# UniCST: Next-scale Latent Prediction for Continuous Spatio-Temporal World Modeling

Anonymous ICCV submission

Paper ID 00011

## Abstract

*Generative world models excel at synthesizing plausible visual sequences but still fall short in capturing the continuous 4D structure of real environments. We introduce* UniCST*, a unified 4D latent world model that jointly learns Continuous Spatio-Temporal representations with minimal inductive bias, enabling seamless, spatio-temporally coherent video generation. Built on a next-scale latent prediction paradigm,* UniCST *constructs its 4D latent hierarchy in a coarse-to-fine fashion thus achieving near real-time speeds. This makes it ideally suitable for controllable 4D generation and downstream embodied tasks. Extensive experiments on large-scale driving datasets demonstrate that* UniCST *outperforms state-of-the-art methods in both visual fidelity and inference latency, establishing a new baseline for practical world modeling in autonomous systems.*

## 1. Introduction

Next-token prediction has become a key recipe in building Artificial General Intelligence (AGI), especially in language domains [9, 30, 37]. However, extending this paradigm to physical intelligence—which underlies embodied agents like autonomous vehicles and robots—poses unique and significant challenges. Unlike textual or symbolic domains, the physical world exists in continuous 4D space and is governed by rigid physical laws, spatial geometric constraints, and multimodal sensory signals. Developing world foundation models (WFMs) [8, 25, 40, 54, 55] that can simulate and reason about such environments demands a coherent understanding of vision, geometry, motion, and interaction.

Recent progress in visual generative models [1, 3, 8, 25, 55] has demonstrated impressive capabilities in synthesizing plausible video sequences from text, motion cues, or visual prompts, primarily enabled by large-scale pretraining on internet-scale corpora. However, these models often lack proper grounding in physical and geometric constraints, limiting their applicability in domains that demand physical plausibility, like robotics and autonomous driving.

To address these shortcomings, recent approaches have introduced explicit geometric conditioning, incorporating structured inputs such as 3D bounding boxes, HD maps, depth maps, 3D occupancy, and LiDAR [31, 33, 49, 68, 71, 78]. While these techniques enhance realism under constrained settings, they typically require accurate annotations and involve complex pre-processing pipelines, which limit their scalability to heterogeneous or unlabeled datasets. Moreover, such models often suffer from limited flexibility due to strong inductive biases—such as assumptions about adjacent camera views [19, 75, 78] or reliance on structured 3D video tokenizer [55, 62]—that reduce generalizability across diverse real-world scenarios. In addition, most existing models depend heavily on video diffusion or pixel-wise autoregressive generation methods. Although these techniques yield high-quality visual outputs, they incur substantial computational costs and suffer from prohibitively slow inference speeds, which significantly hinder their suitability for real-time simulation, interactive applications, or downstream control tasks.

To achieve true spatial intelligence, we argue for a unified architectural framework that minimizes inductive bias while retaining physical meaning. Such a foundation is essential not only for high-fidelity and controllable video synthesis but also for enabling long-tail data generation [2, 31], closed-loop training and evaluation [1, 26, 77], and interactive decision-making within complex physical environments while enjoying the realtime inference speed [5, 35, 90]. Inspired by "next-scale prediction" autoregressive models [27, 65], we propose UniCST which emphasizes:

- **Efficiency**: Leveraging multiscale representations to compress high-dimensional visual information, enabling an efficient transition from coarse abstractions to fine-grained details.
- **Scalability and Generalization**: Accommodating heterogeneous data from diverse sources, enhancing flexibility and improving generalization to unseen scenarios.
- **Unified 4D Latent Representation**: Forming a continu-

