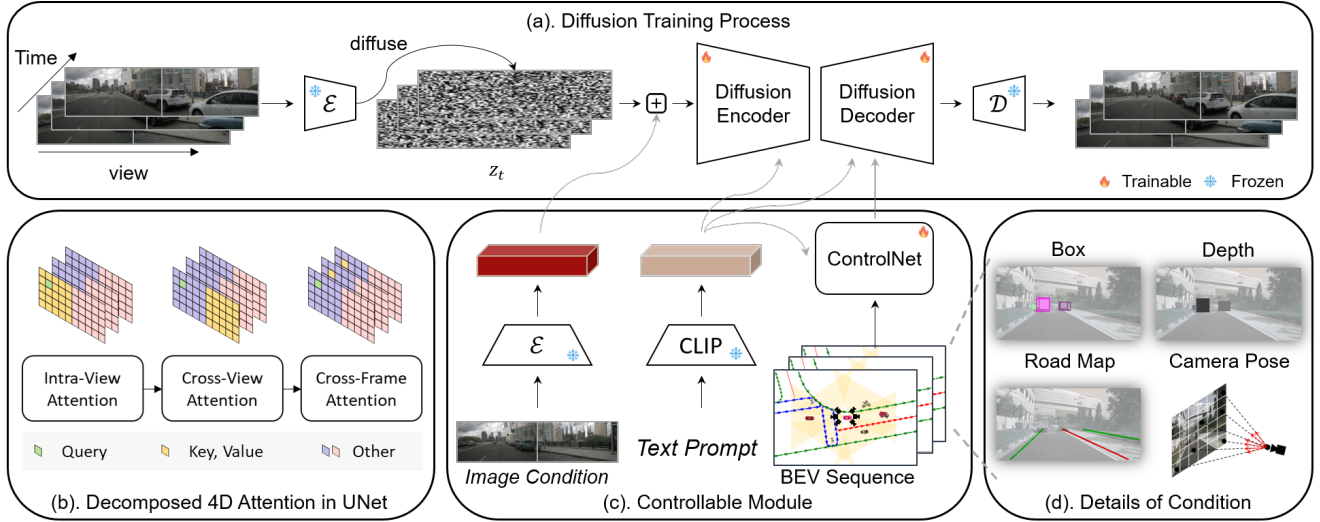
Figure 2. **Overview of Panacea.** (a). The diffusion training process of Panacea, enabled by a diffusion encoder and decoder with the decomposed 4D attention module. (b). The decomposed 4D attention module comprises three components: intra-view attention for spatial processing within individual views, cross-view attention to engage with adjacent views, and cross-frame attention for temporal processing. (c). Controllable module for the integration of diverse signals. The image conditions are derived from a frozen VAE encoder and combined with diffused noises. The text prompts are processed through a frozen CLIP encoder, while BEV sequences are handled via ControlNet. (d). The details of BEV layout sequences, including projected bounding boxes, object depths, road maps and camera poses.

the mean-square error, as detailed in Eq. 2. Once trained, DMs are able to synthesize a new data x_0 from random noise $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ by sampling x_t iteratively, as formulated in Eq. 3. Here μ_θ and Σ_θ are determined through the denoiser model ϵ_θ [13].

$$x_t = \alpha_t x + \sigma_t \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), x \sim p(x) \quad (1)$$

$$\min_\theta \mathbb{E}_{t,x,\epsilon} ||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||_2^2 \quad (2)$$

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (3)$$

Latent diffusion models (LDMs) [34] are a variant of diffusion models that operate within the latent representation space rather than the pixel space, effectively simplifying the challenge of handling high-dimensional data. This is achieved by transforming pixel-space image into more compact latent representations via a perceptual compression model. Specifically, for an image x, this model employs an encoder $\mathcal{E}$ to map x into the latent space $z = \mathcal{E}(x)$. This latent code z can be subsequently reconstructed back to the original image x through a decoder $\mathcal{D}$ as $x = \mathcal{D}(z)$. The training and inference processes of LDMs closely mirror those of traditional DMs, as delineated in Eq. 1-3, except for the substitution of x with the latent code z.

