PHYSICAL INFORMED DRIVING WORLD MODEL

Zhuoran Yang, Yanyong Zhang *
University of Science and Technology of China
shanpoyang@mail.ustc.edu.cn, yanyongz@ustc.edu.cn

ABSTRACT

Autonomous driving requires robust perception models trained on high-quality, large-scale multi-view driving videos for tasks like 3D object detection, segmentation, and trajectory prediction. While world models provide a cost-effective solution for generating realistic driving videos, challenges remain ensuring that these videos adhere to fundamental physical principles, such as relative and absolute motion, spatial relationships like occlusion and spatial consistency, and temporal consistency. To address these, we propose **DrivePhysica**, an innovative model designed to generate realistic multi-view driving videos that accurately adhere to essential physical principles through three key advancements: (1) a Coordinate System Aligner module that integrates relative and absolute motion features to enhance motion interpretation, (2) an Instance Flow Guidance module that ensures precise temporal consistency via efficient 3D flow extraction, and (3) a Box Coordinate Guidance module that improves spatial relationship understanding and accurately resolves occlusion hierarchies. Grounded in physical principles, we achieve state-of-the-art performance in driving video generation quality and downstream perception tasks.

1 INTRODUCTION

Autonomous driving has attracted extensive attention from both industry and academia for decades Shi et al. (2016); Zheng et al. (2024a). To achieve robust perception in autonomous vehicles, models require high-quality, large-scale multi-view driving videos with labeling to train models for tasks like 3D object detection, segmentation, and trajectory prediction. World models Jia et al. (2023); Wang et al. (2023c) have emerged as a promising solution for generating diverse and realistic driving videos. They can simulate complex scenarios while addressing the high costs and labor of labeling real driving data.

However, generating realistic driving videos that strictly adhere to physical principles—such as relative motion and absolute motion understanding, temporal consistency, and spatial relationship awareness—remains a substantial challenge due to the large sampling space and limited control conditions in diffusion models. Specifically: **1)Motion Reference System Understanding:** Models often struggle to accurately interpret both relative and absolute velocities. For instance, as shown in Fig. 1 (a), models like Panacea Wen et al. (2024) fail to understand relative motion. In reality, the parked black car and the parked white car should exhibit slight movement relative to the ego vehicle. However, the black car remains stationary. These limitations in motion comprehension result in unrealistic driving videos, reducing the effectiveness of world models in perception-based tasks. **2)Temporal Consistency:** Preserving stable attributes of moving objects(such as color and texture) over time remains a challenge for numerous driving world models Gao et al. (2024); Rombach et al. (2022); Ho & Salimans (2021); Ho et al. (2020); Song et al. (2020). For instance, as depicted in Fig. 1 (b), DriveWM Wang et al. (2023d) fails to maintain temporal consistency, resulting in the car's color varying unrealistically from frame to frame. **3)Spatial Relation Understanding:** Many world models Wen et al. (2024); Gao et al. (2024) frequently misrepresent spatial relationships, including occlusion hierarchy (correct depth ordering of objects) and cross-view consistency (maintaining coherent structures across multiple camera perspectives). As shown in Fig. 1 (c), Panacea Wen et al. (2024) fails to maintain an accurate occlusion hierarchy within vehicle bounding boxes, and the vehicle generated in the bounding box of the rear parking lot was placed closer to the ego vehicle,

*Corresponding Author

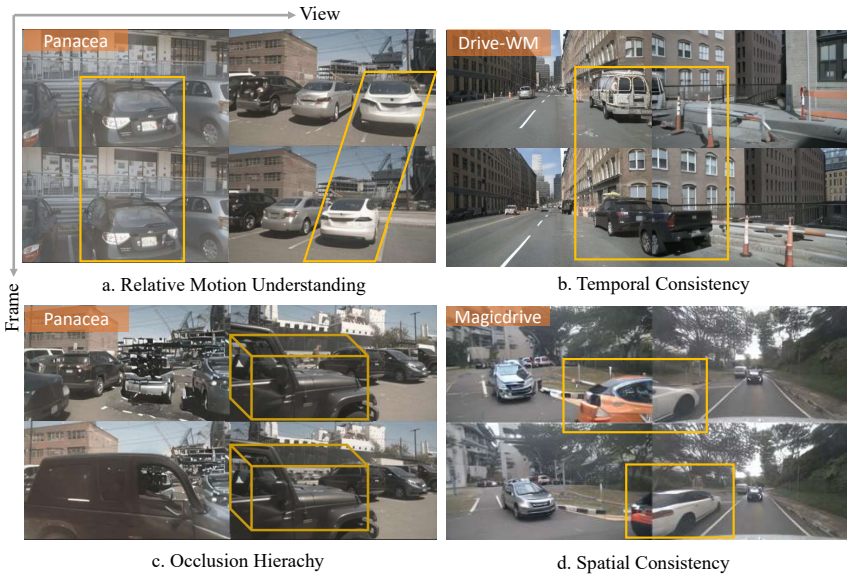


Figure 1: Limitations of previous works in modeling physical laws within driving scenarios. **(a)** Incorrect understanding of relative motion: In reality, the parked black car and the parked white car should exhibit slight movement relative to the ego vehicle. However, the black car remains stationary. **(b)** The color of the vehicle changes over time. **(c)** Incorrect understanding of the occlusion hierarchy: The box condition in the background is incorrectly generated in the foreground. **(d)** The appearance of the same car across two views is inconsistent.

appearing as a moving vehicle on the road, resulting in the wrong occlusion relationship. While in Fig. 1 (d), MagicDrive Gao et al. (2024) struggles to ensure cross-view spatial coherence.

To address these challenges, we propose **DrivePhysica**, a driving world model that effectively adheres to key physical principles, including motion reference system understanding, temporal consistency, and spatial relationship awareness. DrivePhysica achieves state-of-the-art performance in both the quality of generated videos and validation in downstream tasks.

Firstly, to help model accurately interpret the motion reference system, We introduce **C**oordinate **S**ystem **A**ligner (**CSA**) module, which uses camera pose parameters to align different conditions under the ego coordinate system and the absolute world coordinate system. Camera parameters enable the transformation of absolute world coordinates into the ego-relative coordinate system, thereby aligning the two coordinate systems successfully. CSA module provides complementary perspectives, enhancing the model's understanding of relative and absolute motion.

Secondly, to ensure temporal consistency, we introduce **I**nstance **F**low **G**uidance (**IFG**) module, a lightweight 3D flow extractor based on the motion vectors of the surrounding instances between frames, avoiding the complex 2D optical flow design used in DrivingDiffusion Li et al. (2023a).. The instance flow serves as a basis to track and propagate attributes (such as color and texture) over frames, aiding in generating temporally consistent videos. Operating in the full 3D space instead of being restricted to the 2D image plane, our instance flow enables a more accurate temporal understanding of object positioning.

Finally, to help the world model grasp spatial relationships, we propose **B**ox **C**oordinate **G**uidance (**BCG**) module to embed 3D bounding box coordinates. This direct encoding of 3D positioning helps the model capture occlusion hierarchies. We also ensure cross-view consistency through parameter-free spatial view-inflated attention.

