E-4DGS: High-Fidelity Dynamic Reconstruction from the Multi-view Event Cameras

Chaoran Feng[*]
Peking University
chaoran.feng@stu.pku.edu.cn

Zhenyu Tang[*]
Peking University
zhenyutang@stu.pku.edu.cn

Wangbo Yu
Peking University
wangboyu@gmail.com

Yatian Pang
National University of Singapore
yatian.pang@u.nus.edu

Yian Zhao
Peking University
zhaoyian@stu.pku.edu.cn

Jianbin Zhao
Dalian University of Technology
1518272584@mail.dlut.edu.cn

Li Yuan[†]
Peking University
yuanli-ece@pku.edu.cn

Yonghong Tian[†]
Peking University
yhtian@pku.edu.cn

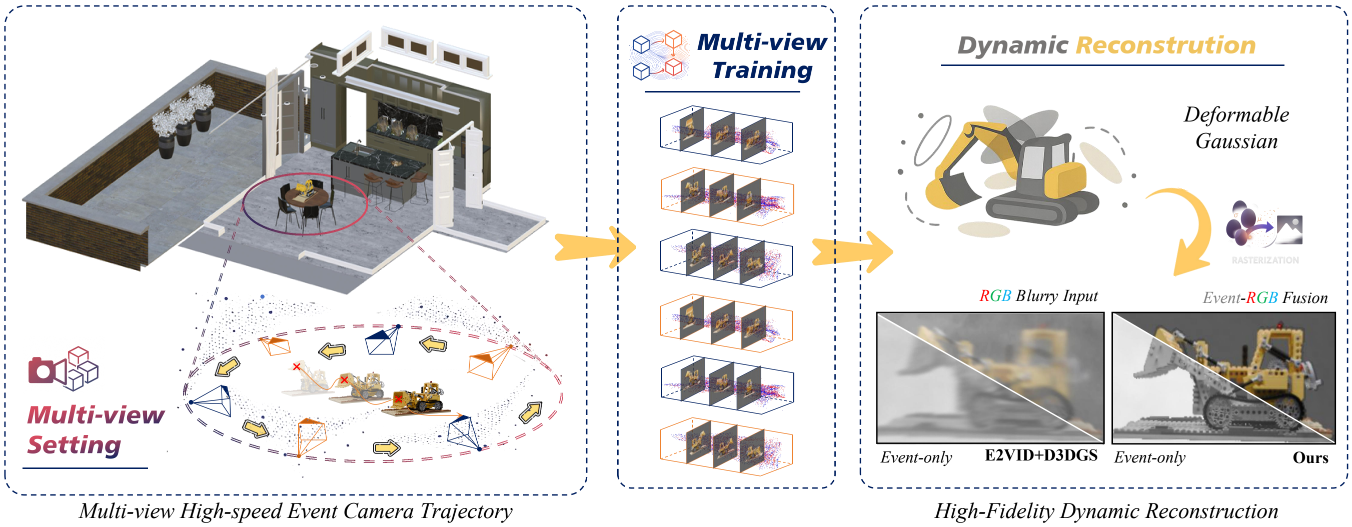


Figure 1: Our E-4DGS reconstructs temporally consistent and photorealistic dynamic scenes using event streams and sparse RGB frames captured from multi-view moving cameras, effectively handling complex motion and lighting variations.

Abstract

Novel view synthesis and 4D reconstruction techniques predominantly rely on RGB cameras, thereby inheriting inherent limitations such as the dependence on adequate lighting, susceptibility to motion blur, and a limited dynamic range. Event cameras, offering advantages of low power, high temporal resolution and high dynamic range, have brought a new perspective to addressing the scene reconstruction challenges in high-speed motion and low-light scenes. To this end, we propose *E-4DGS*, the first event-driven dynamic Gaussian Splatting approach, for novel view synthesis from multi-view event streams with fast-moving cameras. Specifically, we introduce an event-based initialization scheme to ensure stable training and propose event-adaptive slicing splatting for time-aware reconstruction. Additionally, we employ intensity importance pruning to eliminate floating artifacts and enhance 3D consistency, while incorporating an adaptive contrast threshold for more precise optimization. We design a synthetic multi-view camera setup with six moving event cameras surrounding the object in a 360-degree configuration and provide a benchmark multi-view event stream dataset that captures challenging motion scenarios. Our approach outperforms both event-only and event-RGB fusion baselines and paves the way for the exploration of multi-view event-based reconstruction as a novel approach for rapid scene capture.

CCS Concepts

• **Computing methodologies** → **Reconstruction**.

Keywords

Event-driven 4D Reconstruction, 3D Gaussian Splatting, Novel View Synthesis, High-speed Robot Egomotion.

1 Introduction

Novel view synthesis (NVS) and dynamic scene reconstruction are critical for immersive applications such as virtual and augmented

[*] Both authors contributed equally to this research.
[†] Correspongding Auther.

reality (VR/AR) [35, 49, 93], scene understanding [6, 27, 36, 86], 3D content creation [8, 37, 51, 70, 100], and autonomous driving tasks [17, 50, 87, 95]. While Neural Radiance Fields (NeRF) [47] has recently achieved remarkable success in photorealistic rendering of static scenes, their extension to dynamic scenarios remains challenging—primarily due to substantial training time. In contrast, 3D Gaussian Splatting (3DGS) [29] provides notable advantages in real-time rendering and significantly faster training. Yet, existing dynamic extensions of 3DGS struggle to handle scenes with fast motion effectively, primarily due to the inherent limitations of RGB cameras, which, owing to their high latency and limited dynamic range, are prone to motion blur when capturing fast-moving scenes.

Compared to RGB cameras that capture images at fixed intervals, event cameras operate asynchronously by recording brightness changes as event spikes with microsecond-level latency, offering extremely low latency and high dynamic range [13, 65, 73] Owing to such advantageous, event cameras have recently been adopted for novel view synthesis and scene reconstruction tasks [84]. For example, event-driven NeRF methods [26, 33, 60] leverage event accumulation frames and depend on known or estimated camera trajectories to reconstruct NeRF representation. In parallel, event-driven 3DGS approaches [19, 25, 79, 92] utilize the sharp structural information provided by event streams to reconstruct 3DGS representation, enabling efficient rendering and training. However, these methods are primarily designed for static scene reconstruction and are not well-suited for modeling dynamic environments. In the more challenging task of dynamic scene reconstruction, relying solely on a single event camera inherently limits the ability to capture complete scene dynamics—especially in scenarios involving fast motion, large deformations, or severe occlusions. Moreover, the coupling between object and camera motion can often lead to mutual cancellation of contrast changes, resulting in neutralized events [10, 14, 19] that obscure fine-grained geometric details.