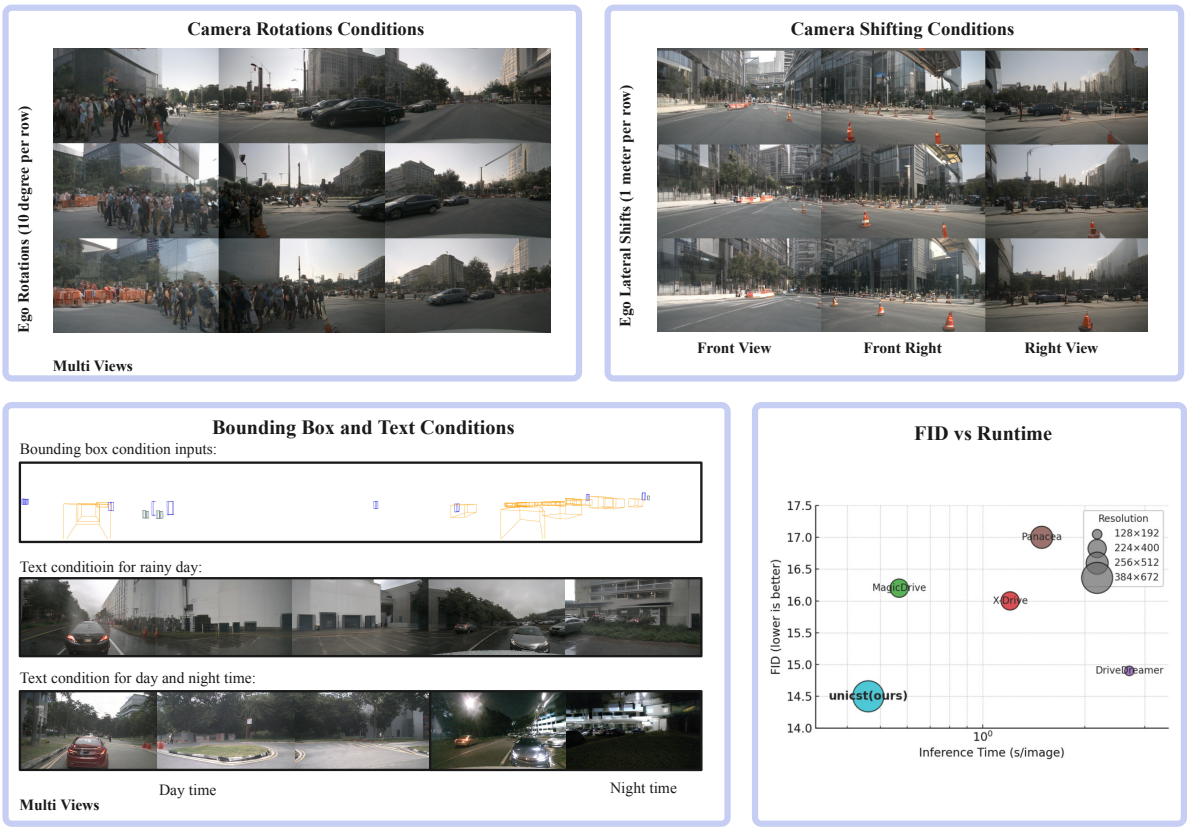


Figure 1. UniCST generates photorealistic, temporally coherent *multi-view* videos with precise *3D* control—supporting continuous camera rotations and translations, high-level text prompts for global appearance, and fine-grained 3D bounding-box cues for object category, position, and scale. Across this diverse conditioning space, UniCST preserves cross-view spatial alignment and temporal stability, delivering state-of-the-art fidelity (FID/FVD) and markedly higher throughput than previous multi-view video generators.

ous and coherent latent space across spatial and temporal dimensions for consistent reasoning.

- **Versatility**: Supporting diverse sensor configurations and conditioning inputs by minimizing handcrafted spatial and temporal biases.

We exploit extensive pretraining on natural images [27], subsequently fine-tuning the model on domain-specific street-view data. Each generated frame jointly incorporates multiview images to ensure spatial consistency, while causal temporal conditions facilitate robust temporal coherence. Our scale-aware conditioning enables the model to capture intricate spatial-temporal correspondences across multiple abstraction levels. Furthermore, by adapting the next-scale autoregressive paradigm, our framework can generate realistic videos with near-real-time throughput.

We benchmarked UniCST on a large-scale driving dataset and demonstrated its superiority in both visual realism and inference time. Capable of effectively scaling to datasets with varying sensor configurations and video frequencies, our method also adeptly supports diverse downstream tasks through flexible input conditioning. The substantial speed-up achieved by our framework enables near-real-time video generation, significantly enhancing its practicality for interactive physical-world applications such as autonomous driving. To the best of our knowledge, we are the first to propose near-real-time multiview 4D video generation based on the next-scale prediction paradigm.

## 2. Related Works

### 2.1. Video Generative Models

Recent advances in generative modeling have pushed the fidelity and diversity of *2D* video synthesis to new heights. Early transformer-based, masked autoregressive approaches operate on discrete tokens to generate high-quality frames in sequence [21, 52, 67, 79, 86]. In parallel, diffusion-based pipelines iteratively denoise latent or pixel representations, yielding vivid motion and appearance [2, 7, 22, 51, 55, 59, 61, 84]. 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 [22, 51, 84]. Some

other models also generate videos conditioned on reference images [60, 73], or reference videos [46].

### 2.2. Spatio-Temporally Grounded Generation

To more accurately simulate the geometry and dynamics of real-world scenes, several approaches leverage explicit 3D priors via reconstruction techniques. Neural rendering methods (e.g., NeRF [38, 53]) and Gaussian splatting variants [17, 56, 81, 91] produce highly consistent multi-view renderings but depend on costly per-scene optimization and often fail to generalize beyond their training trajectories. More recent feed-forward 3D reconstruction models [42, 69, 70, 87] improve efficiency yet focus solely on recovering captured geometry rather than generating fully novel scenarios and handling the complex agent interactions. Building on these explicit representations, a second line of work injects spatial cues—depth maps, 3D bounding boxes, HD maps, or LiDAR scans—into diffusion or autoregressive backbones to enable end-to-end *4D* generation [24, 31, 33, 68, 71, 78]. While these frameworks achieve improved geometric grounding, they introduce substantial inductive biases, require complex preprocessing pipelines, and impose stricter data requirements, ultimately slowing down processing and hindering scalability to internet-scale datasets.

### 2.3. World Models for Autonomous Driving

Autonomous driving serves as an ideal testbed for physically grounded world models due to its stringent demands on modeling complex geometric environments, dynamic actor interactions, and the need for coherent spatiotemporal representations across multi-sensor inputs. Early works leveraged GANs [23] to synthesize realistic driving scenarios [39, 64]. Subsequent approaches, such as DriveDreamer [72] and GAIA-1 [32], demonstrated action-conditioned video generation but were constrained primarily to front-view perspectives.

