EvDiff3D: Event-Aware Diffusion Repair for High-Fidelity Event-Based 3D Reconstruction

Kanghao Chen[1†], Zixin Zhang[1†], Hangyu Li[1], Lin Wang[3], Zeyu Wang[1,2‡]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology
[3] Centre for Advanced Robotics Technology Innovation (CARTIN), School of EEE, Nanyang Technological University
kchen879@connect.hkust-gz.edu.cn, linwang@ntu.edu.sg, zeyuwang@ust.hk

Abstract

Event cameras are bio-inspired sensors that capture visual information through asynchronous brightness changes, offering distinct advantages including high temporal resolution and wide dynamic range. While prior research has investigated event-based 3D reconstruction for extreme scenarios, existing methods face inherent limitations and fail to fully exploit the unique characteristics of event data. In this paper, we present **EvDiff3D**, a novel two-stage 3D reconstruction framework that integrates event-based geometric constraints with an event-aware diffusion prior for appearance refinement. Our key insight lies in bridging the gap between physically grounded event-based reconstruction and data-driven appearance repair through a unified cyclical pipeline. In the first stage, we reconstruct a coarse 3D scene under supervision from event loss and event-based monocular depth constraints to preserve structural fidelity. The second stage finetunes an event-aware diffusion model based on a pretrained video diffusion model as a repair prior to enhance the appearance in under-constrained regions. Based on the diffusion model, our pipeline operates within a reconstruction-generation cycle that progressively refines both geometry and appearance using only event data. Extensive experiments on synthetic and real-world datasets demonstrate that EvDiff3D significantly outperforms existing methods in perceptual quality and structural consistency.

Introduction

Event cameras are bio-inspired vision sensors that asynchronously capture brightness changes at each pixel with microsecond latency. Unlike conventional frame-based cameras that sample scenes at fixed intervals, event cameras respond exclusively to pixel-level intensity changes (i.e., events). These distinctive characteristics endow event cameras with superior temporal resolution and wider dynamic range, facilitating numerous applications across computer vision (Jiang et al. 2020; Chen et al. 2025a; Gallego et al. 2020; Alonso and Murillo 2019; Liang et al. 2024; Chen et al. 2024; Zhou et al. 2025; Chen et al. 2025b), robotics (Chamorro, Sola, and Andrade-Cetto 2022; Mahlknecht et al. 2022; Mueggler et al. 2018), and VR (Chen, Wang, and Wang 2025; Zhang et al. 2024; Morgenstern et al. 2023; Dubeau et al. 2020).

Recent advances have extended event-based methods from low-level vision tasks to high-level applications, including 3D scene reconstruction and neural rendering. Neural representations such as Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) have achieved remarkable success in photorealistic scene synthesis. However, these methods typically require dense, high-quality RGB images and exhibit degraded performance under sparse, noisy, or asynchronous inputs. Therefore, this limitation has motivated the integration of event cameras to leverage their unique advantages in challenging scenarios.

Despite the valuable insights from existing work, integrating event cameras with neural representations remains highly challenging. Purely event-based approaches (e.g., EventNeRF (Rudnev et al. 2023), Ev-NeRF (Hwang, Kim, and Kim 2023), Event-3DGS (Han et al. 2024)) leverage the high temporal resolution of event streams to constrain intensity variations during camera motion. However, relying exclusively on event data fails to capture rich semantic content due to the inherent sparsity of event streams, as shown in Figure 1 (c). In contrast, dual-modality methods that combine event data with RGB frames, e.g., E2NeRF (Qi et al. 2023), Deblur e-NeRF (Low and Lee 2024), Ev-DeblurNeRF (Cannici and Scaramuzza 2024), achieve more detailed appearance reconstruction. Nevertheless, these approaches depend on well-calibrated images and remain constrained by the limitations of RGB cameras under extreme conditions. This dependency reduces real-world adaptability and introduces training redundancy. Consequently, existing methods either fail to effectively recover fine-grained details or underexploit the full potential of event data.

To address these limitations, we introduce **EvDiff3D**, a novel event-based 3D reconstruction framework that integrates event-aware generative priors with event-based geometric constraints within a synergetic framework. Our key insight lies in bridging the gap between physically grounded event-based reconstruction and data-driven appearance repair through a unified cyclical optimization pipeline.

Specifically, we propose a two-stage optimization framework for reconstructing high-fidelity 3D scenes, consisting of an initial structural construction followed by an appear-

[1] † Equal contribution.

[2] ‡ Corresponding author.

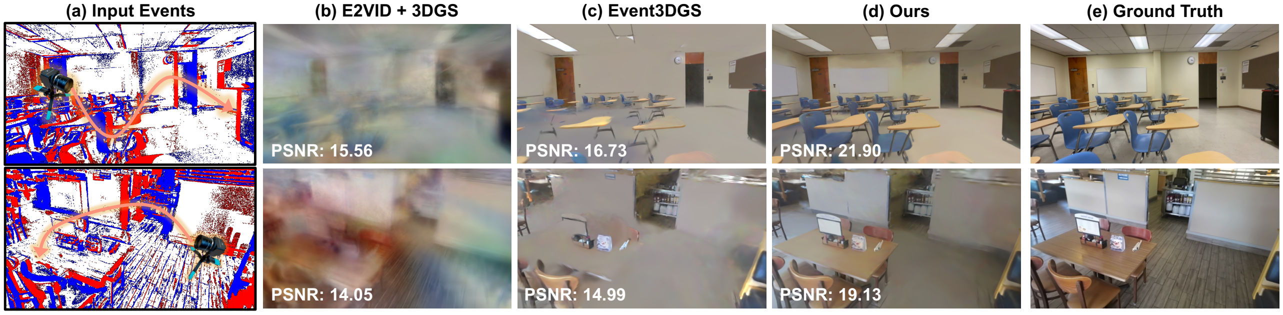

Figure 1: Comparison of Our EvDiff3D with baselines. Our EvDiff3D improves event-based 3DGS reconstruction by integrating event-based geometric constraints and an event-aware diffusion repair model.

