arXiv:2412.19505v3 [cs.CV] 16 Aug 2026

DrivingWorld: Constructing World Model for Autonomous Driving via Video GPT

Xiaotao Hu[1,2⋆], Mingkai Jia[1,2], Xiaoyang Guo[2], Qian Zhang[2], Xiao-xiao Long[3], and Wei Yin[2]

[1] Hong Kong University of Science and Technology, Hong Kong SAR
[2] Horizon Robotics, Shanghai, China
[3] Nanjing University, Nanjing, China

Abstract. Recent successes in autoregressive (AR) generation models, such as ChatGPT, have inspired efforts to extend these techniques to visual tasks. Although various approaches have been proposed to construct video-based world models for autonomous driving, their performance has consistently remained suboptimal. This limitation arises because the classic GPT framework is designed for one-dimensional contextual information and lacks the capacity to model the high-dimensional dynamics essential for video generation. In this paper, we present *DrivingWorld*, a framewise autoregressive world model for autonomous driving that effectively captures both spatial and temporal dynamics to predict high-fidelity future states. Our method employs a next-frame prediction strategy to ensure temporal coherence between consecutive frames and a next-token prediction strategy to capture spatial details within each frame. To further enhance generalization, we introduce balanced attention and masking strategies for token prediction, enabling long-term future prediction and precise control. Experiments demonstrate that our approach produces video clips that are twice as long—up to 40 seconds—while maintaining high fidelity and consistency over previous methods.

Keywords: Autonomous Driving · World Model · Video Generation.

1 Introduction

Predicting future events is crucial for autonomous systems, as it enables informed decision-making for safe and efficient navigation in complex environments, while enhancing adaptability to dynamic conditions and improving both performance and reliability. To this end, researchers are exploring world models, which predict future states from past observations and potential actions. Specifically, a world model takes a sequence of past states (observations, actions, and other relevant variables) as input and outputs the predicted state at the next time step. By learning structured representations of the environment, such models support more effective decision-making in autonomous systems.

⋆ Corresponding author. E-mail: xiaotao.hu@outlook.com

However, most existing world model approaches for autonomous driving rely heavily on extensive annotations (such as high-definition maps, 3D bounding boxes, and linguistic labels), limiting their scalability. The scarcity and slow growth of publicly available annotated datasets hinder progress, while even state-of-the-art methods remain ineffective in handling out-of-distribution scenarios.

Next-token autoregressive (AR) models have achieved remarkable success in Natural Language Processing and image generation, largely owing to their self-supervised training paradigm, which can effectively leverage large-scale unlabeled data. We aim to exploit this advantage to address the data limitations faced by world models in autonomous driving. However, directly adopting a next-token AR framework by discretizing multimodal inputs into a one-dimensional token sequence presents **two critical limitations**:

Control Signal Degradation: The token composition is severely imbalanced—each temporal state contains merely two control tokens (orientation + location) versus hundreds of image tokens. This imbalance introduces two fundamental issues: 1) ***Attention Dilution***: When predicting pose-related tokens, the attention mechanism must process a disproportionate number of image tokens ($> 99\%$ of total tokens). The massive visual features inevitably dominate attention weights, diluting the connection with historical control signals [5,8,18,20,28]. 2) ***Temporal Disconnection***: The sparse distribution of control tokens (2 control tokens vs. 500+ image tokens) hinders effective modeling of vehicle dynamics. The traditional attention mechanism struggles to establish coherent temporal dependencies for control signals across frames, as they get intermittently submerged in visual feature surges (see Section III-B for detailed experiments).

Computational Inefficiency: The traditional attention mechanism requires each predicted token to interact with all preceding tokens, resulting in ($\mathcal{O}(T^2n^2)$) computational complexity proportional to the number of historical frames. This exhaustive cross-token interaction imposes substantial computational overhead. These limitations collectively impair the model's ability to generate accurate vehicle control signals, ultimately compromising the world model's controllability when using pose modalities for trajectory prediction.

The key contributions of this work are threefold:

- We propose *DrivingWorld*, a novel self-supervised, frame-wise world model for autonomous driving that decouples temporal dynamics from spatial structures in world modeling.
- Our framework employs a novel dual-strategy approach: **Inter-Frame Temporal Modeling** and **Internal-Frame Spatial Reasoning**. It explicitly separates temporal modeling from cross-modal spatial reasoning while maintaining their interaction, addressing both the temporal coherence and modality imbalance challenges.
- Experiments show that the proposed model not only achieves good generalization and generates realistic video sequences over 40 seconds long, but also produces effective policies for vehicle control.