3.2. Generating High-Quality Multi-View Videos

Here we describe how we upgrade a pre-trained image LDM [34] for high-quality multi-view video generation.

Our model utilizes a multi-view video dataset p_{data} for training. Each video sequence, encompasses T frames, indicating the sequence length, V different views, and dimensions H and W for height and width, respectively.

Our framework is built on the Stable Diffusion (SD) [34] model, which is a strong pre-trained latent diffusion model for image synthesis. While the SD model excels in image generation, its direct application falls short in producing consistent multi-view videos due to the lack of constraints between different views and frames in the sequence. Therefore, we introduce an innovative architecture: a decomposed 4D attention-based UNet [35], designed to concurrently generate the entire multi-view video sequence. The joint diffused input z is structured with dimensions $H \times (W \times V) \times T \times C$, where C represents the latent dimension. This multi-view video sequence is constructed by concatenating the frames across their width, which aligns with their inherent panoramic nature. Fig. 2 (a) illustrates the overall training framework of the proposed model. Beyond the proposed 4D attention-based UNet, we also introduce a two-stage generation pipeline, which largely boosts the generation quality.

3.2.1 Decomposed 4D Attention

The decomposed 4D attention-based UNet is designed to enable cross-view and cross-frame modeling while ensuring computational feasibility. A naive approach to multi-
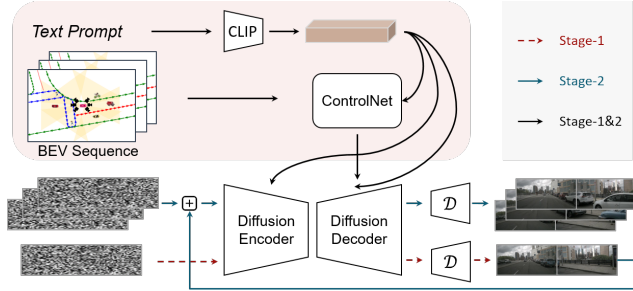
Figure 3. **The two-stage inference pipeline of Panacea.** Its two-stage process begins by creating multi-view images with BEV layouts, followed by using these images, along with subsequent BEV layouts, to facilitate the generation of following frames.

view video generation might employ 4D (HWVT) attention to exhaustively explore the multi-view video representations for coherent sample generation. However, this approach demands excessive memory and computational resources, surpassing the capabilities of even the most advanced A100 GPUs. Therefore, we propose a more efficient architecture called decomposed 4D attention, drawing inspiration from recent advancements in video representation learning [1, 2, 24, 56]. Our model selectively retains the most critical attention operations: the attention between adjacent views and the attention among spatially aligned temporal patches. This leads to the introduction of two new attention modules—cross-view attention and cross-frame attention—alongside the existing intra-view spatial attention. Empirical evaluations in Sec. 4.4 demonstrate that this decomposed 4D attention framework effectively generates coherent multi-view videos, maintaining both network feasibility and efficiency.

Fig. 2 (b) details our decomposed 4D attention mechanism. The intra-view attention retains the design of the original spatial self-attention in the Stable Diffusion (SD) model, as formulated in Eq. 4. To enhance cross-view consistency, we introduce cross-view attention. Our observations indicate that the correlation between adjacent views is paramount, while the correlation among non-adjacent views is comparatively less significant and can be disregarded. This cross-view attention is formulated in Eq. 5. The cross-frame attention, mirroring the design of VLDM [3], focuses on spatially aligned temporal patches. This component is crucial in endowing the model with temporal awareness, a key factor in generating temporally coherent videos.

$$\text{Att}_{iv}(Q, K, V) = \text{softmax}\left(\frac{Q_t^v (K_t^v)^T}{\sqrt{c}}\right) V_t^v \quad (4)$$

$$\text{Att}_{cv}(Q, K, V) = \text{softmax}\left(\frac{Q_t^v \left([K_t^{v-1}, K_t^{v+1}]\right)^T}{\sqrt{c}}\right)$$
$$\cdot [V_t^{v-1}, V_t^{v+1}]$$
$$(5)$$

Here Q_t^v, K_t^v, V_t^v represent the queries, keys, and values within frame t and view v, respectively.