ance refinement. In the first stage, we generate a coarse 3DGS by jointly optimizing an event loss with monocular depth constraints derived from event-based depth estimation (Zhu et al. 2025). While this stage produces a geometrically plausible 3DGS, the resulting appearance and textures often exhibit artifacts and missing content due to the inherent sparsity of event data. To address these limitations, the second stage fine-tunes an event-aware diffusion model, leveraging a pretrained video diffusion model DynamiCrafter (Xing et al. 2024) as a generative prior to repair under-constrained regions and enhance visual fidelity. Concretely, using a large-scale, real-world video dataset (Ling et al. 2024), we first perform 3D scene reconstruction with intensity-difference supervision to simulate event data, compensating for the absence of actual event data. From the reconstructed scenes, we render videos along the same camera trajectories as the reference videos, using them to fine-tune the diffusion model for event-aware artifact repair under supervision from the original reference videos. Additionally, we extract global scene features from randomly sampled event frames (i.e., accumulated events), which provide structural cues for conditioning the diffusion model. In this way, we obtain an event-aware diffusion model that facilitates a unified cyclical pipeline for reconstruction and repair, effectively aligning event-induced structure with generative priors to recover semantically rich and visually coherent scene details, as shown in Figure 1(d).

We validate our method on both synthetic and real-world datasets, demonstrating that EvDiff3D significantly outperforms previous event-based methods. Our approach achieves a 12.1% improvement on PSNR metrics and produces qualitatively superior results with sharper structures and enhanced appearance details.

The key contributions of this work are fourfold:

- We propose EvDiff3D, a novel two-stage framework for event-based 3D reconstruction that integrates event-based geometric constraints with diffusion priors.
- We develop a video diffusion model that efficiently repairs event-aware artifacts and generates plausible content in under-constrained regions.
- We introduce a unified cyclical optimization pipeline that progressively refines artifact-prone views while preserving event-induced structural integrity.
- We conduct comprehensive experiments on both synthetic and real-world datasets, demonstrating state-of-the-art performance in event-based 3D reconstruction.

Related Work

Neural Representations for 3D Reconstruction Neural representations have achieved remarkable success in photorealistic scene synthesis. NeRF (Mildenhall et al. 2021) pioneered neural view synthesis by optimizing fully-connected networks supervised by sparse multi-view observations. Subsequent works improved efficiency and quality: Mip-NeRF (Barron et al. 2021) enhances detail preservation through anti-aliased conical frustum rendering, while Instant-NGP (Müller et al. 2022) achieves acceleration using multiresolution hash tables. 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) represents scenes through explicit Gaussian primitives, enabling real-time rendering via differentiable rasterization. However, these methods require dense, high-quality RGB images and exhibit degraded performance under sparse or asynchronous inputs. Our framework builds upon 3DGS while addressing event-based reconstruction challenges.

Event-Based 3D Reconstruction Event cameras asynchronously capture brightness changes with superior temporal resolution. Recent neural rendering approaches have integrated event data with established frameworks to leverage these unique characteristics. Purely event-based methods, e.g., Ev-NeRF (Hwang, Kim, and Kim 2023), Event-NeRF (Rudnev et al. 2023), exploit the high temporal resolution to constrain intensity changes but struggle with semantic content recovery due to event sparsity, often resulting in artifacts and limited color fidelity. Dual-modality methods (Qi et al. 2023; Cannici and Scaramuzza 2024; Lee and Lee 2025) combine event data with RGB frames to achieve deblurring and temporal consistency. Nevertheless, these approaches are highly dependent on well-calibrated RGB images and remain constrained by the limitations of conventional RGB cameras under challenging or extreme conditions. Consequently, existing methods face inherent tradeoffs: purely event-based approaches lack the capacity to reconstruct fine details, while dual-modality approaches diminish the full potential of event data by relying heavily on RGB supervision. Recently, DiET-GS (Lee and Lee 2025)

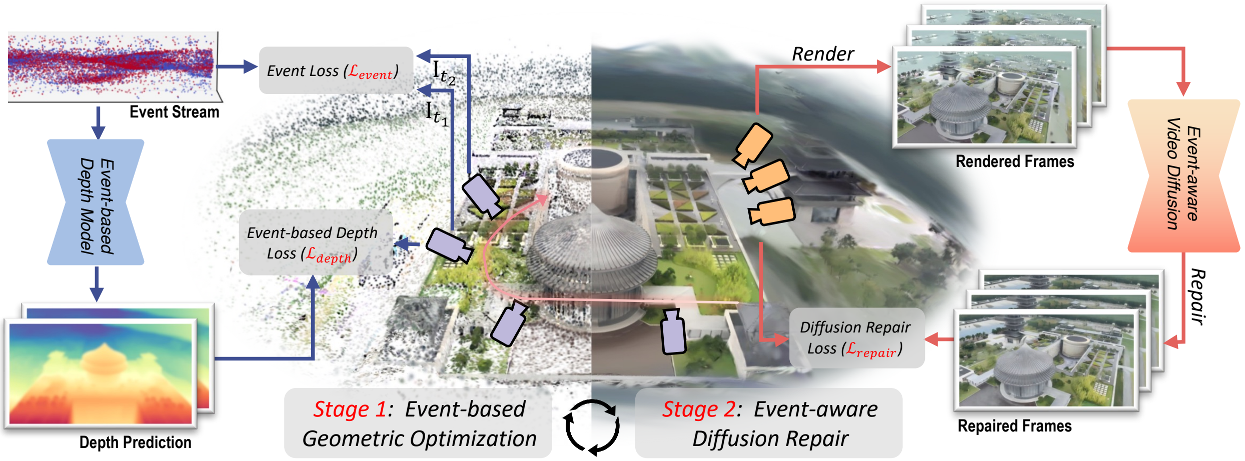
Event Stream
Event-based Depth Model
Depth Prediction
Event Loss ($\mathcal{L}_{event}$)
I_{t_2}
I_{t_1}
Event-based Depth Loss ($\mathcal{L}_{depth}$)
Stage 1: Event-based Geometric Optimization
Stage 2: Event-aware Diffusion Repair
Render
Rendered Frames
Event-aware Video Diffusion
Repair
Repaired Frames
Diffusion Repair Loss ($\mathcal{L}_{repair}$)