Based on the above observation, we aim to investigate the following research question: *How can we efficiently reconstruct a high-fidelity dynamic scenes using multi-view fast-moving event cameras?* With the captured multi-view event streams, a straightforward approach is to adopt a two-stage pipeline: first, reconstructing intensity frames from the event streams using E2VID [10, 59] and obtain Gaussian initialization points from COLMAP [63] ; then, applying an off-the-shelf reconstruction method for futher reconstruction [9, 88, 90, 91]. However, this naïve solution compromises the temporal precision and sparsity of event data by converting it into intensity frames, introducing accumulation error and extensive costs, resulting in degraded reconstruction consistency.

To this end, we propose *E-4DGS*, an end-to-end event-based framework for high-fidelity dynamic 3D reconstruction from multi-view event streams. To address the initialization challenge under sparse event observations, we introduce an event-specific strategy to generate stable Gaussian primitives without relying on RGB-based Structure from Motion (SfM). We further design an event-adaptive slicing mechanism that segments and accumulates event streams for accurate supervision, and propose a multi-view 3D consistency regularization to enhance structural alignment. Additionally, *E-4DGS* supports optional refinement using a few motion-blurred RGB frames. To our knowledge, this is the first event-only framework enabling view-consistent 3D Gaussian reconstruction in dynamic scenes. For evaluation, we introduce a multi-view synthetic event dataset that serves as a benchmark for dynamic scene reconstruction. The dataset encompasses a diverse set of dynamic scenes with simultaneous camera and object motion, ranging from "mild" to "strong". We compare our method against two-stage baselines that utilize E2VID for intensity reconstruction followed by frame-based methods, trained either with event streams alone or with a combination of RGB videos and event sequences. Our approach significantly outperforms all baselines, achieving state-of-the-art results while enabling continuous and temporally coherent reconstruction of dynamic scenes. These results demonstrate that operating directly on raw event data, especially under challenging conditions with camera motion, yields higher-fidelity dynamic scene reconstruction compared to methods relying on reconstructed RGB frames. To summarize, the main contributions are as follows:

- We present *E-4DGS*, the event-driven approach for reconstructing adynamic 3D Gaussian Splatting representation from multi-view event streams.
- We introduce an event-based initialization scheme for stable training, propose event-adaptive slicing splatting and adaptive event threshold for supervision, and design intensity importance pruning to enhance 3D consistency.
- We construct a multi-view synthetic dataset with moving cameras for 4D reconstruction from event streams. Our method achieves state-of-the-art performance, and we will release our work to support future research.

2 Related Work

2.1 Dynamic Reconstruction from RGB Frames

Modeling dynamic scenes from moving RGB cameras alone is still a challenging open task in computer vision. A widely used approach to this problem is to learn coordinate-based neural scene representations allowing rendering novel views and representing dynamic scenes. Previous works such as neural radiance field (NeRF) and its variants D-NeRF [55], KFD-NeRF [98] and more [3, 34, 39, 52, 53, 89] used implicit neural representations in combination with volume rendering. They are based on Multi-Layer Perceptrons (MLPs), which are relatively compact and require minimal storage space once trained. However, they are expensive to optimize and lead to slow training and evaluation which limits its expansion on the real-time rendering and real-world applications. The recently emerging 3D Gaussian Splatting (3DGS) [29] and its variants [5, 18, 69, 96] have reshaped the landscape of dynamic radiance fields due to its efficiency and flexibility. The pioneering work Deformable3DGS (D3DGS) [90] enhances dynamic Gaussian representations with a tiny deformable field for tracking the motion of Gaussian points. Similarly, other methods [21, 43, 64, 72, 76, 78, 91] models Gaussian motion using point-tracking functions for stable point moving. Our approach adopts D3DGS as the dynamic representation due to its simple and efficient structure, and then presents its application to the supervision from event streams. It inherits thereby the advantages of event streams and 3DGS for dynamic view synthesis.

2.2 Dynamic Reconstruction from Event Data

Event cameras have been widely used to reconstruct dynamic scenes from non-blurry RGB Frames of fast motion. Previous works,

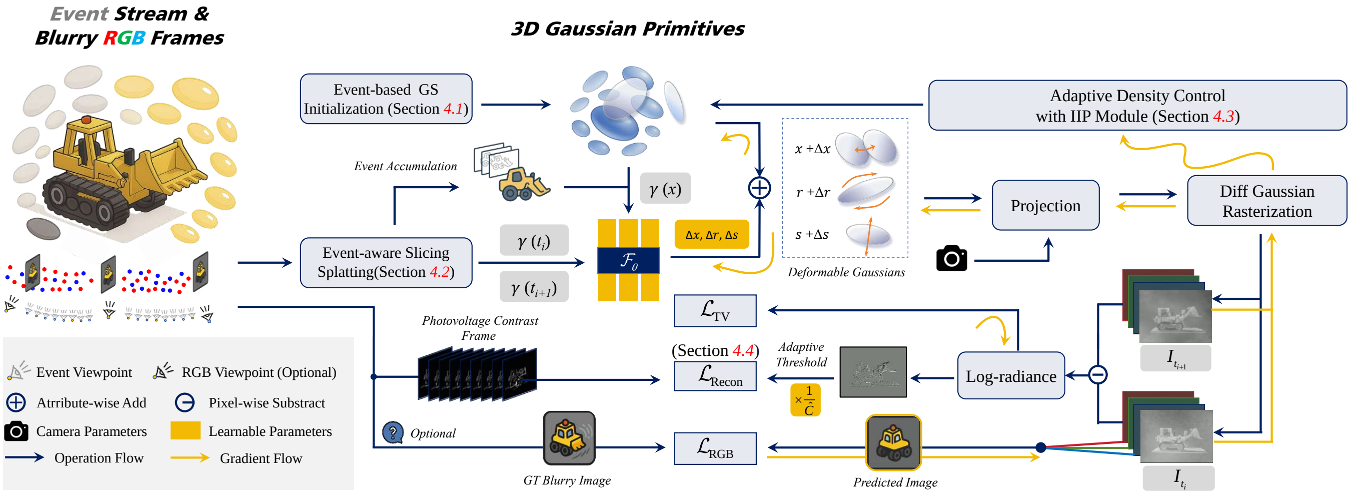


Figure 2: The overview of our method E-4DGS.Our E-4DGS framework establishes temporal-coherent 4D representations through a cascaded processing of event streams: The event-driven initialization (§4.1) constructs spatio-temporal Gaussians via polarity-encoded density fields, followed by differentiable feature distillation (§4.2) where adaptive slicing operators disentangle high-frequency patterns for splatting-based optimization. Cross-view consistency (§4.3) is then imposed through deformable Gaussian reprojection coupled with photometric saliency pruning, while multi-modal alignment (§4.4) ultimately achieves photometric fidelity via kernel-attentive RGB-event synchronization.