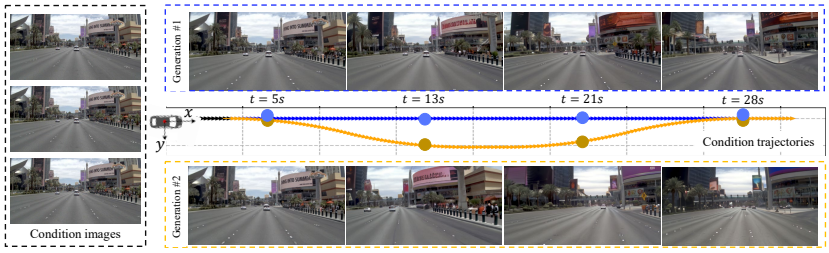


Fig. 1: Controllable generation results of our method. Our method takes a short video clip as input and can generate multiple possible future driving scenarios conditioned on different pre-defined trajectory paths. We use one straightforward path and one curved path as examples, and the results show that our method achieves accurate control and future prediction.

2 Related Work

2.1 World Model

World models capture a comprehensive representation of the environment and forecasts future states based on a sequence of actions [17], extensively explored in both game and laboratory environments. Dreamer [9] trained a latent dynamics model using past experiences to forecast state values and actions within a latent space. Built upon the original Dreamer model, DreamerV2 [10] achieved human-level performance in Atari games. DreamerV3 [11] employed larger networks and successfully learned to acquire diamonds in Minecraft from scratch.

Recently, world models for driving scenarios have garnered significant attention in both academia and industry. Most previous works have been limited to simulators or well-controlled lab environments. Drive-WM [33] explored diffusion-based real-world driving planners. GenAD [37] addressed dynamics in driving scenes with temporal reasoning blocks. Vista [7] explored video generation for driving scenarios using diffusion models. However, reliance on annotated data in previous approaches severely restricts the applicability of world models. GAIA-1 [12] investigated real-world driving planners based on AR models but had large parameters and computational demands. In this paper, we propose an efficient, self-supervised world model within an AR framework for autonomous driving scenarios, offering lower computational cost and superior performance.

2.2 VQVAE

VQVAE [30] learned a discrete codebook representation via vector quantization to model image distributions. VQGAN [6] improved realism by incorporating LPIPS loss [40]. MoVQ [41] tackled spatial normalization issue via variant-aware vector quantization. LlamaGen [26] further fine-tuned VQGAN, showing that a larger codebook size could enhance reconstruction performance. However, these methods focused on single-image processing tasks, lacking the temporal modeling

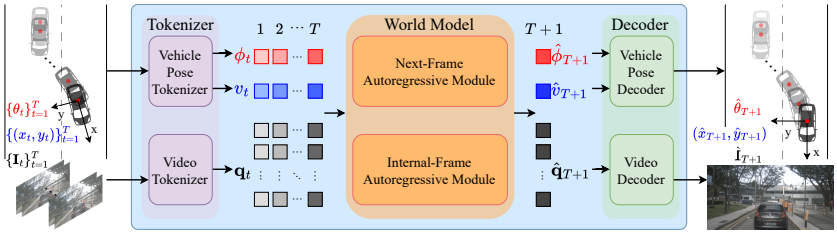


Fig. 2: Pipeline of *DrivingWorld*. The vehicle orientations $\{\theta_t\}_{t=1}^{T}$, ego locations $\{(x_t, y_t)\}_{t=1}^{T}$, and a front-view image sequence $\{\mathbf{I}_t\}_{t=1}^{T}$ are taken as the conditional input, tokenized as latent embeddings. Our proposed multimodal world model attempts to comprehend them and forecast the future states, detokenized to the vehicle orientation $\hat{\theta}_{T+1}$, location $(\hat{x}_{T+1}, \hat{y}_{T+1})$, and the front-view image $\hat{\mathbf{I}}_{T+1}$. Our AR approach generates 40+ second videos.

capabilities necessary to maintain coherent frame-to-frame consistency in video. To address this, we propose a temporal-aware tokenizer and decoder.

2.3 Video Generation

Currently, there are three mainstream video generation models: GAN-based, diffusion-based, and GPT-based methods. GAN-based methods [38] often suffer from mode collapse, limiting output diversity. Additionally, the adversarial learning between the generator and discriminator often causes instability. Diffusion-based methods [1] struggle with precise video control due to inherent stochasticity in the denoising process. Maintaining object consistency during denoising is also a significant challenge. On the other hand, traditional GPT-based methods [36] allow for a certain level of control, but their computational cost grows quadratically with the sequence length, significantly impacting model efficiency. In this paper, we propose a novel dual-strategy world model framework, which ensures controllability while significantly reducing computational cost and improving model efficiency.