Figure 2: EvDiff3D pipeline. Overview of the proposed EvDiff3D framework. The first stage reconstructs a coarse 3DGS using event loss and event-based depth constraints. The second stage fine-tunes an event-aware diffusion model to repair under-constrained regions and enhance appearance details.

also incorporates diffusion priors. However, it still relies on RGB images, and its diffusion prior is directly adapted from image-generation models, limiting its robustness in event-only scenarios. In contrast, our method enables high-fidelity 3D reconstruction using only event data by introducing a specialized event-aware diffusion prior within a cyclical optimization pipeline, fully exploiting the unique strengths of event streams for 3D reconstruction.

Diffusion Models for 3D Reconstruction DreamFusion (Poole et al. 2023) introduced Score Distillation Sampling (SDS), enabling NeRF training through pretrained 2D diffusion models. Improvements include FlowDreamer (Li, Chu, and Shi 2024) with Unified Classifier-Free Guidance and ProlificDreamer (Wang et al. 2024) with Variational Score Distillation. Recently, methods like ReconFusion (Wu et al. 2024), 3DGS-Enhancer (Liu, Zhou, and Huang 2024), and GenFusion (Wu et al. 2025) utilize diffusion priors for sparse-view reconstruction enhancement. However, these approaches are primarily designed for RGB-based scenarios and are ineffective in repairing event-specific artifacts or handling the unique characteristics of event data. In this work, we introduce the first event-aware diffusion prior tailored for event-based 3D reconstruction. Our method synergistically integrates event-aware generative priors with event-driven geometric constraints, effectively addressing the challenges of data sparsity and semantic content recovery inherent in event cameras.

Method

We propose EvDiff3D, a two-stage framework that addresses the challenge of high-fidelity 3D reconstruction based on only event data, as shown in Figure 2. Our approach is guided by an event-aware diffusion prior combined with event-based geometric constraints. The first stage leverages event loss and event-based depth cues to ensure structural fidelity, reconstructing a coarse 3DGS. The second stage fine-tunes an event-aware diffusion model to repair under-constrained regions and enhance semantic detail. A cyclical pipeline is applied to progressively refine both geometry and appearance, fully exploiting the unique advantages of event cameras without requiring RGB images.

Stage 1: Event-Based Geometric Optimization

In the first stage, we reconstruct a coarse 3DGS from event sequences to capture the scene's global structure and geometry. This initialization is essential for the subsequent diffusion-based optimization, as 3DGS may struggle to achieve fine texture reconstruction under the highly stochastic generative guidance of diffusion models, which poses significant convergence challenges for our cyclical optimization pipeline. Through this coarse geometric optimization, we establish a robust structural foundation.

Our approach builds upon existing event-based reconstruction methods (Rudnev et al. 2023; Qi et al. 2023; Hwang, Kim, and Kim 2023) by incorporating a basic event loss. Additionally, we introduce a depth loss produced by event-based depth estimation to better constrain depth information, thereby leveraging event data's inherent advantage in capturing structural fidelity.

Event Loss. During optimization, we render two images $\mathbf{I}_{t_1}$ and $\mathbf{I}_{t_2}$ at timestamps t_1 and t_2. We convert both images to log space and compute their difference. This difference map is then compared with the ground truth event frame to calculate the supervision loss:

$$\mathcal{L}_{event} = \|(\log(\mathbf{I}_{t_1}) - \log(\mathbf{I}_{t_2})) - \mathbf{E}(t_1, t_2)\|_1, \quad (1)$$

where $\mathbf{E}$ represents the event frame with element $E_{x,y}$ at position (x, y) recording the event count.

Event-Based Depth Loss. To further constrain the 3DGS geometry, we leverage monocular depth estimation derived from event data using an off-the-shelf event-based depth estimator (Zhu et al. 2025). Given depth estimation $\mathbf{D}_{\text{event}}$ from the event data, we render a corresponding depth map

$\mathbf{D}_{\text{pred}}$ from the 3DGS. The depth loss is computed as the L_2 distance between the rendered and estimated depth maps:

$$\mathcal{L}_{\text{depth}} = \|(s\mathbf{D}_{\text{pred}} + t) - \mathbf{D}_{\text{event}}\|_2, \tag{2}$$

where s and t are scale and shift parameters used to align the two depth maps, optimized during the training process.

Coarse Optimization. The overall loss function for the first stage combines Equations 1 and 2:

$$\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{event}} + \lambda_d \mathcal{L}_{\text{depth}}, \tag{3}$$

where $\lambda_d = 0.5$ denotes the weighting factor that balances the two loss components.

Stage 2: Event-Aware Diffusion Repair

Following the coarse reconstruction stage, the resulting 3DGS exhibits plausible geometry but often suffers from coarse textures and missing content, limiting overall reconstruction quality, as illustrated in Figure 7. This limitation arises because event cameras primarily capture high-frequency changes (e.g., edges), leading to a lack of fine-grained appearance details and failing to reconstruct low-frequency regions (e.g., homogeneous surfaces such as white walls). To overcome these limitations, we introduce a refinement stage that incorporates a diffusion prior, building on the demonstrated effectiveness of diffusion models in recent generative methods (Poole et al. 2023; Tang et al. 2023; Liu, Zhou, and Huang 2024). Although diffusion models can generate semantically meaningful frames, they always introduce stochasticity in the generative objectives and fail to adapt to the event-based scenarios. This randomness and poor generalization can hinder the 3DGS from maintaining the integrity of event data, as fine-grained optimization becomes challenging under conflicting objectives. To address these issues, we propose fine-tuning an event-aware diffusion repair model and integrating it within a cyclical reconstruction–repair framework, enabling the 3DGS model to iteratively refine both geometry and appearance while maintaining consistency with event data.

