2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

UniFuture: A 4D Driving World Model for Future Generation and Perception

Dingkang Liang[1], Dingyuan Zhang[1], Xin Zhou[1], Sifan Tu[1],
Tianrui Feng[1], Xiaofan Li[2], Yumeng Zhang[2], Mingyang Du[1], Xiao Tan[2], Xiang Bai[†1]

***Abstract*— We present UniFuture, a unified 4D Driving World Model designed to simulate the dynamic evolution of the 3D physical world. Unlike existing driving world models that focus solely on 2D pixel-level video generation (lacking geometry) or static perception (lacking temporal dynamics), our approach bridges appearance and geometry to construct a holistic 4D representation. Specifically, we treat future RGB images and depth maps as coupled projections of the same 4D reality and model them jointly within a single framework. To achieve this, we introduce a Dual-Latent Sharing (DLS) scheme, which maps visual and geometric modalities into a shared spatio-temporal latent space, implicitly entangling texture with structure. Furthermore, we propose a Multi-scale Latent Interaction (MLI) mechanism, which enforces bidirectional consistency: geometry constrains visual synthesis to prevent structural hallucinations, while visual semantics refine geometric estimation. During inference, UniFuture can forecast high-fidelity, geometrically consistent 4D scene sequences (image-depth pairs) from a single current frame. Extensive experiments on the nuScenes and Waymo datasets demonstrate that our method outperforms specialized models in both future generation and geometry perception, highlighting the efficacy of unified 4D modeling for autonomous driving. The code is available at `https://github.com/dk-liang/UniFuture`.**

I. INTRODUCTION

The physical world in which autonomous vehicles operate is inherently four-dimensional, consisting of 3D spatial geometry evolving over the temporal dimension. Consequently, an ideal Driving World Model (DWM) should be capable of simulating this 4D dynamic evolution, enabling the ego-vehicle to anticipate how the 3D surroundings will unfold. While traditional DWMs [1], [2], [3], [4] have made significant strides in simulating future scenarios, they often fall short of capturing the full 4D nature of the environment.

Recent advancements in DWMs, driven by large-scale pre-trained diffusion models, have predominantly focused on fine-grained 2D video generation [2], [5], [6], [7], [8]. As shown in Fig. 1(a), these models excel at synthesizing visually realistic RGB sequences. However, they largely overlook the underlying 3D geometry, such as depth, essentially predicting cinematic hallucinations rather than physical realities. Without explicit geometric modeling, these 2D-based DWMs struggle with spatial reasoning tasks, including handling occlusions and estimating accurate distances. This results in predictions that may be visually plausible but physically inconsistent. Conversely, depth-aware perception models [9], [10], [11] excel at extracting high-level geometric structures but are typically limited to static 3D snapshots of the present or past. As illustrated in Fig. 1(b), they lack the capability to forecast how these 3D structures will evolve, missing the critical temporal dimension. These limitations point to a significant gap: *Can we develop a unified driving world model that integrates appearance, geometry, and dynamics to forecast the authentic 4D evolution of driving scenes?*

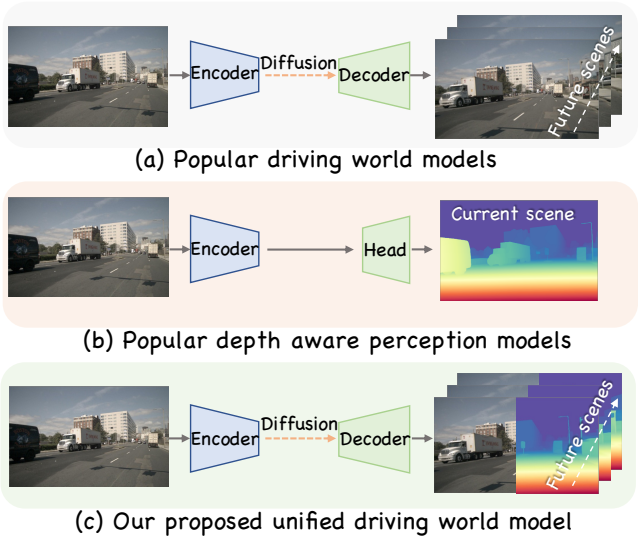

Fig. 1: An intuition comparison between existing methods and our unified world model. Unlike (a) 2D-focused generation and (b) static 3D perception, our UniFuture (c) jointly predicts future appearance and geometry, effectively modeling the 4D evolution of the world.

To bridge this gap, we propose UniFuture, a unified 4D Driving World Model that simultaneously forecasts future scene appearance (RGB) and geometry (Depth), as shown in Fig. 1(c). We posit that an image and its corresponding depth map are merely distinct projections of the same underlying 4D reality. Therefore, modeling them in isolation is suboptimal. To unify them, we introduce a Dual-Latent Sharing (DLS) scheme. Rather than training separate encoders for texture and geometry, DLS maps both modalities into a shared latent space. Implicitly, this constructs a unified spatio-temporal representation. This design not only eliminates the need for additional pre-training but also ensures that the generated geometry and texture are entangled at the feature level, mirroring their physical correlation in the real world.

Building on this unified representation, we propose a Multi-scale Latent Interaction (MLI) mechanism to enforce spatio-

[1] Huazhong University of Science and Technology.
[2]Baidu Inc., China.
†Corresponding author. dkliang@hust.edu.cn

temporal consistency. In a 4D world, geometry constrains visual appearance, while visual cues refine geometric estimation. MLI facilitates this bidirectional information flow within a multi-scale UNet architecture. By iteratively injecting depth latents into the image stream and propagating refined image features back to the depth stream, we ensure that the predicted future is not just a video, but a coherent 4D point cloud sequence (as demonstrated in our experiments). This mutual interaction integrates low-level pixel synthesis with high-level spatial reasoning, resulting in predictions that are geometrically consistent over time.