3 Method

Our proposed world model, *DrivingWorld*, leverages an AR architecture to predict future states with high efficiency, capable of long-horizon predictions, i.e. *beyond* 400 *frames*. Figure 2 presents our pipeline, which consists of tokenizers for encoding vehicle poses and videos into discrete tokens, the world model for future forecasting, and the decoders to reconstruct vehicle poses and videos from discrete tokens. Specifically, *DrivingWorld* focuses on predicting the state at time step $T+1$ based on the historical states from time steps 1 to T, where each state consists of a vehicle pose (θ_t, x_t, y_t) and the corresponding frame $\mathbf{I}_t$ in temporal order. The prediction is performed in two steps: (1) the model first

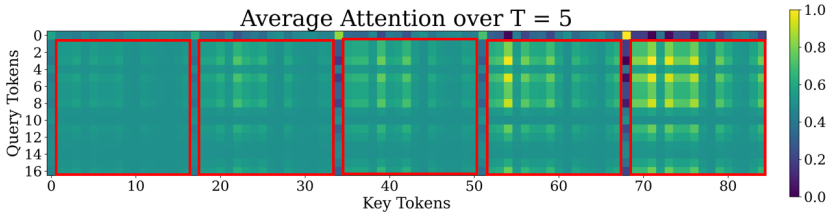


Fig. 3: Attention map visualization. Attention map from the fifth Transformer block of a naive AR model [23]. Red boxes highlight control signal degradation, where image tokens dominate attention over control tokens. Brighter colors toward the right indicate temporal disconnection, with query tokens favoring nearby timesteps.

predicts the vehicle's next action $(\hat{\theta}_{T+1}, \hat{x}_{T+1}, \hat{y}_{T+1})$; (2) this predicted action, together with the historical inputs, is then used to generate the next frame $\hat{\mathbf{I}}_{T+1}$. θ_t and (x_t, y_t) are the vehicle's orientations and locations, while $\mathbf{I}_t$ represent the video frames. Vehicle's poses (θ_t, x_t, y_t) are tokenized to ϕ_t and v_t, and $\mathbf{I}_t$ is tokenized to $\mathbf{q}_t$.

3.1 Tokenizer

Tokenizer aims to transform continuous inputs to discrete tokens. We explore token discretization for two types of signals, frames $\{\mathbf{I}_t\}_{t=1}^{T}$ and vehicle poses $\{(\theta_t, x_t, y_t)\}_{t=1}^{T}$.

Video Tokenizer. We follow VQGAN [6] to quantize each image $\mathbf{I}_t$ into a set of tokens $\mathbf{q}_t \in [K]^{H \times W}$. However, VQGAN [6] lacks temporal processing, causing noticeable inconsistency. In our implementation, we insert a causal attention layer both *before* and *after* VQGAN [6] quantization, where the attention operates on the temporal dimension.

Vehicle Pose Tokenizer. To accurately represent the vehicle's orientation θ and location (x, y), we adopt a coordinate system centered at the ego vehicle, as depicted in Figure 2. Instead of global poses, we adopt relative poses between adjacent time steps. This is because global poses present a significant challenge due to the increasing magnitude of global pose values over long-term sequences. As sequences extend, managing global pose values becomes more difficult, complicating normalization and reducing model robustness.

For the sequence of the vehicle's orientation $\{\theta_t\}_{t=1}^{T}$ and location $\{(x_t, y_t)\}_{t=1}^{T}$, we compute relative values at each time step based on the previous one. The relative location and orientation at the first time step are set to zero. The pose sequence is denoted by $\{\Delta\theta_t\}_{t=1}^{T}$ and $\{(\Delta x_t, \Delta y_t)\}_{t=1}^{T}$. To tokenize them, we discretize the pose's surrounding space. Specifically, we discretize the orientation into α categories, and the X and Y axes into β and γ bins, respectively. The

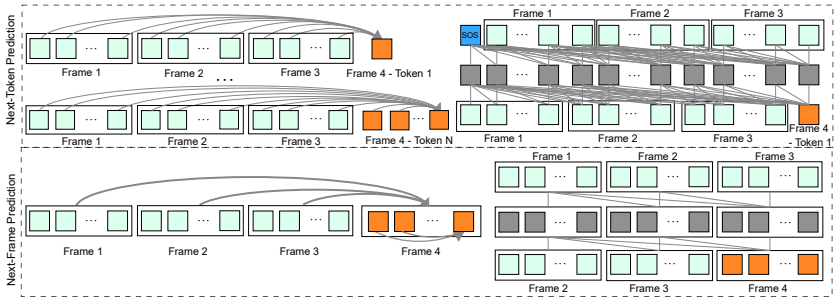


Fig. 4: Illustration of next-token prediction and next-frame prediction. Unlike next-token prediction, next-frame prediction significantly reduces attention operations, decreasing computational load and alleviating attention dilution.

