UniDynamics: Event-RGB Fusion for Unified Future 4D Dynamic Scene Generation

Daikun Liu[1⋆], Xin Zhan[2⋆], Teng Wang[1†], Xiaoping Wang[2†], and Changyin Sun[1]

[1] Southeast University, China
{dkliu, wangteng, cysun}@seu.edu.cn
[2] Huazhong University of Science and Technology, China
{xinzhan, wangxiaoping}@hust.edu.cn

Abstract. We propose **UniDynamics**, a diffusion-based framework for future 4D dynamic scenes (RGB, depth, and optical flow) generation from a single event-RGB pair, without requiring long histories or control priors as in existing methods, while explicitly modeling future motion fields. The core idea is to leverage event streams to offer an alternative motion prior for single-RGB extrapolation, and to enforce geometric and motion constraints throughout generation via multimodal modeling. Specifically, we design an Event Latent Enhancement (ELE) module to align and enhance event latents into diffusion-injectable conditioning features, providing robust initial motion priors and reliable texture/structure cues. We further introduce a Perceptual Dynamics Space (PDS) embedded in the multi-scale U-Net, which decouples and adaptively interacts depth and flow while continuously feeding back constraints to appearance features, improving geometric-motion consistency for physically plausible and spatiotemporally coherent prediction. Experiments on VKitti2 and DSEC demonstrate state-of-the-art performance, producing high-quality, temporally coherent, and 4D-consistent future predictions, especially under challenging high-speed motion blur.

Keywords: Event-RGB fusion · Future RGB-depth-flow prediction· 4D dynamics · Diffusion model

1 Introduction

In autonomous driving scenarios, the ability of infering the evolution of the environment over time and understand future scenes based on current visual observations is critical for autonomous systems. It serves as an essential foundation for motion planning [15], obstacle avoidance [37], and risk-aware decision [33].

Some existing works focus on generative driving world models [8, 13, 14, 24, 40, 43], which leverage the current frame or historical observations and introduce conditions such as trajectories, velocities, or text to improve controllability of

⋆ Equal contribution. † Corresponding authors.

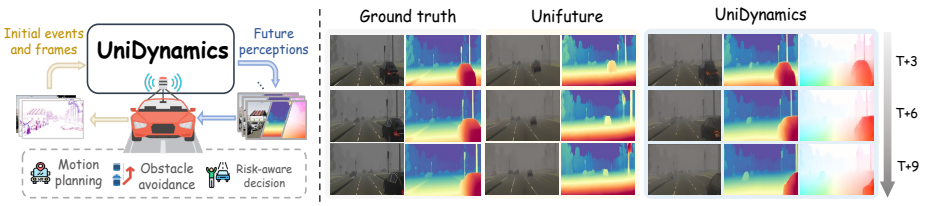

Fig. 1: Our **UniDynamics** infers future 4D scenes using only the current event-RGB pair. Leveraging motion cues and texture priors of events, it jointly models appearance, geometry, and motion, producing more faithful and spatiotemporally coherent predicts, especially under motion blur, where UniFuture [24] with only a blurred RGB fails.

future generations. However, the reliance on additional conditioning signals substantially increases system complexity and adaptation cost. In contrast, vision-only video prediction methods [21] model the temporal evolution of semantic features from pretrained video foundation models, better matching the need for a "plug-and-play" perception front-end, but their fine-grained fidelity is often inferior to generative approaches. More importantly, in real-world open-road deployment, long video histories and consistent structured priors are not always reliably available, making a more practical and transferable setting that uses only a single current frame as the minimal input. Recent Unifuture [24] extrapolates future scenes under single-frame conditioning by leveraging motion-prior conditions (e.g., velocity, acceleration, and trajectory). However, without such priors, scene dynamics are unobservable, and motion blur further degrades appearance cues, leading to erroneous extrapolation (see UniFuture in Fig. 1).

Notably, the inherent temporal sensitivity of event cameras [1, 25] provides directional and boundary motion cues, offering an alternative motion prior for future extrapolation under single-frame conditioning. In particular, event streams offer continuous motion observations without requiring long histories, and remain informative even under fast-motion blur, providing stable initial momentum and structural cues. *We therefore argue that event streams provide a more stable and blur-robust motion prior, enabling reliable future extrapolation even under single-frame conditioning.* In addition, the above methods often restrict prediction to static RGB appearance [8, 14, 40] or depth geometry [13, 21, 24] and lack explicit motion-field modeling, which is crucial for assessing scene dynamics, interaction risks, and potential collisions [16], and serves as a key constraint for physically consistent and temporally coherent future-scene prediction and percption.

To this end, we propose **UniDynamics**, a diffusion-based RGB, depth, and optical flow predicting framework with an event-RGB pair observation, as illustrated in Fig. 1. The core of our approach lies in event-guided condition enhancement and multimodal jointly interacting and predicting. Specifically, we treat the event stream as complementary to the RGB observation and introduce an Event Latent Enhancement (ELE) module to encode it into the event latent. To obtain stable and generalizable event cues, we use the RGB latent as an appearance reference and align/enhance the event latent via cross-attention layers, yielding

a condition that provides initial momentum and texture/boundary cues while remaining compatible with RGB-pretrained SVD [2]. Since appearance, geometry, and motion are heterogeneous views of the same scene, we adopt the latent space of a pretrained image autoencoder as a shared alignment domain, diffuse on the image latent, and derive depth and flow from it, avoiding modality-specific encoders and improving cross-modal alignment. We further embed a Perceptual Dynamics Space (PDS) into the multi-scale U-Net to enable continuous interaction between geometry/motion priors and appearance features. Task-specific adapters and gated interactions in PDS decouple depth and flow to reduce multi-task interference and suppress flow-induced structural perturbations, improving temporal coherence while preserving stable structure.

