RAYNOVA: Scale-Temporal Autoregressive World Modeling in Ray Space

Yichen Xie[1,2*], Chensheng Peng[1,2,*], Mazen Abdelfattah[1], Yihan Hu[1], Jiezhi Yang[1], Eric Higgins[1]
Ryan Brigden[1], Masayoshi Tomizuka[2], Wei Zhan[1,2†]
[1] Applied Intuition, [2] UC Berkeley

Abstract

World foundation models aim to simulate the evolution of the real world with physically plausible behavior. Unlike prior methods that handle spatial and temporal correlations separately, we propose RAYNOVA*, a geometry-agonistic multiview world model for driving scenarios that employs a dual-causal autoregressive framework. It follows both scale-wise and temporal topological orders in the autoregressive process, and leverages global attention for unified 4D spatio-temporal reasoning. Different from existing works that impose strong 3D geometric priors,* RAYNOVA *constructs an isotropic spatio-temporal representation across views, frames, and scales based on relative Plücker-ray positional encoding, enabling robust generalization to diverse camera setups and ego motions. We further introduce a recurrent training paradigm to alleviate distribution drift in long-horizon video generation.* RAYNOVA *achieves state-of-the-art multi-view video generation results on nuScenes, while offering higher throughput and strong controllability under diverse input conditions, generalizing to novel views and camera configurations without explicit 3D scene representation. Our code will be released at https://raynova-ai.github.io/.*

1. Introduction

> Space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.
>
> — *Hermann Minkowski*

World foundation models (WFMs) [4, 18, 28, 39, 40] aim to simulate the evolution of complex scenes within a dynamic real world governed by physical laws. In a continuous 4D spatio-temporal world, a world model is designed to generate physically plausible images or videos captured by cameras positioned at specific locations and timesteps.

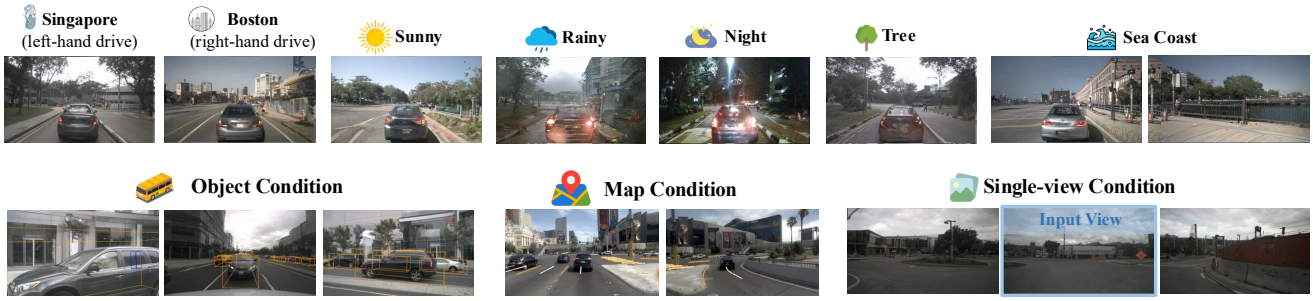
To impose physical plausibility, prior work commonly considers spatial and temporal constraints separately. For spatial reasoning, some methods leverage positional correlations across multiple viewpoints, including camera adjacency [12, 68] and cross-view projection [17]. For temporal coherence, previous works assume strong multi-frame dependencies among images captured by the same cameras [49, 67, 68] and thereby exploit existing techniques from video generation, such as temporal VAEs [3, 74]. However, this decoupled design fundamentally restricts the model's flexibility to handle novel sensor configurations and rapid camera motions. Alternatively, some approaches construct explicit 3D scene representations (*e.g.* point clouds [70, 80] or BEV/volume features [35, 78]), within a unified global coordinate system to enhance spatio-temporal consistency. Although empirically effective within constrained domains, this geometry-forcing approach impedes generalization to unconstrained open-world environments beyond the distribution of training scenarios.

To overcome these limitations, we pursue a universal framework that retains physical plausibility with minimal inductive biases. We propose RAYNOVA, a world foundation model supported by a *dual-causal autoregressive strategy* over both scale and time. The scale-wise autoregression is built up on "next-scale prediction" visual autoregressive model [19, 55], enabling the model to capture intricate visual patterns across multiple levels of abstraction. The temporal autoregression works frame-by-frame in the unified 4D spatio-temporal space to improve temporal coherence.

The core insight of RAYNOVA lies in a novel spatio-temporal conditioning mechanism that represents the physical world with *relative positional encoding in camera ray space*. This ray-level representation is isotropic in continuous 4D space, thereby reducing dependency on specific camera configurations, motion patterns, or view overlaps. By encoding relative rather than absolute positions, it naturally supports extrapolation beyond the training distribution and generalizes to unconstrained real-world conditions. Leveraging this representation, RAYNOVA performs coherent reasoning across multi-view and multi-frame images, accommodating diverse input conditions within a unified

*Equal Contribution. Work done during internship at Applied Intuition.
†Correspondence: wei.zhan@applied.co

(a) RAYNOVA is not bound to any specific sensor setups, allowing flexible video generation in continuous 4D world space.



(b) RAYNOVA takes diverse input conditions with high fidelity.

(c) RAYNOVA can generate long-horizon multi-view videos in a large range.

Figure 1. Demonstrations of RAYNOVA as Versatile World Foundation Model.

4D geometric space. We highlight the following aspects of our method:

- **Versatile World Foundation Model.** Our method supports diverse input and output formats for various use cases with a single model. Video generation can be conditioned on multiple optional control signals (*e.g.*, texts, objects, maps, images), while output formats remain flexible in terms of perspectives, resolutions, and frame rates.
- **Scalable Data-Driven Framework.** Our training paradigm is capable of ingesting heterogeneous training data from diverse sources with different sensor configurations. Unlike prior works with handcrafted biases [7, 17], RAYNOVA does not require auxiliary supervision such as depth maps, point clouds, or optical flow.
- **Extendable Position Embedding.** Our relative ray-level positional encoding supports extrapolation beyond the

training range, theoretically allowing an unbounded spatial extent. We also develop a recurrent training paradigm to mitigate distribution drift in long video generation.

- **Efficient Video Generation.** Our hierarchical multi-scale representation enables rapid progression from coarse abstractions to fine-grained details, and can be seamlessly integrated with existing acceleration techniques for visual autoregressive models [16].

We evaluate RAYNOVA on driving datasets, demonstrating superior performance in both video generation and novel view synthesis. The results demonstrate high visual realism, strong spatio-temporal coherence, and high fidelity to multiple control signals together with substantial computational efficiency. These advantages enhance its practicality for physical-world applications such as autonomous driving simulation.

2. Related Works

Video Generative Models. Recent advances in generative models have pushed the fidelity and diversity of monocular video synthesis to new heights. Early transformer-based, masked autoregressive approaches operate on discrete tokens to generate high-quality frames in sequence [14, 37, 58, 71, 76]. In parallel, diffusion-based pipelines iteratively denoise latent or pixel representations, yielding vivid motion and appearance [1, 2, 15, 36, 40, 43, 47, 74]. In just a few years, these models excel at visual quality, offering realistic and vivid generated videos. Most of these approaches focus on text-to-video tasks, generating videos conditioned on textual prompts [15, 36, 74]. Some other models also generate videos conditioned on reference images [46, 65], or reference videos [32]. In this paper, we go beyond video generation for a versatile world foundation model with flexible multi-view camera setups, ego motions, and input conditions.