DrivePhysica establishes a robust, physically informed foundation for generating realistic, multi-view driving videos, achieving state-of-the-art performance in both video generation quality and downstream perception task validation. Our contributions are three-fold:

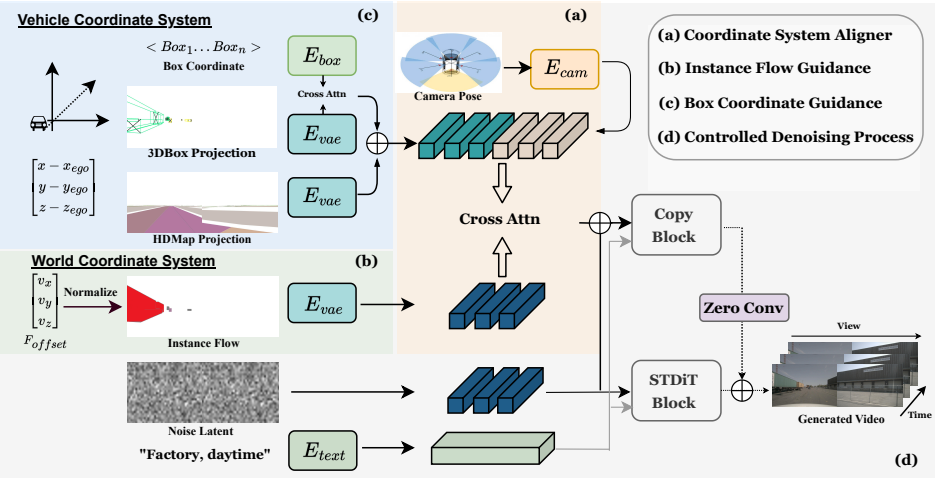


Figure 2: Overview of DrivePhysica. (a) refers to Coordinate System Aligner module, which uses camera pose parameters to align the vehicle coordinate system with the world coordinate system. (b) refers to Instance Flow Guidance module, which utilizes the instance flow to improve temporal consistency. (c) refers to Box Coordinate Guidance module, which encodes the box coordinates to provide spatial relation information. (d) refers to Controlled Denoising Process, enabled by ST-DiT with ControlNet for unified condition control.

- To enhance motion and spatial understanding, we propose a Coordinate System Aligner (CSA) module, enabling the model to better interpret the relationship between absolute and relative motion, addressing a key limitation in current driving world models. By encoding the relative 3D position of each instance in every frame, the model gains enhanced spatial awareness, effectively capturing occlusion hierarchies.
- To maintain stable object attributes over time, we introduce 3D Instance Flow Guidance with a lightweight flow extractor and operating in 3D space rather than within the 2D image plane, our approach allows for a more precise temporal understanding of object positioning.
- DrivePhysica achieves state-of-the-art (SOTA) performance in both the quality of generated videos and downstream perception metrics. DrivePhysica can simulate long-tail but critical driving scenarios, such as sudden braking and lane cutting, by leveraging synthesized conditions from the Carla Simulator.

2 METHOD

In this section, we first define the concept of physical laws that must be satisfied in the context of driving video generation, and then present DrivePhysica, a novel framework for generating realistic multi-view driving videos that faithfully adhere to essential physical principles outlined earlier.

2.1 IMPORTANT PHYSICAL LAWS TO FOLLOW

Driving video generation requires frames to be consistent with real-world physical laws. In particular, a driving world model should satisfy three key properties: *motion reference system understanding*, *temporal consistency*, and *spatial relationship awareness*.

Motion Reference System Understanding. The model must correctly interpret both the world coordinate system and the ego-vehicle coordinate system. Objects stationary in the world coordinate system may exhibit apparent motion in the ego-vehicle frame due to ego motion, and vice versa. Failing to distinguish these reference systems leads to incorrect motion patterns (Fig. 1(a)).

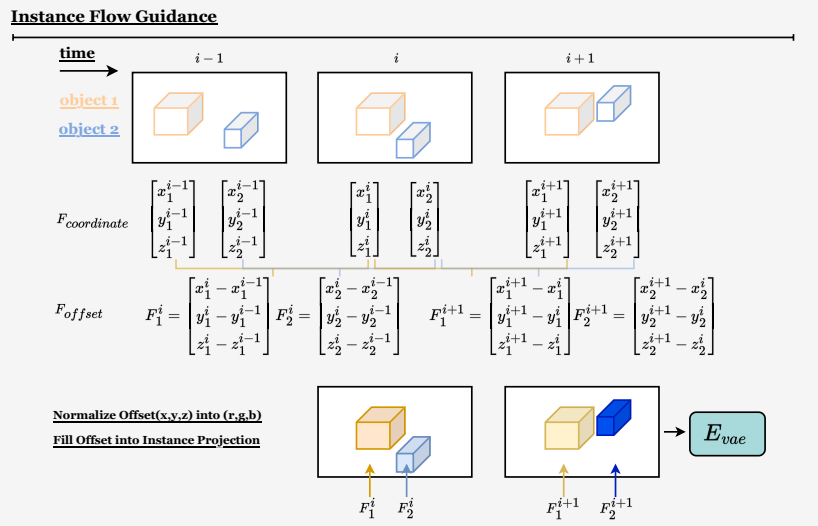


Figure 3: Overview of **Instance Flow** construction. Instance coordinates are tracked in world space, converted to relative motion offsets, projected onto the camera view, and encoded as an offset map integrated into the ST-DiT pipeline for lightweight motion control.

Temporal Consistency. Object appearance and identity should remain stable over time, following physical principles such as material invariance. For example, attributes like color and texture should not change abruptly across consecutive frames (Fig. 1(b)).

Spatial Relationship Awareness. The model should preserve correct spatial relationships, including relative distances and occlusion ordering, based on perspective geometry. Closer objects must consistently occlude farther ones, and spatial layouts should remain coherent across viewpoints (Fig. 1(c,d)).

2.2 DrivePhysica: Physical Law Acquisition

To enable the world model to accurately comprehend and adhere to the fundamental physical principles outlined in Sec.2.1 and thereby facilitate realistic driving video generation, we integrate three-level control conditions: **scene condition** (text and camera pose), **vehicle coordinate system condition** (3D bounding box coordinates, 3D bounding box projection, and road map projection), and **world coordinate system condition** (instance flow). An overview of the DrivePhysica is in Fig. 2.

2.2.1 Coordinate System Aligner

We propose a **C**oordinate **S**ystem **A**ligner (**CSA**) to align motion conditions defined in two complementary coordinate systems: the **vehicle coordinate system** and the **world coordinate system**. Vehicle-based conditions describe spatial layouts relative to the ego vehicle, while world-based conditions encode absolute motion patterns. Considering only vehicle-based conditions limits the model's ability to capture true object motion in the world coordinate system, as shown in Fig.1(a).

CSA explicitly aligns these two condition types using camera pose parameters, enabling the model to jointly reason about relative and absolute motion and thus generate physically consistent videos.

Condition Encoding. Vehicle coordinate system conditions, including 3D bounding box projections and road map projections, are encoded by a shared VAE encoder E_{vae}. Bounding box coordinates are further encoded by E_{box} (Sec. 2.2.3). Their summed embedding is denoted as $h^{vehicle}$. World coordinate system conditions, such as instance flow (Sec. 2.2.2), are encoded by the same VAE to obtain h^{world}.

Camera Pose Embedding. We represent camera pose as $\mathbf{P} = \{\mathbf{K}, \mathbf{R}, \mathbf{T}\}$, where $\mathbf{K} \in \mathbb{R}^{3\times3}$, $\mathbf{R} \in \mathbb{R}^{3\times3}$, and $\mathbf{T} \in \mathbb{R}^{3\times1}$. The parameters are concatenated as $\mathbf{P} = [\mathbf{K}, \mathbf{R}, \mathbf{T}]^T \in \mathbb{R}^{7\times3}$. We

apply Fourier embedding followed by an MLP E_{cam} to obtain the camera pose embedding:

$$h^c = E_{\text{cam}}(\text{Fourier}(\bar{\mathbf{P}})),$$

which matches the video patch embedding dimension.

Condition Alignment. CSA fuses the two condition embeddings via cross-attention:

$$h^{condition} = \text{CrossAttn}(\text{cat}(h^{vehicle}, h^c),\ h^{world}).$$

The resulting unified embedding $h^{condition}$ is fed into ControlNet to produce control signals for the ST-DiT denoiser.

By explicitly modeling camera geometry, CSA enables consistent alignment between world-level motion and ego-centric layouts, improving motion realism and temporal consistency in generated driving videos.

2.2.2 Instance Flow Guidance

To ensure *temporal consistency*, we introduce a lightweight **I**nstance **F**low **G**uidance (**IFG**) module to assist the model in maintaining the stability of object attributes (such as color and texture) over time, illustrated schematically in Fig. 3.