including model-based methods [48, 60] and learning-based methods [22, 59, 71], process event and RGB frames with 2D priors but lack 3D consistency. Other event-based methods address tasks such as detection, tracking, and image/3D reconstruction, including lip reading [62, 67], object tracking [4, 82, 105, 107], and pose estimation [16, 106]. However, these methods still do not incorporate 3D priors to reconstruct scene appearance and are not applicable to represent 3D scenes, which is our goal of proposed *E-4DGS*.

For static scene reconstruction, recent event-based methods [2, 25, 30, 38, 42, 56, 57, 68, 75, 79, 80, 83, 92, 94, 94, 102, 104] have achieved high-fidelity 3D reconstruction and novel-view synthesis (NVS) tasks using supervision from event pixels or event accumulation. These methods primarily rely on consistent event sequences from a single mono-event camera. However, extending static scene representations to dynamic scenes with event streams is a challenging task, as the movement of objects and the simultaneous motion of the event camera can introduce ambiguity in the events. Different from only a single mono-camera setting, our proposed E-4DGS reconstructs the dynamic scene with the multi-view camera setting, providing more multi-view consistency details.

Recently, a growing trend is the use of dynamic neural radiance fields (DNeRF) or Dynamic 3DGS (4DGS) for dynamic scene representation and novel view synthesis. DE-NeRF [44] and EBGS [85] reconstruct dynamic scenes using monocular event streams and RGB frames from a moving camera, modeling deformations in a canonical space. The former is based on DNeRF, while the latter relies on 4DGS. EvDNeRF [1], which is based on canonical volumes, and DynEventNeRF [61], which uses temporally-conditioned MLP-based NeRF, both utilize multi-view event streams to reconstruct dynamic scenes. However, the former does not model appearance, and the latter is trained slowly due to volume rendering. In contrast, our proposed E-4DGS achieves higher-quality reconstruction by accurately capturing complex geometries and lighting effects than NeRF-based models, while also offering fast training and inference speeds for real-time, real-world applications.

3 Preliminaries

3.1 Deformable 3D Gaussian Splatting

Deformable3DGS [90] offers an explicit method for representing a 4D dynamic scene $\mathbb{G}$ with the canonical space and th deformable space based on 3D Gaussian Splatting [29]. In the canonical space, these 3D Gaussian points have the following parameters: mean point μ, covariance matrix Σ, opacity σ, and color $\mathbf{c}$ and a 3D Gaussian point $G(x) \in \mathbb{G}$ is defined as follows:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{1}$$

where, Σ is divided into two learnable components: the quaternion r represents rotation, and the 3D-vector s represents scaling. then, the color of each pixel can be calculated using the following formula:

$$C(x) = \sum_{i\in N(x)} c_i\alpha_i(x)\prod_{j=1}^{i-1}\left(1-\alpha_j(x)\right), \tag{2}$$

where $\alpha_i(x) = \sigma_i \exp\left(-\frac{1}{2}(x-\mu_i^{2D})^T\Sigma^{-1}(x-\mu_i^{2D})\right)$, and N is the number of Gaussian points that intersect with the pixel x.

In the deformable space, Deformable3DGS employ a compact MLP layer to represent motion of Gaussian points. Given timestamp t and center position x of 3D Gaussians as inputs, the deformation MLP produces offsets, which subsequently transform the canonical 3D Gaussians to the deformed space:

$$(\Delta x, \Delta r, \Delta s) = \mathcal{F}_\theta(\gamma(\text{sg}(x))), \gamma(t)) \tag{3}$$

where $sg(\cdot)$ indicates a stop-gradient operation, γ denotes the positional encoding as defined in [90]. Therefore, a dynamic Gaussian point can be represented as $G(x + \Delta x, r + \Delta r, s + \Delta s)$ at timestamp t.

3.2 Event Generation Model

A single event is represented as $e_k = (x_k, y_k, p_k, t_k)$ in the event streams $\mathcal{E}$, denoting a brightness change registered by an event sensor at timestamp t_k, pixel location $\mathbf{u_k} = (x_k, y_k)$ in the event camera frame with polarity $p_k \in \{-1, +1\}$. The change between adjacent timestamps can be calculated from intensity images I.

$$L(\mathbf{u}_k, t_k) - L(\mathbf{u}_k, t_{k-1}) = \sum_{t_{k-1} < t \leq t_k} p_t C^{p_t} \overset{\text{def}}{=} \Delta E_{\mathbf{u}_k}(t_{k-1}, t_k), \quad (4)$$

$$\text{where} \quad L = \log(I). \quad (5)$$

Here, the thresholds $C^p \in \{C^{-1}, C^{+1}\}$ define boundaries for classifying the event as positive or negative, with the polarity of an event indicating a positive or negative change in logarithmic illumination.