With event-based motion encoding and the proposed PDS-driven multimodal interaction and refinement, UniDynamics achieves high-fidelity and spatiotemporally coherent future 4D scene predicting, particularly in challenging scenarios with high-speed motion blur, as illustrated in Fig. 1. Experiments on the synthesized event-based VKitti2 [7] and the zero-shot generalization studies on DSEC [11] demonstrate that our method achieves higher-quality and more spatiotemporally coherent future scene generation and perception. Our main contributions are summarized as follows:

- We propose **UniDynamics**, a diffusion-based future 4D scene generation framework that fuses an event stream with an RGB observation, enabling physically plausible and spatiotemporally consistent modeling of future scene appearance (RGB), depth, and optical flow.
- We introduce an event latent enhancement module that yields generalizable and diffusion-injectable event conditions, providing stable initial motion cues and rich texture/boundary information under an RGB input.
- We design a Perceptual Dynamics Space integrated into the U-Net, where depth and optical flow are decoupled and modeled in a latent representation space aligned with RGB. Through feedback injection, PDS enables bidirectional interaction with appearance latent, improving structural stability and spatiotemporal coherence.
- Our method delivers higher-quality and more spatiotemporally consistent future 4D scene prediction and perception on VKitti2 and DSEC, particularly in challenging scenarios such as high-speed motion blur.

2 Related Work

RGB-based Future Scene Prediction Methods. Future scene prediction from visual observations is critical for autonomous driving. Generic video prediction methods [9, 35, 41] are mostly developed on general datasets and often transfer poorly to driving scenarios. Driving-focused generative world models [8, 14, 18, 23, 40, 43] improve controllability by introducing additional conditions such as actions, text, trajectories, and other priors, and some further model RGB with depth [13,24]. However, the reliance on extra structured conditions increases system complexity and adaptation cost. Vision-only prediction methods

based on CNN/RNN, masked modeling, or foundation models [4,20,21,26] avoid extra priors, but typically have lower fine-grained fidelity and often require long histories, incurring memory and latency. However, in deployment-relevant single-frame settings, scene dynamics are difficult to capture, and appearance cues are often degraded under fast motion. Meanwhile, most methods lack explicit motion-field modeling, limiting effective spatiotemporal consistency constraints. Motivated by these gaps, we introduce event streams as dynamic observations for the deployment-relevant settings, and adopt a diffusion framework to jointly predict RGB, depth, and optical flow for physically plausible and spatiotemporally coherent 4D scene generation and perception.

Event-based Future Scene Prediction Methods. Compared to conventional cameras, event cameras provide reliable motion observations for predicting in dynamic scenes. Recent works [5,22,42] validate the effectiveness of events for temporal extrapolation. E-Motion [42] combines the high temporal resolution of events with video diffusion to predict future event frames, while F^3 [5] learns predictive representations by predicting events and transferring them to tasks such as optical flow and semantic segmentation. However, event-only methods often break down in static or low-texture regions due to the dynamic characteristics and sparsity of events. DESSERT [22] fuses events with RGB for diffusion-based single-frame prediction, but long-horizon prediction suffers from error accumulation. Moreover, these methods also lack explicit motion-field modeling. Therefore, we exploit the complementarity of events and RGB and jointly model 4D scenes to achieve temporally coherent future evolution.

Event-RGB-fusion-based Vsion Tasks Monocular depth and optical flow estimation are core vision tasks for autonomous driving. Since events and frames offer complementary temporal and appearance cues, many works fuse them to improve robustness and accuracy. Early works on event–RGB depth estimation make efforts on using recurrent fusion [10], flow-based compensation [32], and cross-modal consistency with pseudo-supervision [47]. Later methods improve fusion robustness via structural priors [29], high-level pattern/token interactions [27], cross-modal attention with multi-scale interaction [6,19], or frequency-domain fusion [34]. For optical flow, several approaches extend RAFT [36] to fuse events and RGB, while DCEIFlow [38] and BFlow [12] estimate dense continuous-time motion, and STFlow [44] narrows the modality gap with spatially guided temporal aggregation. Diffusion-based fusion [39] is also explored by incorporating events into FlowDiffuser [28] for challenging conditions. Although these methods target current/past perception rather than prediction, they consistently demonstrate the complementarity of RGB–event fusion, motivating our event–RGB scheme for future scene prediction.

3 UniDynamics

3.1 Preliminaries

Stable Video Diffusion (SVD) [2] generates high-fidelity videos via latent diffusion. It encodes frames into a VAE latent space, progressively adds Gaussian

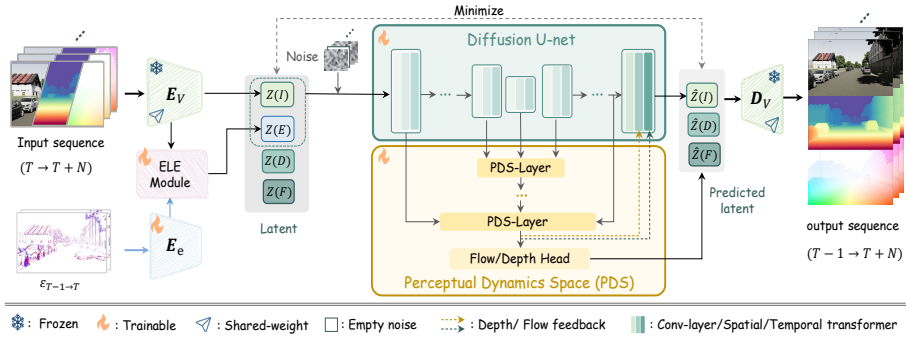

Fig. 2: The training pipeline of UniDynamics. The event stream is encoded by the Event Latent Enhancement (ELE) module into diffusion-injectable features that provide initial motion and texture cues. RGB, depth, and optical flow are derived from the same RGB latent: RGB is generated via diffusion denoising, while depth and flow are jointly learned through an additional Perceptual Dynamics Space (PDS). These modalities interact and are refined continuously across multiple scales, enabling physically plausible and spatiotemporally coherent future scene prediction.

noise to the target latent sequence during training, and optimizes a spatiotemporal U-Net to learn the reverse denoising process conditioned on the input-frame latent. At inference, it starts from Gaussian noise, iteratively denoises under the same conditioning, and decodes the resulting latents into video frames. Leveraging high-fidelity generation of SVD, recent works [8, 13, 24, 40] synthesize realistic future driving scenes and even extend to geometry-aware prediction [13, 24]. However, under the deployment-relevant single-frame setting, where initial motion cues cannot be provided, future extrapolation becomes highly ill-posed, and limited dynamics modeling further restricts temporal coherence.