Our approach shifts from 2D video prediction to 4D world modeling, offering several key advantages. First, the synergy of appearance and geometry enhances prediction quality. Specifically, depth priors stabilize video generation against temporal distortions, while appearance priors refine the details of future depth estimation. Second, our unified framework serves as a powerful foundation for downstream tasks. It enables self-driving systems to simulate different scenarios and make informed decisions [12], while also generating highly consistent annotated data for automated training. Extensive experiments on the nuScenes dataset demonstrate that UniFuture achieves state-of-the-art performance. Compared to the strong baseline Vista [2], we reduce FID by 23.9% while simultaneously outperforming specialized diffusion-based depth estimation methods [13] in future geometric prediction.

The main contributions are summarized as follows: **1)** We propose UniFuture, a novel 4D Driving World Model framework. By seamlessly integrating future generation and perception, we extend world modeling from 2D pixel space to 4D geometric space. **2)** We introduce the Dual-Latent Sharing (DLS) scheme and Multi-scale Latent Interaction (MLI) mechanism. These modules effectively unify heterogeneous modalities in a shared latent space and enforce bidirectional spatio-temporal consistency. **3)** Our method achieves impressive performance in both future scene generation and depth estimation, demonstrating the potential of unified 4D modeling for autonomous driving.

II. RELATED WORK

A. World Models in Autonomous Driving

World models predict scene evolution based on current observations [14], [15] and have become a key focus in autonomous driving. Many approaches [16], [17], [5] excel in forecasting 2D visual representations, demonstrating strong dynamic modeling capabilities. Most methods rely on generative techniques like autoregressive transformers [18], [19] and diffusion models [20], [21], [22] for future video prediction. Specifically, GAIA-1 [18] formulates world modeling as a sequence task using an autoregressive transformer, while ADriver-I [23] integrates multi-modal LLMs and diffusion for control and frame generation. DriveDreamer [1] enforces structured traffic constraints in diffusion-based prediction, with DriveDreamer-2 [24] further incorporating LLMs for customizable video generation. Drive-WM [25] enhances multi-view consistency via view factorization, and GenAD [6] introduces large-scale video datasets to improve zero-shot generalization. HERMES [3] seamlessly integrates 3D scene understanding and future scene evolution (generation) in an MLLM. Based on GenAD, Vista [2] enhances dynamics and preserves structural details with two extra loss functions, achieving high-resolution, high-fidelity, and long-term scene evolution prediction. Subsequent works [26], [20] have made significant advancements in prediction duration [20] and the integration of prediction with planning [19], [7], [27]. Despite advancements, these world models only focus on low-level visual representations and overlook the geometry information, which hinders the ability of spatial reasoning. Instead, the proposed UniFuture integrates visual information with geometry features seamlessly, enabling the geometry-aware reasoning ability of the world model.

B. Monocular Depth Estimation

Monocular depth estimation, which aims to recover the geometry information from a single image or video, has gained significant attention in autonomous driving research [28]. It serves as a fundamental step for various downstream 3D tasks [29], [30], [31], [32], [33]. Conventional works [34], [35], [36] leverage multi-scale feature fusion to combine global and local depth information, achieving remarkable performance within specific domains. MiDaS [37] and ZoeDepth [38] leverage multi-dataset joint training to enhance generalization, while DepthAnything [10], [39] explores large-scale unlabeled and synthetic data to improve detail modeling in complex scenes. Another promising direction tries to incorporate rich priors of generative models trained on vast amounts of wild images. Marigold [13] pioneers the use of Stable Diffusion [40] for affine-invariant depth estimation, followed by Lotus [41], which refines denoising schedules for better adaptation. DepthCrafter [42] extends this framework to open-world videos for consistent long-sequence depth estimation. MERGE [43] proposes a unified model for generation and depth estimation, starting from a fixed pre-trained text-to-image model. While existing depth estimation methods focus on perceiving the current or past environment, our approach integrates future depth estimation with the world model. This synergy leverages the world model's reasoning capabilities for high-quality depth prediction while using precise geometric cues to enhance dynamic scene modeling.

III. METHOD

This paper proposes UniFuture, a unified 4D Driving world model that seamlessly integrates future scene generation with depth-aware perception to simulate the dynamic evolution of driving environments. Built upon an SVD-based video generation framework [2], our approach bridges the gap between appearance (RGB) and geometry (Depth), which are two heterogeneous projections of the same 4D reality. We achieve this through two key components: 1) A **Dual-Latent Sharing (DLS)** scheme, which maps both modalities into a unified spatio-temporal latent space, implicitly entangling texture with geometry; and 2) A **Multi-scale Latent Interaction (MLI)** mechanism, a bidirectional feedback system that