Therefore, given a supervisory event stream $\mathcal{E}$, we can supervise our proposed *E-4DGS* by comparing the predicted brightness change $\Delta\hat{E}(t_{k-1}, t_k)$ and the ground truth $\Delta E(t_{k-1}, t_k)$ by Equation (4) over all image pixels. In general, we substitute intensity frames $\hat{I}_t$ with the rendered results $\hat{C}_t$ and can utilize photo-realistic loss [90] between the predicted intensity frames and the ground-truth event of event-based single integral (ESI) [60]:

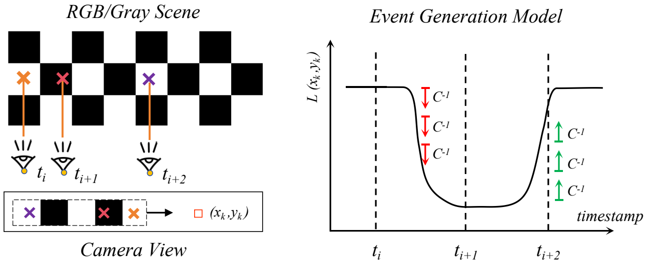
$$\mathcal{L}_{gs} = \sum_{\mathbf{u}_k \in \hat{I}} (\lambda \mathcal{L}_1(\Delta\hat{E}_{\mathbf{u}_k}, \Delta E_{\mathbf{u}_k}) + (1 - \lambda)\mathcal{L}_{D-SSIM}(\Delta\hat{E}_{\mathbf{u}_k}, \Delta E_{\mathbf{u}_k})) \quad (6)$$

4 Method

We propose **E-4DGS**, a method for high-fidelity dynamic scene reconstruction using sparse event camera streams. Given multi-view event data capturing a dynamic scene, E-4DGS reconstructs a 4D model that allows novel view generation at arbitrary times. To address the challenges posed by the sparse nature of event data and the dynamic characteristics of the scene, we introduce an event-based initialization strategy (Section 4.1), an event-aware slicing splatting technique to preserve geometric details (Section 4.2), and multi-view 3D consistency regularization for improved scene fidelity (Section 4.3). Additionally, we utilize adaptive event supervision and color recovery to enhance the reconstruction quality (Section 4.4). The overview of our method is illustrated in Figure 2.

4.1 Event-based Initialization

The Gaussian primitives are initialized using a point cloud derived from Structure-from-Motion (SfM) [45] with RGB frames in the vanilla 3DGS. However, their performance is hindered by inaccurate dynamic Gaussian initialization due to view inconsistencies caused by object motion. Furthermore, applying SfM to extract Gaussian points from event sequences is more challenging than using RGB frames with COLMAP [63], due to the sparse nature of event streams. Some methods [19, 25, 79, 83] randomly initialize Gaussians within a fixed cube without considering unbounded scenes. Other methods perform better than random initialization but are more complex. Elite-3DGS [103] employs a two-stage approach with E2VID [59] to convert events into images, followed by SfM for point cloud initialization, while E-3DGS [92] uses exposure enhancement [66] tricks before obtaining the SfM points.



(a) Visualization of a sharp edge passing a pixel in an event sensor.

(b) Photocurrent or intensity change at the pixel (x_k, y_k).

Figure 3: Demonstration of how time window discretizations can influence the count of events between timestep pairs. The time window (t_i, t_{i+1}) produces two negative events, whereas (t_i, t_{i+2}) produces no events.

Thus, we adopt an event-specific strategy for Gaussian point initialization, balancing performance and efficiency. Specifically, 1) For object scenes, we initialize the point cloud with 100,000 points in a fixed cube, consistent with original 3DGS settings; 2) For medium or large scenes, we employ a dense-to-sparse radiative sphere initialization, mimicking realistic distribution where point density is highest at the center and decreases toward the boundaries. We set sphere's radius to $r = 10.0$ with 200,000 initial points.

We also experimented with initializing the Gaussian primitives using random pointcloud and E2VID+COLMAP, and further details are provided in the supplementary materials. While our approach yielded a slight performance drop than the E2VID+COLMAP's performance, the latter requires more computational complexity.

4.2 Event-adaptive Slicing Splatting

In event-based scene reconstruction pipelines, the slicing strategy for the event stream significantly influences reconstruction quality. As the duration of the event time window (t_i, t_{i+1}) increases, the predicted events become a discretized, aliased representation of the continuous brightness variations in the scene.

For instance, Figure 3 illustrates that measurements recorded by the event sensor between timestamps (t_i, t_{i+1}) produce three negative events at the selected pixel, whereas measurements over the interval (t_i, t_{i+2}) yield no events. This effect is particularly notable in our pipeline, as the process of accumulating polarity inherently neutralizes events. Moreover, existing works [60, 83] have demonstrated that using consistently short windows impedes the propagation of high-level illumination, while consistently long windows often result in a loss of local detail. While they randomly sampled the length of event timestamp window, a drawback is that it does not take into account the camera speed or event rate, causing the sampled windows to contain either too many or too few events. Additionally, Hu et al. [24] and Han et al. [20] revealed that regions with uniform and smooth intensities typically do not trigger any events, leading to spatial sparsity in the event streams used as supervisory signals.

Based on the aforementioned observations, we propose an event-adaptive slicing strategy to address this issue. Specifically, during the training of our *E-4DGS*, we deliberately vary the time window of batched events and incorporate event noise during the event accumulation process. Notably, these modifications lead to an improved generation of finely-sliced events at test time. The detailed process of event-adaptive slicing are as follow:

1) *Event Accumulation Range Setting:* For each timestamp, we randomly sample and slice a target number of events streams within the event count range $[N_{min}, N_{max}]$.

2) *Event Accumulation Jitter:* During our sampling process, we add Gaussian noise to pixels that do not record any events within the whole event timestamp window. This augmentation enhances gradient optimization in smooth regions and increases the overall robustness of the pipeline against noisy events. It serves the same purpose as Event Sampling in [44], and the whole process is defined as follows:

$$\Delta E_{\mathbf{u}}(t_{start}, t_{end}) = \begin{cases} \int_{t_s}^{t_e} p_\tau C^{p_\tau}, d\tau & \text{if } \mathbb{1}_{\text{trig}} \neq 0, \\ \Delta \cdot \mathcal{N}\left(0, \sigma_{\text{noise}}^2\right) & \text{if } \mathbb{1}_{\text{trig}} = 0. \end{cases} \quad (7)$$

where, $\Delta E(\cdot)$ denotes the event frame accumulated from all event polarities triggered at pixel coordinate $\mathbf{u}$ within the current event time window. $\mathbb{1}_{\text{trig}}$ denotes the spiking of the events. t_{start}, t_{end}, and $\Delta t = t_{\text{end}} - t_{\text{start}}$ represent the start timestamp, end timestamp, and the time interval of the event time window, respectively.

This strategy not only guarantees a diverse range of event window lengths, but also curtails the loss of fine details that can occur due to neutralization. Moreover, it helps preserve critical geomerty details, thereby enhancing the overall fidelity of the reconstruction.