3.2 Overview

We propose **UniDynamics**, a diffusion-based framework for future 4D scene generation, as illustrated in Fig. 2. Given the current frame at time T and corresponding event stream $\mathcal{E}_{T-1 \rightarrow T}$, we build on SVD as the generative backbone and aim to predict the future N-frame RGB sequence along with the depth maps and optical flows. Since RGB, depth, and optical flow provide complementary views of the same scene, we adopt the latent space of a pretrained image VAE encoder $\boldsymbol{E}_V$ as a unified representation. The diffusion backbone generates future appearance only in the image latent $Z(I)$, while depth and flow latents are initialized from $Z(I)$ and progressively refined through multi-scale U-Net interactions, enabling joint modeling of appearance, geometry, and motion in a shared latent space. This avoids training extra encoders, promotes cross-modal alignment, and allows lightweight networks to translate between modalities. To provide initial momentum under single-frame conditioning and support joint multimodal generation, our architecture comprises two key com-

ponents. **(1) Event latent guided motion conditioning.** We extract motion cues from the event stream and employ an Event Latent Enhancement module to produce a diffusion-injectable conditioning feature $Z(E)$, providing stable motion signals and rich texture/boundary priors that complement RGB. **(2) Perceptual Dynamics Space for depth and flow modeling.** To capture the coupled evolution of appearance, geometry, and motion in future scenes, we introduce the Perceptual Dynamics Space (PDS), a multi-scale interaction module embedded within the U-Net. PDS jointly models depth and optical-flow latents and performs bidirectional feature exchange with the diffusion backbone to impose multimodal consistency and refine the generated predictions. Finally, the predicted latent codes $\hat{Z}(I)$, $\hat{Z}(D)$, and $\hat{Z}(F)$ are decoded by the shared VAE decoder $\boldsymbol{D}_V$ into RGB frames, depth maps, and optical flow, enabling joint generation of future 4D scenes.

3.3 Event Latent Guided Motion Injection

Event Representation. Event cameras generate event stream $\mathcal{E}_{T-1\to T}$ with N_e events $\{x_i, y_i, t_i, p_i\}_{i=0}^{N_e}$ based on brightness changes, where (x, y), t, and p are coordinates, timestamp, and polarity, respectively. To extract motion cues, we convert events into voxel grids with B temporal bins [46]:

$$\mathcal{V}(b, x, y) = \sum_{i=0}^{N_e-1} p_i \, \delta(x - x_i) \, \delta(y - y_i) \, \max(0, 1 - |b - t_i^*|), \tag{1}$$

$$t_i^* = \frac{(t_i - t_{min})}{t_{max} - t_{min}} \times (B - 1), \tag{2}$$

where $\delta(\cdot)$ denotes the Kronecker delta function, $b \in [0, B-1]$ is the index of the temporal bin, and $\max(0, 1 - |b - t_i^*|)$ represents the linear interpolation kernel along the temporal dimension.

Event Latent Enhancement. We encode event frames into event features F_e using a learnable event encoder $\boldsymbol{E}_e$ comprised of 2D resblocks. To obtain a stable event latent that provides complementary cues for single RGB, we introduce an Event Latent Enhancement (ELE) module, as shown in Fig. 3 (a). Specifically, we first align event features with the RGB latent using bidirectional cross-attention:

$$Z'(E) = \mathrm{BiCA}((F_e, \mathrm{Proj}(Z(I)))) + F_e, \tag{3}$$

$$Z'(I) = \mathrm{BiCA}((\mathrm{Proj}(Z(I)), F_e)) + \mathrm{Proj}(Z(I)), \tag{4}$$

where $\mathrm{Proj}(\cdot)$ denotes a linear projection, and the event and RGB features are alternately used as the queries, keys, and values for each other. We then perform modality fusion to $Z'(E)$ and $Z'(I)$ with convolution and apply cross-modal attention with the RGB latent as key and value to leverage RGB appearance information for enhancing the final event latent $Z(E)$:

$$Z(E) = \mathrm{Proj}(\mathrm{CA}(\mathrm{Fusion}(Z'(E), Z'(I)), Z(I)))) + Z(I). \tag{5}$$

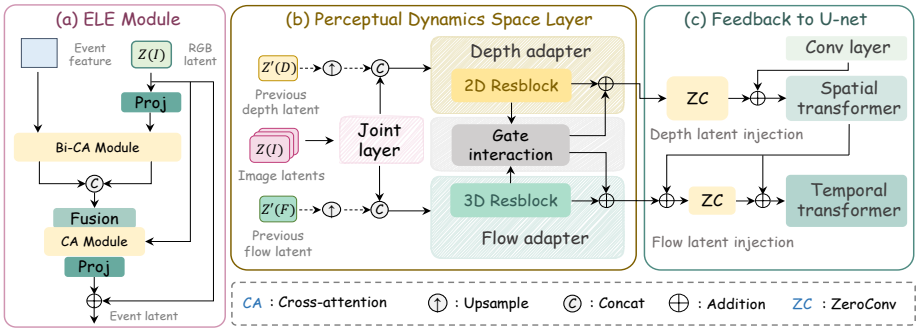

Fig. 3: UniDynamics submodules: (a) ELE module, which exploits the complementarity of events and RGB to produce an event latent with effective initial momentum and rich texture cues; (b) PDS layer, which decouples the depth and flow spaces to enable independent learning with adaptive interaction; and (c) depth geometry and flow motion cues are injected back into the U-Net's spatial and temporal layers, respectively, providing effective constraints on RGB evolution.

This design enriches the event latent with rich appearance of RGB, especially when events are sparse, while ensuring latent-space compatibility, leading to more effective and robust conditioning for the diffusion process.

Since the event latent primarily encodes short-term motion, directly concatenating $Z(E)$ with future noise latents as a strong condition for all future steps (as done for RGB latents) may overly entangle transient cues with global scene evolution, introducing bias and causing structural artifacts, temporal drift, or long-horizon overfitting. Instead, we adopt an SVD-compatible, softer conditioning scheme: We treat $Z(E)$ as the latent of a "historical-state RGB" at $T-1$, add noise to it jointly with the future latent sequence during training, and concatenate the current RGB latent $Z(I)$ to each noisy latent along the channel dimension as the reference. The model then generates frames at $\{T-1, \ldots, T+N\}$. This formulation embeds events as an implicit initial-momentum prior rather than a hard constraint, allowing the spatiotemporal U-Net to adaptively extract and propagate motion cues through multi-scale aggregation, while also providing useful texture/structure guidance when RGB is blurred.