3D Scene Representation Learning. To more accurately simulate the geometry and dynamics of real-world scenes, several approaches leverage explicit 3D priors to enhance the geometric consistency, typified by NeRF [27, 38] and Gaussian splatting [10, 41, 73, 79] variants. Despite high-quality image renderings, their inherent inductive biases limit the adaptability of models to complex scenarios beyond the training distribution. On the other hand, several approaches work on the geometry-free representation with improved generalization ability, most of which resort to camera ray space [22, 25, 26, 50]. However, despite these advances, they still rely on the camera rays in the global coordinate, which serves as another bottleneck of generalization for extrapolation beyond the training distribution. In contrast, we exploit the relative positions in the camera ray space for spatio-temporal grounding with minimal inductive biases, which boosts the video generation in a large range.

World Models for Autonomy. Autonomous driving serves as an ideal testbed for physically grounded world models due to its stringent demands for modeling complex geometric environments, dynamic actor interactions, and the need for coherent spatio-temporal representations. Early efforts [23, 64] demonstrate action-conditioned video generation but are constrained primarily to a single front-view perspective. Later studies [63, 66, 68] have extended video generation to multi-view cameras with diverse input conditions such as textual descriptions, object bounding boxes, and HD maps. To enhance geometric consistency and structural controllability, the community commonly introduces various inductive biases. Several approaches [12, 66, 68] count the spatial consistency of the mutual condition between adjacent cameras, but this design binds the model to some fixed camera configurations. Other efforts incorporate explicit 3D scene representations such as point clouds [70, 78, 80], occupancy grids [35], 3D latent features [13, 81] to ensure spatial alignment in world space. However, the reliability of the above 3D structural priors highly depends on the overlap between camera perspective views. These representations also limit the world models to a constrained 3D space, hindering generalization in the open world. In contrast, our approach uniquely learns a continuous 4D latent representation based on the relative position in the camera ray space that jointly encodes spatial and temporal dimensions, enforcing multi-view spatial consistency and frame-to-frame temporal coherence with minimal hand-crafted biases.

3. Methodology

RAYNOVA, as shown in Fig. 2, is a versatile world foundation model with autoregression. After introducing preliminary "next-scale prediction" work (Sec. 3.1), we formulate the dual-causal autoregressive framework for multi-view video generation in Sec. 3.2. Critical for coherent spatio-temporal reasoning, we construct an isotropic 4D position embedding based on the relative positions between camera rays in Sec. 3.3. This enables the development of a transformer-based geometry-free framework with minimal inductive biases for world modeling that takes diverse control signals in Sec. 3.4. Finally, to alleviate the distribution drift in the autoregressive generation of long videos, we develop a recurrent training paradigm that aligns the distributions in the training and inference stages (Sec. 3.5).

3.1. Preliminary: Next-Scale Prediction

Unlike vanilla autoregressive models [29, 75] that flatten the 2D grids of images into 1D tokens, some recent work [19, 55] shifts from "next-token prediction" to "next-scale prediction" strategy. Each image is quantized by the tokenizer into K multiscale token maps $X_{1:K} = (X_1, X_2, \ldots, X_K)$ with increasingly higher reso-

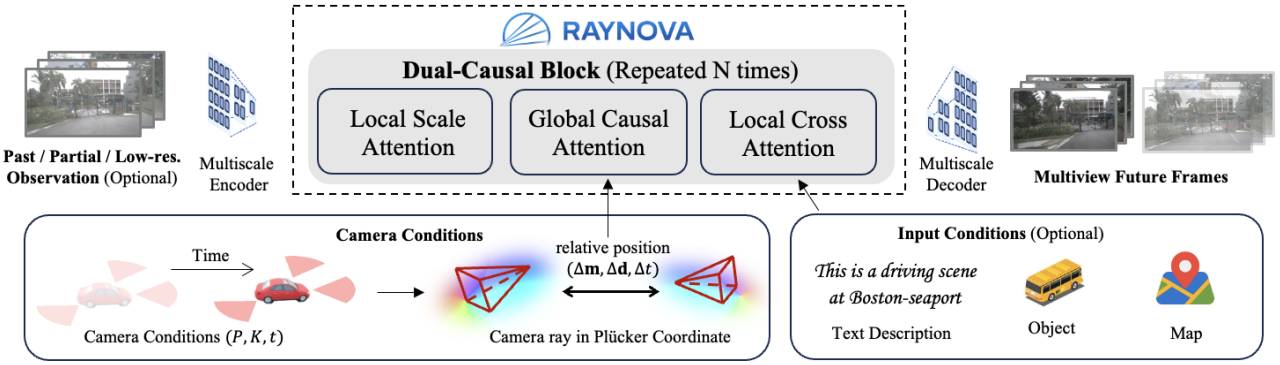


Figure 2. Overview of RAYNOVA Framework. RAYNOVA is composed of dual-casual (scale and time) blocks. The local scale attention and local cross attention works on each image indepedently, while the global causal attention works across multi-view and multiframe images enhanced with a unified ray-level relative position embedding for better spatio-temporal consistency.

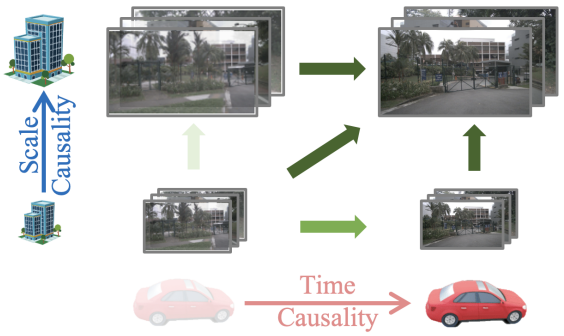


Figure 3. Dual-Causality for Multi-View Video Generation. Green arrows represent the causal dependency, while the darkness indicates the topological order of autoregression (from light to dark).

lutions $h_k \times w_k, k = 1, 2, \ldots, K$. The autoregressive likelihood is formulated as follows.

$$p(X_1, X_2, \ldots, X_K) = \prod_{k=1}^{K} p(X_k | X_1, X_2, \ldots, X_{k-1}) \quad (1)$$

where X_k is the token map at scale k containing $h_k \times w_k$ visual tokens. The sequence $(X_1, \ldots, X_{k-1})$ is the prefix context for the prediction of X_k. During the k-th autoregressive step, all $h_k \times w_k$ tokens are generated in parallel.

3.2. Formulation of Dual-Causal Autoregression

We formulate our proposed RAYNOVA for world modeling via multi-view videos beyond image-level "next-scale prediction". In favor of versatility, each image $I^{v,t}$ is tokenized independently as a multiscale token map $X^{v,t}_{1:K}$ for each frame $t = 1, \ldots, T$ and each camera view $v = 1, \ldots, V$. To model the joint distribution of multi-view and multi-frame images in a unified space, our autoregressive model considers two folds of causality: *scale* and *time* (Fig. 3).

For the scale causality, we start from a single-frame case. The distribution of multi-view images in the same frame is modeled in a joint manner, since multi-view images describe a unified 3D space at a specific time step as a whole. For clarity, we omit the timestep superscripts.

$$p(X_1^{1:V}, \ldots, X_K^{1:V}) = \prod_{k=1}^{K} p(X_k^{1:V} | X_1^{1:V}, \ldots, X_{k-1}^{1:V}). \quad (2)$$

For time causality, the generation of each frame is conditioned on historical information. Different from some prior works [49, 68], we do not assume strong inter-dependency among multi-frame images from the same camera since this inductive bias cannot be maintained with complex ego-motions (*e.g.* turning). In contrast, we condition the multi-view image generation for the current timestep on all the views from past frames.

$$p(X_{1:K}^{1:V,1:T}) = \prod_{t=1}^{T} p(X_{1:K}^{1:V,t} | X_{1:K}^{1:V,1:t-1}). \quad (3)$$

Combining the scale and time causality from Eq. 2 and Eq. 3, we formulate the dual-causal block in RAYNOVA as

$$p\left(X_{1:K}^{1:V,1:T}\right) = \prod_{t=1}^{T} \prod_{k=1}^{K} p\left(X_k^{1:V,t} | X_{1:k-1}^{1:V,1:t}\right). \quad (4)$$

It models the joint distribution of multi-view and multi-frame images with both spatial and temporal consistency.