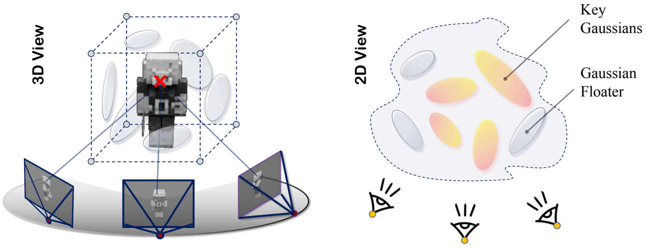
4.3 Intensity Importance Pruning

In the vanilla Gaussian Splatting pipeline, the opacity of all Gaussian points is gradually reduced, and points with low transparency are pruned during the Gaussian pruning stage. However, this method is unsuitable for our event-based approach, as it results in excessive coupling between the canonical and deformation fields and simultaneous camera and object motion, further exacerbating the issue. Therefore, we eliminate the reset opacity operation same as in [11]. and drawing inspiration from LightGaussians [12], which emphasizes a compact representation of static scenes by pruning redundant Gaussians based on spatial attributes such as transparency and volume, we adopt a specialized strategy, *Intensity Importance Pruning* (IIP), to remove floaters across both the canonical and deformable spaces. With this strategy, the importance of each Gaussian point is computed for each training viewpoint at every timestamp. Gaussian primitives with an importance score below a fixed threshold are then pruned, effectively mitigating the floater issue and enhance the 3D consistency from multi-view event streams.

Specifically, for a Gaussian point $g_i \in \mathbb{G}$, the Gaussian importance w_i over the images $\mathcal{I}$ of all training views and timestamps $\mathcal{T}$, is defined as follows:

$$w_i = \underset{\mathbf{x}\in I,\, t\in\mathcal{T}}{\mathrm{Max}} \left(\alpha_i(\mathbf{x} \mid t) \prod_{j=1}^{i-1} \left(1 - \alpha_j(\mathbf{x} \mid t) \right) \right). \quad (8)$$

Here, $I \in \mathcal{I}$ denotes the intensity image. We prune Gaussian points whose importance scores satisfy $w_i < 0.015$, following the approach [12]. As shown in Figure 4, our method effectively removes floating artifacts absent from the training views. In addition, we perform Gaussian cloning and splitting following the 3DGS protocol, ensuring that child Gaussian points inherit the dynamic characteristics of their parent Gaussian points.



(a) Importance Computation from Intensity Images

(b) Intensity Importance Pruning

Figure 4: The process of the intensity importance pruning.

4.4 Event Supervision and Optimization

Adaptive Event Supervision. According to previous work [40], the ground truth of the event contrast inherently contains some errors. Additionally, in the real scenes captured by an event sensor, the event constrast threshold C^p varies due to the environment disturbance, which can make Equation. 3.2 impractical to use in a real-world setup. Thus, if we directly apply the photometric loss with Equation. 6 to compare the rendered intensity frames with those derived from event data, the inherent discrepancies will be strictly penalized during optimization, which may in fact degrade the overall reconstruction quality. To bridge the gap between synthetic and real event data, we introduce learnable threshold parameters $\hat{C}$ and compute the rendered intensity frame as follows:

$$\Delta \hat{E}_{\mathbf{u}_k}(t_{k-1}, t_k) = \hat{L}(\mathbf{u}_k, t_k) - \hat{L}(\mathbf{u}_k, t_{k-1}) \overset{\text{def}}{\simeq} \sum_{t_{k-1} < t \leq t_k} p_t \hat{C}, \quad (9)$$

Here, we can simpfy this process as follows:

$$N_{gt}(\cdot)_{t_{k-1}}^{t_k} = \frac{1}{C}\left(\left(V(\cdot, t_2) - V(\cdot, t_1) \right) \right), \quad (10)$$

$$N_{pred}(\cdot)_{t_{k-1}}^{t_k} = \frac{1}{\hat{C}}\left(\left(\hat{L}(\cdot, t_k) - \hat{L}(\cdot, t_{k-1}) \right) \right), \quad (11)$$

$$\mathcal{L}_{Recon} = \frac{1}{H \times W} \sum_{\mathbf{u} \in \hat{L}} \sqrt{\left(N_{gt}(\mathbf{u}) - N_{pred}(\mathbf{u}) \right)^2 + \epsilon^2} \quad (12)$$

Here, $V(\mathbf{u}, t)$ denotes the photovoltage in event pixel $\mathbf{u}$ at timestamp t and ϵ is a small constant added for numerical stability.

The overall event supervision loss is given by:

$$\mathcal{L}_{\text{Event}} = \lambda_{\text{Recon}} \mathcal{L}_{\text{Recon}} + \lambda_{\text{TV}} \mathcal{L}_{\text{TV}}, \quad (13)$$

where $\mathcal{L}_{\text{TV}}$ is a total variation regularization term encouraging spatial smoothness, and λ_{Recon}, λ_{TV} are weighting factors balancing the contributions of each component.

Combined Gain and Offset Correction. Since event cameras only capture logarithmic intensity differences rather than absolute log-intensity values, the predicted log-intensity $\hat{L}$ from our 4DGS method is determined only up to an additive offset for each color channel. Moreover, there is a scale ambiguity in the reconstructed color balance and illumination of the scene, when only the event contrast threshold is known. Thus, it's necessary to correct and align the color value for every color channel like previous works [42, 60, 97], using the correction formula as follow:

$$\hat{L}(\mathbf{u}_k, t_k) \overset{\text{def}}{=} g_c \cdot \hat{L}(\mathbf{u}_k, t_k) + \Delta c, \tag{14}$$

where, g_c and Δc are the color correction parameters, and derived via ordinary least squares [42] with the ground-truth log-intensity $L(\mathbf{u}_k, t_k)$ as defined in Section 3.2. Notably, the images captured by a separate standard camera are affected by saturation in real-world scenes due to its limited dynamic range, and they are not raw recordings but have undergone lossy in-camera image processing. Moreover, the contrast threshold of real event cameras varies spatially across the image plane and temporally over time [24], making accurate color correction challenging and potentially leading to misalignment in the synthesized views of real scenes.

5 Experiments

5.1 Experimental Setting

5.1.1 Implementation Details. (1) Training Assumption: To reconstruct dynamic scenes using Gaussian Splatting [29] from high-speed, multi-view event cameras, we assume that our method leverages accurate camera intrinsics and high-quality, frequency-consistent extrinsics to enable precise interpolation at arbitrary timestamps. Specifically, we apply linear interpolation for camera positions and spherical linear interpolation (SLERP) for camera rotations For the synthetic event dataset, we adopt the original contrast thresholds C^{+1} and C^{-1} from the v2e simulation settings [24]. In the real-world autonomous driving dataset, we initialize the contrast thresholds using expected values of the event camera settings. This prior assumption provides a stable starting point, leading to more consistent training and improved 3D reconstruction performance.