3.4 Perceptual Dynamics Space for Depth-Flow Modeling

We embed the multi-scale interaction space into four stages $\{1, \frac{1}{2}, \frac{1}{4}\}$ of the U-Net, where two task-specific adapters decouple the unified latent representation into geometry and motion subspaces to model the evolution of depth and flow. A gated bidirectional interaction further regulates information exchange between them, and the resulting geometric and motion cues are injected into the spatial and temporal layers of the U-Net decoder, respectively, to regularize RGB generation and improve structural stability and spatiotemporal coherence.

Depth-Flow Adapter and Gated Bi-interaction. As shown in Fig. 3 (b), each PDS layer first fuses the features from the corresponding U-Net encoder and decoder blocks, together with the output of the previous PDS layer, using a joint layer with linear projection. The fused features are then decoupled, via feature reuse, into depth and flow adapters to alleviate gradient "tug-of-war" in multi-task learning. In addition, each adapter also takes the outputs $Z'(D)$ and $Z'(F)$ from the previous PDS layer as input, enabling continuous cross-layer modeling. The depth adapter uses 2D convolutions to capture spatial structure and outputs F_d, while the flow adapter uses 3D convolutions to model spatiotemporal dependencies and outputs F_f.

Depth and flow are tightly coupled: motion boundaries often align with structural edges, and motion patterns are strongly constrained by scene geometry. To capture this complementarity, we introduce a gated bidirectional interaction module in each PDS layer to enable adaptive information exchange between the depth and flow subspaces. Concretely, we first concatenate them to form a fused representation F_{fu}. Then we use two convolutional branches to generate cross-modal updates, whose injection strengths are adaptively modulated by sigmoid gates and added to the residuals:

$$\bar{Z}(D) = \text{Sigmoid}(\text{Conv}_\text{f}(F_{fu})) * F_f + F_d, \tag{6}$$

$$\bar{Z}(F) = \text{Sigmoid}(\text{Conv}_\text{d}(F_{fu})) * F_d + F_f, \tag{7}$$

The resulting features are then injected back into the deocoder of U-Net, and also upsampled and passed to the next PDS layer. After multiple layers, we obtain the final depth and flow latent representations $\hat{Z}(D)$ and $\hat{Z}(F)$.

Depth and Flow Latent Injection. To impose geometric and motion constraints throughout diffusion generation, we feed back the depth/flow latents modeled in PDS into the decoder of the U-Net backbone, as shown in Fig. 3 (c). Since depth captures static geometry and boundaries that directly regularize spatial details, while optical flow describes inter-frame motion and occlusion evolution that better guides temporal aggregation and dynamics propagation, we inject depth and flow latents X_d and X_f into the U-Net's spatial and temporal layers, respectively. To avoid overly strong constraints early in the training stage that could disrupt pretrained priors, we adopt a zero-initialized convolutional injection strategy for both depth and flow:

$$Y_d = \text{ZeroConv}(X_d), Y_f = \text{ZeroConv}(X_f + X_t), \tag{8}$$

where X_t is the temporal feature in the current U-Net block. Y_d and Y_f are then added to the spatial and temporal features, respectively:

$$X'_s = Y_d + X_s, X'_t = Y_f + X_t, \tag{9}$$

where X_s is the spatial feature of the U-Net block. By initializing the weights of ZeroConv to zero, the network starts with an equivalent of "no injection" and can stably learn to adaptively increase the injection strength during training, achieving a progressive fusion of constraints from weak to strong.

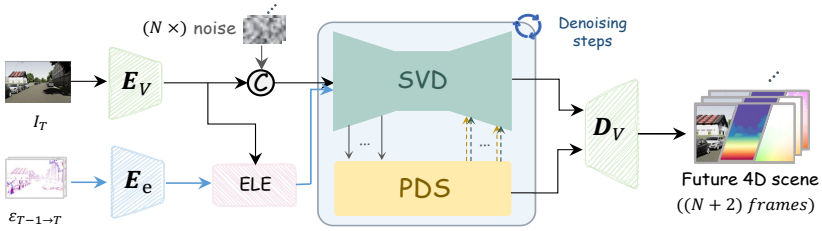

Fig. 4: UniDynamics inference pipeline: given a single event–RGB pair, it simultaneously predicts future RGB, depth, and optical flow. The RGB latent is concatenated with noise and, together with the event latent, fed into the U-Net for denoising to obtain multimodal latents, which are finally decoded by a shared decoder into future RGB frames, depth maps, and optical flow.

3.5 Training and Inference

Our framework forecasts the future N frames together with frames at $T-1$ and T of RGB, depth, and optical flow conditioned on a paired of event-RGB. We use a weight-shared VAE encoder to map multiple modalities into latent representations, and perform unified modeling on the shared RGB latent space to obtain $\hat{Z}(I)$, $\hat{Z}(D)$, and $\hat{Z}(F)$, which are decoded by the same VAE decoder to produce 4D scene predictions.

As shown in Fig. 2, during training, we add noise to the non-conditioning frames and the event-enhanced latent frame, concatenate the conditioning-frame latent, and jointly generate the pseudo-history, current, and future RGB-depth-flow outputs. We optimize generation and perception tasks with a weighted combination of losses. Specifically, for RGB, we compute a latent-space loss $L(I)$ between $Z(I)$ and $\hat{Z}(I)$, including an MSE term, a dynamics-enhanced term, and a structure-preserving term. For depth, we use an analogous loss $L(D)$, and additionally apply a scale- and shift-invariant loss L_{SSI} in pixel space between the predicted and target depth [30]. For optical flow, the loss consists of a latent-space MSE term $L(F)$ and a pixel-space L_1 term. We additionally apply an MSE loss $L(E)$ between $Z(E)$ and the RGB-frame latent at $T-1$ to stabilize training of the ELE module. The overall objective is:

$$L = L(I) + L(D) + L(F) + L(E) + \lambda_1 L_{\text{SSI}} + \lambda_2 L_1, \tag{10}$$

where λ_1, λ_2 are the balance coefficient.

As the inference pipeline in Fig. 4, we concatenate the current RGB latent with the noise-free event-enhanced latent and N Gaussian noise maps, and feed them into the U-Net backbone for iterative denoising. Together with PDS-based joint modeling and multimodal interaction, this enables physically plausible and spatiotemporally coherent future scene prediction and perception.