More recent studies have expanded input modalities, views, and overall versatility. Notably, Drive-WM [74] introduced a diffusion-based, multi-view world model conditioned on images, textual descriptions, layout information, and ego motions, while Vista [20] harnessed pretrained Stable Video Diffusion [6] for high-resolution and versatile scenario conditioning. DriveDreamer [71] further integrated future ego-action predictions to enhance controllability. DrivingGPT [15] unified world modeling and planning tasks within an autoregressive transformer framework, emphasizing the integration of perceptual understanding with decision-making. Copilot4D [88] extended this paradigm to LiDAR sensors using discrete latent diffusion models. Contemporary world models for autonomous driving have advanced in extending generative horizons [13, 50], while enhancing geometric consistency, sensor coherence, and structured controllability. These improvements have been driven by both diffusion-based approaches [36, 48, 63] and autoregressive frameworks [15, 33].

In contrast, our approach uniquely learns a continuous 4D latent representation that jointly encodes spatial and temporal dimensions with minimal hand-crafted biases. By employing a hierarchical next-scale prediction strategy, UniCST progressively constructs its latent representation in a coarse-to-fine fashion, inherently enforcing multi-view spatial consistency and frame-to-frame temporal coherence. Furthermore, our highly parallelizable inference architecture achieves near real-time generation, positioning UniCST optimally for practical simulation and diverse downstream embodied tasks. Comprehensive experiments illustrate UniCST's superior visual fidelity and reduced latency relative to prior methods, establishing a new benchmark for practical and effective world modeling in autonomous systems.

## 3. Methodology

UniCST, as shown in Fig. 2, works as a versatile foundation model to generate driving scenes. In Sec. 3.2, we formulate the next-scale prediction framework for multi-view and multi-frame images. Allowing for various sensor configurations, we unify all the representations and conditions in an isotropic 4D world space (Sec. 3.3). To enhance the spatial consistency and temporal coherence, we develop scale-wise cross-view and inter-frame condition modules in Sec. 3.4.

### 3.1. Preliminary: Next-Scale Prediction

Unlike vanilla auto-regressive models [41, 85] that flatten the 2D grids of images into 1D tokens, recent work [65] shifts from "next-token prediction" to "next-scale prediction" strategy. Each image is quantized by the tokenizer [18] into $K$ multiscale token maps $R_{1:K} = (R_1, R_2, \ldots, R_K)$ with increasingly higher resolutions $h_k \times w_k, k = 1, 2, \ldots, K$. The autoregressive likelihood is formulated as follows.

$$p(R_1, R_2, \ldots, R_K) = \prod_{k=1}^{K} p(R_k | R_1, R_2, \ldots, R_{k-1}) \quad (1)$$

where the $R_k$ is the token map at scale $k$ containing $h_k \times w_k$ visual tokens. The sequence $(R_1, \ldots, R_{k-1})$ serves as the prefixed context for the prediction of $R_k$. During the $k$-th autoregressive step, all $h_k \times w_k$ visual tokens are generated in parallel.

### 3.2. Next-Scale Prediction for multi-view Videos

**Formulation.** Beyond single image, we give a formulation to our proposed UniCST for the generation of multi-view

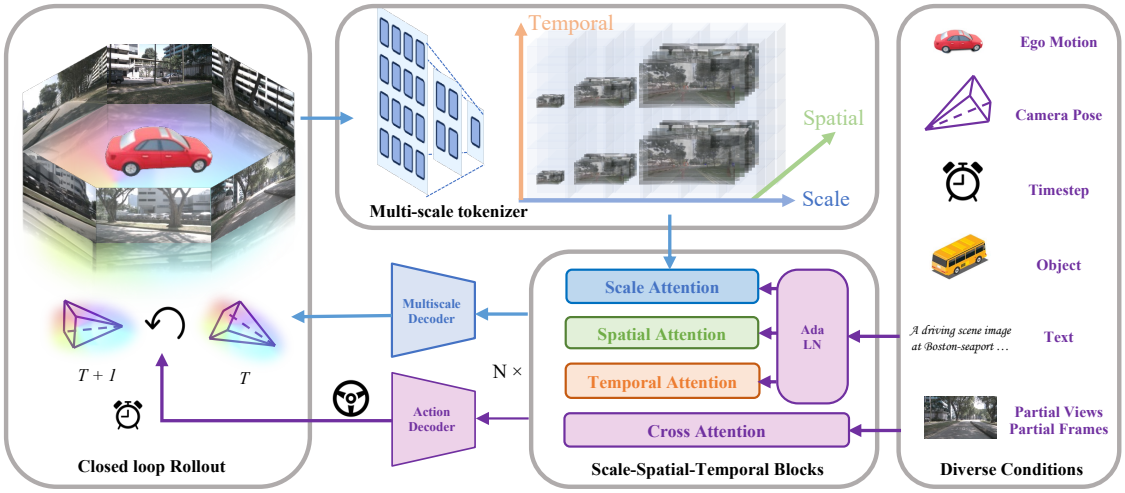


Figure 2. **Overview of UniCST.** A *multi-scale tokenizer* converts the multi-view, multi-frame video stream into a hierarchy of scale–spatial–temporal tokens. These tokens pass through *Scale–Spatial–Temporal (SST) blocks*, which factor scale, spatial, and temporal attention, while *AdaLN* and *cross-attention* layers inject diverse conditioning signals (ego motion, camera pose, text, objects, partial observations, *etc.*). An *action decoder* predicts the ego action; its rigid transform is applied to the 3D Plücker-ray embedding—*visualized by the rainbow-colored frustum*—to update the viewpoint, enabling closed-loop roll-outs with consistent geometry and motion across time.