(2) Training Details: We implemented E-4DGS based on the official code of Deformable3DGS [90], Gaussianflow [41], E-NeRF [33] and Event3DGS [19, 83] with Pytorch and conduct all experiments on a single NVIDIA RTX 4090 GPU. During training, we render at a resolution of 346 × 260 for the synthetic dataset and retain the original resolution 640 × 480 for real-scene data. Events are accumulated into frames using our adaptive slicing strategy (Section 4.2), where the number of events per temporal window is randomly sampled from a predefined range to introduce temporal diversity and enhance robustness. Specifically, we set $[N_{\min}, N_{\max}] = [5 \times 10^3, 10^4]$ for object-level scenes and $[10^5, 10^6]$ for indoor or large-scale scenes. Additionally, Gaussian noise ($\sigma_{\text{noise}} = 0.02$) is injected into event-void pixels during accumulation to improve optimization in textureless regions. Each scene is trained for 50,000 iterations using the Adam optimizer. The overall loss consists of an event-based supervision loss, a total variation regularization term and a RGB reconstruction loss (opt.), weighted by $\lambda_{\text{Recon}} = 1.0$, and $\lambda_{\text{TV}} = 0.005$, $\lambda_{\text{RGB}} = 1.0$, respectively. The stabilization constant ϵ in $\mathcal{L}_{\text{Recon}}$ is set to 0.001. The learnable event contrast threshold $\hat{C}$ is initialized to 0.15 for synthetic scenes and 0.2 for real-scene scenes, and is jointly optimized during training. To prevent interference with dynamic scene modeling, opacity reset is disabled like in [11] and color correction is applied only at inference time in all scenes.

5.1.2 Evaluation Metrics. For synthetic and real-scene datasets, we employ the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [77], and VGG-based Learned Perceptual Image Patch Similarity (LPIPS) [101] to evaluate the similarity between rendered novel views and ground-truth novel views.

5.1.3 Baselines. At the time of writing, the event-based dynamic reconstruction methods Dynamic EventNeRF [61] and EBGS [85] have not been publicly released. Although EvDNeRF [1] is open-sourced, it focuses solely on modeling geometric edges rather than performing holistic scene reconstruction. Consequently, we compare our proposed method against RGB-based baselines that do not utilize event data and are trained either on blurry RGB recordings or on RGB videos reconstructed from events using E2VID [59]. We choose Deformable3DGS [90] and Deblur4DGS [81] as the RGB-based baseline with blurry RGB inputs or event-integral inputs.

5.2 Experimental Evaluation

5.2.1 Synthetic dataset. To generate synthetic data, we render 8 dynamic scenes in Blender [7] at 3000 FPS from six moving viewpoints uniformly distributed around the object at the same height. The rendered sequences are then processed by the event simulator v2e [24] to produce corresponding event streams.

(a) Novel View Synthesis: As demonstrated in Table 1, our proposed *E-4DGS* outperforms the baselines E2VID + D3DGS across all synthetic scenes in all metrics. This result is intuitive, as E2VID benefits from being trained on a large dataset but does not account for 3D consistency, whereas our method explicitly incorporates it. Moreover, EvDNeRF only models the edge of a single object and does not capture the appearance of the dynamic scene, leading to inferior performance compared to the two-stage method and our proposed *E-4DGS*. The qualitative comparison of novel view synthesis in Figure 5 shows that our method produces reconstructed scenes with fewer floaters and more photorealistic rendering results.

(b) Motion Blur Decoupling: Using event sequences for deblurring blurry RGB frames is a common task. In our experiments, we simulate blurry images using Blender [7] by integrating images over the exposure time using LERP and SLERP, which yields realistic, motion-dependent blur. Table 1 show that our method perform better than all 4D reconstruction baselines. The results of our proposed method are better than the two-stage method which is combining E2VID with frame-based D3DGS. Furthermore, our method outperforms the frame-based 4D deblurring baseline [81][1], demonstrating that inherent blur-resistant characteristics of events offer greater advantages than relying solely on blur formation.

(c) Dynamic Reconstruction with Event and Frame Fusion: We combine event sequences and blurry frames by an event-RGB weighted combination, caculated as follows:

$$\mathcal{L}_{Fusion} = \mathcal{L}_{Event} + \lambda_{RGB}\mathcal{L}_{RGB} \tag{15}$$

Here, $\mathcal{L}_{RGB}$ is the original photo-realistic rendering loss of D3DGS [90] with the L1 and D-SSIM loss terms. Due to the discrete nature of events, although event sequences capture sharp edges, they remain noisy in low-light or uniform areas, which results in fog-like artifacts in dynamic scenes. Moreover, the color in a

[1] This work need to motion masks and frames as inputs. Thus, we utilize MonST3R [99] to extract motion masks and train the whole scenes as the original setting.