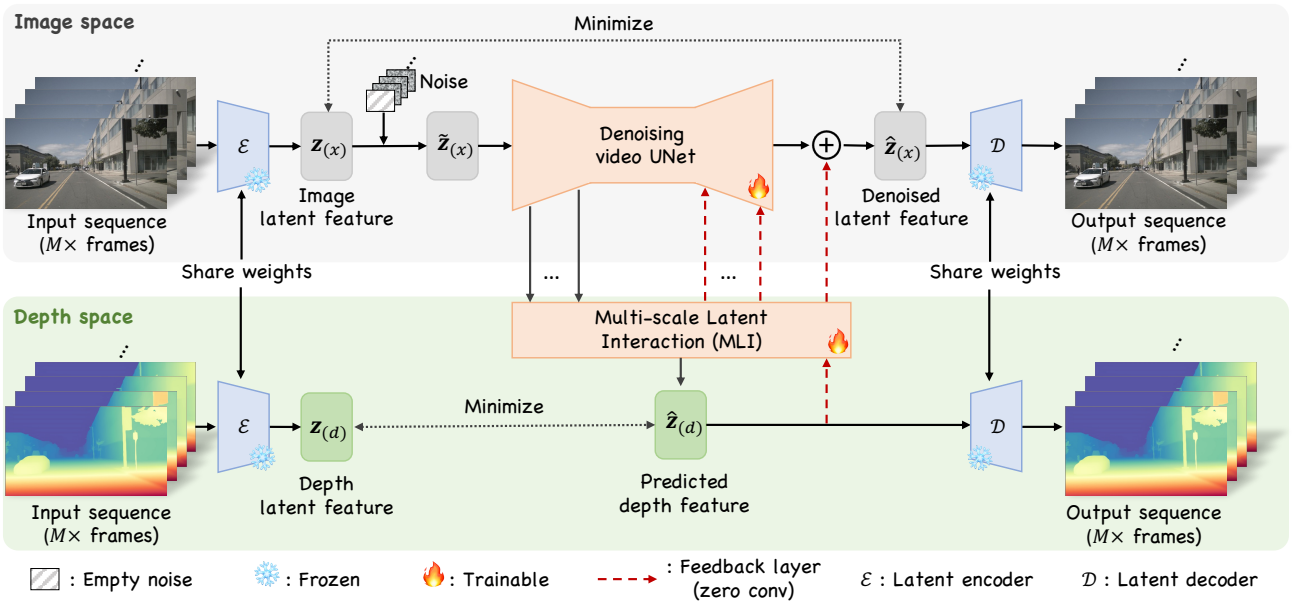

Fig. 2: The training pipeline of UniFuture. To learn a 4D Driving World Model, we introduce the Dual-Latent Sharing (DLS) scheme, which unifies visual appearance (image) and 3D geometry (depth) into a shared latent space without additional pre-training. The image latent undergoes a conditional denoising process, while the depth latent is explicitly predicted via the Multi-scale Latent Interaction (MLI) mechanism. MLI enforces bidirectional spatio-temporal consistency between texture and structure, ensuring coherent 4D scene evolution.

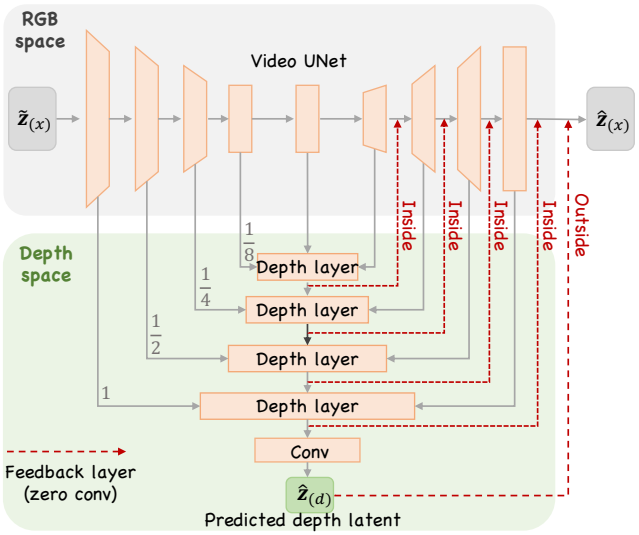

Fig. 3: The details of our proposed Multi-scale Latent Interaction (MLI) mechanism, designed to bridge the gap between visual and geometric feature spaces.

refines features across scales. These components ensure that pixel-level synthesis is geometrically grounded, and spatial reasoning is visually informed, resulting in coherent 4D predictions.

During training (Fig. 2), UniFuture takes an image-depth pair sequence with M frames as input, treating it as a discretized 4D scene representation. The image latent feature

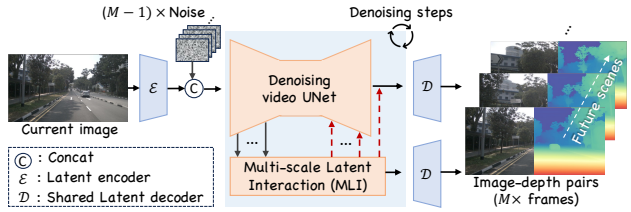

Fig. 4: The inference pipeline of UniFuture. It transforms a single 2D observation into a 4D forecast (future image-depth pairs). The latent representation evolves through the MLI-enhanced denoising UNet, producing geometrically aligned future sequences.

follows the conditional denoising process[1], while the depth latent feature is concurrently optimized. During inference (Fig. 4), given only a single current image, the model predicts future image-depth pairs by concatenating $(M-1)\times$ noise embeddings. This process effectively hallucinates a temporally and geometrically consistent 4D future from a static 3D observation.

A. Dual-Latent Sharing Scheme: Constructing a Unified 4D Latent Space

In the physical world, an RGB image and its corresponding depth map are intrinsically linked descriptions of the same scene. We argue that learning a 4D world model requires a representation where these two modalities are aligned. To this end, we introduce the Dual-Latent Sharing (DLS) scheme. Rather than treating depth estimation as an auxiliary

[1]The first frame serves as a noise-free condition.

task with a separate encoder, DLS unifies image and depth representations within a shared latent space.

