UniDrive-WM: Unified Understanding, Planning and Generation World Model for Autonomous Driving

Zhexiao Xiong[1,2⋆], Xin Ye[1], Burhan Yaman[1], Sheng Cheng[3], Yiren Lu[1,4], Jingru Luo[1], Nathan Jacobs[2], and Liu Ren[1]

[1] Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)
[2] Washington University in St. Louis
[3] Arizona State University
[4] Case Western Reserve University

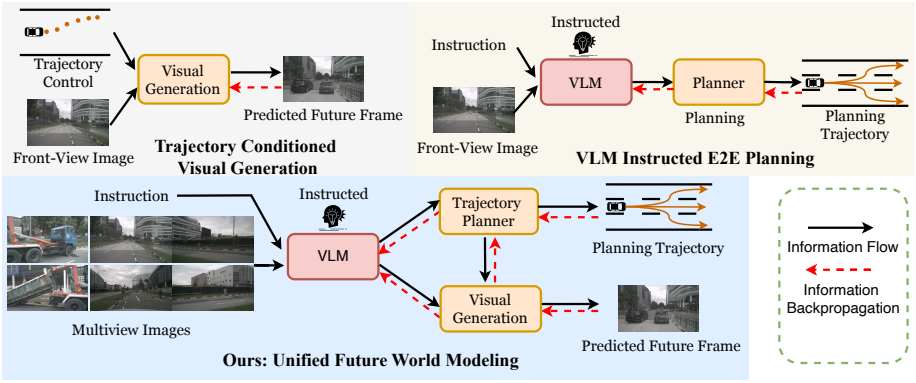

Fig. 1: Top Left: trajectory-conditioned visual generation. Top Right: VLM-instructed E2E Planning. Bottom: our unified future world modeling method. Compared with conditioned future image generation models and VLM-instructed planning method, our framework establishes the connection among the reasoning, action and visual generation spaces via joint VLM-guided trajectory planner and future frame generation.

Abstract. World models have become central to autonomous driving, where accurate scene understanding and future prediction are crucial for safe control. Recent work has explored using vision–language models (VLMs) for planning, yet existing approaches typically treat perception, prediction, and planning as separate modules. We propose **UniDrive-WM**, a unified VLM-based world model that jointly performs driving-scene understanding, trajectory planning, and trajectory-conditioned future image generation within a single architecture. UniDrive-WM's trajectory planner predicts a future trajectory, which conditions a VLM-based image generator to produce plausible future frames. These predictions provide additional supervisory signals that enhance scene understanding and iteratively refine trajectory generation. We further compare

⋆ Work was done during internship at Bosch Research North America.

discrete and continuous output representations for future image prediction, analyzing their influence on downstream driving performance. Experiments on the challenging Bench2Drive benchmark show that UniDrive-WM produces high-fidelity future images and improves planning performance by 7.3% in L2 trajectory error and 10.4% in collision rate over the previous best method. These results demonstrate the advantages of tightly integrating VLM-driven reasoning, planning, and generative world modeling for autonomous driving. The project page is available at `https://unidrive-wm.github.io/UniDrive-WM`.

Keywords: World Models · VLM · Autonomous Driving

1 Introduction

Recent progress in multimodal large language models (MLLMs) has been propelled by the strong perception, reasoning, and instruction-following capabilities of vision–language models (VLMs). In parallel, visual generation has advanced along two complementary lines: autoregressive (AR) token prediction and diffusion-based continuous generation—enabling high-fidelity image synthesis across diverse tasks. Motivated by these trends, a growing body of *unified models* seeks to couple understanding and generation within a single VLM, allowing the system to "think and generate" in a shared semantic space [5, 28, 55].

In autonomous driving, world models [35, 66, 66] are typically defined as learned models that infer a latent representation of the current scene and predict future states conditioned on actions, ideally enabling robust planning and control. While several VLM-based approaches have been explored, many rely on text-only intermediates, first producing natural-language descriptions of future trajectories or scenes, and then invoking a separate stage to decode trajectories or render images. This pipeline introduces an information bottleneck: rich visual–geometric cues are abstracted into text, incurring inevitable loss and compounding errors across stages. Meanwhile, generative models can generate visually plausible frames, but they typically lack explicit state estimation and reasoning. In particular, they do not maintain structured representations of the scene and cannot reliably condition on multi-view cues or high-level instructions, which provides no differentiable bridge to the action space for verifiable planning. Consequently, they struggle to answer causal queries (e.g., "what if the pedestrian accelerates?"), to enforce safety constraints, or to propagate plan-consistent signals back into perception. In reality, a driver's cognition is inherently joint: perceive the current scene, anticipate the future trajectory, and imagine the next visual scene. This motivates a unified formulation in which understanding, planning, and visual prediction are learned end-to-end within a single VLM framework, allowing information to flow bidirectionally between reasoning, action, and generation.

We present **UniDrive-WM**, a unified world model that *jointly* performs scene understanding, trajectory planning, and future image generation within a VLM-centric architecture. As illustrated in Fig. 1, multi-view observations,

temporal history, and perception cues (e.g., subject bounding boxes) are encoded and projected into an LLM reasoning space. A trajectory planner then produces a differentiable latent distribution over waypoints, bridging the reasoning (language–vision) space and the numeric action space. Conditioning on the predicted trajectory, UniDrive-WM further performs *trajectory-conditioned* future image prediction via two complementary designs: (i) a discrete AR pathway that expands the visual codebook and detokenizes with MoVQGAN, and (ii) an AR+diffusion pathway that predicts continuous latent features with a flow-matching objective before pixel decoding. This design alleviates text-only bottlenecks and enables bidirectional coupling—both information flow and gradient flow—between reasoning, action, and generation. Extensive experiments on the challenging Bench2Drive and nuScenes datasets show that UniDrive-WM generates high-fidelity future image frames and outperforms counterpart methods on planning tasks across both open-loop and closed-loop metrics.

Our main contributions are summarized as follows:

- We present *UniDrive-WM*, a unified vision–language world model (VLM) that seamlessly integrates scene understanding, trajectory planning, and future image generation, enabling direct visual reasoning from spatio–temporal observations.
- We develop and analyze two complementary decoding paradigms for trajectory-conditioned future image prediction: a *discrete* autoregressive (AR) pathway and a *continuous* AR+diffusion pathway, revealing their respective advantages and trade-offs for autonomous driving.
- Extensive experiments demonstrate that our unified framework achieves high-fidelity, planning-conditioned visual generation and significantly improves both planning accuracy and perception performance on standard driving benchmarks, verifying the effectiveness of the proposed framework.