4 Experiments

4.1 Datasets and Metrics

Datasets. We train and evaluate on the driving dataset E-VKitti2 [7], and conduct zero-shot testing on real-world DSEC [11], MVSEC [45], and M3ED [3] datasets to assess generalization. VKitti2 provides high-quality synthetic RGB images, depth maps, and optical flow across diverse driving scenarios and appearance conditions (e.g., varying illumination). We use v2e [17] to synthesize E-VKitti2 dataset. DSEC is a real-world event-camera dataset containing RGB frames, event streams, disparity, and optical flow under different lighting and dynamic scenes. We convert the official disparity maps to depth using the provided conversion code. The *outdoor_day*1 sequence of the MVSEC provides relatively sparse events, enabling evaluation of generalization under weak event cues. Moreover, the M3ED features complex and dynamic outdoor scenes. To further evaluate performance under high-speed motion, we apply flow-based motion blur on E-VKitti2 to emulate image blur, where the prediction target is the corresponding deblured RGB.
Metrics. We evaluate RGB generation using Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD) to assess visual quality and spatiotemporal coherence. For depth, we report absolute relative error (AbsRel), accuracy δ_n ($\delta < 1.25^n$, $n = 1, 2, 3$), and root mean squared error (RMSE) to measure accuracy and structural consistency. For optical flow, we use endpoint error (EPE), angular error (AE), and the percentage of pixels with EPE larger than n pixels (nPE, $n = 1, 2, 3$) to quantify accuracy and robustness.

4.2 Implementation details

For event representations, we use a 100 ms event window and split into B=15 temporal bins. We crop and resize inputs to 192×640 for E-VKitti2 and 320×384 for DSEC, MVSEC, and M3ED. We train the model on E-VKitti2 with the AdamW optimizer using an initial learning rate of 5×10^{-5}, batch size 1, on 2×A100 GPUs for 16k steps. In the multi-task loss, we set $\lambda_1 = 2.0$ and $\lambda_2 = 1.0$. To stabilize training, we apply an exponential moving average (EMA) to smooth model weights and adopt classifier-free guidance by randomly dropping the conditioning branch with probability 15% during training. For memory efficiency, we use DeepSpeed ZeRO-2 [31] to reduce optimizer-state and gradient memory. At the evaluation phase, conditioned on a single frame and events, we predict future $N = 9$ RGB frames along with the corresponding depth and optical flow, and compare them with ground truth. More details in the supplementary material.

4.3 Main Results

Results on E-VKitti2. Tab. 1 reports the results on the E-VKitti2 dataset. The compared methods include our proposed UniDynamics and its variants,

Table 1: Quantitative comparison on E-VKitti2 [7]. "↑" and "↓" indicate that higher and lower values are better, respectively. Best results are highlighted in bold. For unsupported tasks, we mark them as "Unsupported".

Method	Generation		Depth					Optical Flow				
	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	RMSE ↓	EPE ↓	AE ↓	1PE ↓	2PE ↓	3PE ↓
Only future generation												
SimVP [9]	137.8	1301.7	Unsupported									
TAU [35]	127.8	1530.4	Unsupported									
SVD [2]	114.5	1244.4	Unsupported									
Vista [8]	101.8	1364.6	Unsupported									
Only future perception												
FLODCAST [4]	Unsupported		0.57	21.4	52.6	70.1	28.8	11.7	55.7	76.4	63.7	52.4
FUTURIST [20]	Unsupported		0.47	29.7	67.7	84.9	44.9	Unsupported				
SVD-Depth [2]	Unsupported		0.39	52.8	84.1	90.8	61.4	Unsupported				
SVD-Flow [2]	Unsupported		Unsupported					8.5	40.3	88.8	74.1	63.3
Joint generation and single perception												
UniFuture-Depth [24]	44.0	584.1	0.38	56.8	83.9	90.1	59.4	Unsupported				
UniFuture-Flow [24]	46.9	615.3	Unsupported					5.9	33.8	78.8	63.1	52.1
Unify future perception and generation												
UniDynamics (w/o events)	46.8	577.1	0.36	56.0	83.9	90.3	59.0	5.3	34.1	79.2	61.5	49.1
UniDynamics (w/o RGB)	79.2	1048.4	0.50	34.3	72.8	86.1	71.7	3.4	18.6	57.5	41.5	31.6
UniDynamics (ours)	**39.1**	**223.6**	**0.27**	**67.1**	**87.8**	**92.7**	**54.7**	**3.0**	**21.0**	**58.1**	**38.5**	**27.7**

non-diffusion-based methods, and diffusion-based methods. For a fair evaluation, we retrain SimVP [9], TAU [35], FLODCAST [4], FUTURIST [20], SVD [2], Vista [8], and UniFuture [24] under the same dataset settings and evaluation protocol. Specifically, *SVD-Dpeth* and *SVD-Flow* denote the SVD-based only depth and flow generation model, respectively. *UniFuture-Depth* and *UniFuture-Flow* correspond to the original RGB-depth generation and RGB-flow generation settings, respectively. For ours, we report the **UniDynamics** and an ablation *UniDynamics (w/o events)* to verify the contribution of event inputs.

The results show that only our method supports the simultaneous generation of future RGB, depth, and optical flow. Moreover, diffusion-based methods show overall superiority over non-diffusion-based methods in terms of both video generation and future perception. Comparing *Vista* with *UniDynamics (w/o events)* verifies that additionally modeling depth and flow to guide RGB synthesis significantly improves generation quality (from 101.8 to 46.8 on FID) and spatiotemporal coherence (from 1364.6 to 577.1 on FVD). Meanwhile, the results of *UniDynamics (w/o RGB)* show that event-only prediction can still outperform Vista. However, due to the absence of rich appearance references, it yields lower generation quality and perception accuracy than the full UniDynamics model. Compared with *UniFuture* variants and *UniDynamics (w/o events)*, our method achieves slightly better FID, delivers a clear gain in RGB temporal coherence, and also yields stronger performance on the perception tasks. These gains stem from leveraging events to provide initial motion cues and rich texture/structure information, together with the PDS module that enables contin-