As shown in Fig. 2, given an image sequence $\boldsymbol{x}$ (appearance) and its corresponding depth maps $\boldsymbol{d}$ (geometry), we process both through a shared pre-trained latent encoder $\mathcal{E}$, obtaining $\boldsymbol{z}_{(x)}$ and $\boldsymbol{z}_{(d)}$. The image latent $\boldsymbol{z}_{(x)}$ undergoes the standard diffusion process to model temporal dynamics. Concurrently, the depth latent $\hat{\boldsymbol{z}}_{(d)}$ is derived via the Multi-scale Latent Interaction (MLI) mechanism. Finally, both are reconstructed back to the pixel/metric space using the shared decoder $\mathcal{D}$.

By forcing depth maps to traverse the same latent space as natural images, we implicitly encode geometry using the rich, pre-trained semantic priors of the video generator. This eliminates the need for additional depth-specific pre-training and enables seamless cross-modal feature flow.

B. Multi-scale Latent Interaction: Enforcing Spatio-Temporal Consistency

To ensure that the generated 4D predictions are physically consistent (i.e., textures adhere to surfaces, and shapes do not deform unrealistically over time), we propose the Multi-scale Latent Interaction (MLI) mechanism. MLI facilitates explicit bidirectional information flow between the appearance stream and the geometry stream. As illustrated in Fig. 3, MLI consists of hierarchical depth layers for alignment, followed by "Inside" and "Outside" feedback loops.

Depth Layers for Feature Alignment. To translate the rich semantic features from the video UNet into geometric representations, we introduce a hierarchical fusion strategy. Specifically, we extract multi-scale features from the UNet encoder and decoder at scales $\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$. Each depth layer aggregates these features with the upsampled output from the preceding layer. This progressive fusion ensures that the depth estimation benefits from both high-level semantic context (essential for understanding object extent) and low-level structural details (essential for boundaries).

Inside Feedback: Geometry-Guided Generation. While image and depth share a latent space, their feature distributions differ. To bridge this, we apply a zero-initialized convolution to the intermediate depth latent feature $\mathbf{X}$ before injecting it back into the video generation stream. As indicated by the red lines in Fig. 3, inside feedback flows from the intermediate depth layers to the corresponding UNet stages:

$$\mathbf{Y} = \texttt{ZeroConv}\left(\mathbf{X}\right), \tag{1}$$

where $\mathbf{Y}$ acts as a geometric condition added to the video UNet features. By initializing weights to zero, the model starts with standard video generation and progressively learns to utilize geometric cues to refine texture synthesis, preventing geometric inconsistencies in the generated video.

Outside Feedback: Texture-Refined Geometry. To further align the final output, we introduce an outside feedback mechanism. The final predicted depth latent $\hat{\boldsymbol{z}}_{(d)}$ is injected into the denoised image latent $\hat{\boldsymbol{z}}_{(x)}$. This step ensures that the final appearance is strictly conditioned on the predicted geometry, reinforcing the structural integrity of the generated 4D scene.

By integrating these components, MLI achieves a closed-loop interaction: geometry constrains appearance, and appearance refines geometry. This allows UniFuture to simultaneously generate high-fidelity image-depth pairs that are temporally coherent and geometrically accurate.

C. Training Objectives

Our training objectives enforce consistency in both the latent and pixel spaces, crucial for 4D modeling. At the latent level, we minimize the reconstruction error of the denoised image latent $\hat{\boldsymbol{z}}_{(x)}$ using a composite loss $\mathcal{L}_{(x)}$ (including MSE and structural losses), following Vista [2]. Similarly, a loss $\mathcal{L}_{(d)}$ constrains the depth latent. To ensure the physical validity of the predicted geometry, we further impose a Scale- and Shift-Invariant loss $\mathcal{L}_{SSI}$ between the reconstructed predicted depth and the ground truth, following Depth Anything [10]. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{(x)} + \mathcal{L}_{(d)} + \lambda \cdot \mathcal{L}_{SSI}, \tag{2}$$

where λ balances the contributions.

D. Inference Phase

Unlike previous approaches that treat prediction and perception as separate tasks, UniFuture leverages a single current frame to forecast a consistent 4D future. As illustrated in Fig. 4, the input image is encoded and concatenated with $(M-1)\times$ Gaussian noise maps. This combined volume is then refined through the MLI-enhanced UNet. The process jointly evolves the appearance and geometry latents over time, which are finally decoded into a sequence of image-depth pairs. This resulting sequence constitutes a comprehensive 4D world representation, maintaining perceptual realism and structural coherence crucial for downstream autonomous driving applications.

IV. EXPERIMENTS

A. Dataset and Evaluation Metric

Datasets. We conduct experiments on the nuScenes dataset [44], a large-scale autonomous driving benchmark that provides multi-modal sensor data, including RGB images, LiDAR point clouds, and scene annotations. It contains 1,000 driving sequences collected in diverse urban environments. Following Vista [2], we use the front-camera RGB frames as the main training data. We also evaluate our model on the Waymo [45] dataset under a zero-shot setting to assess its generalization to unseen driving scenes. Since neither nuScenes [44] nor Waymo [45] provides dense pixel-wise depth annotations, we adopt DepthAnythingV2 [39] to generate depth labels.

Evaluation metrics. For the generation task, we evaluate the quality and realism of generated frames using Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD), which measure the distributional similarity between generated and real images/videos. Lower FID and FVD scores indicate better generation quality. For the depth estimation task, we

TABLE I: Comparison of our method with specialized generation and depth estimation models. ★ marks our baseline method, which is fine-tuned under the same resolution and iteration constraints using the official pre-trained weight of Vista [2]. Our method predicts depth up to 25 future frames, while Marigold [13] supports only single-frame depth estimation. ♦ To enable a reasonable comparison, we train Marigold with next-frame supervision using the 1st and 12th future frames.