relative pose at time t is tokenized as follows:

$$\begin{aligned} \phi_t &= \left\lfloor \frac{\Delta\theta_t - \theta_{min}}{\theta_{max} - \theta_{min}} \alpha \right\rfloor, \\ v_t &= \left\lfloor \frac{\Delta x_t - x_{min}}{x_{max} - x_{min}} \beta \right\rfloor \cdot \gamma + \left\lfloor \frac{\Delta y_t - y_{min}}{y_{max} - y_{min}} \gamma \right\rfloor . \end{aligned} \tag{1}$$

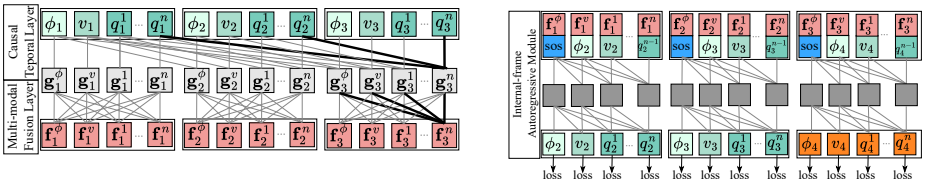
Finally, we process the past T states $\{(\theta_t, x_t, y_t, \mathbf{I}_t)\}_{t=1}^{T}$ and tokenize them into a discrete sequence $\{(\phi_t, v_t, \mathbf{q}_t)\}_{t=1}^{T}$. Specifically, v_t denotes the tokenized representation of the vehicle's 2D location (discretized x and y coordinates).

3.2 Prelimilaries of Next-Token Prediction

While future state prediction can be viewed as a long-context *next-token prediction* task, directly applying language AR models [22, 23] to video generation faces key bottlenecks (see Section I). To probe these limitations, we visualize the attention map from the fifth Transformer block of a naive AR model [23], as shown in Figure 3. For clarity, we use 16 image tokens, 1 control token, and consider a history of T=5 past states. Two notable phenomena can be observed: 1) **Control Signal Degradation**: Image tokens in the query tend to attend predominantly to the abundant image tokens in the key (highlighted by red boxes), while largely ignoring the control token that should also be referenced. This demonstrates that the influence of control signals is diluted in the attention mechanism. 2) **Temporal Disconnection**: Query tokens tend to attend mainly to key tokens from temporally nearby states (indicated by brighter colors toward the right), while neglecting tokens from more distant timesteps. This hinders long-term dependency learning and degrades performance.

3.3 World Model

To mitigate attention dilution, enhance temporal connections, and reduce redundant computations, we propose a novel **next-frame AR module** for au-



(a) Next-frame AR module (b) Internal-frame AR module

Fig. 5: Illustration of next-frame AR module and internal-frame AR module. In next-frame AR module (a), causal temporal layer links temporal dependencies at a location; multi-modal fusion layer integrates spatial information within a timestep. Internal-frame AR module (b) processes tokens in parallel within each timestep, preserving sequential dependencies inside the frame.

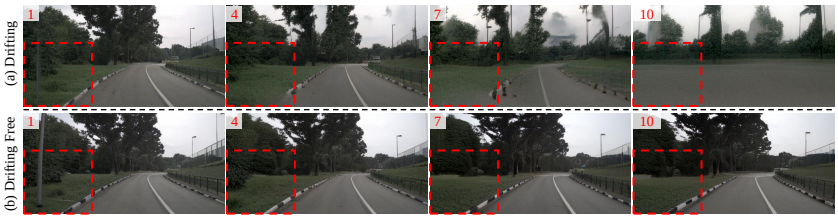


Fig. 6: **Effectiveness of our random masking strategy**. Our masking strategy effectively mitigates AR drifting. Without it, the world model suffers severe drift, causing generated videos to degrade rapidly after the 10th frame.

tonomous driving. Specifically, we decouple the long-sequence attention into two parts: one for temporal information attention and the other for multimodal information attention. The temporal information attention allows each token to focus solely on its past information, while the multimodal information attention enables each token to attend to all relevant information within the current time step, as shown in Figure 4. This operation decouples the sequence, enabling separate handling of temporal and multimodal information. By doing so, it mitigates attention dilution, strengthens temporal connections, and significantly reduces the computational burden. The input is historical T states $\{(\phi_t, v_t, \mathbf{q}_t)\}_{t=1}^{T}$. The goal is to predict the next state $(\phi_{T+1}, v_{T+1}, \mathbf{q}_{T+1})$.

Next-Frame AR Module. The structure is shown in Figure 5a. It is composed of a causal temporal layer and a multimodal fusion layer. This decouples the processing of temporal and multimodal information, which significantly reduces the length of the sequence in attention. Specifically, given past T states $\{(\phi_t, v_t, \mathbf{q}_t)\}_{t=1}^{T}$, we unfold the 2D image tokens into a 1D form in zig-zag order to get $\{(\phi_t, v_t, q_t^1, ..., q_t^n)\}_{t=1}^{T}, n = H \times W$. The causal temporal layer in Figure 5a captures dependencies across different time steps at the same spatial location.