Intensity Difference for 3D Reconstruction. To fine-tune the diffusion model for event-based scenarios, paired data comprising event-driven artifact-prone videos and corresponding high-quality videos is required. For the high-quality videos, we leverage the large-scale DL3DV dataset (Ling et al. 2024), which provides diverse, high-quality RGB video content, to serve as the supervision for fine-tuning our diffusion model. To generate the artifact-prone videos, we reconstruct 3DGS from event data and render videos along the same camera trajectories as those in the high-quality dataset. However, the DL3DV dataset does not include event data, and converting RGB videos to event streams using modern v2e models (Hu, Liu, and Delbruck 2021) is computationally prohibitive at this scale. To address this limitation, we introduce an intensity-difference supervision strategy that guides 3DGS reconstruction to emulate an event camera, inherently capturing brightness changes. This approach enables efficient generation of event-driven artifact-prone video for model adaptation.

Specifically, given two adjacent video frames $\{I_t, I_{t+\Delta t}\}$, we compute a dense per-pixel difference map $\Delta I = |I_{t+\Delta t} - I_t|$. During training, we render two frames from the 3DGS at corresponding adjacent timestamps and supervise the rendered difference $\Delta\hat{I}$ against ΔI using an L_1 loss: $L_1 = ||\Delta\hat{I} - \Delta I||_1$. After intensity difference reconstruction, we render videos along identical camera trajectories as the input sequences, forming paired event-driven artifact-prone and high-quality videos for diffusion model training. After being trained on these synthetic paired videos, the diffusion model demonstrates strong generalization in repairing event-aware artifacts while maintaining computational efficiency by leveraging a large-scale video dataset without requiring computationally expensive v2e conversion. Detailed qualitative results are provided in the Appendix.

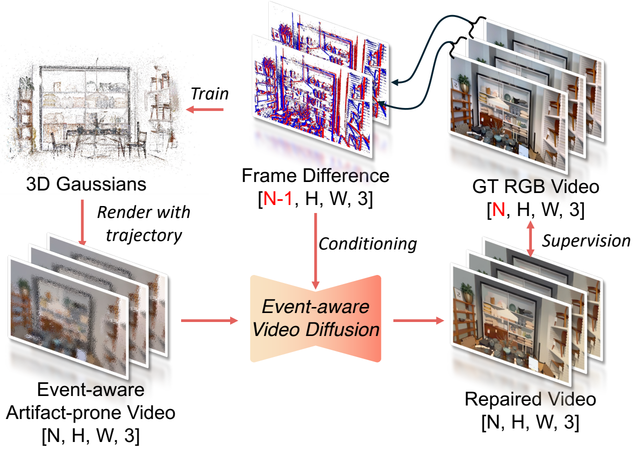

Figure 3: The illustration of our training process of the event-aware video diffusion model.

Fine-Tuning Event-Aware Diffusion Models. We build our event-aware video repair model upon the foundation of pretrained video diffusion i.e., DynamiCrafter (Xing et al. 2024), and adapt it for event-driven artifact repair. Given the paired data from intensity difference reconstruction, we train the diffusion model to repair artifact-prone rendered videos $\mathbf{V}_{\text{rendered}}$ and generate high-quality videos $\mathbf{V}_{\text{gt}}$ that match ground truth captures. The ground truth RGB video $\mathbf{V}_{\text{gt}}$ is encoded into latent space $\mathbf{z}_0 := E(\mathbf{V}_{\text{gt}})$, to which we add noise at timestep t to obtain $\mathbf{z}_t$.

To guide the generation process with event-aware information, we introduce two complementary conditioning signals $\mathbf{c}$. First, the rendered artifact-prone video $\mathbf{V}_{\text{rendered}}$ is encoded and concatenated with the noised latent $\mathbf{z}_t$ to enable sequence-level conditioning. This design leverages the visual information from the rendered videos to provide detailed spatial cues while preserving temporal consistency across frames. Additionally, we incorporate global scene features by extracting CLIP embeddings from randomly sampled accumulated event frames. These features offer high-level semantic conditioning, capturing structural information derived from event data to guide the generative process toward semantically coherent reconstructions. The video denoising network ϵ_θ is optimized using:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_{\text{event}}, \mathbf{c}_{\text{rendered}}, t)\|_2^2\right], \tag{4}$$