Method	Reference	Resolution	Generation		Depth estimation			
			FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Only future depth estimation								
Marigold (0-th frame) [13]	CVPR 24	320×576	Unsupported		20.4	80.3	93.1	96.5
Marigold (1-st frame)♦ [13]	CVPR 24	320×576			21.9	77.7	91.2	95.6
Marigold (12-th frame)♦ [13]	CVPR 24	320×576			39.0	65.8	81.9	89.7
Only future generation								
DriveGAN [16]	CVPR 21	256×256	73.4	502.3	Unsupported			
DriveDreamer [1]	ECCV 24	128×192	52.6	452.0				
WoVoGen [17]	ECCV 24	256×448	27.6	417.7				
DrivingDiffusion [5]	ECCV 24	512×512	15.8	332.0				
GenAD [6]	CVPR 24	256×448	15.4	184.0				
Panacea [8]	CVPR 24	256×512	17.0	139.0				
Drive-WM [7]	CVPR 24	192×384	15.2	122.7				
Vista★ [2] (baseline)	NeurIPS 24	320×576	15.5	101.5				
Unify future generation and depth estimation								
UniFuture (ours)	-	320×576	**11.8**	**99.9**	**8.936**	**91.4**	**97.6**	**98.9**

use Absolute Relative Error (AbsRel) to quantify relative depth differences and threshold accuracy (δ) to measure the proportion of accurate predictions within a specified relative error. Higher δ values and lower AbsRel values indicate better depth estimation performance. It should be noted that to evaluate the consistency between the predicted future depth and the future scene, we utilize DepthAnythingV2 [39] to process the predicted future scene as pseudo-depth labels when computing the aforementioned depth metrics unless otherwise specified.

B. Implementation Details

We adopt the representative framework of Vista [2] as the baseline. Our model is trained with a batch size of 1 on 8 × NVIDIA H20 GPUs. We use the AdamW optimizer with a learning rate of 5×10^{-5}, and the training process runs for 8K iterations to ensure convergence. Besides, following Vista [2], we use the Exponential Moving Average (EMA) strategy to make training stable and drop out each activated action mode with a ratio of 15% to allow classifier-free guidance. To optimize memory usage, we adopt the DeepSpeed ZeRO-2 strategy, which effectively reduces memory overhead. We set the length of the video $M = 25$, and the loss balance weight $\lambda = 0.5$.

C. Main Results

We evaluate UniFuture on both future scene generation and future geometry perception to investigate its capability as a unified 4D world model. Unlike prior works that treat visual synthesis and depth estimation as disjoint objectives, UniFuture aims to simulate the holistic 4D evolution of driving scenes through shared representations and structured interaction. For evaluation, we adopt Vista [2], a state-of-the-art driving world model, as our primary baseline. As shown in Tab. I, our method delivers consistently superior performance, validating our 4D modeling hypothesis.

TABLE II: Zero-shot results on the Waymo dataset. Our method demonstrates superior generalization in both 4D generation and perception.

Method	Generation		Perception			
	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Vista [2] (baseline)	23.8	238.4	Unsupported			
UniFuture (ours)	**16.3**	**227.6**	**9.517**	**89.3**	**97.0**	**98.7**

On future scene generation, UniFuture achieves a significant 3.7-point reduction in FID (from 15.5 to 11.8) and a highly competitive FVD score compared to Vista [2]. These improvements highlight the critical role of geometry-aware synthesis: by explicitly modeling depth in the shared latent space, our model enforces structural constraints on the generated video. This prevents common artifacts like object deformation or temporal flickering seen in pure 2D models [2], [5], [8], which rely solely on pixel-level patterns. In essence, UniFuture does not just "paint" pixels; it "renders" a physically grounded 4D world, leading to superior realism and temporal coherence.

On future geometric perception, UniFuture surpasses the state-of-the-art depth estimator Marigold [13], achieving the lowest AbsRel (8.936) and the highest threshold accuracies. While Marigold is specialized for high-quality static depth estimation, it operates frame-by-frame and lacks temporal foresight. Consequently, its performance degrades drastically at longer prediction horizons (e.g., AbsRel 39.0 at the 12th frame). In contrast, UniFuture benefits from dynamics-informed estimation: by leveraging the video generator's temporal priors, our model anticipates how scene geometry evolves over time. The unified latent space effectively couples pixel dynamics with geometric structural changes, resulting in 4D predictions that are not only sharp but also temporally stable.

TABLE III: Comprehensive ablation studies validating the core components of our 4D world model: (a) optimization paradigms, (b) unified latent representation (DLS), (c) multi-scale interactions, (d) bidirectional feedback mechanisms, and (e) feature alignment strategies.