3.3. Isotropic Spatio-Temporal Representation

Inductive biases, such as multi-view adjacency or 3D latent features, in previous work limit their generalization in open world and adaptivity to heterogeneous data. Alternatively, we propose an isotropic 4D position embedding by extending the relative rotary position embedding [52] to the camera ray space shared by all the multi-view and multi-frame visual tokens to enhance spatial and temporal consistency.

To this end, we start from a token-wise Plücker ray [42] for the multi-scale feature maps. For convenience, we define a global 3D coordinate space whose origin is the ego-vehicle position at the first timestep. The extrinsic parameters for each camera view v at each timestep t are denoted as $\mathbf{P}^{\mathbf{v},\mathbf{t}}$, while the intrinsic parameter is denoted as $\mathbf{K}^{v,t}$.[1] For each visual token $\mathbf{x}_k^{v,t}$ at scale k, it is straightforward to compute the corresponding Plücker ray $\mathbf{p}_k^{v,t}$ based on the camera parameters and its location in the image space. We further abuse this notation by appending an extra timestep t, so the extended Plücker ray is written as

$$\mathbf{p}_k^{v,t} = (\mathbf{m}_k^{v,t} \in \mathbb{R}^3, \mathbf{d}_k^{v,t} \in \mathbb{R}^3, t) \tag{5}$$

where $\mathbf{m}_k^{v,t} = \mathbf{o}^{v,t} \times \mathbf{d}_k^{v,t}$, camera center $\mathbf{o}^{v,t}$ is in the global coordinate system, and unit vector $\mathbf{d}_k^{v,t}$ is the ray direction.

Despite an intuitive approach that attaches the encoded Plücker ray to the corresponding visual token [8, 26], this absolute position embedding has limited extrapolation ability beyond the scene range in the training stage, thereby hindering the model's generalization. To this end, we introduce *relative embedding* that conditions networks on the relative position between camera rays using the modified self-attention blocks, where the attention score logit between two visual tokens is calculated as:

$$a_{i,j} = \left(\mathbf{R}_{k_i}^{v_i,t_i}\mathbf{q}_{k_i}^{v_i,t_i}\right)^T \left(\mathbf{R}_{k_j}^{v_j,t_j}\mathbf{k}_{k_j}^{v_j,t_j}\right) = \mathbf{q}_{k_i}^{v_i,t_i\,T}\mathbf{R}_{\Delta}^{i,j}\mathbf{k}_{k_j}^{v_j,t_j} \tag{6}$$

where $\mathbf{R}_{\Delta}^{i,j} = \mathbf{R}_{k_i}^{v_i,t_i\,T}\mathbf{R}_{k_j}^{v_j,t_j}$ reflects the relative position between visual token i and j in the camera ray space. We derive $\mathbf{R}_k^{v,t}$ by extending RoPE to 7D space based on the Plücker ray (Eq. 5). For clarity, we omit the superscripts and subscripts for scale k, camera view v, and timestep t. The rotary position encoding is written as:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R_m} & 0 & 0 \\ 0 & \mathbf{R_d} & 0 \\ 0 & 0 & \mathrm{RoPE}_{\frac{d}{4}}(t) \end{bmatrix} \tag{7}$$

where $\mathrm{RoPE}_{\frac{d}{4}}(t) \in \mathbb{R}^{\frac{d}{4}\times\frac{d}{4}}$ is the 1D rotary position embedding [52] for time step and d is head dimension in the multi-head self-attention module [57]. $\mathbf{R_m}$ and $\mathbf{R_d}$ refer to the rotary position encoding for $\mathbf{m}$ and $\mathbf{d}$ terms in the Plücker ray. Given their small scales, we multiply a scalar factor to each of them separately.

$$\mathbf{R_m} = \begin{bmatrix} \mathrm{RoPE}_{\frac{d}{8}}(\mathbf{m}_1) & 0 & 0 \\ 0 & \mathrm{RoPE}_{\frac{d}{8}}(\mathbf{m}_2) & 0 \\ 0 & 0 & \mathrm{RoPE}_{\frac{d}{8}}(\mathbf{m}_3) \end{bmatrix}$$
$$\mathbf{R_d} = \begin{bmatrix} \mathrm{RoPE}_{\frac{d}{8}}(\mathbf{d}_1) & 0 & 0 \\ 0 & \mathrm{RoPE}_{\frac{d}{8}}(\mathbf{d}_2) & 0 \\ 0 & 0 & \mathrm{RoPE}_{\frac{d}{8}}(\mathbf{d}_3) \end{bmatrix} \tag{8}$$

This relative position embedding considers visual tokens from all scales, views, and frames in a unified 4D space with minimal inductive biases, enabling generalization over long ranges and horizons. The model can adaptively learn interactions between visual tokens across spatial and temporal domains, and accommodate heterogeneous camera parameters and frame rates in training and inference.

3.4. Dual-Causal Block Architecture Design

We build a transformer-based framework for high-quality multi-view video generation upon a pretrained "next-scale prediction" image generation model [19] with minimal 3D inductive bias. In each transformer block, visual tokens are sequentially passed through an *image-wise self-attention module* for image realism, a *global self-attention module* for spatio-temporal consistency, and an *image-wise cross-attention module* for conditional fidelity, as shown in Fig. 2.

Image-wise self-attention for image realism. The image-wise attention follows the design in Infinity [19]. Each image is processed independently, conditioned on its prefix scales, regardless of frames and camera views. Axial 2D RoPE [20] on the image space is adopted, which is normalized to a specific resolution for multi-scale features.

Global self-attention for spatio-temporal consistency. Previous methods [48, 67, 68] often apply separate spatial and temporal attention to reduce computational complexity. However, this decoupled design includes some prior assumptions about the camera motion and scene dynamics. In contrast, we propose a unified global attention based on the formulation in Eq. 4 where each visual token attends to all its prefix in both scale-wise and temporal topological orders from all the camera views (Fig. 3). The relative position embedding explained in Sec. 3.3 is also employed in this module to serve as a unified spatio-temporal representation. This module is implemented with masked self-attention during training, while the mask can be eliminated through proper KV-cache during inference.

Image-wise condition alignment. We apply an image-wise cross-attention module for input conditions, including text descriptions, 3D object bounding boxes, and HD maps.

[1] Although the intrinsic parameters are commonly fixed along the temporal dimension, we keep the superscript t to tolerate special varying cases.