2 Related Work

World Models for Autonomous Driving. The commonly accepted description of world models is understanding the current state and predicting the future state. For autonomous driving tasks, recent world-modeling-based methods aim to infer the ego status and dynamic environments from past observations to enable future planning and prediction, reducing human control in autonomous driving systems. Existing methods can be broadly categorized by their output modality: some focus on visual scene generation or urban view synthesis [12, 13, 14, 16, 27, 58], others emphasize future trajectory planning, including diffusion-based planning [11, 30, 32, 54], while several works address 3D scene reconstruction and simulation in the form of occupancy [34, 43, 48], LiDAR [36, 47, 51, 68], or point clouds [59, 60]. However, most of these approaches focus on a single prediction modality. Recent world models have also been studied for visual navigation [8, 9] and driver-centric dynamics rollout [6]. In contrast, our model jointly performs trajectory planning and future image

generation within a unified world-model framework, coupling planned motion with predicted visual futures under a shared VLM backbone.

Unified Image Understanding and Generation. Recent studies have emphasized the importance of unifying image understanding and image generation within a single framework. Early works such as Chameleon [40], Show-o [52], Transfusion [67], and Janus [49] adopt discrete visual representations and treat visual synthesis as autoregressive token prediction, where images are quantized into discrete tokens analogous to text, enabling LLMs to interpret and generate visual content in a unified token space. More recent models, including Metamorph [42] and the BLIP-3o family [3, 4], integrate autoregressive generation with diffusion-based continuous decoding, often relying on additional visual encoders such as CLIP [37] or SigLIP [57] to enhance visual understanding. In contrast to these general-purpose unified models, our work advances this paradigm in the autonomous driving domain by jointly predicting future frames and planning trajectories within a single VLM-based world model.

Vision–Language–Action Models. Recently, large vision–language–action (VLA) models have emerged as powerful frameworks for embodied decision-making. Built upon pretrained multimodal large language models (MLLMs), these systems typically predict future actions through either discrete action decoders [2, 25, 61] or continuous diffusion-based policy heads [15, 29]. By leveraging large-scale and diverse training corpora, MLLMs provide strong semantic grounding and generalization capabilities for downstream embodied tasks. Recent driving-oriented VLA/action models further explore video-action priors and geometry-aware action modeling for autonomous driving [33, 53]. In the autonomous driving setting, we treat ego-vehicle trajectories as the continuous action representation. Unlike standard manipulation or navigation environments, however, driving scenes are highly dynamic and visually complex, requiring the VLM to integrate richer contextual information—including temporal history, perception features, and multi-view observations to generate reliable and scene-consistent trajectory plans.

3 Method

In this section, we present our unified understanding, generation and planning framework for autonomous driving. The pipeline is shown in Fig. 2. We begin with the formulation of our method, followed by a detailed analysis of the system architecture. We then analyze the model architecture for planning and image generation. For image generation, we explore both (1) autoregressive image generation in a discrete visual representation and (2) diffusion-based generation in a continuous visual representation, and discuss their performance and impact on autonomous driving tasks. The task is defined as jointly predicting the future scene state and planning trajectory:

$$\{\hat{\mathbf{s}}_{t+n}, \hat{\mathbf{a}}_{t:t+m}\} \sim P_\theta\big(\mathbf{s}_{1:t},\, \mathbf{a}_{1:t-1},\, l\big), \tag{1}$$

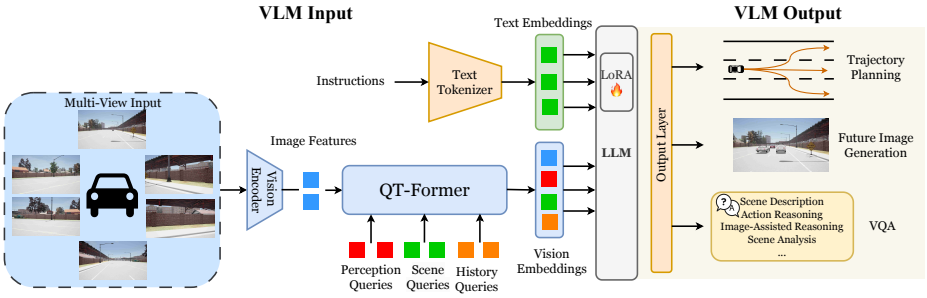

Fig. 2: The pipeline of our UniDrive-WM framework. The pipeline consists of: (1) a QT-Former-based encoder to extract historical context and multi-view visual input; (2) the LLM for performing a reasoning task; and (3) the output layer, which generates the planning trajectory and future image prediction and bridges the gaps between the planning space, image space and reasoning space. For the output layer, we provide a detailed analysis in Fig. 3.

which can be further decomposed as:

$$\begin{aligned} \{\hat{\mathbf{a}}_t, \ldots, \hat{\mathbf{a}}_{t+m}\} &\sim P_\theta(\{\mathbf{a}_t, \ldots, \mathbf{a}_{t+m}\} \mid \mathbf{s}_t, l)\,, \\ \hat{\mathbf{s}}_{t+n} &\sim P_\theta(\mathbf{s}_{t+n} \mid \mathbf{s}_t, \hat{\mathbf{a}}_t, \ldots, \hat{\mathbf{a}}_{t+m}, l)\,, \end{aligned} \tag{2}$$

where $\mathbf{s}_t$ represents the multi-modal state information at time t, including multi-view images, historical context, and perception features; the model predicts the future front-view image $\hat{\mathbf{I}}_{t+n}$ as part of the future state $\hat{\mathbf{s}}_{t+n}$, and $\mathbf{a}_{t+m}$ denotes the predicted trajectory waypoint at time $t+m$; l is the high-level language or instruction condition. The model therefore aims to jointly reason about the future world state and plan consistent trajectories under a unified vision–language world model.

3.1 Vision Language Model

A key component of a world model is understanding the current state and predicting the future state. Therefore, leveraging the Vision Language Model's reasoning and understanding ability, UniDrive-WM understands the current scene and instructs the trajectory planner and image generation module. We build our model upon ORION [11], a VLM-based model for the autonomous driving planning task.