Instance Flow Representation. We introduce the *instance flow* condition, which refers to the motion vectors of surrounding instances in the driving scene. We first capture the spatial coordinates of surrounding objects in the absolute world coordinate system across time:

$$F_{j_{\text{coordinate}}} = \left\{(x_j^i, y_j^i, z_j^i)\right\}_{i=0}^{T-1},$$

where (x_j^i, y_j^i, z_j^i) represents the spatial position of instance j at frame i. To model the motion of the surrounding instances between consecutive frames, we define the *instance flow offset*, which encodes the motion vectors between these frames:

$$F_{j_{\text{offset}}} = \left\{(x_j^i - x_j^{i-1}, y_j^i - y_j^{i-1}, z_j^i - z_j^{i-1})\right\}_{i=1}^{T}.$$

The complete instance flow offset for all N instances at the i-th frame is defined as:

$$F_{\text{offset}}^i = \left\{(x_j^i - x_j^{i-1}, y_j^i - y_j^{i-1}, z_j^i - z_j^{i-1})\right\}_{j=0}^{N},$$

which serves as a basis to track and propagate attributes (such as color and texture) over frames, aiding in generating temporally consistent videos.

Filling Instance Offsets into Pixel Positions. Direct application of frame-to-frame offsets is incompatible with ST-DiT due to its video autoencoder and patchification process. To address this, we transform F_{offset}^i into a trajectory map $h^i \in \mathbb{R}^{H\times W\times 3}$ that aligns with the latent space of video patches. The instance's 3D bounding box is projected onto the camera view to obtain its 2D projection area. Each pixel in this area at frame i is populated with the position offset of the corresponding instance:

$$h^i(h,w) = \begin{cases} F_{j_{\text{offset}}}^i, & \text{if instance } j \text{ projects onto} \\ & (h,w) \text{ at frame } i, \\ 0, & \text{otherwise.} \end{cases}$$

The position offsets of all instances at frame i F_{offset}^i are collectively mapped to the trajectory map $h^i(h,w)$. For the first frame ($i=0$), the map is initialized as a zero matrix: $h^{i=0}(h,w) = 0$.

Normalization and RGB Encoding. The trajectory map h is normalized and converted into RGB space to create a visualized version:

$$h^{vis} = Normalize(h),$$

where the x-offset channel corresponds to the red (R) channel, the y-offset channel to green (G), and the z-offset channel to blue (B).

Using the same video encoder E_{vae} from the OpenSora framework, the visualized trajectory map h^{vis} is encoded into a latent representation:

$$h^{world} \in \mathbb{R}^{T\times H\times W\times 4}.$$

This latent representation integrates seamlessly into the ST-DiT pipeline, enabling robust and lightweight instance flow control.

2.2.3 BOX COORDINATE GUIDANCE

To help the world model understand the *spatial relation*, we introduce a **B**ox **C**oordinate **G**uidance (**BCG**) module, utilizing 3D bounding box coordinates under the vehicle coordinate system as control conditions. This captures the relative distance of instances from the ego-vehicle, providing depth cues that assist the model in understanding occlusion hierarchy.

The driving scene contains a variable number of 3D bounding boxes. We encode each bounding box i in each frame t into a hidden vector $h_t^{b_i}$, with dimensions that match those of the video patches. A 3D bounding box (c_t^i, b_t^i) consists of two types of information: the class label c_t^i and the position of the box b_t^i. For class labels, we follow a method similar to Li et al. (2023b), where the pooled embeddings of the class names, denoted $L_{c_t^i}$, are used as label embeddings. For the box positions $b_t^i \in \mathbb{R}^{8\times 3}$, which are represented by the coordinates of the 8 corner points, we apply Fourier embedding to each point and pass it through an MLP to obtain the encoded position, as described in Equation 2. We then use another MLP to combine both the class and position embeddings into a single hidden vector, as shown in Equation 3. The final hidden states for all the bounding boxes in frame t are represented as $h_{coor_t} = [h_t^{b_1}, \ldots, h_t^{b_{N_t}}]$, where N_t is the number of bounding boxes in frame t.

$$c_t^{b_i} = \text{AvgPool}(E_{\text{text}}(L_{c_t^i})), \tag{1}$$

$$p_t^{b_i} = \text{MLP}_p(\text{Fourier}(b_t^i)), , \tag{2}$$

$$h_t^{b_i} = E_{\text{box}}(c_t^i, b_t^i) = \text{MLP}_b(c_t^{b_i}, p_t^{b_i}). \tag{3}$$

We fuse the three vehicle coordinate system conditions $h_{\text{map}_{\text{proj}}}$, $h_{\text{box}_{\text{proj}}}$, and h_{coor} using cross-attention, where the box embedding serves as inputs to the attention mechanism, and the coordinate embedding serves as the key-value. The fusion process is represented as:

$$h_{\text{vehicle}} = h_{\text{map}_{\text{proj}}} + \text{CrossAttn}(h_{\text{box}_{\text{proj}}}, h_{\text{coor}}),$$

where h_{vehicle} represents the resulting vehicle coordinate system conditions feature after integrating the three input conditions.

This latent representation h_{vehicle} integrates seamlessly into the ST-DiT pipeline, enabling concise control on box coordinate and map projection. Ideally, through training, the model learns to capture instances' spatial relations and their occlusion hierarchy.

3 EXPERIMENT

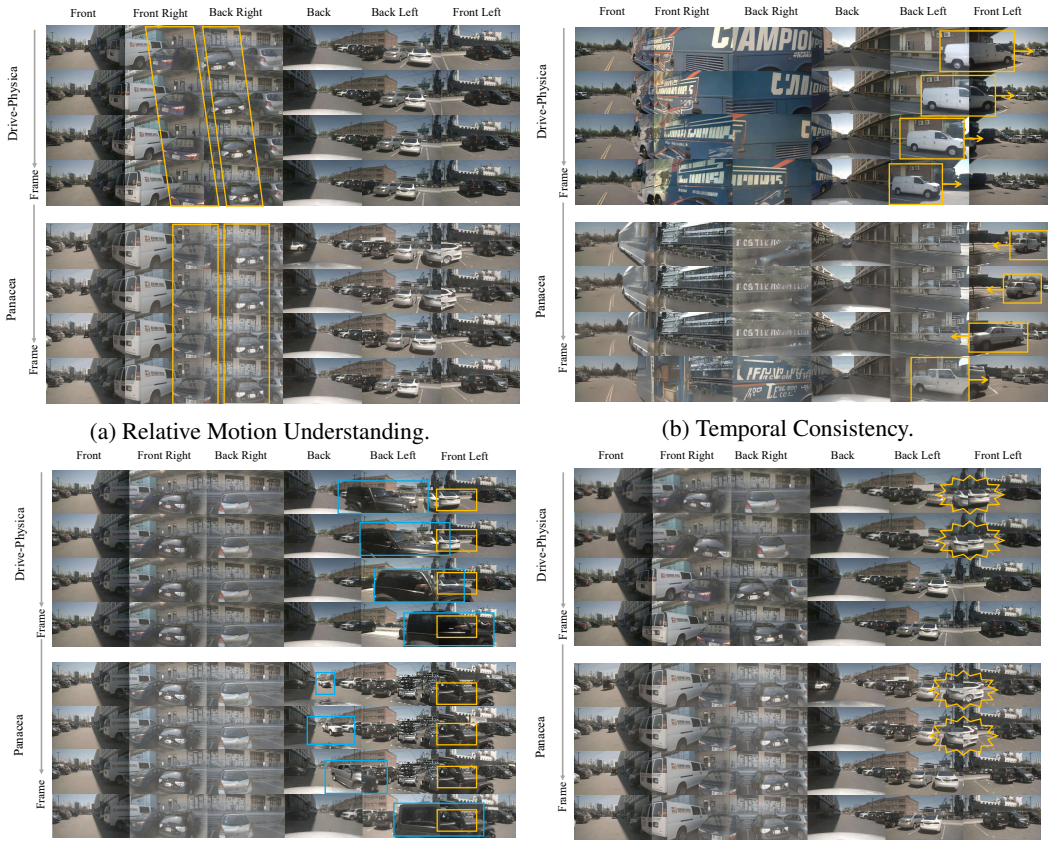
3.1 SETUPS

Metrics. We evaluate the world model's ability to learn physical laws using FID Heusel et al. (2017) and FVD Unterthiner et al. (2018). The *controllability* of DrivePhysica is demonstrated through the alignment between the generated videos and the conditioned BEV sequences. To substantiate this alignment, we evaluate perceptual performance on the nuScenes dataset using metrics such as the nuScenes Detection Score (NDS), mean Average Precision (mAP), mean Average Orientation Error (mAOE), and mean Average Velocity Error (mAVE). We investigate the potential for augmenting the training set to improve model performance. Following Panacea Wen et al. (2024), we use StreamPETR Wang et al. (2023a), a state-of-the-art (SoTA) video-based perception method, as the primary evaluation tool.