This module works on each image independently. The text descriptions are processed through a T5 encoder [45]. For each object bounding box per frame, we project its eight corners to camera image space. It is possible that each box may be projected to one or more camera views. The projected coordinates of eight corners in the image space are encoded with MLP and then concatenated with the object category embedding from T5 encoder [45] to obtain a high-dimensional latent feature. For the HD map, we sample several 3D points for each map element. In contrast to directly encoding the structural map element (*e.g.* polylines or masks), this sampling approach is more flexible when dealing with unstructured or sketched maps. The points are projected to camera image space similar to object bounding box corners. The project coordinates of points belonging to the same map element are encoded with a lightweight PointNet-style network [44], and are then concatenated with the map element category embedding from the T5 encoder [45] to obtain a high-dimensional feature. The latent features for text, object boxes, and HD maps are combined in the cross-attention module, allowing the visual tokens to attend to them simultaneously. For both objects and map conditions, we employ an Axial 2D RoPE [20, 52] between the features and tokens based on their projected centers in the image space. Experiments show that this relative position embedding can notably improve the locality of conditions.

Given the multi-view order-invariant nature of the above modules, we attach a frame-wise local absolute Plücker ray embedding in the beginning of our model to distinguish different views. The image-wise self-attention and cross-attention modules inherit the pretrained weights from Infinity [19], while the global self-attention module is randomly initialized and combined with zero-initialization [77]. This alternating local-global module can help balance the spatio-temporal consistency and per-image realism.

3.5. Recurrent Training for Long-Horizon Videos

Given the GPU memory constraints, our autoregressive model training is limited to a few frames. Although it is extendable during inference through sliding windows, the generation of long-horizon videos suffers from distribution drift due to the gap between training and inference stages.

To bridge this gap, we propose a novel training paradigm to align the training and inference distributions (Alg. 1). It recurrently conducts forward and backward propagation frame by frame, and the gradients are accumulated until the end of each long sequence for model optimization. Obviously, the later frames are conditioned on the earlier frames. We cache the latent features in the global self-attention module since it is the only module operating across frames (Sec. 3.4). While previous works often cache the keys and values separately [11, 51] during inference, we directly store the latent features before KV projection layer instead.

Algorithm 1: Recurrent Training for Long-Horizon Video Generation

Require: Number of video frames T
Require: Maximal cache length M
Require: Multi-view video generation model that returns latent feature cache G_θ^L
Require: KV projection Proj_θ^{KV} (included in G_θ)

loop
  Initialize KV latent feature cache $\mathbf{L}_{KV} \leftarrow []$
  for $t \in 1, \ldots, T$ **do**
    $\tilde{\mathbf{x}}^t = \text{Random\_Bitwise\_Error}(\mathbf{x}^t)$
    # Add random noise in scale-wise features to simulate the prediction errors
    $\mathbf{KV} = \text{Proj}_\theta^{KV}(\mathbf{L}_{KV})$
    $\hat{\mathbf{x}}^t, \mathbf{l}_{KV}^t \leftarrow G_\theta^L(\tilde{\mathbf{x}}^t; \mathbf{KV})$
    if *length*$(\mathbf{L}_{KV}) > M$ **then**
      $\mathbf{L}_{KV}$.pop(0)
    $\mathbf{L}_{KV}$.append $(\text{stop\_grad}(\mathbf{l}_{KV}^t))$
    # Cache latents $\mathbf{L}_{KV}$ instead of $\mathbf{KV}$ to obtain gradient for Proj_θ^{KV} in past frames
    $loss^t = \text{Cross\_Entropy}(\hat{\mathbf{x}}^t, \mathbf{x}^t)$
    Accumulate gradient of $loss^t$
    # Per-frame backpropogation to save VRAM
  Update parameter θ
  Clear accumulated gradient

In this case, we can 1) save half of the GPU memory and 2) ensure the KV projection layer in the computational graph, which is critical for temporal information extraction.

However, there is an error accumulation issue during inference since the "next-scale prediction" framework takes ground-truth visual tokens as inputs during training. To this end, we draw inspiration from Infinity [19] to randomly incorporate some mistakes in the visual token inputs during training to simulate the prediction errors during inference.

4. Experiments

4.1. Experimental Setups

Datasets. RAYNOVA can scale up to heterogeneous training data with varying camera configurations, resolutions, or frame rates. We combine two public datasets (nuScenes [5] and nuPlan [6]), which are converted to a unified ScenarioNet [30] format. NuScenes [5] dataset contains ~5 hours of driving data collected by 6 cameras with bounding box and map annotations, and scene descriptions from OmniDrive [62]. For nuPlan [6] dataset, we filter the entire dataset based on complete sensor coverage and extract ~55 hours of data from 8 cameras with auto-labeled bounding boxes and maps. We apply GPT-4o mini [54] for scene descriptions. RAYNOVA can be further scaled to larger-scale

Table 1. Multi-view Video Generation Performance.

Method	Resolution	Metrics		
		FID↓	FVD↓	Throughput↑
MagicDrive [12]	224×400	16.2	-	1.76
X-Drive [70]	224×400	16.0	-	0.83
DriveDreamer [63]	256×448	14.9	341	0.37
BEVWorld [78]	-	19.0	154	-
Panacea [68]	256×512	17.0	139	0.67
DrivingDiffusion [31]	512×512	15.8	332	-
RayNova	384×672	**10.5**	**91**	**1.96**

data, with details in the supplementary materials.

Evaluation Metrics. To evaluate synthetic image quality, we adopt Frechet Video Distance (FVD) [56] and Frechet Inception Distance (FID) [21]. We also evaluate fidelity to conditions through perception models using NDS for objects and mIoU for map. Unless otherwise specified, we conduct the evaluation on the nuScenes validation set [5].

Implementation Details. We build RayNova on two variants of the pretrained Infinity [19] model with 130M and 2B parameters respectively. Our training pipeline includes three stages: 1) We start from training the model on low-resolution (192 × 336) multi-view short clips 2) The model is further finetuned on short clips of high-resolution (384 × 672) multi-view images. 3) Finally, we adopt the recurrent training in Sec. 3.5 on long sequence of driving scenarios to learn the long-term temporal coherence. All the training is conducted on 32 NVIDIA A100 GPUs.

4.2. Video Generation Performance

Results and Comparisons. In Tab. 1, we compare the multi-view generation performance of RayNova with other existing world models. We do not include in-house data in our model training, which makes our training data scale less than most of other approaches. Results show that we attain FID and FVD much better than all the baselines, demonstrating better image quality and temporal coherence. Notably, RayNova delivers a throughput of **1.96 images/s**, much faster than diffusion baselines, demonstrating the efficiency of our autoregressive architecture. It is also worth mentioning that our model applies an image-based VAE encoder-decoder to minimize the inductive bias relevant to camera setups and ego-motions, although this design choice may hurt the FID and FVD metric. We also provide some qualitative visualizations in Fig. 1. It can be observed that our model can generate under different conditions. Using dual-causal autoregression, RayNova not only matches or exceeds the performance of models trained on orders of magnitude more private data, but also operates at higher frame rates.

Object & Map Conditions. Our model takes optional object and map inputs as extra local geometric conditions to control the multi-view video generation. In addition to qualitative visualization in Fig. 1, we provide some additional quantitative results of object and map condition fidelity. We apply some state-of-the-art perception models [33, 61, 69] to our synthetic images in nuScenes validation set. Both vision-based and multi-modal (combined with ground-truth point clouds) perception models are considered in the evaluation to comprehensively evaluate both the semantic and geometric realism. It is worth mentioning that StreamPETR takes long sequences as input, requiring not only object realism in each single frame but also temporal coherence over long sequences. In Tab. 2 and 3, our synthetic images show great fidelity to both objects and maps notably better than baselines. For map condition, the projection of coordinates to camera space is not perfect due to the lack of height information in the map annotation, but RayNova still shows robust performance. For object condition, our synthetic images can reach a performance similar to real images.