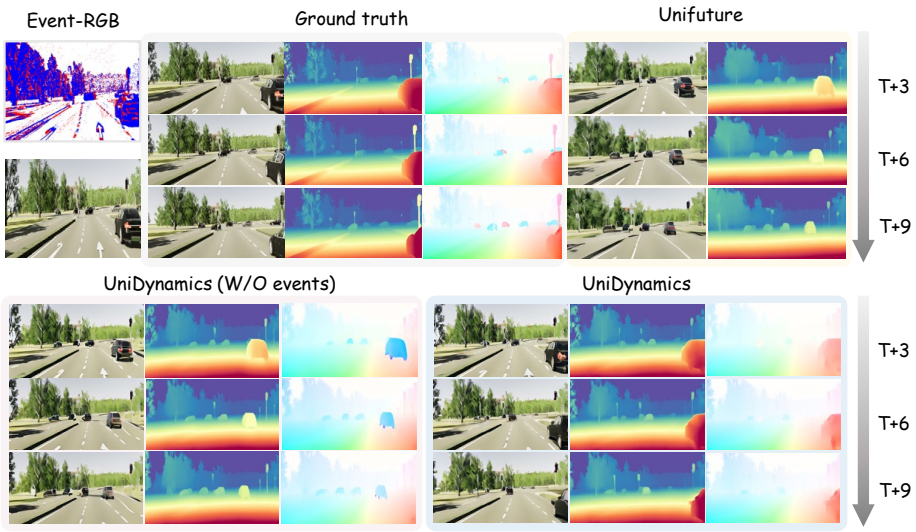

Fig. 5: Qualitative comparisons of future 4D scene prediction on E-VKitti2. UniDynamics preserves spatiotemporal coherence and yields more accurate perception results.

uous depth–flow–RGB interaction. Overall, our method delivers higher-fidelity (the lowest FID of 39.1) and more coherent RGB generates (the lowest FVD of 223.6), more accurate and structurally consistent depth, and lower-error, more robust optical flow, demonstrating the effectiveness of event-guided multimodal modeling for future scene generation and perception.

As shown in Fig. 5, **UniDynamics** generates RGB frames that better follow realistic scene evolution, leading to more accurate depth and flow predictions that capture scene geometry and motion more faithfully. In contrast, the baselines produce RGB frames with incorrect motion, resulting in less accurate depth and optical flow. These results demonstrate the effectiveness of our event-guided joint RGB–depth–flow modeling for coherent future forecasting and perception. Moreover, Fig. 6 visualizes the 3D point clouds reconstructed from the predicted RGB and depth, as well as the scene flow derived from the predicted optical flow. The reconstructions and scene flow closely match the ground truth in overall structure and motion trends, indicating that **UniDynamics** has the potential to enable 4D scene reconstruction of future worlds. More visualizations are provided in the supplementary material.

Robustness under Motion Blur. We evaluate robustness to motion blur on E-VKitti2 by simulating image degradation under high-speed motion. Tab. 2 compares our method with the strongest baseline *UniFuture* and our ablation *UniDynamics (w/o events)*. With a single blurred RGB input, the UniFuture family and *UniDynamics (w/o events)* perform poorly on all generation and perception metrics, consistent with the qualitative comparison in Fig. 1. Such challenging scenarios lack a reliable appearance reference, which degrades gen-

Table 2: Performance comparison under **motion blur** conditions on the E-VKitti2 [7].

Method	Generation		Depth					Optical Flow				
	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	RMSE ↓	EPE ↓	AE ↓	1PE ↓	2PE ↓	3PE ↓
UniFuture-Flow [24]	61.4	644.3	Unsupported					5.67	32.3	78.4	62.6	51.4
UniFuture-Depth [24]	56.1	668.7	0.391	57.9	84.1	90.1	59.5	Unsupported				
UniDynamics (w/o events)	59.8	617.7	0.288	50.2	79.9	84.2	58.5	5.40	33.8	78.6	60.9	48.7
UniDynamics (ours)	**53.6**	**272.8**	**0.265**	**68.1**	**88.0**	**92.9**	**54.4**	**2.86**	**19.5**	**55.4**	**36.0**	**25.9**

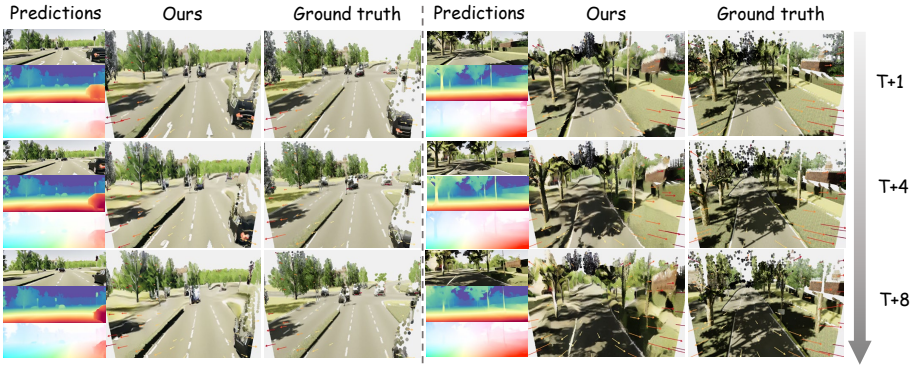

Fig. 6: The reconstructed 3D point cloud and scene flow, obtained from the future RGB, depth, and optical flow, demonstrate our ability to predict future 4D scenes.

eration quality as well as depth/flow estimation accuracy. These results indicate that event signals can provide effective texture and structural cues to compensate for blurred RGB in high-speed scenes, and that multimodal joint modeling is crucial for robust future prediction.

Zero-shot on real-world datasets. We evaluate cross-domain and sim-to-real generalization by directly transferring the model pretrained on E-VKitti2 to three real-world datasets. As shown in Tab. 3, although our FID is slightly inferior to UniFuture-Depth on DSEC and MVSEC, the introduction of events and optical flow substantially improves FVD performance. Moreover, our method achieves clearly better generation quality on the more complex M3ED dataset, indicating more temporally coherent and physically plausible scene evolution. In addition, our perception results improve substantially, demonstrating the effectiveness of PDS in modeling geometry and motion cues. Overall, **UniDynamics** generalizes well across domains and produces more coherent and accurate future 4D scene forecasts in real scenes.

4.4 Ablation Study

We conduct ablation studies on E-VKitti2 [7] to verify the contribution of the key components of UniDynamics, and report the results in Tab. 4.

Event Inputs. We remove the event input (*UniDynamics (w/o events)*) and condition the model only on a single RGB frame. The results show a substantial

Table 3: Zero-shot results on the real-world datasets. "-" indicates that the result is either not reported or the dataset does not support the computation of the metric.