3.2 MAIN RESULTS

3.2.1 QUANTITATIVE ANALYSIS

Quality of Videos and Adherence to Physical Laws. To verify the high fidelity of our generated results, we compare our approach with various state-of-the-art driving video generation methods. For fairness, we generate the entire validation set without applying any post-processing strategies



(a) Relative Motion Understanding.

(b) Temporal Consistency.

(c) Occlusion Hierarchy.

(d) Spatial Consistency.

Figure 4: Qualitative comparison with Panacea. **(a)** Relative motion understanding: our model correctly renders background and foreground cars moving backward relative to the ego vehicle, while Panacea fails to capture the relative motion. **(b)** Temporal consistency: our model preserves the white car's shape and orientation over time, whereas Panacea shows noticeable identity changes. **(c)** Occlusion hierarchy: our model correctly places the closer moving car in front of the farther stationary car, while Panacea violates the expected occlusion order. **(d)** Spatial consistency: our model maintains consistent geometry across views, while Panacea produces inconsistent shapes.

Method	Multi-View	Multi-Frame	FVD↓	FID↓
BEVGen Swerdlow et al. (2023)	✓			25.54
BEVControl Yang et al. (2023)	✓		-	24.85
WoVoGen Lu et al. (2024)	✓	✓	417.7	27.6
Drive-WM Wang et al. (2023d)	✓	✓	122.7	15.8
DriveDreamer Wang et al. (2023c)	✓	✓	452	52.6
Panacea Wen et al. (2024)	✓	✓	139	16.96
DriveDreamer2 Zhao et al. (2024)	✓	✓	55.7	11.2
DrivePhysica	✓	✓	38.06	3.96

Table 1: Comparing FID and FVD metrics with SoTA methods on the validation set of the nuScenes dataset. We generate the entire validation set without applying any post-processing strategies to select specific samples.

to select specific samples. As shown in Tab. 1, our approach demonstrates significantly superior generation quality, achieving an FVD of 38.06 and an FID of 3.96. These metrics substantially outperform those of all counterparts, including both video-based methods like DriveDreamer-2 Zhao et al. (2024) and image-based solutions such as BEVControl Yang et al. (2023). This highlights that the videos produced by our model exhibit both higher visual quality and better temporal consistency, thereby demonstrating the model's strong capabilities in motion reference system understanding, temporal consistency, and spatial relationship awareness.

Method	Real	Generated	NDS↑
Oracle	✓	-	46.90
Panacea	-	✓	32.10 (68.00%)
DrivePhysica	-	✓	40.51 (86.38%)

Table 2: Comparison of the generated data with real data on the nuScenes validation set in (T+I)2V scenarios, employing a pre-trained perception model StreamPETR. Our model DrivePhysica achieves a relative performance of 86.38% on the nuScenes Detection Score (NDS), underscoring a robust alignment of the generated samples.

Controllability Quality for Autonomous Driving. The controllability of our method is quantitatively assessed through the perception performance metrics derived from the StreamPETR Wang et al. (2023a) framework. The relative performance metrics, compared to the perception scores of real data, serve as indicators of the control alignment between the generated videos and the control conditions. ① **Perception Validation Performance.** We generate the entire validation set of the nuSences by DrivePhysica. As depicted in Tab. 2, DrivePhysica achieves a relative performance of 86.38% on the nuScenes Detection Score (NDS), employing a pretrained perception model StreamPETR Wang et al. (2023a), underscoring a robust alignment between the generated videos and the control conditions. ② **Perception Training with Data Augmentation.** DrivePhysica can generate large-scale labeled data for training autonomous driving perception models. To evaluate its effectiveness, we synthesize a training dataset for nuScenes and retrain StreamPETR using different data configurations: real only, generated only, and their combination. All models are evaluated on the real nuScenes validation set. We first retrain StreamPETR using only real data as the baseline (Tab. 3, line 4). Compared with Panacea (line 1), our re-implementation achieves stronger performance, and DrivePhysica further improves results under this stronger setting. Training solely on generated data achieves 35.5% mAP and 43.67% NDS, corresponding to 92.69% and 90.41% of the performance obtained using only real data, respectively. This indicates that the generated dataset alone is sufficient to train competitive perception models. Combining generated data with real data yields further gains, improving NDS to 51.9, a $+3.6$ increase over training with real data only. These results demonstrate that DrivePhysica-generated data serves as an effective data augmentation strategy for perception training.

3.2.2 Qualitative Analysis

We present four qualitative comparisons of videos generated by our model and Panacea, evaluating them against the four physical laws defined in Sec. 2.1.

Relative Motion Understanding. In Fig.4(a), as the ego vehicle moves forward, the background and foreground cars should appear to move backward relative to it. In Panacea, the black car fails to exhibit correct relative motion and does not move backward as expected relative to the ego vehicle. In contrast, our model accurately captures the relative motion of each instance, demonstrating a precise understanding of both the vehicle coordinate system and the world coordinate system.

Temporal Consistency. In Fig.4(b), in Panacea, the white car's shape and orientation (e.g., the direction of the car's front head) change over time. In contrast, our model preserves the white car's attributes throughout the frame, demonstrating superior temporal consistency.

Occulusion Hierarchy. In Fig.4(c), the stationary car (controlled by the orange box condition) is positioned farther from the ego vehicle, while the moving car (controlled by the blue box condition) is closer. In Panacea, the generated video incorrectly places the farther stationary car in front, obstructing the closer moving car, therefore violating the expected occlusion hierarchy. In contrast, our model correctly renders the closer moving car in front, with the farther stationary car appropriately occluded, demonstrating a superior understanding of occlusion hierarchy.

Spatial Consistency. In Fig.4(d), in Panacea, the white car exhibits different shapes in different views, reflecting spatial inconsistency. In contrast, our model maintains a consistent spatial representation across views, ensuring coherence throughout the view.

3.2.3 Long-tail Scenarios Simulation.

Our DrivePhysica can simulate a variety of long-tail driving scenarios. We can change the weather and time of the scenes by modifying the text prompts, as shown in Fig.7.

Method	Real	Generated	mAP↑	mAOE↓	mAVE↓	NDS↑
Panacea	✓	-	34.5	59.4	29.1	46.9
Panacea	-	✓	22.5	72.7	46.9	36.1
Panacea	✓	✓	37.1 (+2.6%)	54.2	27.3	49.2 (+2.3%)
DrivePhysica (Re-Implement)	✓	-	38.3	62.1	28.8	48.3
DrivePhysica (Ours)	-	✓	35.5	59.7	29.4	43.67
DrivePhysica (Ours)	✓	✓	42.0 (+3.7%)	53.2	26.8	51.9 (+3.6%)

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

Settings	FVD↓	FID↓
DrivePhysica	38.06	3.96
w/o Coordinate System Aligner	47.34 (+9.28)	5.14 (+1.18)
w/o Instance Flow Guidance	58.86 (+20.8)	6.28 (+2.32)
w/o Box Coordinate Guidance	41.21 (+3.15)	4.22 (+0.26)

Table 4: Ablation study results in (T+I)2V scenarios on the generated nuScenes validation set.