Vision Encoder For the vision encoder, we adopt the QT-Former used by previous methods [11]. Specifically, two sets of learnable queries are processed through self-attention (SA) to exchange information and interact with image features F_m with 3D positional encoding in the cross-attention module. After that, the perception queries are fed into multiple auxiliary heads for object detection, including critical objects, lanes, traffic states and motion prediction of dynamic objects. Additionally, we use a set of history queries $Q_h \in \mathbb{R}^{(N_h \times C_q)}$, where N_h denotes the number of history queries and C_q represents the feature

dimension, and a memory bank $M \in \mathbb{R}^{(N_h \times n) \times C_q}$ to retrieve and store historical information from the past n frames. The QT-Former is formulated as:

$$Q_h = \mathrm{CA}\left(Q_h, M + P_t, M + P_t\right), \quad \hat{Q}_h = \mathrm{CA}\left(Q_h, Q_s, Q_s\right). \tag{3}$$

where CA denotes the Cross-Attention operation, and P_t denotes the relative timestamp embedding at the current timestep t. The updated history queries $\hat{Q}_h$ are stored in the memory bank M, formulated as:

$$M = \left[\hat{Q}_h^{t-n+1}, \cdots, \hat{Q}_h^{t-1}, \hat{Q}_h^{t}\right] \tag{4}$$

Finally, a two-layer Multi-layer Perceptron (MLP) is used to convert the updated history queries $\hat{Q}_h$ and current scene features Q_s to history tokens x_h and scene tokens x_s in the reasoning space of the LLM.

Large Language Model. The large language model (LLM) serves as the reasoning core of our framework, enabling text-based understanding and high-level reasoning tasks such as scene description, visual question answering, and action reasoning. We fine-tune the LLM using LoRA [17] to efficiently adapt it to domain-specific reasoning and instruction-following within autonomous driving scenarios.

3.2 Trajectory Planner

The trajectory planner establishes a differentiable connection between the semantic reasoning space of the VLM and the numerical action space of trajectory prediction through distribution learning in a latent space. We formulate trajectory generation as a conditional distribution problem, where the planner models a multi-modal distribution over future trajectories conditioned on high-level reasoning embeddings from the VLM:

$$p_\theta(\mathbf{a}_{t:t+m} \mid \mathbf{s}_t, \mathbf{h}_{\mathrm{VLM}}), \tag{5}$$

where $\mathbf{a}_{t:t+m}$ denotes the predicted trajectory waypoints, $\mathbf{s}_t$ represents the current state embedding, and $\mathbf{h}_{\mathrm{VLM}}$ denotes the high-level reasoning embedding (hidden representation) extracted from the VLM, which encodes the fused multi-view observations, temporal history, and textual instructions. A latent variable $\mathbf{z}_a$ is introduced to capture the stochasticity of future motion in a Gaussian latent space, and the planner learns to decode diverse yet semantically consistent trajectories from $\mathbf{z}_a$. In practice, we omit the explicit KL regularization term used in conventional VAE objectives, as we empirically find it may destabilize training in large-scale multimodal setups. This design provides a stable and differentiable bridge between language-based reasoning and continuous trajectory prediction.

3.3 Future Image Generation

Since driving scenes are continuous in nature, autonomous driving datasets naturally contain dense future frames, allowing us to leverage abundant video data

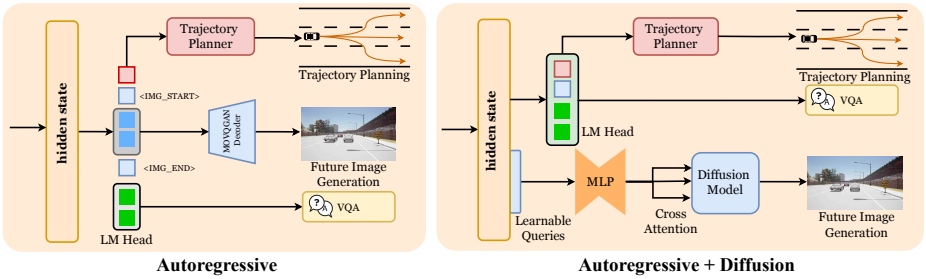

Fig. 3: The two design choices for image generation in a unified multimodal model. For future image prediction, we use both Autoregressive and Autoregressive+Diffusion architectures. (a) Left: Autoregressive architecture; (b) Right: AR+Diffusion architecture.

to improve image generation quality. Inspired by recent advances in unifying image understanding and generation, we adopt two different image generation architectures for future image prediction, as illustrated in Fig. 3. We discuss the design choices involved in building the understanding, planning, and image generation modules within a unified VLM framework, and analyze their impact on planning performance.

Formally, the future image prediction task can be formulated as:

$$p_\theta(\hat{\mathbf{I}}_{t+n} \mid \mathbf{s}_t, \mathbf{h}_{\text{VLM}}), \tag{6}$$

where $\mathbf{s}_t$ denotes the multi-modal state information (including multi-view observations, temporal history, and perception features), and $\mathbf{h}_{\text{VLM}}$ is the high-level reasoning embedding extracted from the VLM that encodes both visual and textual context. The model aims to predict the future front-view frame $\hat{\mathbf{I}}_{t+n}$ conditioned on these multi-modal context embeddings. This unified probabilistic formulation covers both the discrete **autoregressive (AR)** and the continuous **AR+Diffusion** architectures, which instantiate different parameterizations of the same conditional distribution $p_\theta(\cdot)$. We next describe the two paradigms instantiated under this formulation.

Autoregressive Generation(Discrete Representation). As shown in Fig. 3(a), for discrete image generation, we first activate the VLM's visual generation ability by enlarging its codebook. Assume that the language–vision model uses a joint codebook $Z = Z_L + Z_I$. Through enlarging this codebook, the VLM can autoregressively perform multi-modal next-token prediction in both language and vision spaces. We introduce special tokens $\langle image\_start \rangle$ and $\langle image\_end \rangle$ to indicate the boundaries of visual token sequences and when to use the vision head. Specifically, based on the input visual tokens and the planning token $\mathbf{a}_{t:t+m}$, the model autoregressively performs next-token prediction, represented

as:

$$P_\theta(q_{1:HW} \mid \mathbf{s}_t, \mathbf{a}_{t:t+m}) = \prod_{i=1}^{HW} P_\theta(q_i \mid q_{<i}, \mathbf{s}_t, \mathbf{a}_{t:t+m}), \tag{7}$$

where q_i are the visual tokens, $\mathbf{s}_t$ represents the current multi-modal state (including visual observations, temporal history, and perception features), and $\mathbf{a}_{t:t+m}$ denotes the corresponding planning token that conditions the generation of subsequent visual tokens.