Motion Cues in Synthetic Videos. In addition to visual quality and condition fidelity, we further evaluate whether

Table 2. Fidelity to Object Condition

(a) Camera (StreamPETR[61])

Method	NDS↑
Oracle	46.9
Panacea [68]	32.1 (68%)
RayNova	**41.9** (89%)

(b) Multisensor (SparseFusion[69])

Method	NDS↑
Oracle	72.8
X-Drive [70]	69.6 (95%)
RayNova	**72.0** (99%)

Table 3. Fidelity to Map Condition

(a) Camera (BEVFusion-C[33])

Method	mIoU↑
Oracle	57.1
MagicDrive [12]	28.9 (51%)
RayNova	**34.9** (61%)

(b) Multisensor (BEVFusion[33])

Method	mIoU↑
Oracle	63.0
MagicDrive [12]	47.0 (75%)
RayNova	**49.9** (79%)

Table 4. Motion Cues in Generated Videos

Method	Motion Planning (L_2 ↓)		
	$1s$	$2s$	$3s$
Oracle	0.41	0.70	1.05
RayNova	**0.42** (98%)	**0.73** (96%)	**1.09** (96%)

Table 5. Novel View Synthesis Performance with Camera Shifts.

Method	Shift $1m$		Shift $2m$		Shift $4m$	
	FID↓	FVD↓	FID↓	FVD↓	FID↓	FVD↓
StreetGaussian [72]	32.12	153.45	43.24	256.91	67.44	429.98
OmniRe [9]	31.48	152.01	43.31	254.52	67.36	428.20
FreeVS [60]	51.26	431.99	62.04	497.37	77.14	556.14
RAYNOVA	**14.11**	**117.20**	**15.58**	**122.60**	**17.48**	**143.48**

Table 6. Ablation Study on Camera Ray.

S.-T. Module	Pos. Emb.	FID↓	FVD↓
✗	-	18.7	214
✓	Absolute	**17.2**	138
✓	Relative	**17.2**	**124**

Table 7. Ablation Study on Scale Causality and Model Size.

Past-frame Condition	Size	FID↓	FVD↓
same scale	130M	20.7	151
all scales	130M	60.4	454
prefix scales	130M	17.2	124
prefix scales	2B	**10.5**	**91**

our synthetic videos support plausible driving decisions by applying an end-to-end planner (VAD [24]) pretrained on real scenes to our generated videos. Tab. 4 shows that the pretrained planner derives actions from synthetic videos consistent with real scenes. This reflects not only visual realism but also physical and motion plausibility in our generated videos.

Zero-shot to Unseen Camera Configurations. The ray-level relative position embedding allows RAYNOVA to generalize to unseen camera configurations in a zero-shot manner. An example of sensor setups different from nuScenes and nuPlan is visualized in Fig. 6 of the appendix.

4.3. Novel View Synthesis

The relative position embedding in camera ray space allows our model to synthesize images for novel views. We follow [60] to evaluate the performance of RAYNOVA for novel view synthesis task on nuScenes validation set. Different shifts $\{1m, 2m, 4m\}$ are applied to the camera viewpoints. We measure the FID and FVD metrics between the synthetic shifted multi-view videos and the original real videos. In Tab. 5, our FID and FVD metrics only drop slightly after viewpoint shift. Compared to baselines, RAYNOVA exhibits a significant performance gain even if it does not include any 3D inductive bias such as radiance field or point clouds. These results demonstrate the great capability of RAYNOVA in camera controllability and novel view synthesis in a geometry-free data-driven manner.

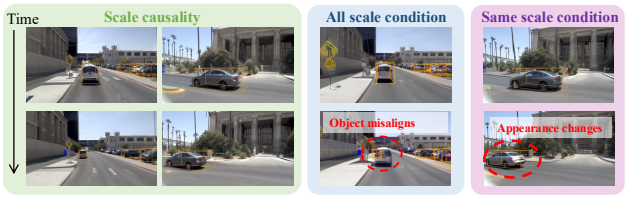


Figure 4. Ablation Study on Scale Causality. Conditioning on all scales in history hurts the modeling of dynamics, while conditioning only on same scale is insufficient for temporal coherence.

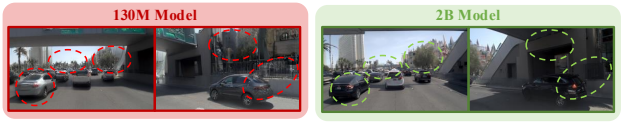


Figure 5. Ablation Study on Model Size. Large scale model can bring significantly better visual quality.

4.4. Ablation Study

Scale Causality. RAYNOVA is built up on scale and time dual-causality. The temporal causality is inevitable if we would like to extend to long video autoregression. For scale causality (Eq. 4), different from low-scale condition in previous frames, we consider two alternatives conditions: 1) all scales or 2) same scale features in previous frames. Results are shown in Tab. 7 and Fig. 4. Higher scales features of previous time steps cause future frames simply copying the details without simulating the dynamics, and same scale condition is far from enough for stable spatio-temporal modeling.

Relative Ray Embedding. One of our key contributions lies in the relative position encoding in the camera ray space, which is critical for the spatio-temporal consistency. To justify our design, we consider several alternatives in Tab. 6. We find that this relative position embedding can perform better than absolute Plücker ray embedding, let alone none spatio-temporal reliance.

Recurrent Training. The recurrent training stage is greatly helpful for temporal coherence in video generation. It improves FVD from 100 to **91** within hundreds of iterations.

Model Size. Our minimal reliance on inductive bias provides a data-driven framework. Large model can exhibit significantly superior performance. We compare two variants of our model with 130M and 2.8B parameters. In Fig. 5

and Tab. 7, with the same training data, larger model enables much better image quality.

5. Conclusion

This paper introduces RAYNOVA, a versatile autoregressive 4D world foundation model. Built from dual-causal blocks, RAYNOVA performs unified 4D spatio-temporal reasoning by autoregressing along both the *scale* and *time* topological orders. To enhance spatio-temporal consistency, we design a novel isotropic relative position embedding in the continuous 7D Plücker ray space, which represents multi-view, multi-frame, and multi-scale visual features in a unified manner. To mitigate distribution drift in long-horizon video generation, we further adopt a recurrent training paradigm that better aligns the distributions of training and inference. Experiments demonstrate that RAYNOVA can generate realistic multi-view videos with high visual fidelity, strong spatio-temporal coherence, and low latency under diverse conditioning signals, while generalizing to novel views and unseen camera configurations.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3

[2] 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*, 2023. 3

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 1

[4] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 1

[5] 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*, pages 11621–11631, 2020. 6, 7

[6] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles, 2022. 6

[7] Anthony Chen, Wenzhao Zheng, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Kurt Keutzer, and Shanghang Zhang. Geodrive: 3d geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421*, 2025. 2

[8] Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024. 5

[9] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024. 8

[10] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction, 2025. 3

[11] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 6

[12] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 1, 3, 7, 13

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

[14] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *ECCV*, 2022. 3

[15] 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, 2024. 3

[16] Hang Guo, Yawei Li, Taolin Zhang, Jiangshan Wang, Tao Dai, Shu-Tao Xia, and Luca Benini. Fastvar: Linear visual autoregressive modeling via cached token pruning. *arXiv preprint arXiv:2503.23367*, 2025. 3, 14

[17] Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, and Hao Zhao. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation, 2025. 1, 2

[18] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 1

[19] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 1, 3, 5, 6, 7, 13

[20] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024. 5, 6

[21] 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 equilib-

rium. *Advances in neural information processing systems*, 30, 2017. 7

[22] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[23] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving, 2023. 3