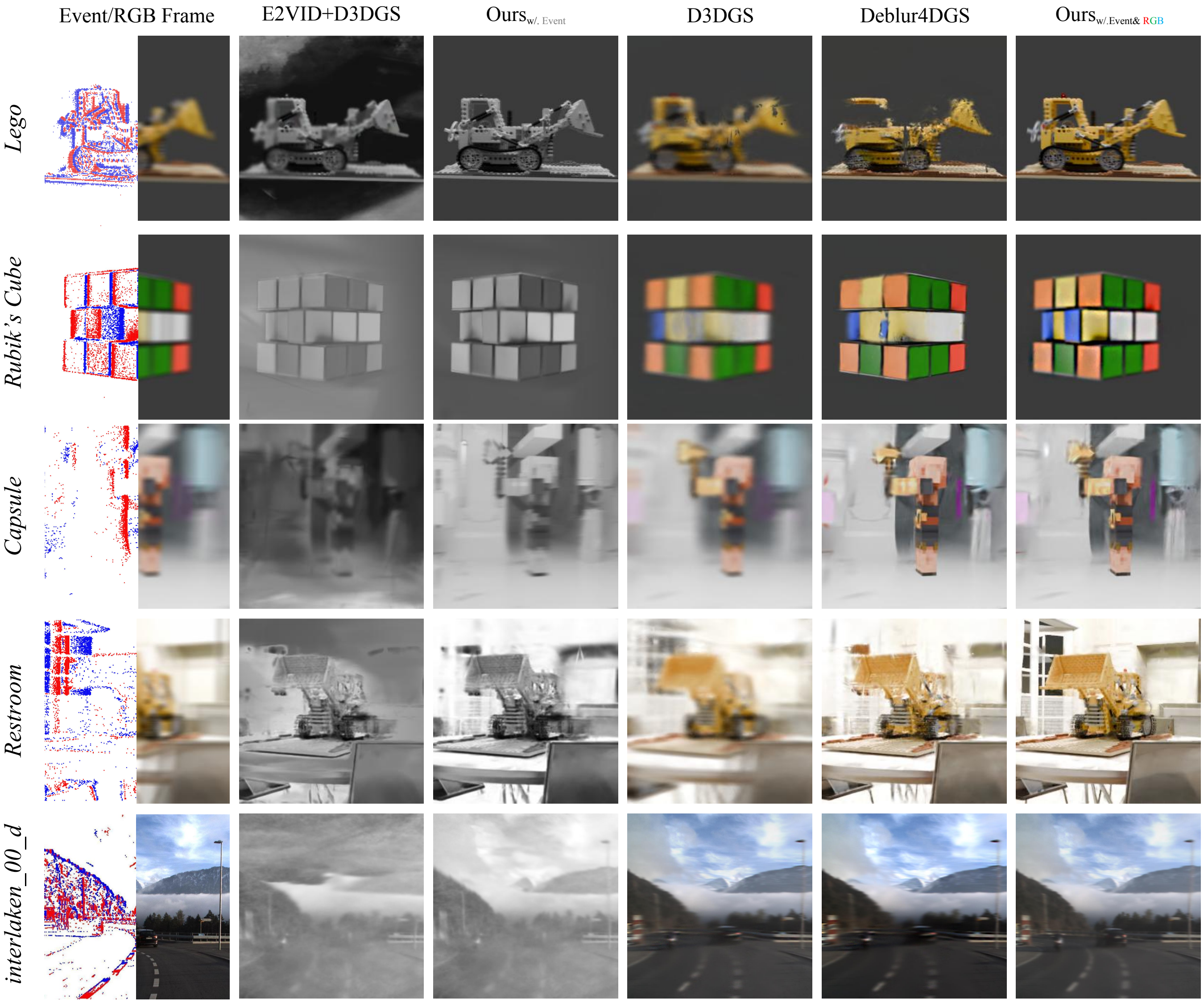


Figure 5: Qualitative results of novel view synthesis. Compared with 4D reconstruction-based methods [81, 90], our approach produces more realistic rendering results with fine-grained details in the synthetic and real scenes.

Table 1: Quantitative comparison of different methods for novel view synthesis from event streams. The best and second-best results are highlighted in bold and underlined, respectively. The average value is computed across 8 synthetic scenes.

Method	Lego			Rubik's Cube			Capsule			Restroom			Average		
	↑PSNR	↑SSIM	↓LPIPS	↑PSNR	↑SSIM	↓LPIPS	↑PSNR	↑SSIM	↓LPIPS	↑PSNR	↑SSIM	↓LPIPS	↑PSNR	↑SSIM	↓LPIPS
D3DGS$_{w/o\ blur}$	26.47	0.910	0.098	20.30	0.868	0.207	31.23	0.956	0.077	28.05	0.935	0.074	23.81	0.861	0.173
D3DGS$_{w/\ blur}$	23.62	0.821	0.250	18.12	0.804	0.351	27.51	0.905	0.181	26.46	0.908	0.160	21.73	0.797	0.296
E2VID + D3DGS	20.57	0.765	0.347	16.16	0.752	0.404	26.06	0.851	0.268	24.87	0.856	0.247	19.88	0.728	0.397
Deblur4DGS	23.17	0.813	0.265	17.68	0.786	0.375	28.06	0.908	0.176	26.35	0.900	0.162	21.66	0.797	0.291
E-4DGS$_{event\text{-}only}$	26.85	0.912	0.084	20.97	0.882	0.185	31.85	0.959	0.071	28.83	0.942	0.069	25.38	0.896	0.134
E-4DGS$_{event\&\ RGB}$	**27.23**	**0.925**	**0.078**	**21.23**	**0.895**	**0.172**	**32.41**	**0.963**	**0.068**	**29.02**	**0.949**	**0.067**	**25.62**	**0.903**	**0.129**

Table 2: Ablation study of each component.

Method Components					Synthetic Datasets		
$\mathcal{L}_{recon}$	$\mathcal{L}_{tv}$	ESS	AES	IIP	PSNR↑	SSIM↑	LPIPS↓
✓	✓	✓	✓	✓	**25.38**	**0.896**	**0.134**
✓	–	–	–	–	23.68	0.858	0.178
✓	✓	–	–	–	23.89	0.865	0.171
✓	✓	✓	–	–	24.71	0.876	0.153
✓	✓	–	✓	–	23.97	0.863	0.172
✓	✓	–	–	✓	25.13	0.881	0.142

Table 3: Ablation study on the robustness of deblurring. The best and second results are bold and underlined, respectively.

Blur Degree Metrics	Mild blur PSNR↑/SSIM↑/LPIPS↓	Medium blur PSNR↑/SSIM↑/LPIPS↓	Strong blur PSNR↑/SSIM↑/LPIPS↓
D3DGS	19.05 / 0.62 / 0.41	18.89 / 0.61 / 0.41	16.98 / 0.52 / 0.57
E2VID+D3DGS	17.79 / 0.55 / 0.49	17.69 / 0.54 / 0.49	17.93 / 0.55 / 0.49
Deblur4DGS	19.23 / 0.64 / 0.38	18.92 / 0.61 / 0.42	16.66 / 0.50 / 0.59
E-4DGS$_{event\text{-}only}$	**24.81** / **0.87** / **0.17**	**24.32** / **0.86** / **0.19**	**21.59** / **0.76** / **0.28**
E-4DGS$_{event\&\ RGB}$	**24.95** / **0.88** / **0.17**	**24.78** / **0.87** / **0.17**	**22.06** / **0.80** / **0.26**

shaded area might be slightly off and require correction [42, 60, 61], as it is not directly measured but inferred from derivative-like data. However, incorporating RGB frames helps address these issues by preserving the low-frequency and texture details from the frames while retaining the sharp, high-frequency features from the event sequences. As shown in Figure 5 , our method reconstructs a sharp dynamic scene with accurate colors, achieving the best performance as reported in Table 1. Consequently, the color event data from a color event camera like *DVS346C* is unnecessary, as the predicted color values of the event rays can be directly mapped to grayscale.