AR+Diffusion Generation (Continuous Representation). Apart from fully autoregressive generation, we also design an **Autoregressive + Diffusion** architecture for improved visual quality, as shown in Fig. 3(b).

The diffusion-based generation can be formulated as a conditional denoising process:

$$p_\theta(\hat{\mathbf{I}}_{t+n} \mid \mathbf{s}_t, \mathbf{h}_{\text{VLM}}), \tag{8}$$

where the decoder is conditioned on the high-level reasoning embedding $\mathbf{h}_{\text{VLM}}$ and the latent image features produced by the autoregressive module. Starting from Gaussian noise $\mathbf{X}_0 \sim \mathcal{N}(0, 1)$, the diffusion process gradually refines $\mathbf{X}_0$ toward the latent embedding $\mathbf{X}_1$ of the ground-truth future image.

After extracting continuous visual embeddings, we employ an autoregressive transformer to generate corresponding latent image features conditioned on $\mathbf{h}_{\text{VLM}}$. Given an input instruction, the prompt is tokenized and mapped into a sequence of text embeddings $\mathbf{C} = [c_1, \ldots, c_n]$ before the LM head. To enable visual latent inference, we append a learnable query token $\mathbf{Q}$ to the sequence, where $\mathbf{Q}$ is randomly initialized and updated throughout training. The transformer processes the combined sequence $[\mathbf{C}; \mathbf{Q}]$, during which $\mathbf{Q}$ attends to the semantic context encoded in $\mathbf{h}_{\text{VLM}}$ and aggregates features relevant for image synthesis. The output query token, denoted as $\mathbf{Q}^*$, serves as the predicted visual latent representation and is supervised to match the ground-truth image embedding $\mathbf{X}$ extracted by the vision encoder. To align $\mathbf{Q}^*$ with $\mathbf{X}$, we use a flow-matching objective that models the continuous feature distribution. Given the ground-truth image feature $\mathbf{X}_1$ and the text-conditioned latent query $\mathbf{Q}$, we sample a timestep $t \sim \mathcal{U}(0, 1)$ and a noise vector $\mathbf{X}_0 \sim \mathcal{N}(0, 1)$. A latent point along the interpolation path is computed as

$$\mathbf{X}_t = t\mathbf{X}_1 + (1 - t)\mathbf{X}_0, \tag{9}$$

and the corresponding target velocity is

$$\mathbf{V}_t = \mathbf{X}_1 - \mathbf{X}_0. \tag{10}$$

The diffusion transformer predicts the velocity $\mathbf{V}_\theta(\mathbf{X}_t, \mathbf{Q}, t)$, and the flow-matching loss is

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\|\mathbf{V}_\theta(\mathbf{X}_t, \mathbf{Q}, t) - \mathbf{V}_t\|^2\right]. \tag{11}$$

Notably, we freeze the vision encoder weights and fine-tune only the diffusion decoder. Although the decoder adopts a diffusion architecture, it is trained with a

deterministic reconstruction loss rather than probabilistic sampling objectives. Consequently, during inference, the model performs deterministic reconstruction, which reduces diversity but ensures stable and accurate prediction of future frames in autonomous driving scenarios. We additionally apply CLIP-based supervision between the decoded future image and the ground-truth image to maintain semantic alignment, represented as:

$$\mathcal{L}_{\text{CLIP}} = 1 - \frac{\langle E_{\text{clip}}(I_{\text{pred}}), E_{\text{clip}}(I_{\text{gt}}) \rangle}{\|E_{\text{clip}}(I_{\text{pred}})\|_2 \, \|E_{\text{clip}}(I_{\text{gt}})\|_2}, \tag{12}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product, $E_{\text{clip}}(\cdot)$ denotes the embedding produced by the CLIP image encoder, I_{pred} is the predicted image, and I_{gt} is the ground-truth image. This diffusion pathway complements the discrete autoregressive generation branch by providing a continuous, geometry-aware latent space for future frame synthesis, leading to stable and semantically consistent visual predictions within the unified world-model framework.

3.4 Joint Image Generation and Planning

To achieve joint image generation and planning, we place the planning token immediately before the image tokens, so that the generation of each image token is conditioned on the generated planning representation. In this way, image generation is jointly conditioned on the current-state visual embeddings and the predicted trajectory of the future state, enabling the synthesis of consistent future frames aligned with planned motion.

For the autoregressive generation, image token prediction is treated analogously to language generation and supervised through teacher-forcing with a cross-entropy loss. We adopt the pretrained QT-Former from ORION [11] and freeze its detection head during training. For the planning branch, following the implementations of VAD [24] and ORION [11], we train only the planning head with:

$$\mathcal{L}_{\text{plan}} = \mathcal{L}_{\text{col}} + \mathcal{L}_{\text{bd}} + \mathcal{L}_{\text{mse}}, \tag{13}$$

where $\mathcal{L}_{\text{col}}$ denotes the collision loss, $\mathcal{L}_{\text{bd}}$ the boundary loss, and $\mathcal{L}_{\text{mse}}$ the mean squared error loss.

For the autoregressive architecture, the overall objective is represented as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{plan}}, \tag{14}$$

while for the autoregressive + diffusion architecture, where a learnable latent query in the latent space serves as the conditional embedding for image generation, the full objective becomes:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{plan}} + \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{CLIP}}, \tag{15}$$

where $\mathcal{L}_{\text{FM}}$ is the flow-matching loss and $\mathcal{L}_{\text{CLIP}}$ is the CLIP-based semantic alignment loss.

4 Experiments

4.1 Training Details

We follow ORION [11] for the detection training setup and conduct all experiments on 8×NVIDIA H200 GPUs. Following OmniDrive [44], we use EVA-02-L [10] as the vision encoder and adopt Vicuna 1.5 [62] as the base LLM, fine-tuned with LoRA (rank=16, α=16). The numbers of scene, perception, and history queries are set to 512, 600, and 16, respectively, with a memory bank size of 16 frames.

Input images are augmented and resized to 640×640. For future frame prediction, we use a 192×128 resolution for the autoregressive branch and 512×1024 for the AR+Diffusion branch to balance quality and speed. The diffusion decoder uses 64 latent learnable query tokens. Additional hyperparameters and implementation details are provided in the Appendix.