We further leverage control conditions from the **Carla** simulator Dosovitskiy et al. (2017) to generate long-tail and safety-critical scenarios, such as sudden braking and lane cutting. Carla provides configurable conditions including 3D bounding box projections, lane line projections, and textual scene descriptions. Using these conditions, we synthesize complex driving scenarios (e.g., multi-vehicle intersections, narrow streets, and sudden obstacles) that are rare in real-world data. As shown in Fig. 8, DrivePhysica successfully generates long-tail videos consistent with the provided conditions, demonstrating its ability to generalize to diverse and challenging scenarios and supporting its applicability to autonomous driving research.

3.3 Ablation Study

We validate the effectiveness of three key modules of DrivePhysica using FID and FVD metrics in the scenario with the first frame visible (**(T+I)2V**), as shown in Tab.4.

Coordinate System Aligner. To evaluate the impact of the Coordinate System Aligner module, we conducted an ablation study by removing the camera pose injection. In (T+I)2V scenario in Tab. 4, omitting the camera pose results in a significant degradation of 9.28 in FVD and 1.18 in FID. This underscores the crucial role of the camera pose in aligning different coordinate systems. This alignment is essential for enabling the model to accurately understand and represent the position of instances within the scene. The degradation in both FVD and FID metrics highlights how the absence of this module impairs the model's ability to properly encode spatial and temporal relationships between objects, leading to a loss of motion coherence in the generated videos.

Instance Flow Guidance. To assess the effect of the Instance Flow Guidance module, we conduct an ablation study by eliminating the instance flow injection. In (T+I)2V scenario in Tab. 4, removing the instance flow results in a noticeable drop of 20.8 in FVD and 2.32 in FID, highlighting its essential role in maintaining temporal consistency.

Box Coordinate Guidance. To evaluate the influence of the Box Coordinate Guidance module, we perform an ablation study by removing the 3D bounding box coordinate condition. In (T+I)2V scenario in Tab. 4, omitting the 3D bounding box condition leads to a significant degradation of 3.15 in FVD and 0.26 in FID, emphasizing its vital role in understanding spatial relationships.

4 Conclusion

We propose **DrivePhysica**, a driving world model that integrates key physical principles to achieve SOTA performance in both video generation and downstream perception tasks. Our contributions include the Coordinate System Aligner, 3D Instance Flow Guidance, and Box Coordinate Guidance. They enable DrivePhysica to generate high-quality, multi-view driving videos while effectively capturing motion, occlusion hierarchies, and spatial relationships. Our model outperforms existing methods, demonstrating the benefits of incorporating physical principles into driving world models.

REFERENCES

Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

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. 2023b.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

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.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

Ruiyuan Gao, Kai Chen, Enze Xie, Hong Lanqing, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *ICLR*, 2024.

Xi Guo, Zhicheng Wang, Qin Yang, Weifeng Lv, Xianglong Liu, Qiong Wu, and Jian Huang. Gan-based virtual-to-real image translation for urban scene semantic segmentation. *Neurocomputing*, 394:127–135, 2020.

Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, and Qin Yang. Learning video representations of human motion from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20197–20207, 2022.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model, 2023a. URL https://arxiv.org/abs/2310.07771.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. 2023b.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation, 2024. URL https://arxiv.org/abs/2312.02934.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. *URL https://arxiv. org/abs/2204.06125*, 7, 2022.

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.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

Xiaogang Shi, Bin Cui, Gillian Dobbie, and Beng Chin Ooi. Uniad: A unified ad hoc data processing system. *ACM Transactions on Database Systems (TODS)*, 42(1):1–42, 2016.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2107–2116, 2017.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.

Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *arXiv preprint arXiv:2301.04634*, 2023.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 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:1812.01717*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *CVPR*, pp. 3621–3631, 2023a.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation, 2023b.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023c.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving, 2023d. URL https://arxiv.org/abs/2311.17918.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.

Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*, 2023.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. 2023.

David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions, 2024.

Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3801–3809, 2018.

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. *arXiv preprint arXiv:2403.06845*, 2024.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv:2402.11502*, 2024a.

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, March 2024b. URL https://github.com/hpcaitech/Open-Sora.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023.

A More Experimental Details

A.1 Architecture

Building on OpenSora V1.1 Zheng et al. (2024b), we employ a Variational Auto-Encoder (VAE) for video encoding, T5 Raffel et al. (2020) for text encoding, and Spatial-Temporal Diffusion Transformer (ST-DiT) as the foundational model for the denoising process. We reshape the input from $\mathbb{R}^{v\times t\times h\times w\times c}$ to $\mathbb{R}^{t\times h\times (wv)\times c}$ and treat wv as the frame width to improve the consistency of the cross-view.

A.2 Datasets and Baselines

We train and evaluate our model using the nuScenes dataset Caesar et al. (2020). We compare our model with image-based solutions (BEVGen Swerdlow et al. (2023), BEVControl Yang et al. (2023)) and video-based solutions (DriveDreamer Zhao et al. (2024), Panacea Wen et al. (2024)). Our method considers 10 object classes and 10 road classes , surpassing the baseline models in diversity.

A.3 Training Details

Our method is implemented based on OpenSora Zheng et al. (2024b). Initially, we train for 20k iterations on the front-view videos from the NuScenes training set. Next, to adapt to multi-view positional encoding, we froze the backbone and fine-tuned the patch embedder for 2k iterations. Finally, we added all the control modules and trained the entire model for 100k iterations with a mini-batch size of 1. All training inputs were set to 16x256x448 and conducted on 8 A100 GPUs. Additionally, during training, we set a 0.2 probability of not adding noise to the first frame and assigned a timestep of 0 to the first frame, enabling the model to have image-to-video generation capability. As a result, during testing, the model can autoregressively iterate. Experimental results show that our method can stably generate over 200 frames.

B More Quantitative Results

We employ StreamPETR Wang et al. (2023a), a state-of-the-art (SoTA) video-based perception method, as our main evaluation tool. We compare the validation performance of our generated data against real data.

Metrics. The controllability of **DrivePhysica** 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).

B.1 Perception Validation Performance.

The controllability of our method is quantitatively assessed based on the perception performance metrics obtained using StreamPETR. We generate the entire validation set of the nuSences by **DrivePhysica**. The relative performance metrics, compared to the perception scores of real data, serve as indicators of the alignment between the generated samples and the control conditions. As depicted in Tab. 2, **DrivePhysica** achieves a relative performance of 86.38% on the nuScenes Detection Score (NDS), underscoring a robust alignment of the generated samples.

B.2 Ablation Study on StreamPETR.

We further validate the effectiveness of the three key components of **DrivePhysica** through perception performance metrics obtained using StreamPETR, evaluated under the (T+I)2V scenario. As shown in Tab. 5, the perception metrics confirm the contributions of **Coordinate System Aligner**, **Instance Flow Guidance**, and **Box Coordinate Guidance**. Among these, **Instance Flow Guidance** demonstrates the greatest improvement in perception performance, indicating that this module

Stage	Real	Generated	NDS↑
	✓	-	46.90
DrivePhysica	-	✓	40.51 (86.38%)
w/o Coordinate System Aligner	-	✓	38.64 (82.39%)
w/o Instance Flow Guidance	-	✓	37.31 (79.55%)
w/o Box Coordinate Guidance	-	✓	40.23 (85.78%)

Table 5: Ablation study results in (T+I)2V scenarios, assessed based on the perception performance metrics obtained using StreamPETR.

plays the most significant role in enabling the model to generate controllable driving videos. This conclusion is consistent with the findings in Tab.4, which also highlights the critical importance of **Instance Flow Guidance** in enhancing the model's generative capabilities.

C More Visualization Results

Here, we provide additional visualization results to showcase our model's strong ability to generate high-fidelity, realistic, and diverse multi-view driving videos. We sample 8 frames from each generated video as a demo to save space in the paper. Our model is capable of generating high-quality, long-duration driving videos through iterative processing. We provide a web page in the supplementary materials for additional results.

C.1 Prompt Edit

DrivePhysica enables video editing by modifying only the text prompt condition while keeping all other conditions fixed.