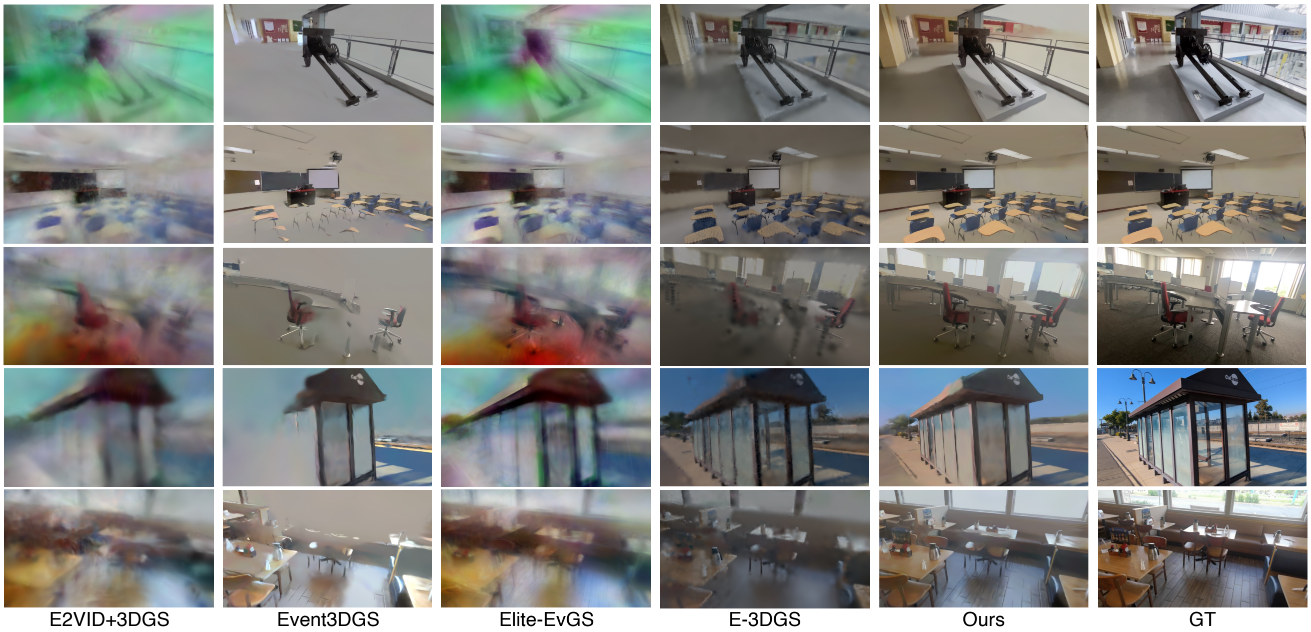

Figure 4: Qualitative comparison of novel view synthesis on Ev-DL3DV dataset

where $\mathbf{c}_{\text{event}}$ represents CLIP features extracted from accumulated event frames, $\mathbf{c}_{\text{rendered}}$ denotes the encoded event-driven artifact-prone video condition, and the optimization objective aims to reconstruct the original noise $\boldsymbol{\epsilon}$.

Cyclical Reconstruction-Repair. Once the event-aware diffusion model is trained, we employ a cyclical optimization pipeline that alternates between 3DGS reconstruction and diffusion-based repair. This iterative process progressively refines both geometry and appearance while ensuring consistency with event-derived constraints.

In each cycle, we first render video sequences from the current 3DGS along camera trajectories sampled based on event camera poses. These rendered videos are then processed by our event-aware diffusion model, producing repaired versions with enhanced semantic content and reduced artifacts. The repaired videos serve as direct supervision targets for further optimizing the 3DGS using a standard RGB reconstruction loss. We update the 3DGS through multiple gradient steps, jointly leveraging the repaired-video reconstruction loss and the coarse event-based loss from the first stage. Formally, given artifact-prone rendered video sequences $\mathbf{V}_{\text{rendered}}$ from the current 3DGS, we repair them through the event-aware diffusion model to obtain enhanced video frames $\mathbf{V}_{\text{repaired}}$. The repaired frames serve as RGB supervision targets for 3DGS optimization. We formulate a reconstruction loss between newly rendered frames from updated 3DGS parameters $\mathbf{V}_{\text{new}}$ and the repaired frames:

$$\mathcal{L}_{\text{repair}} = \|\mathbf{V}_{\text{new}} - \mathbf{V}_{\text{repaired}}\|_1 + \lambda_{\text{lpips}}\mathcal{L}_{\text{LPIPS}}(\mathbf{V}_{\text{new}}, \mathbf{V}_{\text{repaired}}), \tag{5}$$

where $\mathcal{L}_{\text{LPIPS}}$ is the perceptual loss ensuring semantic consistency, and $\lambda_{\text{lpips}} = 0.2$ balances pixel-level and perceptual losses. The final objective in the second stage combines event-based constraints with repair supervision:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{event}} + \lambda_d\mathcal{L}_{\text{depth}} + \lambda_{\text{repair}}\mathcal{L}_{\text{repair}}, \tag{6}$$

where $\lambda_{\text{repair}} = 0.5$ is the balancing weight.

Subsequently, artifact-prone videos are re-rendered from the updated 3DGS, and the process is repeated. This cyclical procedure continues until convergence, enabling the model to iteratively refine scene geometry and appearance by integrating diffusion priors with event-driven geometric constraints for high-fidelity reconstruction.

Experiments

Datasets. We evaluate our approach on two benchmark datasets to demonstrate effectiveness across both synthetic and real-world scenarios: the synthetic **Ev-DL3DV** dataset (Ling et al. 2024) and the real-world **TUM-VIE** dataset (Klenk et al. 2021). For the synthetic dataset construction, we select 24 diverse scenes from the DL3DV-Benchmark, encompassing both indoor and outdoor environments under various illumination conditions, which are not included in the DL3DV training dataset. Event streams are synthesized using the v2e simulator (Hu, Liu, and Delbruck 2021) with Bayesian filtering to emulate realistic colorful event data. The TUM-VIE dataset comprises real-world recordings captured using a Prophesee Gen4 event sensor, with RGB views provided by an externally calibrated camera for reference. Camera extrinsics are tracked at a 120 Hz frequency. Following the evaluation protocol established by E-3DGS (Han et al. 2024), we utilize the *mocap-1d-trans* and *mocap-desk2* sequences for comparative analysis.

Evaluation Metrics. We employ a comprehensive set of metrics to quantitatively assess reconstruction quality. For standard image quality assessment, we utilize three widely-adopted metrics: PSNR, SSIM, and LPIPS (Zhang et al.

Metrics	E2VID+3DGS	Event3DGS (Xiong et al. 2025)	Elite-EvGS (Zhang, Chen, and Wang 2025)	E-3DGS (Han et al. 2024)	EvDiff3D (Ours)
PSNR↑	13.76	16.82	13.63	14.27	**18.86**
SSIM↑	0.5534	0.7481	0.5481	0.7805	**0.7841**
LPIPS↓	0.7691	0.3584	0.6221	0.5162	**0.3243**
MUSIQ↑	19.17	55.11	21.494	39.13	**59.80**
BRISQUE↓	75.413	67.67	66.67	74.55	**57.20**
NIQE↓	10.72	9.14	8.97	8.44	**5.84**

Table 1: Quantitative comparison on the synthetic dataset of Ev-DL3DV. The best results are highlighted in **bold**, and the second-best is highlighted in underline.