4.2 Dataset

We train and evaluate UniDrive-WM on the Bench2Drive [22] dataset and the nuScenes [1] dataset. For Bench2Drive, following prior work, we use the base split of 1000 driving scenes: 950 for training and 50 for open-loop validation. Each scene covers roughly 150 m of continuous driving in a distinct traffic scenario. For closed-loop evaluation, we follow the official protocol consisting of 220 short routes spanning 44 interactive scenarios (five routes per scenario). For nuScenes, we adopt the official train/validation split for all experiments.

4.3 Training Pipeline

To enable planning and image generation while preserving the VQA capability of the underlying VLM, we adopt a two-stage training strategy.

Stage 1: Joint Planning and Image Generation We train the full model end-to-end, with the LLM updated via LoRA. For the autoregressive generator, the image-start token is placed immediately after the planning token to enforce planning-conditioned autoregressive decoding. For the AR+Diffusion architecture, the 64 learnable latent query tokens are appended after the planning features in the latent space to condition the diffusion decoder.

Stage 2: Joint Planning, Image Generation, and VQA We continue training with mixed VQA and driving data while keeping the same architectural setup as Stage 1. This stage reinforces the alignment of the vision–language–planning space by jointly optimizing VQA, trajectory planning, and future image prediction, enabling unified multimodal reasoning within a single framework.

4.4 Results

Evaluation of Trajectory Planning Results We evaluate our results on the base validation set of Bench2Drive and nuScenes. We report both open-loop and

Table 1: Closed-loop and Open-loop Results of E2E-AD Methods on the Bench2Drive base set. C/L refers to camera/LiDAR. Avg. L2 is averaged over a 2-second prediction horizon sampled at 2 Hz, similar to UniAD. * denotes expert feature distillation. NC: navigation command, TP: target point, DS: Driving Score, SR: Success Rate.

Method	Condition	Modality	Closed-loop Metric				Open-loop Metric
			DS↑	SR(%)↑	Efficiency↑	Comfortness↑	Avg. L2 ↓
TCP* [50]	TP	C	40.70	15.00	54.26	47.80	1.70
TCP-ctrl*	TP	C	30.47	7.27	55.97	51.51	-
TCP-traj*	TP	C	59.90	30.00	76.54	18.08	1.70
TCP-traj w/o distillation	TP	C	49.30	20.45	78.78	22.96	1.96
ThinkTwice* [21]	TP	C	62.44	31.23	69.33	16.22	0.95
DriveAdapter* [20]	TP	C&L	64.22	33.08	70.22	16.01	1.01
AD-MLP [56]	NC	C	18.05	0.00	48.45	22.63	3.64
UniAD-Tiny [18]	NC	C	40.73	13.18	123.92	47.04	0.80
UniAD-Base [18]	NC	C	45.81	16.36	129.21	43.58	0.73
VAD [24]	NC	C	42.35	15.00	157.94	46.01	0.91
MomAD [39]	NC	C	44.54	16.71	170.21	48.63	0.87
GenAD [64]	NC	C	44.81	15.90	-	-	-
DriveTransformer-Large [23]	NC	C	63.46	35.01	100.64	20.78	0.62
ORION [11]	NC	C	77.74	54.62	151.48	17.38	0.68
Ours(AR)	NC	C	**79.22**	**56.36**	158.44	28.01	0.64
Ours(AR+Diffusion)	NC	C	**79.31**	**56.42**	158.65	27.93	0.63

Table 2: Open-Loop Planning and perception metrics on Bench2Drive. Lower is better for planning errors; higher is better for mAP/NDS.

Method	L2 (m)↓			Box + Collision(%)↓			mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	NDS↑
	1s	2s	3s	1s	2s	3s						
BEVFormer [31]	–	–	–	–	–	–	0.616	0.372	0.079	0.044	0.808	0.642
UniAD [18]	0.521	1.265	2.140	0.770	3.87	7.91	0.121	0.518	0.170	0.096	0.977	0.316
VAD [24]	0.454	0.912	1.477	0.102	0.202	0.296	0.509	0.385	0.0854	0.0308	0.594	0.604
Orion [11]	0.268	0.631	1.129	0.213	0.467	0.743	0.646	0.366	0.0765	0.0271	0.258	0.723
Ours(AR)	**0.247**	**0.598**	**1.079**	**0.198**	**0.435**	**0.668**	**0.663**	**0.335**	**0.0740**	**0.0262**	**0.249**	**0.746**
Ours(AR+Diff)	**0.241**	**0.589**	**1.066**	**0.192**	**0.426**	**0.657**	**0.675**	**0.328**	**0.0733**	**0.0254**	**0.243**	**0.755**

closed-loop evaluation results. As shown in Tab. 1, for closed-loop evaluation results on Bench2Drive, our method outperforms all the other methods that rely on target point and navigation command as conditions. Results show that our method achieves better performance compared with previous end-to-end methods and VLM-guided planning methods, which demonstrates the effectiveness of the image generation modality in boosting the performance of the planning task.

We also evaluate the open-loop performance on Bench2Drive in Tab. 2 and nuScenes in Tab. 3. Notably, UniAD [18] computes L2 metrics and collision rate at each timestep, whereas VAD [24] and ORION [11] consider the average of all previous time steps. We follow the evaluation method of VAD and ORION to evaluate our performance. Compared with previous methods that perform either VLM-guided planning or end-to-end planning, our unified method achieves better performance on both the planning and detection tasks, which shows the effectiveness of future image prediction in improving the performance of the planning task. These consistent gains across both Bench2Drive (synthetic) and

Table 3: Further Open-Loop evaluation on nuScenes.

Metric	VAD-Base [24]	Doe-1 [65]	Drive-VLM [41]	OmniDrive [44]	ORION [11]	FSDrive [55]	**Ours(AR)**	**Ours(AR+Diff)**
Avg. L2 (m)↓	1.25	0.70	0.40	0.84	0.34	0.60	**0.30**	**0.29**
Avg. col (%)↓	1.09	0.21	0.27	0.94	0.37	**0.19**	0.31	0.31

Table 4: Future frame generation results on the Bench2Drive [22] and nuScenes [1] datasets. AR: Autoregressive. AR+Diff: Autoregressive+Diffusion. Lower FID indicates better visual fidelity.