Dataset	Method	Generation		Depth					Optical Flow				
		FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	RMSE ↓	EPE ↓	AE ↓	1PE ↓	2PE ↓	3PE ↓
DSEC	TAU [35]	126.5	1052.9					Unsupported					
	FLODCAST [4]	Unsupported		0.54	26.5	63.2	74.6	10.3	15.3	25.8	-	-	-
	FUTURIST [20]	Unsupported		0.40	38.1	66.9	85.8	**8.0**			Unsupported		
	UniFuture-Depth	**46.2**	285.7	0.35	39.7	71.2	89.5	8.9			Unsupported		
	UniFuture-Flow	54.3	314.8			Unsupported			12.4	32.4	96.3	88.1	79.1
	UniDynamics	50.9	**224.0**	**0.30**	**47.9**	**78.1**	**92.9**	8.1	**6.8**	**23.4**	**93.9**	**80.0**	**65.6**
MVSEC	TAU [35]	136.9	1335.5					Unsupported					
	UniFuture-Depth	**69.8**	694.8	0.23	34.3	**92.7**	99.8	-			Unsupported		
	UniFuture-Flow	88.9	722.7			Unsupported			17.9	51.9	90.9	83.2	76.7
	UniDynamics	77.2	**427.0**	**0.21**	**48.8**	91.3	**99.9**	-	**11.5**	**45.2**	**86.8**	**75.8**	**67.5**
M3ED	TAU [35]	156.1	1047.9					Unsupported					
	UniFuture-Depth	88.6	705.4	0.23	64.4	86.5	94.1	10.3			**Unsupported**		
	UniFuture-Flow	88.1	650.9			Unsupported			-	-	-	-	-
	UniDynamics	**68.5**	**508.3**	**0.22**	**66.5**	**87.3**	**94.5**	**10.1**	-	-	-	-	-

drop in generation quality and coherence (FID increases from 39.1 to 46.8, and FVD from 223.6 to 577.1), along with large degradations in depth and optical-flow accuracy. This confirms that events provide crucial motion cues and rich texture/structure information for guiding future scene evolution as well as geometry and motion modeling. Moreover, removing the ELE module, i.e., the "hard in" setting, and directly concatenating event features to the noisy latent variables of future frames as RGB features leads to performance degradation. This suggests that ELE effectively inject the useful motion cues and texture information of events to the diffusion process while preserving the inherent static appearance reference from RGB.

Flow Branch. We remove the optical-flow branch (*UniDynamics-Depth*) and keep only the depth branch. Without flow, PDS no longer needs to account for flow-related gradient constraints, and since depth and RGB are both static scene descriptors, the depth branch can focus more on geometric regression, making a slight improvement reasonable. However, the model simultaneously loses an explicit motion-field constraint, so the diffusion process must infer complex long-horizon dynamics mainly from single-frame appearance and local event cues. As a result, spatiotemporal consistency and physically plausible motion propagation become harder to maintain, leading to degraded RGB FID/FVD.

Depth Branch. Removing the depth branch (*UniDynamics-Flow*) leads to decreases in both RGB quality (FID/FVD) and optical-flow accuracy, due to the depth latent being injected as a geometry prior into the U-Net's spatial layers, and plays a crucial role in stabilizing scene structure and boundaries throughout the diffusion process. Moreover, reliable optical flow estimation also relies on stable static structure and well-aligned motion boundaries. When geometric constraints are absent, the motion branch must estimate flow solely from the RGB latent, leading to higher EPE/AE and worse nPE metrics. Overall, depth

Table 4: Ablation studies of four key components on E-VKitti2 [7].

Setting	Generation		Depth					Optical Flow				
	FID ↓	FVD ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	RMSE ↓	EPE ↓	AE ↓	1PE ↓	2PE ↓	3PE ↓
UniDynamics (w/o events)	46.8	577.1	0.36	56.0	83.9	90.3	59.0	5.32	34.1	79.2	61.5	49.1
UniDynamics (hard in)	43.3	417.6	0.33	60.2	85.8	91.4	56.7	4.63	28.2	71.8	53.5	41.4
UniDynamics-Depth (w/o flow)	42.0	264.2	**0.27**	**68.6**	**88.5**	**93.0**	**53.9**	Unsupported				
UniDynamics-Flow (w/o depth)	44.1	268.3	Unsupported					3.64	24.8	64.4	45.3	34.1
UniDynamics (w/o adapters)	50.8	384.9	0.30	62.6	85.6	91.6	56.0	4.60	30.3	72.3	53.5	41.9
UniDynamics (ours)	**39.1**	**223.6**	**0.27**	67.1	87.8	92.7	54.7	**3.00**	**21.0**	**58.1**	**38.5**	**27.7**

and flow are complementary: depth stabilizes spatial structure, which benefits both appearance synthesis and motion estimation.

Task Adapters in PDS. We further remove the task adapters and the gated bidirectional interaction in PDS to verify the effectiveness of task decoupling in PDS, and instead perform 3D convolutions directly in the joint layer to jointly model depth and optical flow. The resulting unified feature is then injected into the U-Net decoder via zero-conv. The results show clear degradations in generation quality and coherence, as well as in the perception metrics. This is because jointly modeling geometry and motion cues within a joint space tends to induce a gradient "tug-of-war", and injecting such entangled features can disturb the static appearance representations. In contrast, the task adapters in PDS explicitly decouple depth and flow to alleviate gradient conflicts, while the gated bidirectional interaction enables necessary information exchange without sacrificing task-specific modeling. Moreover, the two features are injected into the spatial and temporal layers according to their physical roles, further improving the accuracy and spatiotemporal consistency of 4D scene generation.

5 Conclusion

Toward practical deployment, we propose **UniDynamics**, a diffusion model conditioned on an event–RGB pair to generate future RGB, depth, and optical flow. We introduce event streams into the diffusion process via an Event Latent Enhancement (ELE) module, using them as complementary cues to RGB that provide effective initial momentum and reliable appearance guidance for scene extrapolation. We further decouple depth and flow in a Perceptual Dynamics Space (PDS) and tightly couple it with the U-Net denoising process to enforce physically plausible and spatiotemporally coherent 4D scene evolution. Experiments on both synthetic and real datasets demonstrate high-quality, temporally coherent appearance prediction and accurate geometric and motion perception.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62273093, 62236002).