[24] Bo et al. Jiang. Vad: Vectorized scene representation for efficient autonomous driving. In *ICCV*, 2023. 8

[25] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025. 3

[26] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024. 3, 5

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

[28] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 1

[29] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 3

[30] Quanyi Li, Zhenghao Mark Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in neural information processing systems*, 36:3894–3920, 2023. 6

[31] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pages 469–485. Springer, 2024. 7

[32] Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stablizing shape consistency in video-to-video editing, 2024. 3

[33] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 7

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 13

[35] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*, 2023. 1, 3

[36] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation, 2025. 3

[37] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *ICLR*, 2023. 3

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[39] Nvidia. World foundation model. https://www.nvidia.com/en-us/glossary/world-models/, 2024. 1

[40] OpenAI. Sora: Creating video from text, 2024. 1, 3, 13

[41] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes. *arXiv preprint arXiv:2411.11921*, 2024. 3

[42] Julius Plucker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, pages 725–791, 1865. 5

[43] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv*, 2024. 3

[44] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 6

[45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6

[46] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation, 2024. 3

[47] Runway. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha, 2024. 3

[48] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025. 5

[49] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025. 1, 4

[50] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd Van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in neural information processing systems*, 35:9512–9524, 2022. 3

[51] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using

model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. 6

[52] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4, 5, 6

[53] 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. 13

[54] OpenAI Team. Gpt-4o mini: advancing costefficient intelligence, 2024. 6

[55] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 1, 3, 13

[56] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[58] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023. 3

[59] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024. 14

[60] Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. *arXiv preprint arXiv:2410.18079*, 2024. 8

[61] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3621–3631, 2023. 7

[62] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22442–22452, 2025. 6

[63] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 3, 7

[64] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving, 2023. 3

[65] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation, 2024. 3

[66] 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, 2023. 3

[67] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024. 1, 5, 13

[68] 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*, pages 6902–6912, 2024. 1, 3, 4, 5, 7, 13

[69] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17591–17602, 2023. 7

[70] Yichen Xie, Chenfeng Xu, Chensheng Peng, Shuqi Zhao, Nhat Ho, Alexander T Pham, Mingyu Ding, Masayoshi Tomizuka, and Wei Zhan. X-drive: Cross-modality consistent multi-sensor data synthesis for driving scenarios. *arXiv preprint arXiv:2411.01123*, 2024. 1, 3, 7

[71] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using vq-vae and transformers. In *arXiv*, 2021. 3

[72] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024. 8

[73] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting, 2024. 3

[74] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. 1, 3, 13

[75] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3

[76] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: Masked Generative Video Transformer. In *CVPR*, 2023. 3

[77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 6

[78] Yumeng Zhang, Shi Gong, Kaixin Xiong, Xiaoqing Ye, Xiao Tan, Fan Wang, Jizhou Huang, Hua Wu, and Haifeng Wang. BEVWorld: A multimodal world model for autonomous driving via unified BEV latent space, 2024. 1, 3, 7

[79] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes, 2024. 3

[80] Xin Zhou, Dingkang Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. *arXiv preprint arXiv:2501.14729*, 2025. 1, 3

[81] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6772–6781, 2025. 3

Table 8. Ablation Study on Spatio-Temporal Module.

S.-T. Module	FID↓	FVD↓
decoupled	15.6	140
global	**10.5**	**91**

Table 9. Ablation Study on Random Errors in Training.

Random Errors	FID↓	FVD↓
✗	19.8	142
✓	**17.2**	**124**

In this supplementary material, we present additional implementation details and results that could not be included in the main paper due to space constraints. In Sec. A, we provide further details on the model architecture and training procedure. In Sec. B, we report additional ablation studies to support our design choices. In Sec. C, we include more qualitative results. Finally, we discuss the limitations of our work in Sec. D.

A. Implementation Details

As mentioned in Sec. 4.1, RAYNOVA model is initialized from the pretrained weights of Infinity [19]. Although video VAEs [40, 74] are widely recognized for improving temporal coherence and yielding better FVD scores, we instead directly adopt the multi-scale image tokenizer used in Infinity [19] to better accommodate drastic camera motions and varying frame rates.

RAYNOVA includes two model variants with approximately 130M and 2B parameters. Unless otherwise specified, all experiments use the 2B model. The 130M model is only used in several ablation studies, including Tab. 6, Fig. 5, and Tab. 9.

The training of RAYNOVA includes three stages as follows. In the first stage, we first train the model on short multi-view video clips containing 4 frames at a resolution of 192×336. The frame intervals are independently sampled between 0.1s and 1s. We use a batch size of 64 and a learning rate of *5e-5*, and train for 4k iterations. In the second stage, we then train the model on the same short clips but at a higher resolution of 384×672. The batch size is reduced to 32, and the learning rate is decreased to *2.5e-5*. This stage also runs for 4k iterations. In the last stage, we adopt the recurrent training scheme described in Alg. 1. The model is trained on long multi-view image sequences with 20 frames at 384×672 resolution. We use a batch size of 32 and a learning rate of *2.5e-6* for 300 iterations. All training is conducted on 32 NVIDIA A100 GPUs using the AdamW optimizer [34].

B. Additional Ablation Study

Global Causal Attention. In Sec. 3.4, we introduce a global self-attention module within the dual-causal block that jointly attends over all cameras and frames to ensure spatio-temporal consistency. Following prior work [12, 67, 68], we also experiment with an alternative decoupled design that applies two sequential self-attention modules: one for cross-view interactions per frame and one for cross-frame interactions per camera. As shown in Tab. 8, the global attention module significantly outperforms the decoupled spatio–temporal design. The decoupled approach implicitly assumes strong temporal correlation across frames captured by the same camera, thereby injecting strong ego-motion priors. In contrast, the global attention module incorporates minimal inductive bias and therefore generalizes better.

Random Errors in Training. RAYNOVA inherits the teacher-forcing autoregressive training scheme used in VAR [55] and Infinity [19], which can lead to error propagation due to the mismatch between training and inference. To mitigate this issue, we follow the idea introduced in Infinity [19] and inject random errors into the bit-wise multi-scale tokens X_k across all three training stages by randomly flipping a portion of bits in each token. As shown in Tab. 9, this simple strategy improves video generation quality by narrowing the gap between training and inference.

C. Additional Qualitative Results

Zero-Shot to Novel Camera Configurations. The ray-level relative position embedding enables RAYNOVA to generalize to unseen camera configurations. In Fig. 6, we generate multi-view videos under the camera setup of the Waymo Open Dataset [53], which is never seen during training. Nevertheless, RAYNOVA is able to synthesize multi-view videos with reasonable visual quality and temporal coherence.

Camera Rotation & Shifting. In Fig. 7, RAYNOVA is able to simulate camera rotations and translations. This highlights another advantage of our ray-level position embedding: it enables RAYNOVA to effectively function as a feed-forward novel view synthesis model.

Multi-View Video Generation. In Fig. 8 and Fig. 9, we present additional examples of multi-view video generation across diverse camera configurations and environments. We also provide several synthetic multi-view videos in the *videos/* folder using different frame rates (with object and map conditions visualized). RAYNOVA demonstrates strong condition fidelity and high visual realism.

D. Limitation and Future Work

Our current training data are limited to driving scenarios, which constrains the generalization ability of RAYNOVA in non-driving environments. In future work, we plan to incorporate more diverse datasets (*e.g.*, robotics or UAV) into the training pipeline. In addition, our evaluation primarily focuses on multi-view video generation. We aim to explore broader applications of RAYNOVA, such as closed-loop simulation. Recent advances in visual autoregressive models [16, 59] can also be readily integrated into the RAYNOVA framework to further enhance its capability and efficiency.