Table 1: Comparisons on the NuScenes [3]. Blue: NuScenes seen in training. Orange: zero-shot.

Metric	DriveDreamer [31]	WoVoGen [21]	Drive-WM [34]	Vista [7]	*DrivingWorld* Ours	DriveGAN [24]	GenAD (OpenDV) [37]	*DrivingWorld* Ours
FID ↓	52.6	27.6	15.8	6.9	**6.5**	73.4	15.4	7.4
FVD ↓	452.0	417.7	122.7	89.4	**86.0**	502.3	184.0	90.9
Max Duration	4s	2.5s	8s	15s	**40s**	N/A	4s	**40s**
Max Frames	48	5	16	150	**400**	N/A	8	**400**

$$\begin{aligned}\mathbf{g}_t^{\phi} &= \mathcal{F}_a(q_1^{\phi}, ..., q_t^{\phi}),\ t \in [1, T],\\ \mathbf{g}_t^{v} &= \mathcal{F}_a(q_1^{v}, ..., q_t^{v}),\ t \in [1, T],\\ \mathbf{g}_t^{j} &= \mathcal{F}_a(q_1^{j}, ..., q_t^{j}),\ t \in [1, T], j \in [1, n].\end{aligned} \tag{2}$$

The multi-modal fusion layer in Figure 5a focuses on fusing information across different spatial locations at the same time step:

$$\{\mathbf{f}_t^{\phi}, \mathbf{f}_t^{v}, \mathbf{f}_t^{1}, ..., \mathbf{f}_t^{n}\} = \mathcal{F}_b(\mathbf{g}_t^{\phi}, \mathbf{g}_t^{v}, \mathbf{g}_t^{1}, ..., \mathbf{g}_t^{n}),\ t \in [1, T]. \tag{3}$$

However, we find that only using next-frame AR module yields suboptimal results. Recently, multiple image-generation works [26] propose that an AR pipeline for next-token prediction generates better images, and even outperforms diffusion methods. Inspired by this, we propose an **internal-frame AR module** to enhance image quality while also improving the model's controllability.

Internal-Frame AR Module. As shown in Figure 5b, given temporal-multimodal output features $\{\mathbf{f}_t^{\phi}, \mathbf{f}_t^{v}, \mathbf{f}_t^{1}, ..., \mathbf{f}_t^{n}\}$, we add them with prefix internal-state tokens $\{[sos], \phi_{T+1}, v_{T+1}, q_{T+1}^{1}, ..., q_{T+1}^{n-1}\}$ to predict next token:

$$q_{T+1}^{j} = \mathcal{F}_c([sos] + \mathbf{f}_T^{\phi}, ..., q_{T+1}^{j-1} + \mathbf{f}_T^{j}), j \in [1, n]. \tag{4}$$

The internal-frame AR module performs autoregression within each single time step, processing tokens in parallel to improve efficiency while still maintaining sequential dependencies within the frame. However, each front view image is discretized into 512 tokens, while only 2 tokens represent the pose (orientation and location). This imbalance can cause the model to overlook pose signals, leading to unsatisfactory controllable generation. To address this, we propose a novel **balanced attention** that adds regularization to achieve more precise control by prioritizing ego state tokens in the attention mechanism, rather than attending to all tokens equally. Specifically, we have m modalities with token counts $\{n_1, ..., n_m\}$. If position i belongs to modality j, the balanced weight $\hat{z}_i$:

$$\hat{z}_i = z_i + \frac{1}{mn_j}, \texttt{softmax}(\hat{z}_i) = \frac{e^{\hat{z}_i}}{\sum_k e^{\hat{z}_k}}, \tag{5}$$

where $\texttt{softmax}(\hat{z}_i)$ is the final weight.

Random Masking Strategy. Teacher forcing [35] is widely used in current next-token prediction models. Models use past ground-truth tokens to predict

the immediate next token. However, current next-token models struggle with stability when handling continuous data. For instance, when generating a video autoregressively—unlike text [2]—small errors in frame-to-frame predictions gradually accumulate beyond the training horizon, causing the model to diverge. To address this, we propose a random masking strategy (RMS), where some tokens from ground-truth tokens are randomly dropped out. There is a ϵ probability that the sequence performs random masking, and a $(1-\epsilon)$ probability that all ground-truth tokens are used. Each masked token has a η probability of being randomly replaced with an arbitrary value.