3.2.2 Two-Stage Pipeline

To enhance the generation quality, we further adopt a two-stage training and inference pipeline. By bypassing the temporal-aware modules, our model could also operate as a multi-view image generator, which enables a unified architecture for two-stage video generation.

During training, we first train a separate set of weights dedicated to multi-view image generation. Then, as illustrated in Fig. 2, we train the second stage video generation weights, by concatenating a conditioned image alongside the diffused input. This conditioned image is integrated only with the first frame, while subsequent frames employ zero padding. Notably, in this second stage training, we employ ground truth images instead of the generated ones as condition. This approach equips our training process with an efficiency comparable to that of a single-stage video generation scheme.

During inference, as shown in Fig. 3, we first sample multi-view frames using the weights of the first stage. This is followed by the generation of a multi-view video, which is conditioned on the initially generated frames, employing the weights of the second stage. This two-stage pipeline significantly enhances visual fidelity, a result attributable to the decomposition of spatial and temporal synthesis processes. The efficacy of this approach and its impact on visual quality will be further demonstrated in Sec. 4.4.

3.3. Generating Controllable Driving Scene Videos

In our Panacea model, designed for the advancement of autonomous driving systems, the controllability of synthesis samples emerge as a pivotal attribute. Panacea integrates two categories of control signals: a coarse-grained global control, encompassing textual attributes, and a fine-grained layout control, which involves BEV layout sequence.

The coarse-grained global control endows the Panacea model to generate diverse multi-view videos. This is achieved by integrating CLIP-encoded [32] text prompts into the UNet, a method analogous to that used in Stable Diffusion. Benefiting from the Stable Diffusion pre-trained model, Panacea synthesizes specific driving scenes in response to textual prompts, as demonstrated in Fig. 1 (c)

The Panacea model's fine-grained layout control facilitates the generation of synthesis samples that align with annotations. We use BEV layout sequences as the condition. Specifically, for a BEV sequence of duration T, we

convert them into a perspective view and extract the control elements as object bounding boxes, object depth maps, road maps, and camera pose embeddings. Fig. 2 (d) illustrates this process, where we employ different channels, represented by distinct colors, to delineate these segmented elements. This results in layout-controlling images with 19 channels: 10 for depth, 3 for bounding boxes, 3 for road maps, and 3 for camera pose embeddings. These 19-channel images are then integrated into the UNet using the ControlNet [54] framework.

It is noteworthy that the camera pose essentially represents the direction vector [5, 41, 58], which is derived from the camera's intrinsic and extrinsic parameters. Detailed construction of the camera pose is elaborated in the supplementary. This camera pose condition is incorporated to facilitate precise control over the viewpoint.

4. Experiment

4.1. Datasets and Evaluation Metrics

We evaluate the generation quality and controllability of Panacea on the nuScenes [4] dataset.

nuScenes Dataset. The nuScenes dataset, a public driving dataset, comprises 1000 scenes from Boston and Singapore. Each scene is a 20 seconds video with about 40 frames. It offers 700 training scenes, 150 validation scenes, and 150 testing scenes with 6 camera views. The camera views overlap each other, covering the whole 360 field of view.

Generation Quality Metrics. To evaluate the quality of our synthesised data, we utilize the frame-wise Fréchet Inception Distance (FID) [12] and the Fréchet Video Distance (FVD) [43], where FID reflects the image quality and FVD is a temporal-aware metric that reflects both the image quality and temporal consistency.