(a) Optimization paradigms

Setting	Generation		Depth estimation			
	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Image-only	15.5	101.5	Unsupported			
Depth-only	271.0	3143.2	24.196	62.0	84.5	92.4
Detach-grad	**11.2**	104.9	12.351	85.2	95.2	97.8
Joint training	11.8	**99.9**	**8.936**	**91.4**	**97.6**	**98.9**

(b) Depth decoders (Validating DLS)

Decoder	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Convention	15.6	121.1	11.874	86.4	95.9	98.1
Our DLS	**11.8**	**99.9**	**8.936**	**91.4**	**97.6**	**98.9**

(c) Multi-scale interactions

Scale	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑
1	12.9	114.4	9.677	91.1
1, $\frac{1}{2}$, $\frac{1}{4}$	12.6	106.9	9.277	**91.4**
1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$	**11.8**	**99.9**	**8.936**	**91.4**

(d) Inside vs. outside feedback

In	Out	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑
–	–	13.3	123.3	9.034	91.5
✓	–	12.2	110.1	9.034	**91.6**
–	✓	13.5	119.0	9.100	91.4
✓	✓	**11.8**	**99.9**	**8.936**	91.4

(e) Feedback layer types

Setting	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑
Direct add	223.2	2474.6	53.177	49.9
Random-conv	51.2	615.1	26.262	61.7
Zero-conv	**11.8**	**99.9**	**8.936**	**91.4**

D. Zero-Shot Generalization

To evaluate the robustness of our 4D representations, we test UniFuture on the Waymo dataset without fine-tuning. As shown in Tab. II, our method outperforms Vista [2] in future generation with a 7.5-point lower FID (16.3 vs. 23.8) and improved FVD, demonstrating stronger visual coherence in unseen domains. Crucially, while Vista lacks perception capabilities, UniFuture provides accurate zero-shot depth estimation (AbsRel = 9.517), benefiting from the entangled learning of texture and geometry. These results confirm that our unified approach captures fundamental 4D world dynamics that generalize across different driving environments.

E. Ablation Studies

This section dissects the key components of our 4D world model, validating the design choices for unified learning and spatio-temporal interaction.

Impact of Optimization Paradigms (Unified vs. Separate). We first study how the coupling of modalities affects 4D modeling. As shown in Tab. IIIa: **1)** Image-only training generates reasonable textures but lacks geometric understanding. **2)** Depth-only training fails to capture complex scene dynamics. **3)** Separate optimization (Detached-grad) improves depth but degrades generation quality (higher FVD) because geometric constraints are not propagated to the visual stream. **4)** Our Joint training strategy yields the best performance across all metrics. This confirms that 4D modeling is not a zero-sum game; instead, depth modeling stabilizes visual structure, while visual context refines geometric details, creating a mutually beneficial loop.

Effectiveness of the DLS Scheme (Unified 4D Latent Space). We evaluate the core idea of mapping image and depth into a shared manifold. As shown in Tab. IIIb, replacing a conventional convolutional depth decoder with our shared VAE decoder (DLS) significantly improves both generation (FVD: 99.9 vs. 121.1) and perception. This validates that sharing the latent space enforces a tighter coupling between appearance and geometry, which is essential for consistent 4D prediction.

Effectiveness of Multi-scale Interactions (MLI). To ensure fine-grained alignment between visual and geometric features, MLI introduces interactions at multiple scales. Tab. IIIc shows that multi-scale feedback significantly outperforms single-scale interaction. This suggests that 4D consistency requires alignment at both high-level semantic scales (for object dynamics) and low-level structural scales (for boundary precision).

Role of Bidirectional Feedback. We analyze the directionality of feature flow in MLI. As shown in Tab. IIId: **1)** Inside feedback (Geometry-to-Texture) significantly reduces FVD, proving that geometric cues help stabilize video generation. **2)** Combined with Outside feedback (Texture-to-Geometry), we achieve the best balance. This validates our closed-loop design where geometry constrains appearance and appearance refines geometry.

Impact of Feedback Layer Initialization. Finally, Tab. IIIe shows that using Zero-conv initialization is critical. Direct addition or random initialization disrupts the pre-trained latent features, leading to collapse. Zero-conv allows the model to gradually learn the mapping between the heterogeneous image and depth spaces, enabling a smooth transition to 4D modeling.

F. Qualitative Analysis: 4D World Reconstruction

We present qualitative comparisons in Fig. 5, showcasing UniFuture's superiority in generating realistic and geometrically accurate future scenes. Beyond standard metrics, the ultimate test of a 4D world model is its ability to reconstruct the dynamic 3D world. As illustrated in Fig. 6, we project our predicted image-depth pairs into 3D space to form 4D Point Clouds. The reconstructed scenes exhibit temporal continuity and structural integrity, with dynamic objects (e.g., moving vehicles) and static backgrounds (e.g., roads, buildings) evolving consistently. This visualization confirms that UniFuture successfully transcends 2D video generation, effectively functioning as a robust simulator for the 4D physical world.