References

1. Berner, R., Brandli, C., Yang, M., Liu, S.C., Delbruck, T.: A 240×180 10mw 12us latency sparse-output vision sensor for mobile applications. In: 2013 Symposium on VLSI Circuits. pp. C186–C187 (2013)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Chaney, K., Cladera, F., Wang, Z., Bisulco, A., Hsieh, M.A., Korpela, C., Kumar, V., Taylor, C.J., Daniilidis, K.: M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4016–4023 (2023)
4. Ciamarra, A., Becattini, F., Seidenari, L., Del Bimbo, A.: Flodcast: Flow and depth forecasting via multimodal recurrent architectures. Pattern Recognition **150**, 110337 (2024)
5. Das, R., Daniilidis, K., Chaudhari, P.: Fast feature field (F^3): A predictive representation of events (2025)
6. Devulapally, A., Khan, M.F.F., Advani, S., Narayanan, V.: Multi-modal fusion of event and rgb for monocular depth estimation using a unified transformer-based architecture. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2081–2089 (2024)
7. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4340–4349 (2016)
8. 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. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
9. Gao, Z., Tan, C., Wu, L., Li, S.Z.: Simvp: Simpler yet better video prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3170–3180 (June 2022)
10. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robotics and Automation Letters **6**(2), 2822–2829 (2021)
11. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3), 4947–4954 (2021)
12. Gehrig, M., Muglikar, M., Scaramuzza, D.: Dense continuous-time optical flow from event cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(7), 4736–4746 (2024)
13. Hassan, M., Stapf, S., Rahimi, A., Rezende, P.M.B., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., Cannici, M., Aljalbout, E., Ye, B., Wang, X., Davtyan, A., Salzmann, M., Scaramuzza, D., Pollefeys, M., Favaro, P., Alahi, A.: Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22404–22415 (June 2025)
14. 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 **abs/2309.17080** (2023)

15. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
16. Hu, Y., Zhang, Y., Song, Y., Deng, Y., Yu, F., Zhang, L., Lin, W., Zou, D., Yu, W.: Seeing through pixel motion: Learning obstacle avoidance from optical flow with one camera. IEEE Robotics and Automation Letters **10**(6), 5871–5878 (2025)
17. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1312–1321 (2021)
18. Jia, F., Mao, W., Liu, Y., Zhao, Y., Wen, Y., Zhang, C., Zhang, X., Wang, T.: Adriver-i: A general world model for autonomous driving. arXiv preprint arXiv:2311.13549 (2023)
19. Jing, L., Shi, D., Liu, Z., Jin, S., Qiu, C., Qiao, Z., Li, Y., Xia, J.: Unict depth: Event-image fusion based monocular depth estimation with convolution-compensated vit dual sa block. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 1287–1295. International Joint Conferences on Artificial Intelligence Organization (8 2025)
20. Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Advancing semantic future prediction through multimodal visual sequence transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 3793–3803 (June 2025)
21. Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: DINO-foresight: Looking into the future with DINO. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
22. Kong, J., Kim, J.H., Lee, J.S.: Dessert: Diffusion-based event-driven single-frame synthesis via residual training. ArXiv **abs/2512.17323** (12 2025)
23. Li, Q., Jia, X., Wang, S., Yan, J.: Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2). In: ECCV (2024)
24. Liang, D., Zhang, D., Zhou, X., Tu, S., Feng, T., Li, X., Zhang, Y., Du, M., Tan, X., Bai, X.: Seeing the future, perceiving the future: A unified driving world model for future generation and perception. In: IEEE International Conference on Robotics and Automation (ICRA) (2026)
25. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits **43**(2), 566–576 (2008)
26. Lin, Z., Sun, J., Hu, J.F., Yu, Q., Lai, J.H., Zheng, W.S.: Predictive feature learning for future segmentation prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7365–7374 (2021)
27. Liu, H., Qu, S., Lu, F., Bu, Z., Roehrbein, F., Knoll, A., Chen, G.: Pcdepth: Pattern-based complementary learning for monocular depth estimation by best of both worlds. pp. 11187–11194 (2024)
28. Luo, A., Li, X., Yang, F., Liu, J., Fan, H., Liu, S.: Flowdiffuser: Advancing optical flow estimation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19167–19176 (2024)
29. Pan, T., Cao, Z., Wang, L.: Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 10695–10702 (2024)

30. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022)
31. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 3505–3506 (2020)
32. Shi, D., Jing, L., Li, R., Liu, Z., Wang, L., Xu, H., Zhang, Y.: Improved event-based dense depth estimation via optical flow compensation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4902–4908 (2023)
33. 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: Computer Vision – ECCV 2024. pp. 256–274. Springer Nature Switzerland (2025)
34. Sun, P., Jiang, J., Yao, Y., Chen, Y., Zhao, W., Jiang, K., Liu, X.: Fuse: Label-free image-event joint monocular depth estimation via frequency-decoupled alignment and degradation-robust fusion. pp. 17231–17238 (2025)
35. Tan, C., Gao, Z., Wu, L., Xu, Y., Xia, J., Li, S., Li, S.Z.: Temporal attention unit: Towards efficient spatiotemporal predictive learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18770–18782 (2023)
36. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020)
37. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8406–8415 (2023)
38. Wan, Z., Dai, Y., Mao, Y.: Learning dense and continuous optical flow from an event camera. IEEE Transactions on Image Processing **31**, 7237–7251 (2022)
39. Wang, H., Zhou, H., Liu, H., Yan, L.: Injecting frame-event complementary fusion into diffusion for optical flow in challenging scenes. arXiv preprint arXiv:2510.10577 (2025)
40. 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)
41. Wang, Y., Long, M., Wang, J., Gao, Z., Yu, P.S.: PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In: Advances in Neural Information Processing Systems. pp. 879–888 (2017)
42. Wu, S., Zhu, Z., Hou, J., Shi, G., Wu, J.: E-motion: Future motion simulation via event sequence diffusion. Advances in Neural Information Processing Systems **37**, 105552–105582 (2024)
43. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (2025)
44. Zhou, Q., Hou, J., Yang, M., Deng, Y., Li, Y., Xiong, J.: Spatially-guided temporal aggregation for robust event-rgb optical flow estimation. ArXiv **abs/2501.00838** (2025)
45. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)

46. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
47. Zhu, J., Liu, L., Jiang, B., Wen, F., Zhang, H., Li, W., Liu, Y.: Self-supervised event-based monocular depth estimation using cross-modal consistency. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7704–7710 (2023)