We also use a matching-based [39] consistency score, View-Matching-Score (VMS), to measure the cross-view consistency of generated videos. This metric draws inspiration from a similar concept, the view-consistency-score (VCS), previously employed in BEVGen [41]. However, due to the unavailability of the original evaluation, we developed our own implementation of the VMS instead of directly using VCS.

Controllability Metrics. The controllability of Panacea is reflected by the alignment between the generated videos and the conditioned BEV sequences. To substantiate this alignment, we assess the perceptual performance on the nuScenes dataset, utilizing metrics such as the nuScenes Detection Score (NDS), mean Average Precision (mAP), mean Average Orientation Error (mAOE), and mean Average Velocity Error (mAVE). Our evaluation is two-fold: firstly, we compare the validation performance of our generated data against real data using a pre-trained perception model. Secondly, we explore the potential of augmenting

Method	Multi-View	Multi-Frame	FVD↓	FID↓
BEVGen[41]	✓			25.54
BEVControl[52]	✓		-	24.85
DriveDreamer[46]		✓	452	52.6
Panacea	✓	✓	139	16.96

Table 1. Comparing FID and FVD metrics with SoTA methods on the validation set of the nuScenes dataset.

training set as a strategy for performance enhancement.

We employ StreamPETR, a state-of-the-art (SoTA) video-based perception method, as our main evaluation tool. For the assessment of image-based generation approaches, we utilize StreamPETR-S, the single-frame variant of StreamPETR.

4.2. Implementation Details

We implement our approach based on Stable Diffusion 2.1 [34] . Pre-trained weights are used to initialize the spatial layers in UNet. During our two-stage training, the image weights of the first stage is optimized for 56k steps, and the video weights of the second stage is optimized for 40k steps. For inference, we utilize a DDIM [38] sampler configured with 25 sampling steps. The video samples are generated at a spatial resolution of 256×512, with a frame length of 8. Correspondingly, our evaluation model, StreamPETR, building upon a ResNet50[10] backbone, is trained at the same resolution of 256×512. More details can be found in supplementary material.

4.3. Main Results

4.3.1 Quantitative Analysis

Generation Quality. To verify the high fidelity of our generated results, we conduct a comparison of our approach with various state-of-the-art driving scene generation methods. For fairness, we generate the whole validation set without using any post-processing strategies to select samples. Demonstrated in Tab. 1, our approach, Panacea, showcases remarkably superior generation quality, achieving an FVD of 139 and an FID of 16.96. These metrics significantly exceed those of all counterparts, encompassing both video-based method like DriveDreamer and image-based solutions such as BEVGen and BEVControl.

Controllability for Autonomous Driving. The controllability of our method is quantitatively assessed based on the perception performance metrics obtained using StreamPETR [45]. We first generate the entire validation set of the nuSences by Panacea. Then, the perception performance is derived using a pre-trained StreamPETR model. The relative performance metrics, compared to the perception scores of real data, serve as indicators of the alignment between the generated samples and the conditioned BEV sequences. As depicted in Tab. 2, Panacea achieves a relative performance

of 68%, underscoring a robust alignment of the generated samples. Furthermore, we present the first-stage results of our approach, which attains a relative performance of 72%.

Beyond the evaluation on the validation set, the more important feature of Panacea is its ability of generating an unlimited number of annotated training samples. Capitalizing on this, we synthesized a new training dataset for nuScenes, named Gen-nuScenes, to serve as an auxiliary training resource for the StreamPETR model. Intriguingly, the StreamPETR model trained exclusively on Gen-nuScenes achieved a notable nuScenes Detection Score (NDS) of 36.1%, amounting to 77% of the relative performance compared to the model trained on the actual nuScenes training set, as shown in Tab. 3. More importantly, integrating generated data with real data propels the StreamPETR model to an NDS of 49.2, surpassing the model trained only on real data by 2.3 points. Additionally, Fig. 4 illustrates that Gen-nuScenes consistently bolsters the performance of StreamPETR across various real data ratios. These results collectively affirm that our Panacea model is adept at generating controllable multi-view video samples, constituting a valuable asset for autonomous driving systems.