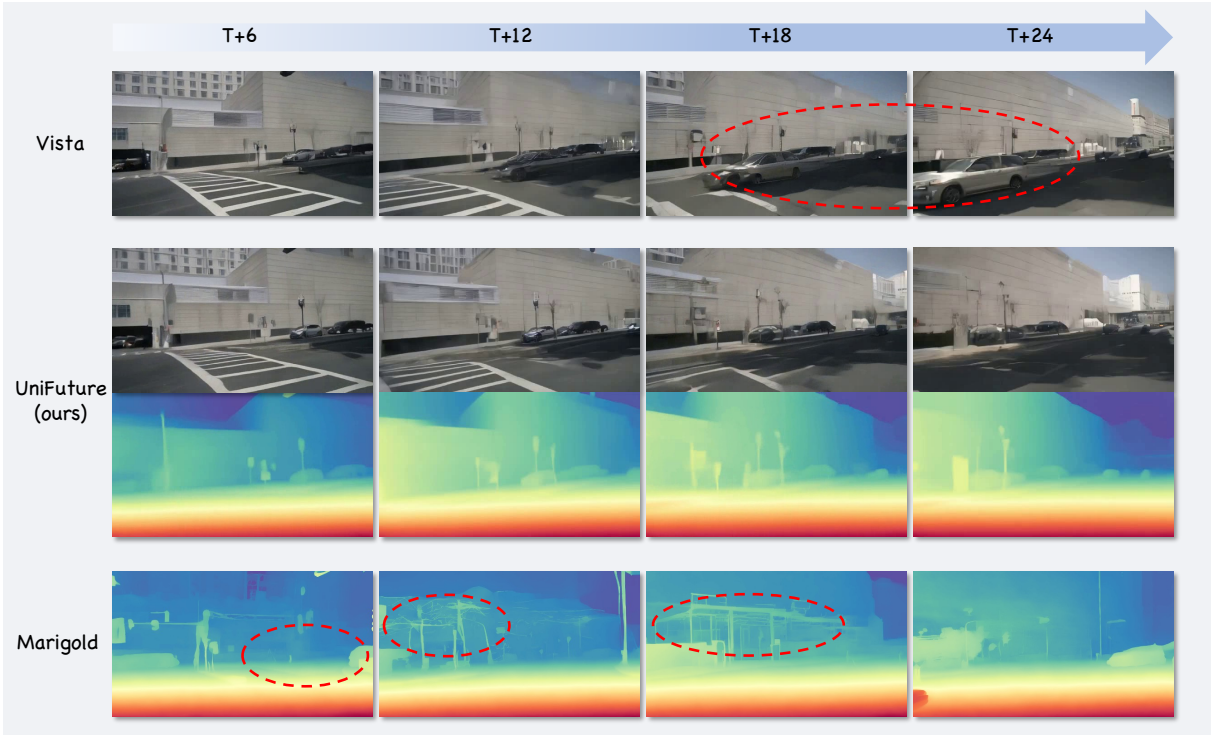

Fig. 5: Qualitative comparisons. Existing DWMs (Vista [2]) generate plausible videos but lack geometric grounding. Specialized depth estimators (Marigold [13]) fail to forecast future geometry. In contrast, UniFuture delivers coherent 4D predictions, maintaining structural integrity over time.

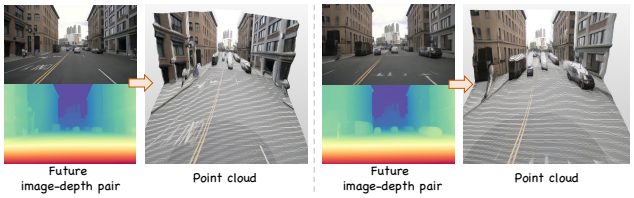

Fig. 6: Visualization of the 4D Point Clouds reconstructed from our predicted future image-depth sequences. This demonstrates UniFuture's capability to simulate not just 2D videos, but the dynamic evolution of the 3D world.

G. Controllable Future Scene Evolution

A crucial characteristic of world models in autonomous driving is their ability to predict future scene evolution based on control signals. Only with this capability can the world model provide a realistic simulation environment, thereby supporting the development of end-to-end reinforcement learning models. Thus, in this section, we present the results of UniFuture for controllable future scene evolution, shown in Fig. 7. Starting from the same condition frame, our method can generate different future scenes corresponding to given commands (e.g., go straight, turn right) with high-quality geometry outputs (i.e., depth). These results demonstrate the potential of our method to support downstream tasks, showing the superiority of UniFuture.

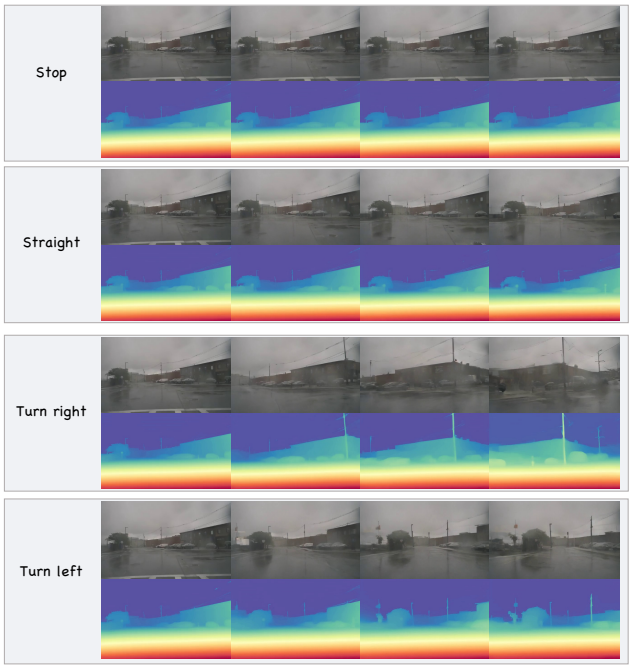

Fig. 7: Controllable future scene evolution with UniFuture. Given the same starting frame, our method generates diverse future trajectories based on different control commands (e.g., stop, go straight, turn right, turn left) while maintaining high-quality geometric consistency.

V. CONCLUSIONS

We propose UniFuture, a 4D driving world model that integrates future scene generation and depth-aware percep-

tion. Through Dual-Latent Sharing (DLS) and Multi-scale Latent Interaction (MLI), our method enables effective cross-modal representation learning and bidirectional refinement, enhancing structural consistency. Experiments show that our method achieves impressive performance, producing high-consistency image-depth pairs with improved realism and geometric alignment. By unifying the future generation and perception, we offer a valuable alternative to world modeling for autonomous driving.

Acknowledgement. This work was supported by the NSFC (623B2038 and 62225603).

REFERENCES

[1] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *ECCV*, 2024.

[2] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," in *NeurIPS*, vol. 37, 2024.

[3] X. Zhou, D. Liang, S. Tu, X. Chen, Y. Ding, D. Zhang, F. Tan, H. Zhao, and X. Bai, "Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation," in *ICCV*, 2025.