and multi-frame videos. For flexible generation and conditions, image $I^{v,t}$ is tokenized independently as a multiscale token map $R^{v,t}_{1:K}$ for each frame $t = 1, 2, \ldots, T$ and each camera view $v = 1, 2, \ldots, V$. UniCST aims to model the joint distribution of multi-view and multi-frame data. Despite different camera perspectives and timesteps, we consider all images from different cameras and timesteps in a unified 4D world representation space to enhance the consistency. Intuitively, each scale multi-view image generation is conditioned on both their previous scales and past frames. For each frame $t_0$, the generation of multi-view images is written as:

$$p\left(R^{1:V,t_0}_{1:K}\right) = \prod_{k=1}^{K} p\left(R^{1:V,t_0}_{k} | R^{1:V,t_0}_{1:k-1}, R^{1:V,1:t_0-1}_{1:K}\right). \quad (2)$$

However, it is obvious that Eq. 2 establishes a high time and memory complexity. To reduce the time and memory cost, we take a further step to decouple the scale, spatial and temporal reliance. In this case, we can formulate the scale-wise generation for each image separately as follows.

$$p\left(R^{v_0,t_0}_{1:K}\right) = \prod_{k=1}^{K} p\left(R^{v_0,t_0}_{k} | R^{v_0,t_0}_{1:k-1}, R^{v\neq v_0,t_0}_{k}, R^{v_0,1:t_0-1}_{k}\right). \quad (3)$$

where: **1) scale reliance** ($R^{v_0,t_0}_{1:k-1}$) refers to the prefix scales of the same image; **2) spatial reliance** ($R^{v\neq v_0,t_0}_{k}$) represents the same scale of other views for current frame; **3) temporal reliance** means the same scale of the same view image in for historical frames.

**Architecture.** We construct the model according to Eq. 3. Each input is quantized to multiscale tokens independently similar with [27]. In the next-scale prediction block, each image attends to prefix scales independently. Extra decoupled spatial and temporal condition modules enhance spatial consistency and temporal coherence with masked self-attention layers. To handle the text and object conditions, we also insert additional cross-attention and AdaLN [57] layers in each block.

### 3.3. Isotropic 4D World Representation Space

Spatial and temporal inductive biases, such as multi-view adjacency or 3D video tokens, in previous work limit their adaptivity to heterogeneous training data and versatile applications. Alternatively, UniCST shares a continuous 4D world representation space for all the modules in the framework including visual tokens and conditions, as shown in Fig. 3.

To this end, we compute a token-wise Plücker ray embedding [58] for the feature map at each scale. We start from a discrete meshgrid of $w_k \times h_k \times D$ in the camera frustum of the token map $R^v_k$. Each point is the meshgrid is represented as $\mathbf{p}^{cam}_{k,j} = (u_j \times d_j \times s^w_k, v_j \times d_j \times s^h_k, d_j, 1)$, where $(u_j, v_j)$ is the 2D coordinate on the token map of scale $k$ and $s^h_k = \frac{h_K}{h_k}, s^w_k = \frac{w_K}{w_k}$ are the downsampling ratio of scale $k$. These points associated to different views can be transformed into a unified 3D coordinate space as

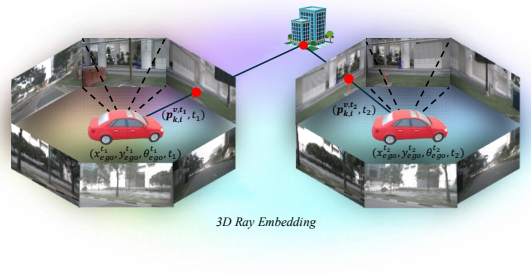


Figure 3. **3D Plücker–Ray Embedding.** Camera intrinsics and extrinsics lift each token into a unified, continuous 4D space–time frame. Because this embedding is *shared across all cameras and timesteps*, any physical point in the 3d space (e.g. the red landmark or the building corner) is mapped to the *same latent position*. This view- and time-consistent representation lets UniCST reason jointly over multi-view, multi-frame observations while keeping object identities and geometry coherent throughout the rollout.

follows like PETR [47].

$$\mathbf{p}_{k,j}^{v,0} = K_v^{-1}\mathbf{p}_{k,j}^{cam} \tag{4}$$

where $K_v$ is the transformation matrix to project points from 3D ego coordinate to camera coordinate. We take a step further to unify the representation of multiple frames in a joint 4D space by considering the motion of each camera.

$$\mathbf{p}_{k,j}^{v,t} = T_{v,t}\mathbf{p}_{k,j}^{v,0} \tag{5}$$

The transformation matrix $T_{v,t}$ represents the relative position of the $v$-th camera at $t$-frame in the world coordinate. Unless otherwise specified, we set the ego-vehicle coordinate of the first frame as the world coordinate. The transformed the points of the original $w_k \times h_k \times D$ meshgrid are transposed as $P_k^{v,t} = \{\hat{\mathbf{p}}_{k,i}^{v,t} \in \mathbb{R}^{(D\times 3)\times C} | i = 1, 2, \ldots, w_k \times h_k\}$ with token-wise correspondence to the token map $R_k^{v,t}$, where each $\hat{\mathbf{p}}_{k,i}^{v,t}$ is associated with a visual token $\mathbf{r}_{k,i}^{v,t}$. An additional time dimension it also attached to the ray embedding as $(\hat{\mathbf{p}}_{k,i}^{v,t}, t)$. This position embedding considers all the visual tokens from different camera views and frames in a continuous 4D world space.