Method	**Dataset**	DriveGAN [26] [CVPR21]	DriveDreamer [45] [ECCV24]	Drive-WM [46] [CVPR24]	GEM [14] [CVPR25]	Doe-1 [65] [arXiv24]	FSDrive [55] [NeurIPS25]	**Ours(AR)**	**Ours (AR+Diff)**
Type	–	GAN	Diffusion	Diffusion	Diffusion	AR	AR	AR	AR+Diffusion
FID ↓	Bench2Drive	62.3	42.8	17.8	13.2	18.6	9.3	**7.2**	**6.6**
	nuScenes	73.4	52.6	15.8	10.5	15.9	10.1	**7.8**	**7.3**

NuScenes (real-world) further indicate the strong generalization ability of our method.

Evaluation of Image Generation Results We evaluate UniDrive-WM on both the autoregressive and AR+Diffusion image generation branches. Qualitative results are shown in Fig. 4 and Fig. 5, and quantitative results are provided in Tab. 4. Across both settings, our method produces visually coherent future frames that closely match the ground-truth scene evolution and remain consistent with the predicted trajectory. These properties translate into competitive or superior image generation quality, as reflected by low FID scores.

A particularly relevant comparison is FSDrive [55], a recent unified vision–planning framework that conditions planning on image features. Unlike FSDrive, our approach conditions image generation directly on the planning tokens, enabling planning-conditioned future frame prediction in both the AR and AR+Diffusion branches. This tighter coupling between planning and visual prediction improves trajectory alignment and yields future frames that more accurately follow the intended motion, contributing to our improved FID performance.

Evaluation on VQA Following prior work, we report our Visual Question Answering (VQA) results in Tab. 5, using DriveLM's GVQA [38] dataset. Our model achieves competitive performance against prior state-of-the-art methods on the understanding task, demonstrating its effectiveness in driving-domain-specific understanding.

4.5 Comparison of AR and AR+Diffusion Architecture

Speed The AR branch is intentionally lightweight (discrete tokens, shallow decoder); the AR+Diff branch is heavier (diffusion decoder) and mainly intended

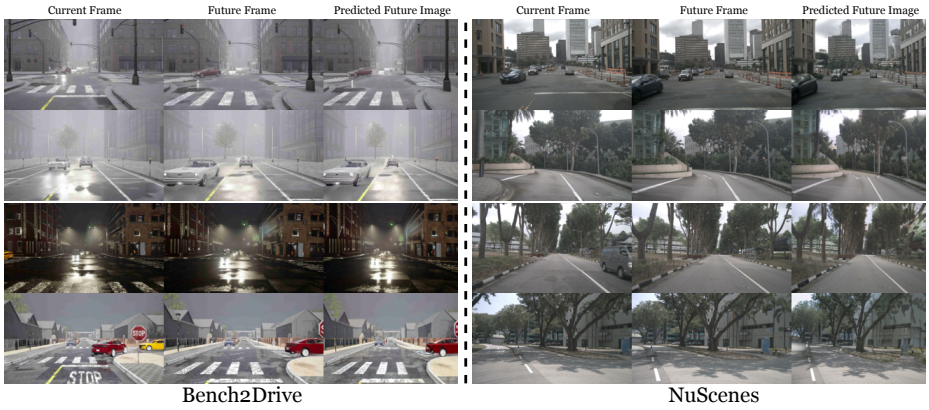

Fig. 4: Qualitative future image prediction results on Bench2Drive and nuScenes using the autoregressive (AR) architecture.

Fig. 5: Qualitative future image prediction results on nuScenes using the AR+Diffusion architecture.

for high-fidelity generation. In open-loop evaluation on Bench2Drive, the inference speed of AR is 2 fps and AR+Diff is 0.4 fps on an A100. We regard speeding up the AR+Diff inference as future work.

Resolution and Compression Trade-offs In the AR architecture, spatial resolution is tightly coupled to sequence length. Generating an $H \times W$ image with a downsampling factor f requires $\frac{H}{f} \times \frac{W}{f}$ discrete tokens. Scaling up resolution quadratically increases the token count, which not only hits the LLM's $O(N^2)$ context limit but also severely slows down training convergence. In contrast, our AR+Diffusion branch decouples resolution from token length. The VLM only needs to predict a compact sequence of continuous latents, which the diffusion decoder maps to high-resolution outputs (e.g., 512×1024). This enables high-fidelity geometric forecasting without overwhelming the reasoning context window or hindering training efficiency.

Image Generation and Complementary Roles Beyond resolution constraints, the two paradigms exhibit distinct characteristics in representation learning (Tab. 4). The discrete quantization in the AR branch acts as an information bottleneck that provides strong semantic regularization. This makes its lightweight decoder highly robust for real-time, latency-sensitive closed-loop control. In contrast, the continuous pathway of the AR+Diffusion branch avoids

Table 5: Results on the DriveLM [38] GVQA benchmark.

Method	Accuracy ↑	ChatGPT ↑	BLEU_1 ↑	ROUGE_L ↑	CIDEr ↑	Match ↑	Final Score ↑
DriveLM baseline [38]	0.00	0.65	0.05	0.08	0.10	0.28	0.32
Cube-LLM [7]	0.39	**0.89**	0.16	0.20	**0.31**	0.39	0.50
TrackingMeetsLMM [19]	0.60	0.58	0.72	0.72	0.04	0.36	0.52
SimpleLLM4AD [63]	0.66	0.57	0.76	0.73	0.15	0.35	0.53
OmniDrive [44]	0.70	0.65	0.52	0.73	0.13	0.37	0.56
FSDrive [55]	0.72	0.63	0.76	0.74	0.17	0.39	0.57
UniDrive-WM(ours)	**0.73**	0.67	**0.77**	**0.76**	0.20	**0.40**	**0.59**

Table 6: Ablation study of the detection head, planning head, and image generation module for our AR architecture on Bench2Drive [22].

Detection head	Planning head	Image Gen	L2 (m)↓			Box + Collision (%)↓			mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	NDS↑
			1s	2s	3s	1s	2s	3s						
	✓	✓	0.482	1.135	2.146	0.387	0.898	1.77	–	–	–	–	–	–
✓		✓	–	–	–	–	–	–	0.638	0.336	0.0726	0.0289	0.283	0.718
✓	✓		0.269	0.632	1.130	0.214	0.469	0.746	0.642	0.367	0.0767	0.0274	0.258	0.721
✓	✓	✓	**0.247**	**0.598**	**1.079**	**0.198**	**0.435**	**0.668**	**0.663**	**0.335**	**0.0740**	**0.0262**	**0.249**	**0.746**

Table 7: Ablation study on open-loop evaluation on Bench2Drive. For the AR architecture, we analyze the effect of swapping the order of the planning and generation heads.