Methods	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	BRISQUE↓
mocap-1d-trans					
Event3DGS	8.16	0.3978	0.4414	33.45	88.87
E-3DGS	7.65	0.4385	0.5374	42.69	86.28
EvDiff3D (Ours)	**12.06**	**0.6258**	**0.4072**	**45.02**	**82.78**
mocap-desk2					
Event3DGS	10.90	0.5673	0.3801	37.43	85.97
E-3DGS	11.72	0.5853	0.4833	35.91	86.59
EvDiff3D (Ours)	**13.47**	**0.6105**	**0.3465**	**38.04**	**82.75**

Table 2: Quantitative comparison on two sequences of the real-world dataset TUM-VIE (Klenk et al. 2021). The best results are highlighted in **bold**, and the second-best is highlighted in underline.

2018). Given that our method introduces enhanced semantic content in event-based reconstruction, we additionally incorporate no-reference image quality assessment metrics: MUSIQ (Ke et al. 2021), BRISQUE (Mittal, Moorthy, and Bovik 2012), and NIQE (Mittal, Soundararajan, and Bovik 2013). These metrics provide complementary evaluation of perceptual quality without requiring reference images.

Synthetic Results

Table 1 and Figure 4 present quantitative and qualitative comparisons on the synthetic Ev-DL3DV dataset, respectively. As shown in Table 1, our EvDiff3D method demonstrates superior performance across all evaluation metrics, achieving state-of-the-art results in event-based 3D reconstruction. Compared to the advanced baseline Event3DGS, EvDiff3D achieves substantial improvements: a 2.04 dB (12.1%) gain in PSNR and a notable 0.034 (9.5%) reduction in LPIPS. The no-reference quality metrics further validate our approach's effectiveness in enhancing semantic content. Figure 4 showcases the qualitative superiority of our method across diverse synthetic scenes. Prior event-only baselines struggle with texture recovery and semantic detail preservation due to the inherent sparsity of event data. For instance, Event3DGS produces blurry reconstructions with prominent artifacts and missing regions (e.g., the floor), underscoring the limitations of relying solely on high-frequency information from event streams. In contrast, EvDiff3D achieves reconstructions with significantly sharper textures and fewer artifacts. Our approach effectively captures fine-grained details such as surface patterns, material properties, and complex geometric structures that are poorly represented or entirely absent in baseline methods, demonstrating its capability to overcome the challenges of event-only 3D reconstruction.

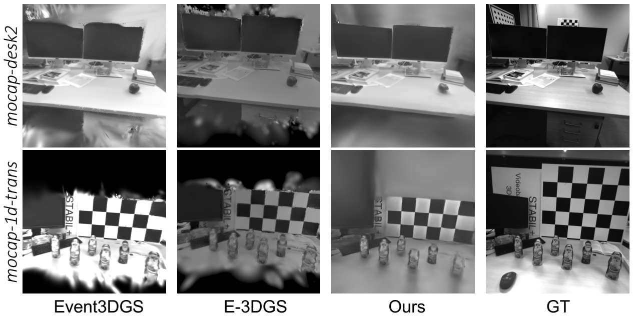

Figure 5: Qualitative comparison on the real-world TUM-VIE (Klenk et al. 2021) datasets.

Real-World Results

Table 2 and Figure 5 present the quantitative and qualitative results on the real-world TUM-VIE dataset (Klenk et al. 2021). Notably, we optimized the 3DGS model using an event camera with a resolution of 1280×720, while evaluation was conducted using a calibrated camera with a resolution of 1024×1024. This discrepancy in resolution and field of view often leads to pronounced artifacts in prior methods that lack diffusion repair, such as black margins outside the event camera's field of view, which severely impact evaluation metrics. Leveraging our cyclical reconstruction and repair pipeline, EvDiff3D effectively eliminates these artifacts and fills the margins with plausible structure, producing high-quality results. The quantitative comparison in Table 2 demonstrates that EvDiff3D significantly outperforms existing methods across all metrics. Figure 5 further illustrates the qualitative improvements achieved by EvDiff3D. Unlike previous methods, which struggle with artifacts and incomplete reconstructions, our framework successfully restores missing details and enhances texture fidelity, demonstrating its effectiveness in real-world scenarios.

Ablation Study

Impact of Event-Based Depth Loss. To assess the impact of the event-based depth loss, we compare our method with

Ablation	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	BRISQUE↓
3DGS Baseline	16.78	0.7468	0.3596	54.76	62.02
+ Depth Loss	17.92	0.7768	0.3311	56.25	58.37
+ Diffusion Repair	18.86	0.7841	0.3243	59.80	57.20

Table 3: Ablation study on the impact of different components in our framework. All the results are evaluated on the Ev-DL3DV dataset.

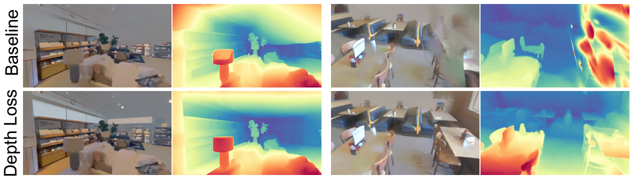

Figure 6: Ablation study on the event-based depth loss.

a baseline that does not incorporate the depth loss over the optimization pipeline by setting $\lambda_d = 0$. Table 3 (2^{nd} row) shows the quantitative results, where the event-based depth loss achieves a 1.14 dB (6.8%) improvement in PSNR compared to the baseline (1^{st} row). Figure 6 demonstrates the superiority of the depth loss in preserving structural fidelity.