We use cross-entropy loss for training:

$$\mathcal{L}_{WM} = -\sum_{t=1}^{T+1}\sum_{i=1}^{n}\log p(q_t^i|\mathbf{q}_{<t}, q_t^{j<i}, \phi_{\leq t}, v_{\leq t}). \tag{6}$$

3.4 Decoder

The next-state tokens $(\phi_{T+1}, v_{T+1}, \mathbf{q}_{T+1})$ are predicted using the world model. Then we can leverage the decoder to generate the corresponding relative orientation $\Delta\theta_{T+1}$, relative location $(\Delta x_{T+1}, \Delta y_{T+1})$, and the reconstructed image $\mathbf{I}_{T+1}$for that state. This process allows us to map the predicted latent representations back into physical outputs, including both spatial and visual data.

Video Decoder. For the predicted image tokens $\mathbf{q}_{T+1}$, we retrieve the corresponding features from the codebook in video tokenizer. The feature is fed into a causal attention layer and several convolution layers to get the video frames.

Vehicle Pose Decoder. For the predicted relative orientation token ϕ_{T+1} and relative location token v_{T+1}, the corresponding values are obtained via the inverse of Eq. 1 as follows:

$$\begin{aligned}
\Delta\theta_t &= \theta_{min} + \frac{\phi_t}{\alpha}\left(\theta_{max} - \theta_{min}\right), \\
\Delta x_t &= x_{min} + \frac{1}{\beta}\left\lfloor\frac{v_t}{\gamma}\right\rfloor\left(x_{max} - x_{min}\right), \\
\Delta y_t &= y_{min} + \left(\frac{v_t}{\gamma}v_t - \left\lfloor\frac{v_t}{\gamma}\right\rfloor\right)\left(y_{max} - y_{min}\right).
\end{aligned} \tag{7}$$

4 Experiments

4.1 Implementation Details

Tokenizer and Decoder. The video tokenizer has $70M$ parameters. The size of adopted codebook is set to $16,384$. The model is trained for $1,000K$ steps with a total batch size of 128 on 32 NVIDIA 4090 GPUs, using images from Openimages [15], COCO [19], YoutubeDV [39], and NuPlan [4] datasets. We

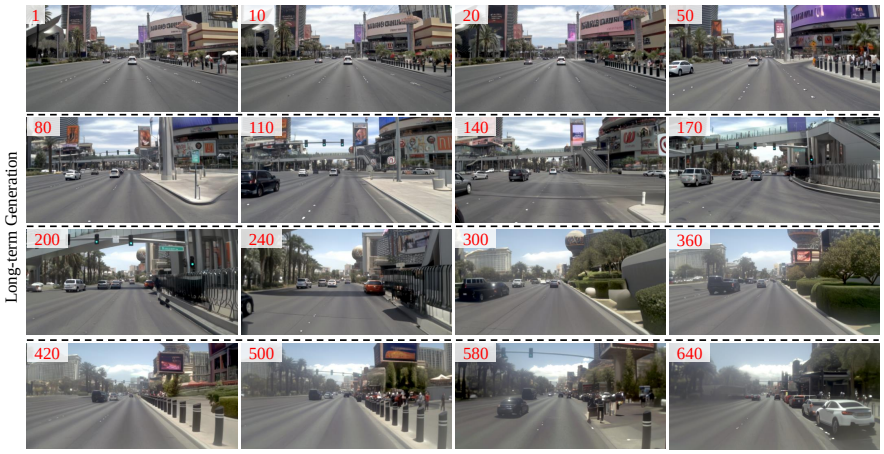


Fig. 7: Long duration video generation. We present some videos generated by our method, each with 640 frames at 5Hz, i.e. 128 seconds. Please notice the coherent 3D scene structures captured by our method across different frames.

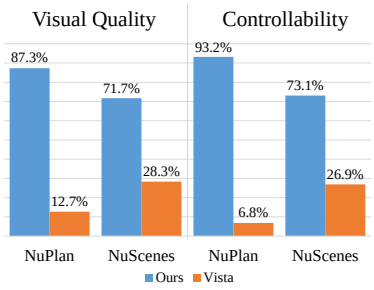


Fig. 8: User study. The value represents the percentage of times one model is favored over the other. Our model surpasses Vista [7] across all metrics.

train the temporal-aware VQVAE using a combination of three loss functions: charbonnier loss [16], perceptual loss [40], and codebook loss [30].

World Model. The world model has $1B$ parameters and is trained on video sequences. The model is conditioned on 15 frames to predict the next frame. The values of ϵ and η in RMS are set to 0.5 and 0.3, respectively. It is trained on over 3456 hours of human driving data. 120 hours come from the publicly available NuPlan [4] dataset, and 3336 hours consist of private data. Training is conducted over 12 days, spanning $450K$ iterations with a batch size of 64, distributed across 64 NVIDIA A100 GPUs. The model is fine-tuned on NuScenes [3] for one day.