We also demonstrate another application scenario of our Panacea model, which could elevate image-only datasets to video datasets by using their real images as condition. This elevation allows for the application of advanced video-based technologies. As indicated in Tab. 4, this data elevation process yields a significant improvement, evidenced by a 5.8 point increase in NDS. Unfortunately, we also note a 2.3 point decrease in mAP, which we hypothesize is due to the inferior quality of the generated samples compared to the realistic ones and a domain mismatch between the generated training data and the real validation data. We hope that future advancements in improving the quality of generated samples will ameliorate this observed degradation.

4.3.2 Qualitative Analysis

Temporal and View Consistency. As depicted in Fig. 1 (a-b), Panacea shows the ability to generate the realistic multi-view videos directly from BEV sequences and text prompts. The generated videos exhibit notable temporal and cross-view consistency. For instance, as seen in Fig. 1 (a), the car in the front view maintains its appearance while approaching. Similarly, the content across different views is coherent, and the newly generated frames align seamlessly with the conditional frames (see Fig. 1 (b)).

Attribute and Layout Control. Fig. 1 (c) illustrates the attribute control capabilities, where modifications to text prompts can manipulate elements like weather, time, and scene. This allows our approach to simulate a variety of rare driving scenarios, including extreme weather conditions such as rain and snow, thereby greatly enhancing the

Stage	Image Size	Real	Generated	NDS↑
	512×256	✓	-	34.3
Panacea S1	512×256	-	✓	24.7 (72%)
	512×256	✓	-	46.9
Panacea	512×256	-	✓	32.1 (68%)

Table 2. Comparison of the generated data with real data on the validation set, employing a pre-trained perception model. The first stage (S1) generates the single-frame image, while the second stage (S2) outputs the multi-frame video. The evaluation of image is carried out by StreamPETR-S (single frame), whereas the video data is assessed through StreamPETR.

Real	Generated	mAP↑	mAOE↓	mAVE↓	NDS↑
✓	-	34.5	59.4	29.1	46.9
-	✓	22.5	72.7	46.9	36.1
✓	✓	37.1 (+2.6%)	54.2	27.3	49.2 (+2.3%)

Table 3. Comparison involving data augmentation using synthetic data. We attempt training exclusively using synthetic data and also explore integrating it with real data.

Multi-Frame	Real	Generated	mAP↑	mAOE↓	mAVE↓	NDS↑
-	✓	-	28.6	67.9	101.6	34.3
✓	-	✓	26.3 (-2.3%)	61.1	42.9	40.1 (+5.8%)

Table 4. Effect of elevating image-only data into video dataset.
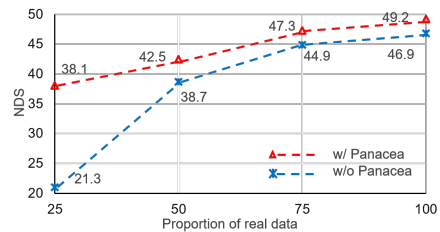

Figure 4. The detailed comparison of synthetic data augmentation across various real data ratios. We select portions of the real data at ratios of 25%, 50%, 75%, and 100%.

diversity of the data. Additionally, Fig. 5 depicts how cars and roads align precisely with the BEV layouts while maintaining excellent temporal and view consistency.