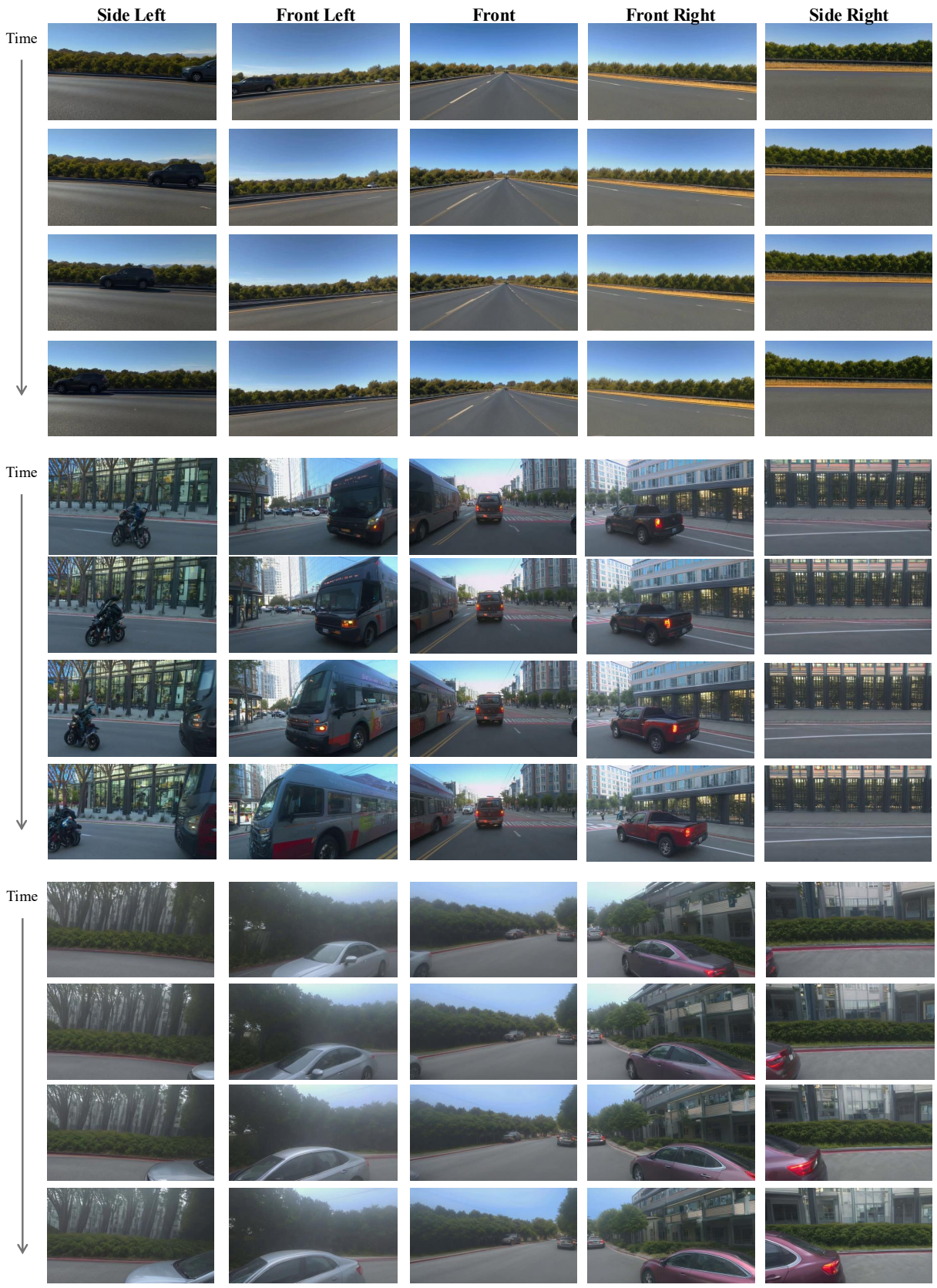
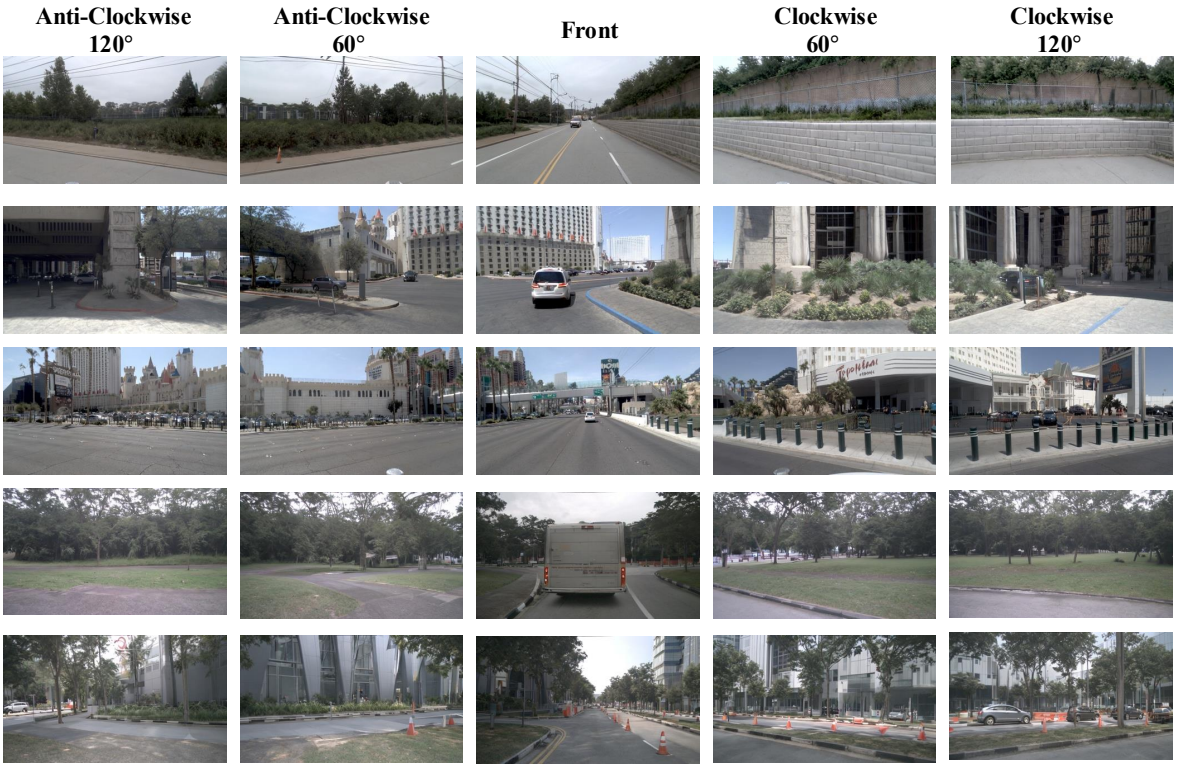


Figure 6. Zero-Shot to Unseen Waymo Open Dataset Camera Configuration.



(a) Camera Rotation.

(b) Camera Shifting.

Figure 7. Novel View Synthesis with Camera Rotation and Shifting.