Evaluation Dataset and Metrics. We use 200 video clips from the NuPlan [4] test dataset as our test set. Additionally, we include 150 video clips from the NuScenes [3] test set as part of our evaluation following Vista [7]. The quality of video generation is assessed using the Frechet Video Distance (FVD), and we

Table 2: Quantitative comparison of VQVAE methods on NuPlan [4].

VQVAE Methods	FVD_{12} ↓	FID ↓	PSNR ↑	LPIPS ↓
VAR [27]	164.66	11.75	22.35	0.2018
VQGAN [6]	156.58	8.46	21.52	0.2602
Llama-Gen [26]	57.78	5.99	22.31	0.2054
Llama-Gen [26] Finetuned	20.33	5.19	22.71	0.1909
Temporal-aware (Ours)	**14.66**	**4.29**	**23.82**	**0.1828**

also report the Frechet Inception Distance (FID) to evaluate image generation quality. All generated videos used for FVD and FID evaluation have the same frame length, following the setting in Vista [7].

4.2 Comparison of Generalization

We provide quantitative comparisons on the NuScenes [3] dataset in Table 1. *DrivingWorld* trained with NuScenes achieves the best performance. *DrivingWorld* trained without it reaches performance comparable to Vista. Additionally, our model can generate the longest video sequences—up to 40 seconds. We also adopt human evaluation for more reliable analysis [32]. As shown in Figure 8, our model not only generates higher visual quality but also effectively adapts to control inputs compared with the SoTA model Vista [7].

4.3 Results of Long-term Video Generation

One of the key advantages of our method is its ability to generate long-duration videos. As shown in Figure 7, we visualize one long-duration video generated by our model. By conditioning on just 15 frames, our model can generate up to 640 future frames at 10 Hz, resulting in 64-second videos with strong temporal consistency. These results demonstrate that our model maintains high video fidelity and preserves 3D structural integrity across the generated frames. In contrast, previous methods often struggle with drifting or degradation in long-duration videos. The ability to generate extended video sequences underscores our model's potential for tasks that require long-term predictions, such as autonomous driving or video synthesis in complex dynamic environments.

4.4 Quantitative Comparison of Image Tokenizers

We further evaluate our temporal-aware image tokenizer against those proposed in other works. Since the image tokenizer is part of a VQVAE, we assess the encoding-decoding performance of these VQVAEs instead. The experiments, conducted on the NuPlan [4] dataset, are summarized in Table 2. The VQVAE models from VAR [27] and VQGAN [6] demonstrate reasonable image quality in terms of PSNR and LPIPS scores, but both fall short on FID and FVD metrics. In contrast, Llama-Gen's VQVAE [26] shows significant improvements in

Table 3: End-to-end motion planning performance on the NuScenes [3] dataset. * represents only using the front camera as input.

Method	Input	Auxiliary Supervision	L2 (m) ↓				Collision Rate (%) ↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3 [13]	Camera	Map & Box & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [25]	Camera	Map & Box & Motion & Tracklets & Occ	0.48	0.96	1.65	1.03	0.05	**0.17**	**0.71**	**0.31**
OccNet [29]	Camera	3D-Occ & Map & Box	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
OccWorld [42]	Camera	3D-Occ	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87
VAD-Tiny [14]	Camera	Map & Box & Motion	0.60	1.23	2.06	1.30	0.31	0.53	1.33	0.72
VAD-Base [14]	Camera	Map & Box & Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
GenAD [43]	Camera	Map & Box & Motion	**0.36**	**0.83**	**1.55**	**0.91**	0.06	0.23	1.00	0.43
Ours	Camera*	None	0.68	1.20	1.90	1.26	**0.03**	0.22	0.90	0.38

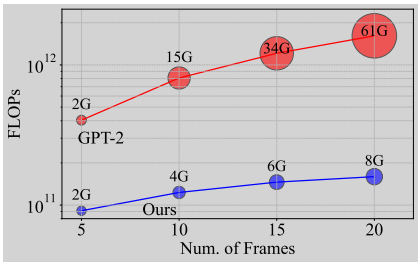


Fig. 9: Comparison of GPT-2 and *DrivingWorld* (ours).

FID and FVD scores. After fine-tuning it on driving scenes, we observe further enhancements in FVD performance. Ultimately, our temporal-aware VQVAE outperforms all others, enhancing temporal consistency and achieving the best scores across all four metrics.

4.5 Ablation Study

Comparison of BA and RMS. As shown in Table 4, removing either Balanced Attention (BA) or Random Masking Strategy (RMS) degrades performance, while using both together achieves the best results. This indicates that BA and RMS are complementary: BA helps preserve control information, and RMS improves long-term temporal modeling, leading to significantly lower FVD.
Discussion on Next-Token Prediction. We compare the memory usage and FLOPs of our *DrivingWorld* with the next-token prediction architecture, specifically GPT-2 [23], which processes tokens sequentially across all frames. As shown in Figure 9, GPT-2's memory usage grows rapidly with sequence length, making it inefficient for long sequences. In contrast, our method decouples temporal and multimodal dependencies for more efficient computation.