In the same time, we also represent conditions in the same world coordinate. For example, the object bounding boxes are represented by the corner coordinates as $\mathbf{p}_{obj,i}^t = \{(x_{cor,i}, y_{cor,i}, z_{cor,i}), t | i = 1, 2, \ldots, 8\}$ in addition to the semantic label $l_{obj,i}$.

This unified 4D representation space considers everything in an isotropic and continuous manner without any handcrafted inductive biases. This representation allows the model to adaptively learn the interaction between different visual token across views and frames and handle diverse camera perspective view in the continuous 4D space.

### 3.4. Decoupled Spatio-Temporal Condition

To enhance the spatio-temporal alignment, we insert two masked self-attention modules to each next-scale prediction block. For better 3D-awareness across views and frames, we attach the ray position embeddings in Sec. 3.3, which is encoded by a shared light-weighted MLP, to each visual token before each block.

For multi-view condition, we perform self-attention on all the visual tokens from all perspective views of the same frame, which is written as follows.

$$R_{k,out}^{v,t} = R_{k,in}^{v,t} + \text{Masked-SA}\left(R_{k,in}^{v,t}, R_{k,in}^{1:V,t}\right) \tag{6}$$

For multi-frame condition, we perform casual self-attention on visual tokens from each view separately. Each visual token attends to other tokens from the same camera in historical frames at the same scale.

$$R_{k,out}^{v,t} = R_{k,in}^{v,t} + \text{Masked-SA}\left(R_{k,in}^{v,t}, R_{k,in}^{v,1:t-1}\right) \tag{7}$$

It is worth mentioning that both spatial and temporal conditions are performed in a scale-wise manner, *i.e.* each visual token only attends to other visual tokens from the same scale. This design not only reduces the computational cost but also allow the model to learn the multi-level cross-view and cross-frame relationship in a coarse-to-fine manner.

### 3.5. Joint World Modeling and Motion Planning

As a world model, our method can plan the motion of the ego vehicle along with generating future camera images. To this end, we add a learnable action token $\mathbf{r}_{act}^t$ to the end of each frame's highest scale. Similar to [45], we slice all the ego vehicle trajectories into segments of length $T_{act}$. These segments are divided into $N$ clusters via K-Means based on the flattened waypoints in the ego-vehicle coordinate space at each time step.

To predict the future ego motion, the action token is passed through the same network as the visual tokens. Consistent with Sec. 3.3, at each frame, the current position $\mathbf{p}_{ego}^t = (x_{ego}^t, y_{ego}^t, \theta_{ego}^t, t)$ in the same world coordinate is embedded and attached to the action token. In each block, the action token would skip the scale attention block and share the rest spatial condition module, temporal condition module, and feed-forward network with visual tokens. In the spatial condition module, the action token attends to the highest scale visual tokens as follows.

$$\mathbf{r}_{act,out}^t = \mathbf{r}_{act,in}^t + \text{Masked-SA}\left(\mathbf{r}_{act,in}^t, R_{K,in}^{1:V,t}\right) \tag{8}$$

For temporal condition module, the action token attends to the action tokens in the previous frames.

$$\mathbf{r}_{act,out}^t = \mathbf{r}_{act,in}^t + \text{Masked-SA}\left(\mathbf{r}_{act,in}^t, \mathbf{r}_{act,in}^{1:t-1}\right) \tag{9}$$

Finally, a classification head attached to the action token would output the predicted cluster based on the action token, based on which we can query the future ego trajectory.

# 4. Experiments

## 4.1. Experimental Setups

### 4.1.1. Datasets.

UniCST is trained on two multimodal driving datasets: nuScenes [11] and nuPlan [12]. We adopt the ScenarioNet [43] format to segment each continuous log into fixed-length 20-second video sequences.

**nuScenes.** The nuScenes [11] dataset was recorded at 2 Hz in two geographically and structurally diverse urban areas—the Seaport district of Boston and the One North, Queenstown, and Holland Village districts of Singapore. Applying our 20-second segmentation yields approximately 4.7 hour of data, from which we allocate 700 sequences (around 3.9 hours) for training and 150 sequences (0.8 h) for evaluation.

**nuPlan.** The nuPlan [12] dataset spans 1,500 h of continuous driving data captured at 10 Hz across four cities—Boston, Pittsburgh, and Las Vegas (USA), and Singapore—covering diverse urban, suburban, and highway scenarios . We filter the whole dataset based on complete sensor coverage, and extract 10,000 20-second sequences, around 55.6 hours in total for training.