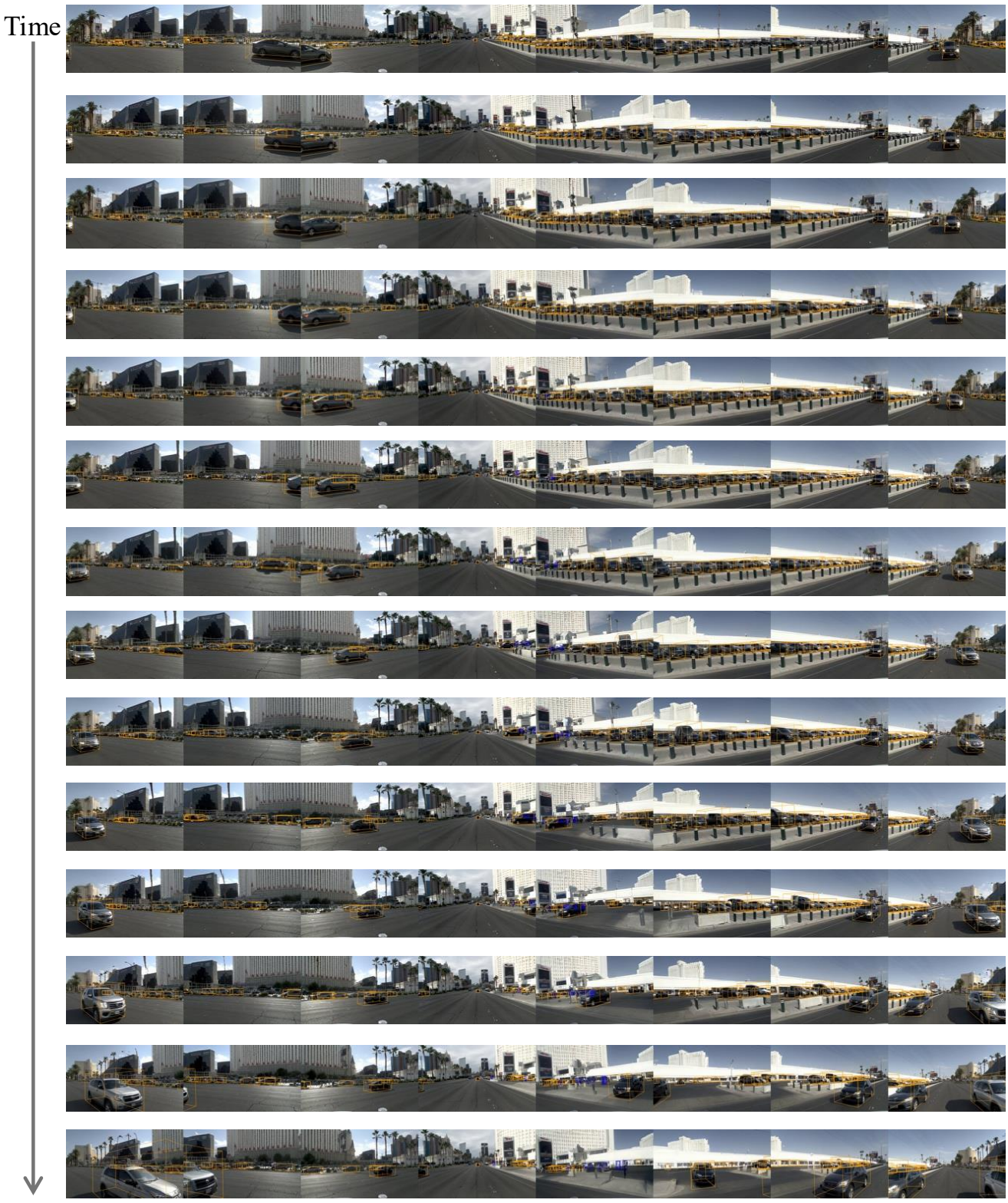


Figure 8. Multi-View Video Generation with Eight Cameras.

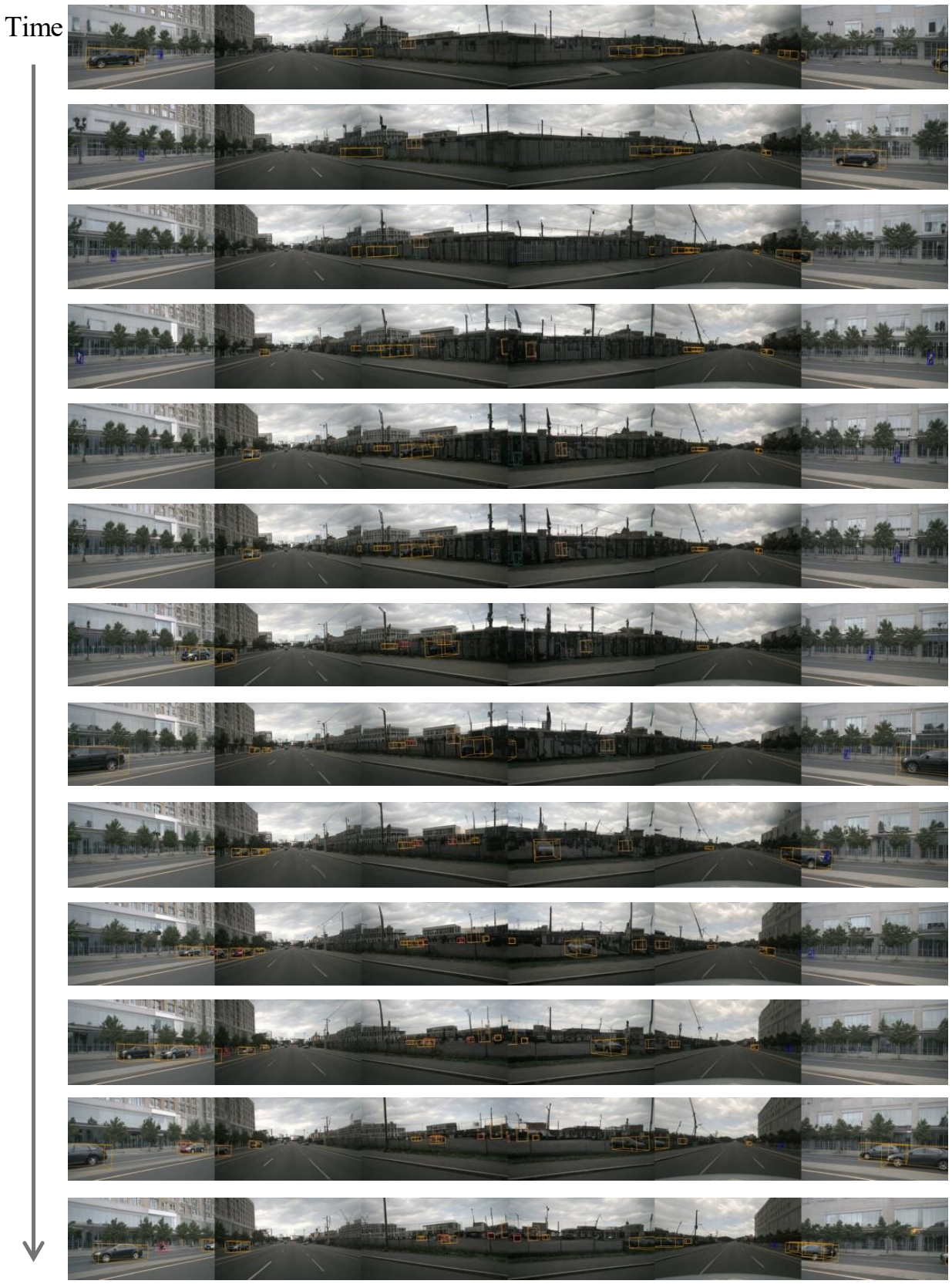


Figure 9. Multi-View Video Generation with Six Cameras.