Impact of Diffusion Repair. As shown in Table 3, incorporating diffusion prior (3^{nd} row) achieves a 0.94 dB (5.2%) improvement in PSNR compared to the baseline without diffusion repair (2^{nd} row), which demonstrates the effectiveness of the diffusion prior to enhance semantic content and reduce artifacts. Qualitative results in Figure 7 also show that our method can effectively repair texture details. Additionally, we ablate the impact of conditioning on the event CLIP feature and observe a 0.11 PSNR drop when it is removed.

Analysis Study

Sensitivity Study of Parameter λ_{repair}. We provide further analysis of the parameter λ_{repair}, which balances the reconstruction loss and the diffusion repair loss. As shown in Figure 8 (a), the performance of our method is relatively stable when λ_{repair} is set between 0.5 to 2.0, with the best performance achieved at $\lambda_{\text{repair}} = 0.5$. When λ_{repair} is too large, the diffusion prior dominates the optimization, leading to oversmoothing and loss of structural details.

Reconstruction from Sparse Event Views. We further analyze the impact of the number of event views on reconstruction quality. To this end, we conduct an analysis study by reducing the number of event views and analyzing the impact on the reconstruction quality. We gradually reduce the number of event views to ratios of 1/1.5, 1/2, and 1/4 of the original event data. As shown in Figure 8 (b), our method maintains reasonable reconstruction quality even with 1/2 of the original event data, achieving a PSNR of 18.27 dB. However, as the number of event views decreases further, the reconstruction quality deteriorates.

Training Time and Rendering Speed. We further analyze the computational complexity of our two-stage framework in comparison to existing methods. Table 4 reports the training time and FPS. Due to the inclusion of the diffusion repair process, our method incurs longer training times than Event3DGS, while maintaining comparable inference speeds. E-3DGS also exhibits increased training time, primarily attributed to its additional pose optimization step.

	Event3DGS	E-3DGS	EvDiff3D (Ours)
Training Time	15 min	50min	40 min
Inference (FPS)	250	250	250

Table 4: Computational Complexity Analysis. All the results are evaluated on a single A40 GPU.

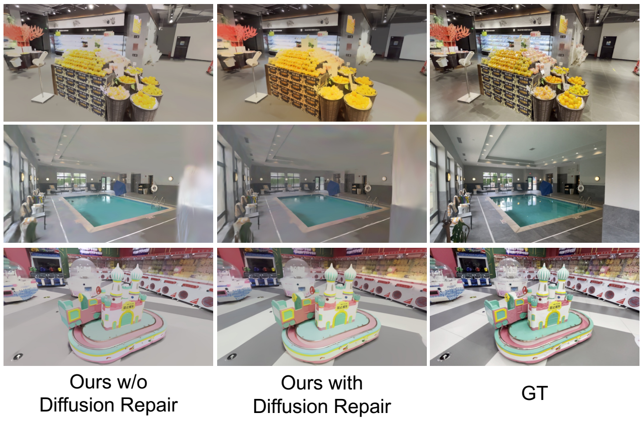

Figure 7: Ablation study on the repair loss.

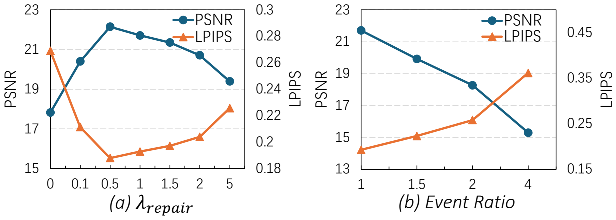

Figure 8: Ablation study on the (a) λ_{repair} and (b) ratio of event view. These two analysis studies were conducted on a single sequence in the Ev-DL3DV dataset.

Conclusion

We have presented EvDiff3D, a novel two-stage framework for high-fidelity 3D scene reconstruction using only event data. Our approach introduces a cyclical optimization pipeline that progressively refines both geometry and appearance by jointly leveraging event-based geometric constraints and a specialized event-aware diffusion prior. Extensive experiments demonstrate that EvDiff3D achieves substantial improvements over existing event-based methods. These results highlight the effectiveness of incorporating diffusion models for enhancing semantic content and repairing event-driven artifacts, underscoring the potential of event-only 3D reconstruction without reliance on images.

Acknowledgments

This work is supported by the National Research Foundation (NRF) Medium Sized Centre (MSC) - Centre for Advanced Robotics Technology Innovation (CARTIN) and the MOE AcRF Tier 1 (Call 2/2025) Grant under Grant No. RG160/25.

References

Alonso, I.; and Murillo, A. C. 2019. EV-SegNet: Semantic Segmentation for Event-Based Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1624–1633.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Cannici, M.; and Scaramuzza, D. 2024. Mitigating Motion Blur in Neural Radiance Fields with Events and Frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9286–9296.

Chamorro, W.; Sola, J.; and Andrade-Cetto, J. 2022. Event-Based Line SLAM in Real-Time. *IEEE Robotics and Automation Letters*, 7(3): 8146–8153.

Chen, K.; Li, H.; Zhou, J.; Wang, Z.; and Wang, L. 2024. LaSe-E2V: Towards Language-guided Semantic-aware Event-to-Video Reconstruction. *Advances in Neural Information Processing Systems*, 37: 70406–70430.

Chen, K.; Liang, G.; Lu, Y.; Li, H.; and Wang, L. 2025a. EvLight++: Low-Light Video Enhancement with an Event Camera: A Large-Scale Real-World Dataset, Novel Method, and More. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, K.; Wang, Z.; and Wang, L. 2025. ExFMan: Rendering 3D Dynamic Humans With Hybrid Monocular Blurry Frames and Events. *IEEE Robotics and Automation Letters*.