In summary, our training set comprises about 59.5 hours of 20-second video sequences drawn from four cities across North America and Asia, and our evaluation set adds another 0.8 hours reserved exclusively from nuScenes.

### 4.1.2. Evaluation Metrics.

For evaluation of the generated future frames, we adopt the Frechet Video Distance (FVD) [66] and the Frechet Inception Distance (FID) [28]. We use the 150 validation sequences from nuScenes for evaluation. To demonstrate the efficiency of our autoregressive framework, we also report the Frames Per Second (FPS) metric (*i.e.* number of generated images per second).

### 4.1.3. Baselines.

We evaluate our method against several notable baselines. Specifically, they can classified into three categories, GAN-based [39], diffusion-based methods like Vista [20], GenAD [89], WoVoGen [48], DriveDreamer [71], *etc.* and autoregressive approaches such as DrivingWorld [33]. Among all these baselines, MagicDrive [19], X-Drive [78] can generate single-frame multi-view images, therefore we only report their FID score.

## 4.2. Implementation

We first trained our 3B model with the resolution of 192 × 336 for 50 epochs using the mixed data of nuscenes and nuplan. The batch size is 4. We randomly drop the bounding boxes and camera views to improve the robustness of our model. Then we finetune our model on high resolution setting (384 × 672) with a batch size of 1. The training is conducted on 64 NVIDIA A100 GPUs and we evaluate the throughput per seconds on one A100 GPU. Our trained model can generate arbitrary numbers of views. For comparison with those single-view models, we generate 6 views and select the corresponding camera view for evaluation.

Table 1. Comparison under single-view setting. DrivingWorld* is trained without Private data. Results are sourced from [34]

| Method | Data Scale | FID ↓ |
|---|---|---|
| DriveGAN [39] | 160h | 73.4 |
| GenAD [83] | 2000h | 15.4 |
| Vista [20] | 1740h | 6.9 |
| WoVoGen [48] | 5h | 27.6 |
| DrivingWorld* [34] | 120h | 16.4 |
| DrivingWorld [34] | 3456h | 7.4 |
| Ours (single-view) | 60h | **4.5** |

## 4.3. Video Generation

We evaluate UniCST on the nuScenes validation split and report Fréchet Inception Distance (FID), Fréchet Video Distance (FVD), and generation throughput (images per second) in Table 8 and Table 2.

**Single-View Comparison.** We first compare against state-of-the-art models that generate a single camera view in Table 8, namely DriveGAN [39], GenAD [4], Vista [4], WoVoGen [48], and DrivingWorld [33]. Note that most of these baselines are trained on hundreds to thousands of hours of data, even including *private* driving dataset. By contrast, UniCST is trained on only 60h of *public* data at 384 × 672 resolution. Despite this modest training budget, our model achieves an FID of **4.5**, outperforming DrivingWorld's FID of 7.4 (trained on 3,456h) and Vista's 6.9 (trained on 1,740h).

**Multi-View Comparison.** Then, we evaluate the multi-view image or video generation ability of UniCST by comparing to MagicDrive [19], X-Drive [78], DriveDreamer [71], DrivingDiffusion [44], and Panacea [76] in Table 2. Our multi-view model trained on 60h public data reaches an FID of **14.5**, improving over Panacea's 17.0 and MagicDrive's 16.2. We also attain an FVD of 134, substantially lower than DriveDreamer's 341 and DrivingDiffusion's 332, demonstrating better temporal coherence. Notably, UniCST deliver a throughput of **2.17 images/s**, much

Table 2. **Comparison with real-world driving world models.**

| Method | Model Setups | | Metrics | | |
|---|---|---|---|---|---|
| | Resolution | Type | FID | FVD | Throughput |
| MagicDrive [19] | 224×400 | Diffusion | 16.2 | - | 1.76 |
| X-Drive [78] | 224×400 | Diffusion | 16.0 | - | 0.83 |
| DriveDreamer [71] | 256×448 | Diffusion | 14.9 | 341 | 0.37 |
| DrivingDiffusion [44] | 512×512 | Diffusion | 15.8 | 332 | - |
| Panacea [76] | 256×512 | Diffusion | 17.0 | 139 | 0.67 |
| Ours | 384×672 | Next-scale AR | **14.5** | **134** | **2.17** |