Future Image Prediction	Avg. L2 (m)↓	Avg. col(%)↓	FID↓
Generation + Planning	0.67	0.45	9.1
Planning + Generation (**Ours**)	**0.64**	**0.43**	**7.2**

token-by-token exposure bias and inherently preserves high-frequency geometric details. This yields higher-fidelity future frames and more precise long-term trajectory alignment (e.g., lower L2 errors and higher NDS). Implemented as parallel decoding heads on top of the same VLM planning backbone, the two branches offer a natural deployment split: AR is suited to fast, reactive planning where semantic stability is critical, while AR+Diffusion is preferable for offline evaluation or high-fidelity geometric forecasting in complex scenes. Overall, the two heads are complementary rather than competing, and together broaden the applicability of UniDrive-WM.

4.6 Ablation Study

We ablate the detection supervision and the image generation module to assess their contributions to planning, shown in Tab. 6. Removing the image generation head noticeably degrades trajectory accuracy and increases collision cases, indicating that future frame prediction provides useful auxiliary signals for planning. Disabling

Table 8: Additional ablation of VQA score on Chat-B2D.

Image Generation	CIDEr ↑	BLEU ↑	ROUGE-L↑
-	65.7	52.4	77.5
✓	**66.7**	**53.5**	**78.6**

detection supervision further harms planning performance by weakening perception quality. Together, these findings show that both perception and future frame prediction play complementary roles, and that jointly optimizing all components within a unified world model yields more reliable driving behavior.

We show experiments regarding changing the order of the planning and image generation modules in the AR architecture in Tab. 7. **Planning → Generation (Ours)** shows improvements over **Generation → Planning** on planning-related metrics and a much larger gain in visual quality, suggesting that the prediction order matters; conditioning future prediction on planned trajectories aligns with the action→observation causal direction and provides a natural mechanism to produce plan-consistent futures that better support planning.

We evaluate VQA performance on Bench2Drive [22] using the Chat-B2D QA annotations [11], shown in Tab. 8. Compared with the model variant without the image-generation module, incorporating image generation consistently improves all VQA metrics. This indicates that the future frame prediction provides additional structural cues that benefit question answering. Intuitively, better modeling of future visual states enhances the model's understanding of the current scene, which aligns with human perception. We also show the visualization of the VQA result in open-loop evaluation in the supplementary material. These results demonstrate that our model not only comprehends the current scene, but also performs effective future-state reasoning, enabling it to produce reliable, decision-oriented answers even in complex driving scenarios and challenging weather conditions.

5 Conclusion and Future Work

We presented UniDrive-WM, a unified framework that integrates scene understanding, trajectory planning, and visual generation into a single world-model–driven pipeline. Leveraging a vision–language model (VLM) as the backbone, our method predicts future image frames from current and historical multi-view observations together with the planning tokens, and the planning-conditioned visual predictions in turn enhance trajectory planning by providing a visual forecast of the expected future scene. Experimental results demonstrate that our approach not only achieves high-fidelity, planning-conditioned image generation, but also yields significant gains in planning accuracy and collision reduction. By conditioning the VLM on rich multi-view inputs, temporal history, and perception features, UniDrive-WM establishes a coherent bridge between reasoning, action, and visual imagination. Looking ahead, we plan to extend the framework to more interactive and long-horizon driving scenarios, paving the way toward next-generation world models for autonomous driving.

Bibliography

[1] 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)

[2] Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)

[3] Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., Xue, L., Xiong, C., Xu, R.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset (2025), https://arxiv.org/abs/2505.09568

[4] Chen, J., Xue, L., Xu, Z., Pan, X., Yang, S., Qin, C., Yan, A., Zhou, H., Chen, Z., Huang, L., et al.: Blip3o-next: Next frontier of native image generation. arXiv preprint arXiv:2510.15857 (2025)

[5] Chern, E., Hu, Z., Chern, S., Kou, S., Su, J., Ma, Y., Deng, Z., Liu, P.: Thinking with generated images. arXiv preprint arXiv:2505.22525 (2025)

[6] Chi, H., Qiu, D., Su, H., Liu, H., Li, Z., Zhang, H., Lv, C.: Driver-wm: A driver-centric traffic-conditioned latent world model for in-cabin dynamics rollout. arXiv preprint arXiv:2605.05092 (2026)

[7] Cho, J.H., Ivanovic, B., Cao, Y., Schmerling, E., Wang, Y., Weng, X., Li, B., You, Y., Krähenbühl, P., Wang, Y., et al.: Language-image models with 3d understanding. arXiv preprint arXiv:2405.03685 (2024)

[8] Dong, Y., Wu, F., Chen, G., Cheng, Z.Q., Hu, Q., Zhou, Y., Sun, J., He, J.Y., Dai, Q., Hauptmann, A.G.: Unified world models: Memory-augmented planning and foresight for visual navigation. arXiv preprint arXiv:2510.08713 (2025)

[9] Dong, Y., Wu, F., Dai, Y., Kong, L., Chen, G., Zhu, X., Hu, Q., Wang, T., Garnica, J., Liu, F., et al.: Language-conditioned world modeling for visual navigation. arXiv preprint arXiv:2603.26741 (2026)

[10] Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing **149**, 105171 (2024)

[11] Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X.: Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. arXiv preprint arXiv:2503.19755 (2025)

[12] 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. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)

[13] Han, X., Jia, Z., Li, B., Wang, Y., Ivanovic, B., You, Y., Liu, L., Wang, Y., Pavone, M., Feng, C., et al.: Extrapolated urban view synthesis benchmark.