[4] S. Tu, X. Zhou, D. Liang, X. Jiang, Y. Zhang, X. Li, and X. Bai, "The role of world models in shaping autonomous driving: A comprehensive survey," *arXiv preprint arXiv:2502.10498*, 2025.

[5] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model," in *ECCV*, 2024.

[6] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo, J. Zhang, A. Geiger, Y. Qiao, and H. Li, "Generalized predictive model for autonomous driving," in *CVPR*, 2024.

[7] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *CVPR*, 2024.

[8] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang, "Panacea: Panoramic and controllable video generation for autonomous driving," in *CVPR*, 2024.

[9] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "Unidepth: Universal monocular metric depth estimation," in *CVPR*, 2024.

[10] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[11] Y. Duan, X. Guo, and Z. Zhu, "Diffusiondepth: Diffusion denoising approach for monocular depth estimation," in *ECCV*, 2024.

[12] X. Li, C. Wu, Z. Yang, Z. Xu, Y. Zhang, D. Liang, J. Wan, and J. Wang, "Driverse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment," in *ACM MM*, 2025.

[13] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *CVPR*, 2024.

[14] D. Ha and J. Schmidhuber, "World models," in *NeurIPS*, 2018.

[15] Y. Guan, H. Liao, Z. Li, J. Hu, R. Yuan, G. Zhang, and C. Xu, "World models for autonomous driving: An initial survey," *IEEE TIV*, 2024.

[16] S. W. Kim, J. Philion, A. Torralba, and S. Fidler, "Drivegan: Towards a controllable high-quality neural simulation," in *CVPR*, 2021.

[17] J. Lu, Z. Huang, Z. Yang, J. Zhang, and L. Zhang, "Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation," in *ECCV*, 2024.

[18] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[19] Y. Chen, Y. Wang, and Z. Zhang, "Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers," 2025.

[20] X. Guo, C. Ding, H. Dou, X. Zhang, W. Tang, and W. Wu, "Infinity-drive: Breaking time limits in driving world models," *arXiv preprint arXiv:2412.01522*, 2024.

[21] J. Jiang, G. Hong, L. Zhou, E. Ma, H. Hu, J. Xiang, F. Liu, K. Yu, H. Sun, K. Zhan, *et al.*, "Dive: Dit-based video generation with enhanced control," in *ECCV workshop*, 2024.

[22] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang, "Panacea: Panoramic and controllable video generation for autonomous driving," in *CVPR*, 2024.

[23] F. Jia, W. Mao, Y. Liu, Y. Zhao, Y. Wen, C. Zhang, X. Zhang, and T. Wang, "Adriver-i: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023.

[24] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," in *AAAI*, 2025.

[25] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *CVPR*, 2024.

[26] M. Hassan, S. Stapf, A. Rahimi, P. Rezende, Y. Haghighi, D. Brüggemann, I. Katircioglu, L. Zhang, X. Chen, S. Saha, *et al.*, "Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control," in *CVPR*, 2025.

[27] D. Liang, C. Zhang, X. Xu, J. Ju, Z. Luo, and X. Bai, "Cook and clean together: Teaching embodied agents for parallel task execution," in *AAAI*, 2026.

[28] U. Rajapaksha, F. Sohel, H. Laga, D. Diepeveen, and M. Bennamoun, "Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–51, 2024.

[29] J. Li, Z. Liu, J. Hou, and D. Liang, "Dds3d: Dense pseudo-labels with dynamic threshold for semi-supervised 3d object detection," in *ICRA*, 2023.

[30] D. Liang, X. Zhou, W. Xu, X. Zhu, Z. Zou, X. Ye, X. Tan, and X. Bai, "Pointmamba: A simple state space model for point cloud analysis," in *NeurIPS*, 2024.

[31] X. Zhou, D. Liang, W. Xu, X. Zhu, Y. Xu, Z. Zou, and X. Bai, "Dynamic adapter meets prompt tuning: Parameter-efficient transfer learning for point cloud analysis," in *CVPR*, 2024.

[32] D. Liang, T. Feng, X. Zhou, Y. Zhang, Z. Zou, and X. Bai, "Parameter-efficient fine-tuning in spectral domain for point cloud learning," *IEEE TPAMI*, 2025.

[33] D. Zhang, D. Liang, Z. Zou, J. Li, X. Ye, Z. Liu, X. Tan, and X. Bai, "A simple vision transformer for weakly semi-supervised 3d object detection," in *ICCV*, 2023.

[34] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using laplacian pyramid-based depth residuals," *IEEE TPAMI*, 2021.

[35] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *CVPR*, 2021.

[36] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "Neural window fully-connected crfs for monocular depth estimation," in *CVPR*, 2022.

[37] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE TPAMI*, vol. 44, no. 3, pp. 1623–1637, 2020.

[38] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[39] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *NeurIPS*, vol. 37, 2024.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[41] J. He, H. Li, W. Yin, Y. Liang, L. Li, K. Zhou, H. Zhang, B. Liu, and Y.-C. Chen, "Lotus: Diffusion-based visual foundation model for high-quality dense prediction," in *ICLR*, 2025.

[42] W. Hu, X. Gao, X. Li, S. Zhao, X. Cun, Y. Zhang, L. Quan, and Y. Shan, "Depthcrafter: Generating consistent long depth sequences for open-world videos," in *CVPR*, 2025.

[43] H. Lin, D. Liang, M. Du, X. Zhou, and X. Bai, "More than generation: Unifying generation and depth estimation via text-to-image diffusion models," in *NeurIPS*, 2025.

[44] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[45] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.