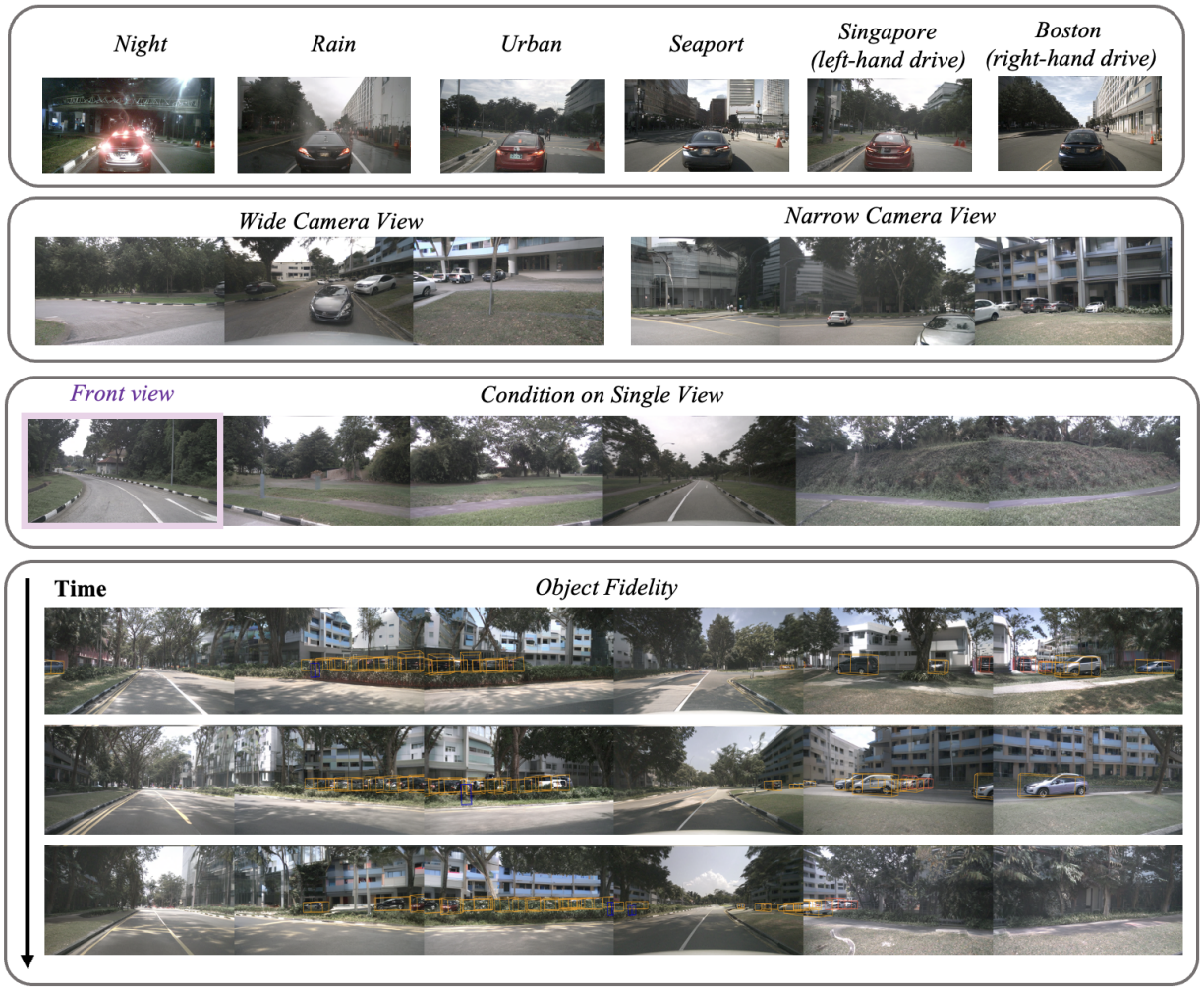


Figure 4. Qualitative Results. Our model can generate under different weather conditions, understand driving rules in different areas (1st row). The camera view can be manipulated by adjusting the camera intrinsics (2nd row). We can also generate multi-view images from single camera view (3rd row) and bounding boxes (4th row).

more than the throughput of the diffusion baseline (MagicDrive at 1.76 images/s), demonstrating the efficiency of our next-scale autoregressive architecture.

We also provide some qualitative visualizations in Fig. 4. It can be observed that our model can generate under different conditions: text, front view, and bounding boxes.

These results show that by leveraging a scale-wise autoregressive, UniCST not only matches or surpasses the fidelity of models trained on orders of magnitude more private data, but also operates at higher frame rates. This combination of *data efficiency*, *generation quality*, and *throughput* makes our approach especially well-suited for down-

Table 3. Ablation on View Embedding.

| View Embed | FID ↓ | FVD ↓ |
|---|---|---|
| None | 24.7 | 280.5 |
| Leanrable | 23.4 | 248.8 |
| Ray | 21.7 | 240.8 |

Table 4. Ablation on time embedding.

| Time Embed | FID ↓ | FVD ↓ |
|---|---|---|
| None | 23.2 | 257.7 |
| Learnable | 25.1 | 261.3 |
| Continuous | 21.7 | 240.8 |

Table 5. Ablation on s.-t. condition.

| S. Cond. | T. Cond. | FID ↓ | FVD ↓ |
|---|---|---|---|
| ✗ | ✓ | 26.0 | 247.8 |
| ✓ | ✗ | 21.9 | 270.3 |
| ✓ | ✓ | 21.7 | 240.8 |

stream driving applications requiring both realism and real-time performance.

### 4.4. Ablation Study

To verify the efficacy of our proposed components, we conducted a series of ablation studies on the nuscenes dataset. To save the training time, we use the base model with fewer parameters (*i.e.* 200M) for ablation, and the model is trained on nuscenes dataset.

**View Embedding.** The ray embedding in an isotropic world space is critical for the interaction across views. As shown in Table 3, without view embedding, the model struggles to learn the inter-relationship across views. With a leanrable embedding $E \in R^{v \times c}$, where $v$ is the number of views, and $c$ is the embedding dimension, the FID, FVD both improve incrementally. However, the model with our proposed ray embedding reports the best performance in terms of FID and FVD since it can benefit from the unified space.