In Fig. 7, we demonstrate the model's editing capability by altering the weather and time of day in the text prompt. Specifically, we add "Sunny," "Rainy," and "Night" to the original text prompt, while maintaining other conditions such as camera pose, 3D bounding box coordinates, 3D bounding box projections, road map projections, and instance flow unchanged. The generated videos showcase high quality and effective editing:

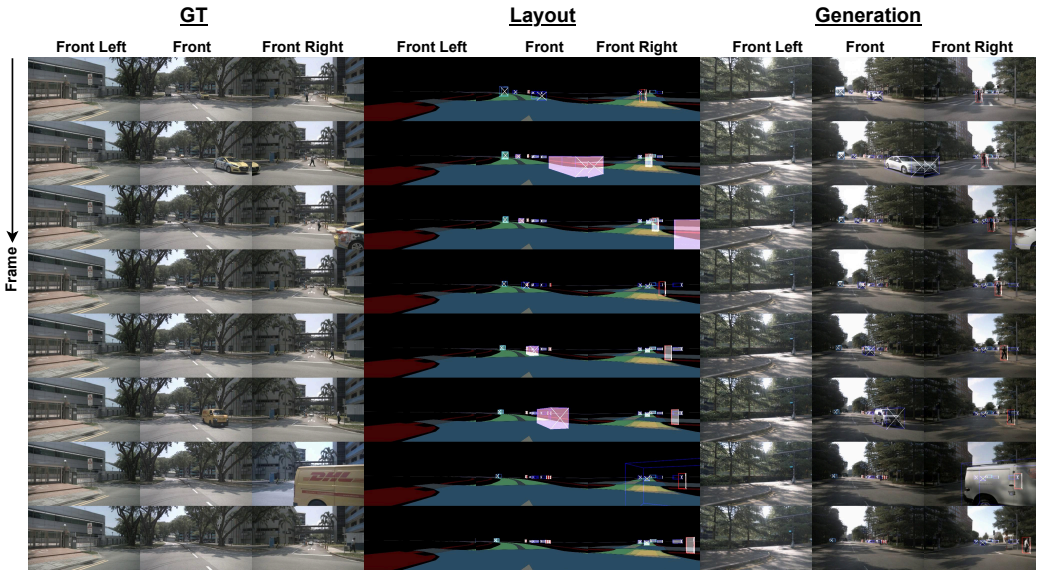
- **Sunny**: Displays clear skies with sunlight shining on the scene, reflecting bright and vivid environmental details.
- **Rainy**: Captures wet road surfaces and blurred camera views caused by raindrops, adding realistic weather dynamics.
- **Night**: Depicts dimly lit scenes with streetlights and reduced visibility, accurately simulating nighttime driving conditions.

These results emphasize the strong editing capability of **DrivePhysica**, producing diverse and realistic driving videos with minimal changes to the input conditions.

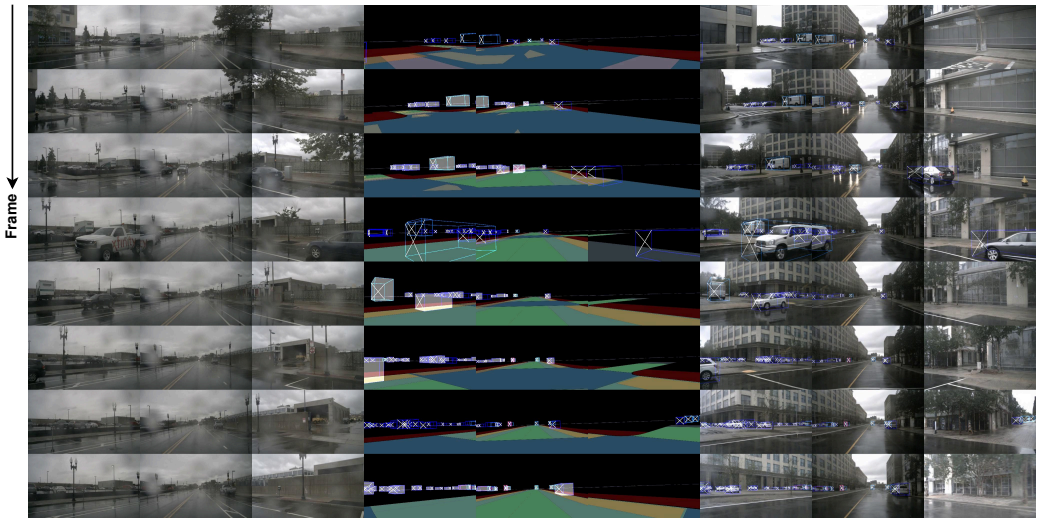
C.2 Control Precision

DrivePhysica excels at generating videos that adhere closely to various control conditions, including 3D bounding box coordinates, 3D bounding box projections, road map projections, and instance flow. In Fig. 5, we overlay the 3D bounding box projections onto the generated videos to illustrate the precision of our control mechanisms. The precision of control is reflected in:

- **Object alignment with bounding box projections**: Objects in the scene are accurately placed and sized to align with their *projected bounding boxes*, as shown in Fig. 5 (a)(b).
- **Road and pedestrian area fidelity**: Drivable areas, sidewalks, and zebra crossings are faithfully generated following the *road map projections*, as shown in Fig. 5 (a)(b).
- **Precision in dense and small objects**: Small and densely packed objects are precisely rendered at their correct locations, following *3D bounding box coordinates*, as shown in Fig. 5 (b).



(a) Objects track their attributes maintaining temporal consistency.



(b) Small and densely packed objects rendered at correct locations.

Figure 5: Precise control mechanisms. We overlay the 3D bounding box projections onto the generated videos. The precision of control is reflected in: **(1)** Objects in the scene are accurately placed and sized to align with their *projected bounding boxes*, as shown in 5a and 5b. **(2)** Drivable areas, sidewalks, and zebra crossings are faithfully generated following the *road map projections*, as shown in 5a and 5b. **(3)** Objects track their previous attributes as guided by the *instance flow*, ensuring temporal consistency across frames. As shown in Figure 5a, the pink-rendered instance flow directs the model to generate the white sedan, maintaining its consistent attributes over time. **(4)** Small and densely packed objects are precisely rendered at their correct locations, following *3D bounding box coordinates*, as shown in 5b.

- **Temporal consistency through instance flow**: Objects track their previous attributes as dictated by the *instance flow*, enabling consistent temporal consistency across frames, as shown in Fig. 5 (a).

These results highlight the superior control and fidelity of **DrivePhysica** in generating realistic and controllable driving videos.

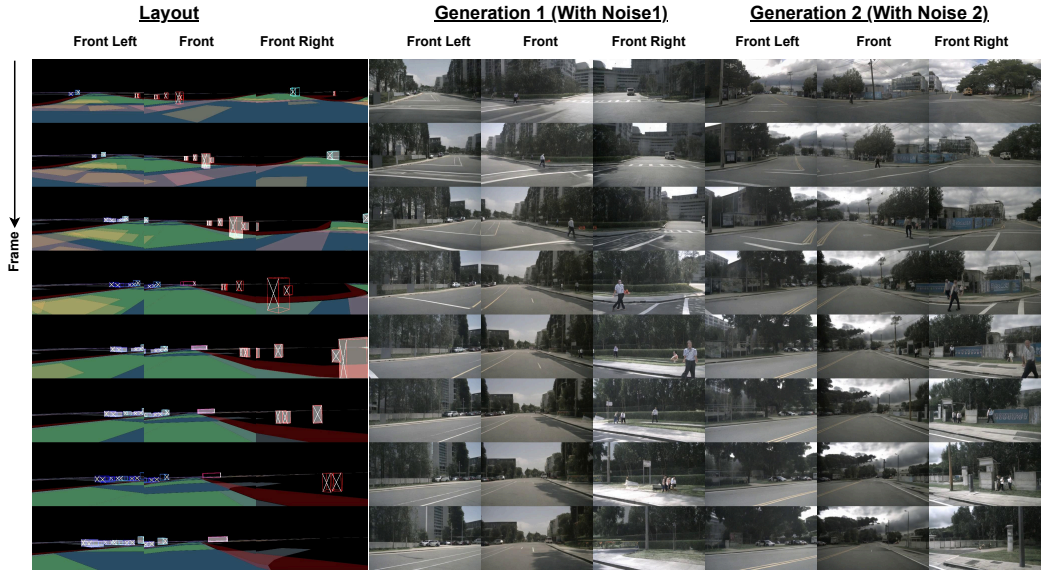


Figure 6: Diverse videos using **varying noise inputs** and **the same control conditions**. By introducing stochastic noise while maintaining consistent control signals—such as 3D bounding box coordinates, lane line projections, and instance flow—our model can produce a variety of videos that adhere to the defined constraints.