Chen, K.; Zhang, Z.; Liang, G.; Jiang, L.; Wang, Z.; and Chen, Y.-C. 2025b. Event-Guided Consistent Video Enhancement with Modality-Adaptive Diffusion Pipeline. In *Annual Conference on Neural Information Processing Systems*.

Dubeau, E.; Garon, M.; Debaque, B.; de Charette, R.; and Lalonde, J.-F. 2020. RGB-D-E: Event Camera Calibration for Fast 6-DOF object Tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, 127–135. IEEE.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 154–180.

Han, H.; Li, J.; Wei, H.; and Ji, X. 2024. Event-3DGS: Event-based 3D Reconstruction Using 3D Gaussian Splatting. *Advances in Neural Information Processing Systems*, 37: 128139–128159.

Hu, Y.; Liu, S.-C.; and Delbruck, T. 2021. v2e: From Video Frames to Realistic DVS Events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1312–1321.

Hwang, I.; Kim, J.; and Kim, Y. M. 2023. Ev-NeRF: Event Based Neural Radiance Field. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 837–847.

Jiang, Z.; Zhang, Y.; Zou, D.; Ren, J.; Lv, J.; and Liu, Y. 2020. Learning Event-Based Motion Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3320–3329.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-Scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Klenk, S.; Chui, J.; Demmel, N.; and Cremers, D. 2021. TUM-VIE: The TUM Stereo Visual-Inertial Event Dataset. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 8601–8608. IEEE.

Lee, S.; and Lee, G. H. 2025. DiET-GS: Diffusion Prior and Event Stream-Assisted Motion Deblurring 3D Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21739–21749.

Li, H.; Chu, X.; and Shi, D. 2024. DreamCouple: Exploring High Quality Text-to-3D Generation Via Rectified Flow. *arXiv preprint arXiv:2408.05008*.

Liang, G.; Chen, K.; Li, H.; Lu, Y.; and Wang, L. 2024. Towards Robust Event-guided Low-Light Image Enhancement: A Large-Scale Real-World Event-Image Dataset and Novel Approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23–33.

Ling, L.; Sheng, Y.; Tu, Z.; Zhao, W.; Xin, C.; Wan, K.; Yu, L.; Guo, Q.; Yu, Z.; Lu, Y.; et al. 2024. DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22160–22169.

Liu, X.; Zhou, C.; and Huang, S. 2024. 3DGS-Enhancer: Enhancing Unbounded 3D Gaussian Splatting with View-consistent 2D Diffusion Priors. *Advances in Neural Information Processing Systems*, 37: 133305–133327.

Low, W. F.; and Lee, G. H. 2024. Deblur e-NeRF: NeRF from Motion-Blurred Events under High-speed or Low-light Conditions. In *European Conference on Computer Vision*, 192–209. Springer.

Mahlknecht, F.; Gehrig, D.; Nash, J.; Rockenbauer, F. M.; Morrell, B.; Delaune, J.; and Scaramuzza, D. 2022. Exploring Event Camera-Based Odometry for Planetary Robots. *IEEE Robotics and Automation Letters*, 7(4): 8651–8658.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1): 99–106.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Morgenstern, W.; Gard, N.; Baumann, S.; Hilsmann, A.; and Eisert, P. 2023. X-maps: Direct Depth Lookup for Event-based Structured Light Systems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 4007–4015. IEEE Computer Society.

Mueggler, E.; Gallego, G.; Rebecq, H.; and Scaramuzza, D. 2018. Continuous-Time Visual-Inertial Odometry for Event Cameras. *IEEE Transactions on Robotics*, 34(6): 1425–1440.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*.

Qi, Y.; Zhu, L.; Zhang, Y.; and Li, J. 2023. E2NeRF: Event Enhanced Neural Radiance Fields from Blurry Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13254–13264.

Rudnev, V.; Elgharib, M.; Theobalt, C.; and Golyanik, V. 2023. EventNeRF: Neural Radiance Fields from a Single Colour Event Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4992–5002.

Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; and Chen, D. 2023. Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22819–22829.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *Advances in Neural Information Processing Systems*, 36.

Wu, R.; Mildenhall, B.; Henzler, P.; Park, K.; Gao, R.; Watson, D.; Srinivasan, P. P.; Verbin, D.; Barron, J. T.; Poole, B.; et al. 2024. ReconFusion: 3D Reconstruction with Diffusion Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21551–21561.

Wu, S.; Xu, C.; Huang, B.; Geiger, A.; and Chen, A. 2025. GenFusion: Closing the Loop between Reconstruction and Generation via Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6078–6088.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; and Wong, T.-T. 2024. DynamiCrafter: Animating Open-Domain Images with Video Diffusion Priors. In *European Conference on Computer Vision*, 399–417. Springer.

Xiong, T.; Wu, J.; He, B.; Fermuller, C.; Aloimonos, Y.; Huang, H.; and Metzler, C. 2025. Event3DGS: Event-Based 3D Gaussian Splatting for High-Speed Robot Egomotion. In *Conference on Robot Learning*, 4100–4118. PMLR.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, T.; Shen, Y.; Zhao, G.; Wang, L.; Chen, X.; Bai, L.; and Zhou, Y. 2024. Swift-Eye: Towards Anti-blink Pupil Tracking for Precise and Robust High-Frequency Near-Eye Movement Analysis with Event Cameras. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, Z.; Chen, K.; and Wang, L. 2025. Elite-EvGS: Learning Event-based 3D Gaussian Splatting by Distilling Event-to-Video Priors. In *IEEE International Conference on Robotics and Automation*, 13972–13978. IEEE.

Zhou, J.; Chen, K.; Zhang, L.; and Wang, L. 2025. PASS: Path-Selective State Space Model for Event-based Recognition. In *Annual Conference on Neural Information Processing Systems*.

Zhu, J.; Pan, T.; Cao, Z.; Liu, Y.; Kwok, J.; and Xiong, H. 2025. Depth Any Event Stream: Enhancing Event-based Monocular Depth Estimation via Dense-to-Sparse Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.