**Time Embedding.** For time embedding, we also consider three settings in Table 4. The performance of learnable embedding is unsatisfactory, even worse than the model without time embedding, since we adopt random time interval in the training. In contrast, the continuous embeddings allow us to adapt to data sampled with different frequencies.

**Spatial Temporal Condition.** Finally, we ablate on the decoupled spatial and temporal attention modules, which both help to improve the spatial consistency and temporal coherence. In contrast, the full attention has a high space and time complexity, which is computationally prohibitive.

## 5. Conclusion

In this work, we have identified the key obstacles in extending next-token prediction from language to physical intelligence: the continuous, multimodal, and physics-constrained nature of real-world environments. We have argued that existing visual world foundation models, despite impressive generative power, lack the necessary grounding in geometry and physics and often rely on expensive diffusion or pixel-wise autoregressive decoders. To address these issues, we introduced a unified architectural framework that (i) minimizes hand-crafted inductive biases, (ii) incorporates explicit geometric and physical conditioning in a flexible manner, and (iii) leverages a next-scale autoregressive decoder for realtime inference. Our approach achieves high-fidelity, physically plausible video synthesis without resorting to heavy pre-processing or proprietary annotations, and it scales gracefully to diverse, unlabeled datasets. Through extensive experiments on driving benchmarks, we demonstrate that our model outperforms state-of-the-art baselines in both visual quality and computational efficiency.

## References

[1] 1X. 1X world model. https://www.1x.tech/discover/1x-world-model, 2024. 1

[2] 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. 1, 2

[3] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. 1

[4] Alexander Amini, Tsun-Hsuan Wang, Igor Gilitschenski, Wilko Schwarting, Zhijian Liu, Song Han, Sertac Karaman, and Daniela Rus. Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2419–2426. IEEE, 2022. 6

[5] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024. 1

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi,

Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 3

[7] 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. 2

[8] 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

[9] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. 1

[10] 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. 2

[11] 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, 2020. 6

[12] 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, 2

[13] 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, 2025. 3

[14] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv preprint arXiv:2311.18561*, 2023. 2

[15] Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers, 2024. 3

[16] 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. 2

[17] 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

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[19] 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, 6, 7

[20] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability, 2024. 3, 6, 2

[21] 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. 2

[22] 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. 2

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3

[24] 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. 3

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

[26] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025. 1

[27] 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, 2, 4

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[30] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1

[31] 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. *arXiv preprint arXiv:2309.17080*, 2023. 1, 3

[32] 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, 2

[33] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructingworld model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024. 1, 3, 6

[34] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructing world model for autonomous driving via video gpt, 2024. 6

[35] Yihan Hu, Siqi Chai, Zhening Yang, Jingyu Qian, Kun Li, Wenxin Shao, Haichao Zhang, Wei Xu, and Qiang Liu. Solving motion planning tasks with a scalable generative model. In *European Conference on Computer Vision*, pages 386–404. Springer, 2024. 1

[36] Junpeng Jiang, Gangyi Hong, Lijun Zhou, Enhui Ma, Hengtong Hu, Xia Zhou, Jie Xiang, Fan Liu, Kaicheng Yu, Haiyang Sun, et al. Dive: Dit-based video generation with enhanced control. *arXiv preprint arXiv:2409.01595*, 2024. 3

[37] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

[38] 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

[39] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation, 2021. 3, 6

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

[41] 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

[42] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 3

[43] Quanyi Li, Zhenghao Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling, 2023. 6

[44] 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. 6, 7

[45] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 5

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

[47] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 5

[48] 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. 3, 6

[49] Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. *arXiv preprint arXiv:2412.03934*, 2024. 1

[50] Enhui Ma, Lijun Zhou, Tao Tang, Zhan Zhang, Dong Han, Junpeng Jiang, Kun Zhan, Peng Jia, Xianpeng Lang, Haiyang Sun, Di Lin, and Kaicheng Yu. Unleashing generalization of end-to-end autonomous driving with controllable long video generation, 2024. 3

[51] 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. 2

[52] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *ICLR*, 2023. 2

[53] 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

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

[55] OpenAI. Sora: Creating video from text, 2024. 1, 2

[56] 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

[57] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[58] Julius Plucker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, pages 725–791, 1865. 4

[59] 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. 2

[60] 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

[61] Runway. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha, 2024. 2

[62] 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. 1

[63] 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. 3

[64] Eder Santana and George Hotz. Learning a driving simulator, 2016. 3

[65] 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

[66] 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. 6

[67] 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. 2

[68] 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. 1, 3, 2

[69] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state, 2025. 3

[70] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy, 2024. 3

[71] 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. 1, 3, 6, 7

[72] 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

[73] 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

[74] 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

[75] 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

[76] 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. 6, 7

[77] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023. 1

[78] 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, 6, 7

[79] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using vq-vae and transformers. In *arXiv*, 2021. 2

[80] 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. 2

[81] 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

[82] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. 2

[83] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Genad: Generalized predictive model for autonomous driving, 2024. 6, 2

[84] 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. 2

[85] 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

[86] 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. 2

[87] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion, 2025. 3

[88] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*, 2023. 3

[89] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving, 2024. 6

[90] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. 1

[91] 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