4.6 Comparison of Trajectory Planning

To demonstrate the applicability of our framework in autonomous driving, we evaluate its vehicle control capability through trajectory planning on the NuScenes

Table 4: Ablation study of Balanced Attention (BA) and Random Masking Strategy (RMS) on NuPlan [4].

Methods	$FVD_{10} \downarrow$	$FVD_{25} \downarrow$	$FVD_{40} \downarrow$
w/o BA & *w/o* RMS	463.21	603.88	665.15
w/o BA	240.25	319.35	361.94
w/o RMS	214.69	283.71	324.58
Ours	**122.37**	**144.26**	**159.46**

benchmark [3], following prior works [13, 25] with L2 error and collision rate as the evaluation metrics, as shown in Table 3. Requiring only front-view images, our method delivers competitive SOTA-level performance and robustness with high practicality for perception-limited scenarios.

5 Conclusion and Future Work

In conclusion, *DrivingWorld* addressed two key limitations of previous video generation models in autonomous driving by leveraging a next-token prediction framework to produce longer, high-fidelity video predictions without control signal degradation or computational inefficiency. Unlike traditional methods that struggled with coherence in long sequences, *DrivingWorld* generated realistic, structured video sequences while enabling precise action control. Our proposed next-frame AR module adopted a next-frame prediction strategy to model temporal coherence between consecutive frames and spatial reasoning within frames. We also implemented an internal-frame AR module to capture spatial information in each frame. Future work will integrate multimodal and multi-view data to improve understanding and control, advancing autonomous driving capabilities.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv:2311.15127 (2023)
2. Brown, T.B.: Language models are few-shot learners. arXiv:2005.14165 (2020)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv:2106.11810 (2021)
5. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv:1904.10509 (2019)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)

7. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. arXiv:2405.17398 (2024)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752 (2023)
9. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv:1912.01603 (2019)
10. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. arXiv:2010.02193 (2020)
11. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv:2301.04104 (2023)
12. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv:2309.17080 (2023)
13. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: European Conference on Computer Vision. pp. 533–549. Springer (2022)
14. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)
16. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2599–2613 (2018)
17. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
18. Li, A., Gong, B., Yang, B., Shan, B., Liu, C., Zhu, C., Zhang, C., Guo, C., Chen, D., Li, D., et al.: Minimax-01: Scaling foundation models with lightning attention. arXiv:2501.08313 (2025)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
21. Lu, J., Huang, Z., Zhang, J., Yang, Z., Zhang, L.: WoVoGen: World Volume-Aware Diffusion for Controllable Multi-Camera Driving Scene Generation. arXiv:2312.02934 (2023)
22. Radford, A.: Improving language understanding by generative pre-training (2018)
23. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
24. Santana, E., Hotz, G.: Learning a Driving Simulator. arXiv:1608.01230 (2016)
25. Shi, X., Cui, B., Dobbie, G., Ooi, B.C.: Uniad: A unified ad hoc data processing system. ACM Transactions on Database Systems (TODS) **42**(1), 1–42 (2016)

26. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv:2406.06525 (2024)
27. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. arXiv:2404.02905 (2024)
28. Tiezzi, M., Casoni, M., Betti, A., Guidi, T., Gori, M., Melacci, S.: Back to recurrent processing at the crossroad of transformers and state-space models. Nature Machine Intelligence pp. 1–11 (2025)
29. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8406–8415 (2023)
30. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
31. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-driven world models for autonomous driving. arXiv:2309.09777 (2023)
32. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision pp. 1–20 (2024)
33. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: 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. pp. 14749–14759 (2024)
34. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
35. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural computation **1**(2), 270–280 (1989)
36. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv:2104.10157 (2021)
37. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., Zhang, J., Geiger, A., Qiao, Y., Li, H.: Generalized Predictive Model for Autonomous Driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
38. Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.W., Shin, J.: Generating videos with dynamics-aware implicit generative adversarial networks. arXiv:2202.10571 (2022)
39. Zhang, Q., Peng, Z., Zhou, B.: Learning to drive by watching youtube videos: Action-conditioned contrastive policy pretraining. In: European Conference on Computer Vision. pp. 111–128. Springer (2022)
40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
41. Zheng, C., Vuong, T.L., Cai, J., Phung, D.: Movq: Modulating quantized vectors for high-fidelity image generation. Advances in Neural Information Processing Systems **35**, 23412–23425 (2022)
42. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
43. Zheng, W., Song, R., Guo, X., Zhang, C., Chen, L.: Genad: Generative end-to-end autonomous driving. In: European Conference on Computer Vision. pp. 87–104. Springer (2024)