In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28718–28728 (2025)
[14] Hassan, M., Stapf, S., Rahimi, A., Rezende, P., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., et al.: Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22404–22415 (2025)
[15] He, H., Zhang, Y., Lin, L., Xu, Z., Pan, L.: Pre-trained video generative models as world simulators. arXiv preprint arXiv:2502.07825 (2025)
[16] 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 preprint arXiv:2309.17080 (2023)
[17] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
[18] Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
[19] Ishaq, A., Lahoud, J., Khan, F.S., Khan, S., Cholakkal, H., Anwer, R.M.: Tracking meets large multimodal models for driving scenario understanding. arXiv preprint arXiv:2503.14498 (2025)
[20] Jia, X., Gao, Y., Chen, L., Yan, J., Liu, P.L., Li, H.: Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: ICCV (2023)
[21] Jia, X., Wu, P., Chen, L., Xie, J., He, C., Yan, J., Li, H.: Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21983–21994 (June 2023)
[22] Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. Advances in Neural Information Processing Systems **37**, 819–844 (2024)
[23] Jia, X., You, J., Zhang, Z., Yan, J.: Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. arXiv preprint arXiv:2503.07656 (2025)
[24] 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. ICCV (2023)
[25] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
[26] Kim, S.W., Philion, J., Torralba, A., Fidler, S.: Drivegan: Towards a controllable high-quality neural simulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5820–5829 (2021)

[27] Lee, J.Y., Liu, Y.R., Tsai, S.R., Chang, W.C., Wu, C.H., Chan, J., Zhao, Z., Lin, C.H., Liu, Y.L.: Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. arXiv preprint arXiv:2510.15869 (2025)
[28] Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542 (2025)
[29] Li, S., Gao, Y., Sadigh, D., Song, S.: Unified video action model. arXiv preprint arXiv:2503.00200 (2025)
[30] Li, Y., Shang, S., Liu, W., Zhan, B., Wang, H., Wang, Y., Chen, Y., Wang, X., An, Y., Tang, C., et al.: Drivevla-w0: World models amplify data scaling law in autonomous driving. arXiv preprint arXiv:2510.12796 (2025)
[31] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arXiv preprint arXiv:2203.17270 (2022)
[32] Liao, Y., Sun, Z., Qiu, X., Zhao, Z., Tang, W., He, X., Zheng, X., Zhang, T., Huang, X., Han, X.: Work zones challenge vlm trajectory planning: Toward mitigation and robust autonomous driving. arXiv preprint arXiv:2510.02803 (2025)
[33] Liu, M., Zhang, D., Liu, J., Cui, J., Xie, H., Chen, G., Ye, H., Yang, M.Y., Nex, F., Cheng, H.: Driveva: Video action models are zero-shot drivers. arXiv preprint arXiv:2604.04198 (2026)
[34] Liu, T., Zhao, S., Pourkeshavarz, M., Li, W., Rhinehart, N.: Occsim: Multi-kilometer simulation with long-horizon occupancy world models. arXiv preprint arXiv:2603.28887 (2026)
[35] Liu, T., Zhao, S., Rhinehart, N.: Towards foundational lidar world models with efficient latent flow matching. arXiv preprint arXiv:2506.23434 (2025)
[36] Liu, T., Zhao, S., Rhinehart, N.: Towards foundational lidar world models with efficient latent flow matching. Advances in Neural Information Processing Systems **38**, 155959–155994 (2026)
[37] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
[38] Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: European conference on computer vision. pp. 256–274. Springer (2024)
[39] Song, Z., Jia, C., Liu, L., Pan, H., Zhang, Y., Wang, J., Zhang, X., Xu, S., Yang, L., Luo, Y.: Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22432–22441 (2025)
[40] Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
[41] Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)

[42] Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164 (2024)
[43] Wang, L., Zheng, W., Ren, Y., Jiang, H., Cui, Z., Yu, H., Lu, J.: Occsora: 4d occupancy generation models as world simulators for autonomous driving. arXiv preprint arXiv:2405.20337 (2024)
[44] Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22442–22452 (2025)
[45] Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-driven world models for autonomous driving. arXiv preprint arXiv:2309.09777 (2023)
[46] 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. arXiv preprint arXiv:2311.17918 (2023)
[47] Wei, H., Lu, F., Zhu, Y., Zheng, Z., Xue, W., Shao, L., Zhang, X., Wu, Y., Fu, R., Chen, G.: Lidardraft: Generating lidar point cloud from versatile inputs. arXiv preprint arXiv:2512.20105 (2025)
[48] Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W.: Occllama: An occupancy-language-action generative world model for autonomous driving. arXiv preprint arXiv:2409.03272 (2024)
[49] Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
[50] Wu, P., Jia, X., Chen, L., Yan, J., Li, H., Qiao, Y.: Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. Advances in Neural Information Processing Systems **35**, 6119–6132 (2022)
[51] Wu, Z., Ni, J., Wang, X., Guo, Y., Chen, R., Lu, L., Dai, J., Xiong, Y.: Holodrive: Holistic 2d-3d multi-modal street scene generation for autonomous driving. arXiv preprint arXiv:2412.01407 (2024)
[52] Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
[53] Yao, J., Kurra, D.D., Lampo, T., Cheng, Z., Guo, D., Yaman, B.: Vlga: Vision-language-geometry-action models for autonomous driving. arXiv preprint arXiv:2606.12396 (2026)
[54] Yu, H., Jin, Y., Canevaro, A., Schmidt, J., Jordan, J., Li, P., Kaufeld, M., Lindner, S., Betz, J., Stork, W.: G2dp: Diffusion planning with spatio-temporal grid guidance. arXiv preprint arXiv:2606.26017 (2026)
[55] Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025)

[56] Zhai, J.T., Feng, Z., Du, J., Mao, Y., Liu, J.J., Tan, Z., Zhang, Y., Ye, X., Wang, J.: Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. arXiv preprint arXiv:2305.10430 (2023)
[57] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
[58] Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. arXiv preprint arXiv:2506.24113 (2025)
[59] Zhang, L., Xiong, Y., Yang, Z., Casas, S., Hu, R., Urtasun, R.: Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. arXiv preprint arXiv:2311.01017 (2023)
[60] Zhang, Y., Gong, S., Xiong, K., Ye, X., Li, X., Tan, X., Wang, F., Huang, J., Wu, H., Wang, H.: Bevworld: A multimodal world simulator for autonomous driving via scene-level bev latents. arXiv preprint arXiv:2407.05679 (2024)
[61] Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
[62] Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
[63] Zheng, P., Zhao, Y., Gong, Z., Zhu, H., Wu, S.: Simplellm4ad: An end-to-end vision-language model with graph visual question answering for autonomous driving. arXiv preprint arXiv:2407.21293 (2024)
[64] 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)
[65] Zheng, W., Xia, Z., Huang, Y., Zuo, S., Zhou, J., Lu, J.: Doe-1: Closed-loop autonomous driving with large world model. arXiv preprint arXiv: 2412.09627 (2024)
[66] Zheng, Y., Yang, P., Xing, Z., Zhang, Q., Zheng, Y., Gao, Y., Li, P., Zhang, T., Xia, Z., Jia, P., et al.: World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28632–28642 (2025)
[67] Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)
[68] Zyrianov, V., Che, H., Liu, Z., Wang, S.: Lidardm: Generative lidar simulation in a generated world. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6055–6062. IEEE (2025)