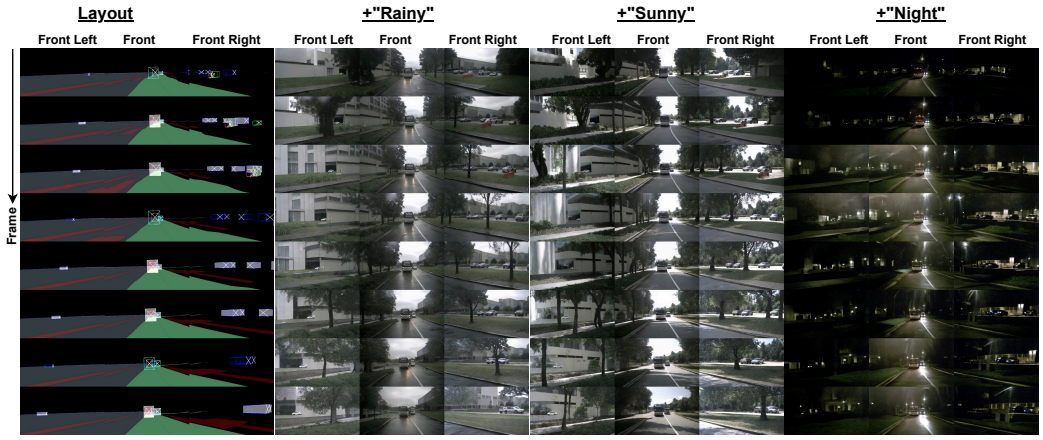
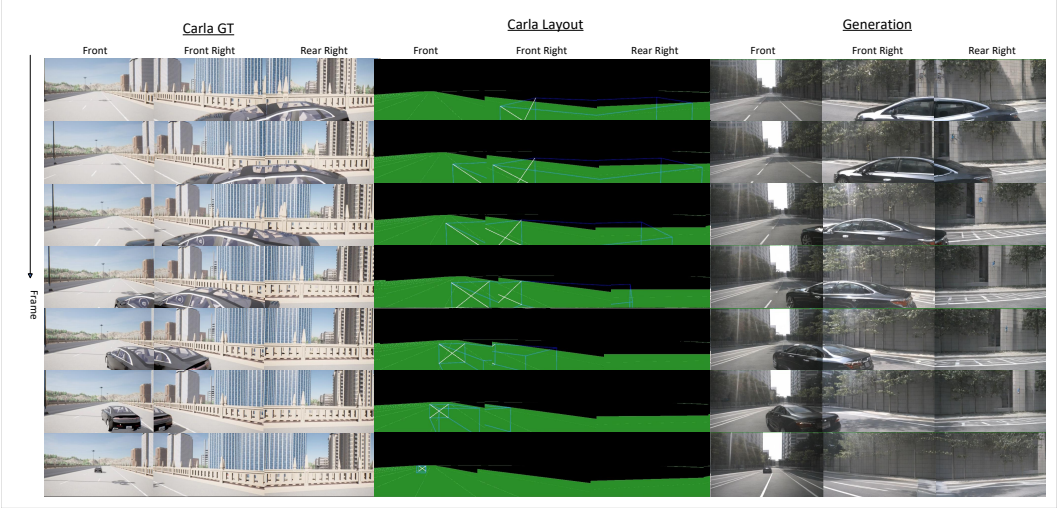


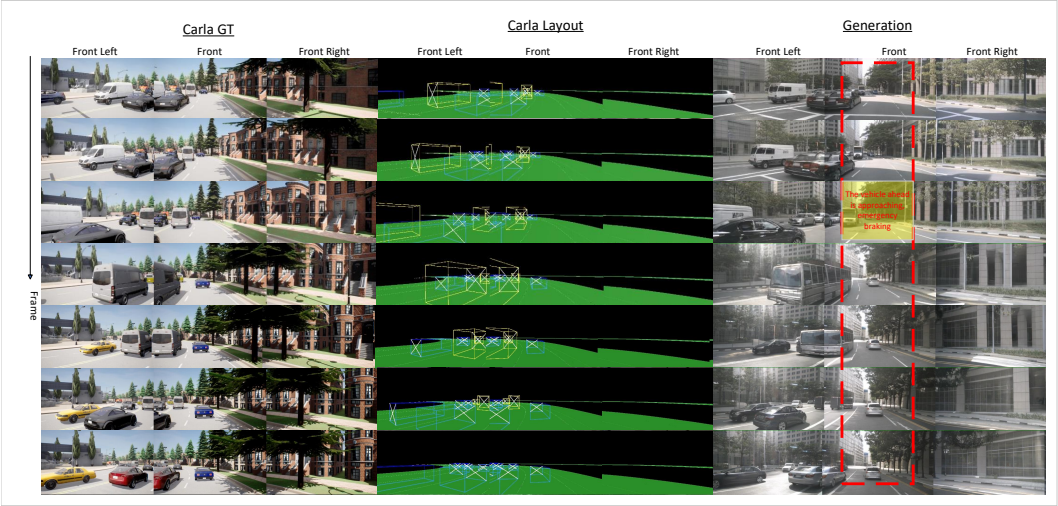
Figure 7: DrivePhysica's **editing** capability by adjusting the weather and time of day. By adding "Rainy," "Sunny," and "Night" to the original text prompt, while keeping other conditions (such as camera pose, 3D bounding box coordinates, 3D bounding box projections, road map projections, and instance flow) unchanged, our model represents strong ability to edit videos effectively. (a) "Rainy": Captures wet road surfaces and blurred camera views caused by raindrops, adding realistic weather dynamics. (b) "Sunny": Displays clear skies with sunlight shining on the scene, reflecting bright and vivid environmental details. (c) "Night": Depicts dimly lit scenes with streetlights and reduced visibility, accurately simulating nighttime driving conditions.

C.3 CARLA-GENERATED LAYOUT CONTROL

DrivePhysica demonstrates the ability to generate high-quality driving videos based on layout conditions provided by the **Carla** simulator Dosovitskiy et al. (2017), which include 3D bounding box projections, lane line projections, and scene description text prompts. The use of Carla-generated layouts addresses a critical limitation in real-world driving video datasets: the lack of diversity in scene types, especially for rare but critical events like lane cutting and sudden braking. By leveraging



(a) Lane cutting scenario.



(b) Sudden braking scenario.

Figure 8: DrivePhysica's ability to generate **rare but critical driving scenarios** based on layout conditions provided by the Carla simulator. The use of Carla-generated layouts addresses a critical limitation in real-world driving video datasets: the lack of diversity in scene types, especially for rare or challenging corner cases.

Carla's highly configurable simulation environment, we can create synthetic layouts that represent complex and diverse driving scenarios, such as multi-vehicle intersections, narrow streets, or sudden obstacles, which are difficult to capture in real-world data.

In Fig. 8, we showcase our model's ability to generate rare videos corresponding to these layouts. The generated videos highlight **DrivePhysica**'s capacity to faithfully adhere to the control signals while producing realistic outputs. Moreover, by effectively handling corner cases, our approach bridges the gap in scene diversity, making it a valuable tool for training and validating driving models under challenging scenarios. The results demonstrate that **DrivePhysica** can not only replicate realistic conditions but also adapt seamlessly to a wide range of complex layouts generated by Carla, further enhancing its applicability in autonomous driving research.