4.4. Ablation Studies

This section validates two important designs in Panacea: decomposed 4D attention and the two-stage pipeline.
Decomposed 4D Attention. We first investigate the impact of the cross-view attention mechanism in decomposed 4D attention. As illustrated in Tab. 5, the exclusion of the cross-view module results in a degradation of 108 and 5.11 in FVD and FID, respectively. This indicates the crucial role of the cross-view module in improving video quality. Additionally, to assess view consistency more precisely, we evaluate the VMS metric. Incorporating the cross-view module
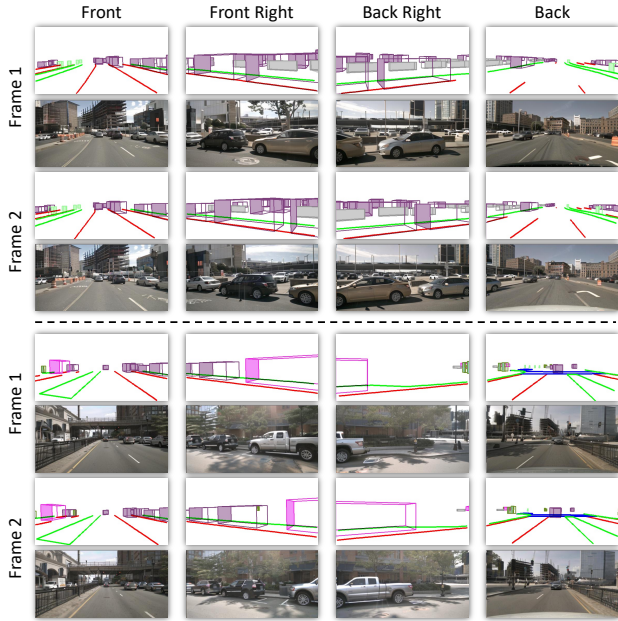
Figure 5. Controllable multi-view video generation. The column shows 4 different views and the rows shows adjacent frames aligned with BEV control.

Settings	FVD↓	FID↓
Panacea	139	16.96
w/o Cross-view	247 (+108)	22.07 (+5.11)
w/o Cross-frame	214 (+75)	20.43 (+3.47)
w/o Two stage	305 (+166)	36.61 (+19.65)

Table 5. Ablation studies on different settings. 'w/o Two stage' denotes the method where a video generation model is directly trained and inferred without an initial stage of image generation.

results in a 0.8 point enhancement, corroborating its efficacy in improving multi-view consistency, as showcased in the right of Fig. 6.

The effectiveness of the cross-view module is also illustrated in the left of Fig. 6. Without the cross-view module, the appearance of cars are inconsistent across different views. Conversely, with the integration of the cross-view module, there is a significant improvement in maintaining the consistency of cars and scenes across views.

To evaluate the temporal attention, we perform an ablation study by eliminating this component. As shown in Tab. 5, a notable degradation of 75 points in FVD is observed when the temporal module is removed, highlighting its crucial role in maintaining temporal consistency. Furthermore, Fig. 7 concisely demonstrates that without the cross-frame attention module, the model fails to retain temporal consistency, as evident in the variations of the car's appearance across frames.

Two Stage Pipeline. To confirm that our two-stage pipeline enhances generation quality, we conduct a comparison

Figure 6. Ablation on the efficiency of cross-view attention (CVA). The cars in the generated frames with or without cross-view attention are highlighted by green and pink dashed boxes, respectively.

Figure 7. Ablation study on the effects of Cross-Frame Attention (CFA). The sequential frames are displayed from left to right.

against the single-stage schema. The single-stage approach yields an FID of 36.61 and an FVD of 305, markedly inferior to the results achieved by Panacea. This comparison unequivocally underscores the critical impact of the two-stage pipeline in substantially elevating the quality of the generated videos.

5. Conclusion

We propose Panacea, a cutting-edge generator meticulously designed to create manipulable panoramic videos for driving scenarios. Within this innovative framework, we incorporate a decomposed 4D attention module to ensure temporal and cross-view consistency, facilitating the generation of realistic multi-view images and videos. Besides, a two-stage training strategy is employed to further enhance the generation quality. Significantly, Panacea is adept at handling a variety of control signals to produce videos with precise annotations. Through extensive experiments, Panacea has demonstrated its proficiency in generating high-quality, well-annotated panoramic driving-scene videos. These videos are invaluable, serving not only in BEV perception but also hold promise in real-world driving simulations. Looking ahead, we aspire to delve into the expansive potential of Panacea in real-world simulation, and integrate control signals with more diversity.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE, 2021. 5

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pages 813–824. PMLR, 2021. 5

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575. IEEE, 2023. 2, 3, 5

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628. Computer Vision Foundation / IEEE, 2020. 2, 6

[5] Dian Chen, Jie Li, Vitor Guizilini, Rares Ambrus, and Adrien Gaidon. Viewpoint equivariance for multi-view 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9213–9222. IEEE, 2023. 6

[6] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, pages 550–567. Springer, 2022. 2

[7] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *CoRR*, abs/2302.08908, 2023. 2, 3

[8] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 3

[9] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. *CoRR*, abs/2305.10474, 2023. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 6

[11] Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S Sukhatme. Neuralsim: Augmenting differentiable simulators with neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9474–9481. IEEE, 2021. 2

[12] 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. In *NIPS*, pages 6626–6637, 2017. 6

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

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. 3

[15] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 17853–17862. IEEE, 2023. 3

[16] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *CoRR*, abs/2112.11790, 2021. 2

[17] Shaofei Huang, Zhenwei Shen, Zehao Huang, Zi-han Ding, Jiao Dai, Jizhong Han, Naiyan Wang, and Si Liu. Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17451–17460, 2023. 2

[18] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: vectorized scene representation for efficient autonomous driving. *CoRR*, abs/2303.12077, 2023. 3

[19] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. *CoRR*, abs/2308.09616, 2023. 3

[20] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1042–1050, 2023. 2

[21] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *CoRR*, abs/2309.00398, 2023. 3

[22] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485. AAAI Press, 2023. 3

[23] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV (9)*, pages 1–18. Springer, 2022. 2, 3

[24] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen CLIP models are efficient video learners. In *ECCV (35)*, pages 388–404. Springer, 2022. 5

[25] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 3

[26] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 2

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *CoRR*, abs/2302.08453, 2023. 3

[28] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804. PMLR, 2022. 3

[29] Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, and Yu-Xiong Wang. Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 17928–17938. IEEE, 2023. 3

[30] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris M. Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and A baseline for temporal multi-view 3d object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 3

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 5

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 2, 3, 4, 6

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, pages 234–241. Springer, 2015. 4

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3

[37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 2, 3, 6

[39] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8922–8931. Computer Vision Foundation / IEEE, 2021. 6

[40] 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*, pages 2446–2454, 2020. 2

[41] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *CoRR*, abs/2301.04634, 2023. 2, 3, 6

[42] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *CoRR*, abs/2307.01097, 2023. 3

[43] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018. 6

[44] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *CoRR*, abs/2308.06571, 2023. 3

[45] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *CoRR*, abs/2303.11926, 2023. 2, 3, 6

[46] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *CoRR*, abs/2309.09777, 2023. 6

[47] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, pages 180–191. PMLR, 2021. 3

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *CoRR*, abs/2212.11565, 2022. 3

[49] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. LAMP: learn A motion pattern for few-shot-based video generation. *CoRR*, abs/2310.10769, 2023. 3

[50] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, Yuxin Huang, Xiaoyu Ye, Zike Yan, Yongliang Shi, Yiyi Liao, and Hao Zhao. MARS: an instance-aware, modular and realistic simulator for autonomous driving. In *CICAI (1)*, pages 3–15. Springer, 2023. 2

[51] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021. 2

[52] Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling streetview elements with multi-perspective consistency via BEV sketch layout. *CoRR*, abs/2308.01661, 2023. 2, 3, 6

[53] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 2

[54] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *CoRR*, abs/2302.05543, 2023. 2, 3, 6

[55] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. MUTR3D: A multi-camera tracking framework via 3d-to-2d queries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 4536–4545. IEEE, 2022. 3

[56] Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel C Codella, Lu Yuan, and Zheng-Jun Zha. T2d: Spatiotemporal feature learning based on triple 2d decomposition, 2023. 5

[57] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22490–22499. IEEE, 2023. 2, 3

[58] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13750–13759. IEEE, 2022. 6

[59] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *CoRR*, abs/2211.11018, 2022. 3