C.4 DIVERSITY WITH VARYING NOISE

DrivePhysica demonstrates the ability to generate diverse driving videos from identical control conditions with **varying noise inputs**, as illustrated in Fig. 6. By introducing stochastic noise while maintaining consistent control signals—such as 3D bounding box coordinates, lane line projections, and instance flow—our model produces a variety of plausible video outputs that adhere to the defined constraints.

D PRELIMINARY

Latent Video Diffusion Model (LVDM). The LVDM enhances the stable diffusion model Ramesh et al. (2022) by integrating a 3D U-Net, thereby empowering efficient video data processing. This 3D U-Net design augments each spatial convolution with an additional temporal convolution and follows each spatial attention block with a corresponding temporal attention block. It is optimized by employing a noise-prediction objective function:

$$l_\epsilon = ||\epsilon - \epsilon_\theta(z_t, t, c)||_2^2, \tag{4}$$

Here, $\epsilon_\theta(\cdot)$ signifies the 3D U-Net's noise prediction function. The condition c is guided into the U-Net using cross-attention for adjustment. Meanwhile, z_t denotes the noisy hidden state, evolving like a Markov chain that progressively adds Gaussian noise to the initial latent state z_0:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{5}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_t)$ and β_t is a coefficient that controls the noise strength in step t.

Spatial-Temporal DiT (ST-DiT). The ST-DiT Peebles & Xie (2023) introduces a novel architecture that merges the strengths of diffusion models with transformer architectures Vaswani et al. (2017). This integration aims to address the limitations of traditional U-Net-based latent diffusion models (LDMs), improving their performance, versatility, and scalability. While keeping the overall framework consistent with existing LDMs, the key shift lies in replacing the U-Net with a transformer architecture for learning the denoising function $\epsilon_\theta(\cdot)$, thereby marking a pivotal advance in the realm of generative modelling.

The ST-DiT architecture incorporates two distinct block types: the Spatial DiT Block (S-DiT-B) and the Temporal DiT Block (T-DiT-B), arranged in an alternating sequence. The S-DiT-B comprises two attention layers, each performing Spatial Self-Attention (SSA) and Cross-Attention sequentially, succeeded by a point-wise feed-forward layer that serves to connect adjacent T-DiT-B block. Notably, the T-DiT-B modifies this schema solely by substituting SSA with Temporal Self-Attention (TSA), preserving architectural coherence. Within each block, the input, upon undergoing normalization, is concatenated back to the block's output via skip-connections. Leveraging the ability to process variable-length sequences, the denoising ST-DiT can handle videos of variable durations.

During processing, a video autoencoder Yu et al. (2023) is first employed to diminish both spatial and temporal dimensions of videos. To elaborate, it encodes the input video $X \in \mathbb{R}^{T\times H\times W\times 3}$ into video latent $z_0 \in \mathbb{R}^{t\times h\times w\times 4}$, where L denotes the video length and $t = T, h = H/8, w = W/8$. z_0 is next "patchified", resulting in a sequence of input tokens $I \in \mathbb{R}^{t\times s\times d}$. Here, $s = hw/p^2$ and p denote the patch size. I is then forwarded to the ST-DiT, which models these compressed representations. In both SSA and TSA, standard Attention is performed using Query (Q), Key (K), and Value (V) matrices:

$$Q = W_Q \cdot I_{norm}; K = W_K \cdot I_{norm}; \tag{6}$$

Here, I_{norm} is the normalized I, W_Q, W_K, W_V are learnable matrices. The textual prompt is embedded with a T5 encoder and integrated using a cross-attention mechanism.

E RELATED WORKS

Controllable Generation. The emergence of diffusion models Zhang et al. (2024) has driven substantial advancements in text-to-video generation An et al. (2023); Blattmann et al. (2023b); Guo et al. (2023); He et al. (2022); Ho et al. (2022); Blattmann et al. (2023a); Singer et al. (2022); Wang et al. (2023b); Zhou et al. (2023). Among these, Video LDM Blattmann et al. (2023b) leverages

a latent diffusion framework that performs denoising in the image latent space, significantly accelerating the generation process. However, text prompts alone are insufficient for precise video control. Subsequent approaches introduced image blocks in conjunction with textual prompts for the denoising network Zhang et al. (2024). In our work, we target the generation of highly realistic street-view videos, which present unique challenges due to their complex environments, including intricate street layouts and dynamic vehicles. To achieve fine-grained control, we go beyond text and image inputs by incorporating road maps, 3D bounding boxes, and BEV keyframes, enabling detailed and accurate video generation.

Multi-View Video Generation. Multi-view video generation is challenged by the need for both multi-view consistency and temporal consistency. MVDiffusion Tang et al. (2023) addresses multi-view consistency through a correspondence-aware attention module that aligns information across views. Tseng et al. Tseng et al. (2023) utilize epipolar geometry to enforce view-to-view regularization. Similarly, MagicDrive Gao et al. (2024) enhances consistency by leveraging priors such as camera poses, bounding boxes, and road maps, along with an additional cross-view attention block. However, these methods primarily focus on generating multi-view images rather than videos and often depend on supplementary data, such as camera poses, which may not be readily available. In contrast, our approach is designed to tackle video generation without reliance on such constraints, offering a more streamlined and efficient solution.

Street-View Generation. Street-view generation methods commonly rely on 2D layouts, including BEV maps, 2D bounding boxes, and semantic segmentation. BEVGen Swerdlow et al. (2023) encodes all semantic information in BEV layouts for street-view generation, while BEVControl Yang et al. (2023) employs a two-stage pipeline to produce multi-view urban scene images. BEVControl's controller generates foreground and background objects, while its coordinator ensures cross-view visual consistency. However, the projection of 3D information into 2D layouts results in the loss of geometric details, making these methods prone to temporal inconsistencies when extended to video generation. To address this, we condition generation on 3D bounding boxes, preserving geometric fidelity across frames. While DrivingDiffusion Li et al. (2023a) adopts a multi-stage pipeline involving multiple models and extensive post-processing, our approach simplifies the workflow through an efficient, end-to-end framework, ensuring both temporal coherence and computational efficiency.

Simulation-to-Real Visual Translation. Recent advancements in leveraging synthetic data for real-world visual tasks have seen significant progress across various domains. Notably, methods like GAN-based translation Guo et al. (2020) and domain randomization Tobin et al. (2017) have bridged the gap between synthetic and real-world data distributions. Synthetic datasets such as Synthia Ros et al. (2016) and Virtual KITTI Cabon et al. (2020) have provided scalable benchmarks for semantic segmentation and autonomous driving. Adversarial training approaches Shrivastava et al. (2017); Zhang et al. (2018) have enhanced domain adaptation by reducing distribution discrepancies. Furthermore, human motion representation learning Guo et al. (2022) has demonstrated the utility of synthetic data in video understanding and biomechanics. These works collectively illustrate the potential of synthetic-to-real transfer in improving model robustness and addressing data scarcity challenges in visual tasks. Unlike these methods, we only extract proxy data such as 3D bounding boxes and road map from the graphics system. Then, utilizing the DrivePhysica model, we can generate more realistic and diverse videos.

F LIMITATIONS

Our work establishes a robust, physically informed framework for generating high-quality, multi-view driving videos, achieving state-of-the-art performance in both video generation quality and downstream perception task validation. However, certain limitations remain, mainly due to time and resource constraints.

Currently, our model's design has not been exhaustively optimized, leaving room for improvement in the quality of the generated videos. For example, the training process is conducted at a relatively low spatial resolution of 256×448, which constrains visual fidelity. Scaling to higher resolutions would require fine-tuning the position embeddings to ensure compatibility, an aspect not yet addressed in this work.

Future research could explore the integration of more advanced generative models, such as SD-XL Podell et al. (2023), and develop more efficient methods to produce high-fidelity videos at larger spatial resolutions. Additionally, the computational cost of inference for DrivePhysica is relatively high, which presents another avenue for improvement. Enhancing the efficiency of DrivePhysica will be a key focus in future developments to make the model more practical for real-world applications.