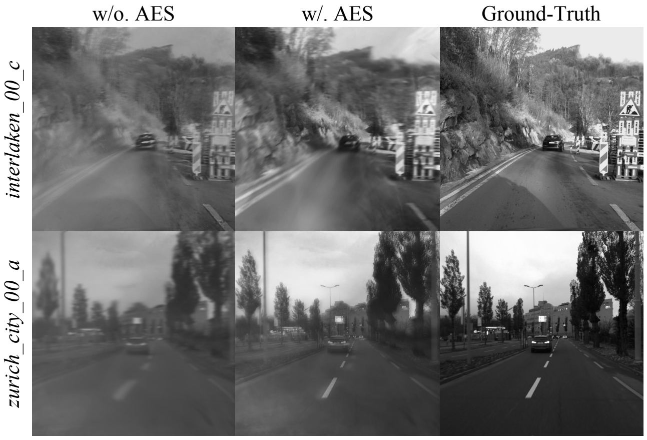


Figure 6: The Performance of the adaptive event supervision on the real-scene of the DSEC dataset.

5.2.2 *Real-scene dataset.* The real-world experiments are conducted on the *interlaken_00_c*, *interlaken_00_d*, and *zurich_city_00_a* sequences from the autonomous driving dataset DSEC [15] captured from a modern, high-resolution event sensor—Prophesee Gen3.1. However, the real-world experiments primarily serve as a qualitative benchmark, as the existing datasets [23, 32, 54], are not specifically designed for the task of NVS and lacks multi-view event streams with settings comparable to our synthetic dataset. This limitation is partly due to the fact that the target novel-view images are captured using a single standard RGB camera, which suffers from saturation effects because of its relatively limited dynamic range. Moreover, these images are not raw sensor outputs but have undergone in-camera image processing, often lossy in nature. In addition, the spectral response curve of the event camera is not publicly available, making color correction potentially inaccurate when aligning synthesized views with real images. Consequently, the dataset does not support accurate quantitative NVS evaluation.

For real-scene evaluation of dynamic reconstruction with event and frame fusion, the E2VID + D3DGS baseline recovers more visual details overall. However, the proposed E-4DGS exhibits fewer artifacts, particularly around foreground objects. While Deblur4DGS achieves improved reconstructions compared to the frame-based method D3DGS, both approaches struggle to recover fine details such as distant lettering as shown in Figure 5. Furthermore, E-4DGS delivers reconstructions with better high-frequency details and better geometry in comparison to E2VID + D3DGS. None of them achieve fully photorealistic quality, primarily due to the limitations of single-view supervision. Nevertheless, such quality is often unnecessary for many robotics applications and may be impractical given the complexity of the scenarios under consideration.

5.3 Ablation Evaluation

To evaluate the impact of each individual component, we conduct extensive qualitative and quantitative ablation studies. We primarily train different variants of our method on both the proposed synthetic and real-world sequences, focusing on the effects of event-adaptive slicing splatting (ESS), adaptive event supervision (AES), and intensity importance pruning (IIP) in the following sections.

Effect of Different Components. For the evaluation without ESS, we use the fixed event sampling number for accumulation instead of the specific strategy and then we use the fixed event threshold value for the evaluation without adaptive event supervision. In Table 2, we observe a clear performance gain from incorporating the ESS and IIP strategies. ESS effectively addresses the non-uniform spatial-temporal distribution of event data, while IIP reinforces multi-view consistency, jointly contributing to improved reconstruction performance. While the addition of the adaptive event supervision component slightly reduces performance on the synthetic dataset, it demonstrably improves texture fidelity and temporal consistency in real-world scenarios in Figure 6.

Effect of Motion Blur at Different Levels In our experiments, to assess the robustness of the deblurring performance, we simulate blurry images with varying degrees of motion blur—mild, medium, and strong—by integrating RGB frames over the exposure time in Blender [7], resulting in realistic, motion-dependent blur patterns. We choose the synthetic scene *Garage* for evaluation. As shown in Table 3, our proposed *E-4DGS* consistently outperforms all baselines across all levels of motion blur, achieving the highest performance. Moreover, our method reconstructs sharper scene details and more accurate object boundaries compared to baselines, especially under strong motion blur, demonstrating its superior capability in preserving both spatial structure and temporal consistency. More details are in the supplementary materials.

About Frame Interpolation Comparisons. We compare our method with the event-based video interpolation approach CBMNet [31] on synthetic indoor scenes. Since there is no ground truth for frame interpolation, we assess the performance of the methods using recent no-reference metrics: CLIPIQA [74] and MUSIQ [28]. * indicates that sharp RGB frames with event sequences are used as input, whereas motion-blurred frames and event data are used in other settings. The results are shown in Table 4.

Table 4: Quantitative comparisons for frame interpolation.

Methods	Input	**CLIPIQA↑**	**MUSIQ↑**
CBMNet* [31]	*sharp*	0.235	61.87
CBMNet [31]	*motion-blur*	0.169	43.95
Deblur4DGS [81]	*motion-blur*	0.193	53.88
E-4DGS	*motion-blur*	**0.208**	**54.62**

6 Conclusion

In this paper, we propose **E-4DGS**, a novel paradigm for real-time dynamic view synthesis based on dynamic 3DGS using multi-view event sequences. We design a synthetic multi-view camera setup with six moving event cameras surrounding an object in a 360-degree configuration and provide a benchmark multi-view event stream dataset that captures challenging motion scenarios. Our approach outperforms both event-only and event-RGB fusion baselines, paving the way for the exploration of multi-view event-based reconstruction as a novel approach for rapid scene capture. Future work will focus on addressing the challenges of handling larger-scale dynamic scenes and improving computational efficiency for real-world applications such as autonomous driving and immersive virtual environments.

References

[1] Anish Bhattacharya, Ratnesh Madaan, Fernando Cladera, Sai Vemprala, Rogerio Bonatti, Kostas Daniilidis, Ashish Kapoor, Vijay Kumar, Nikolai Matni, and Jayesh K Gupta. 2024. Evdnerf: Reconstructing event data with dynamic neural radiance fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5846–5855.

[2] Marco Cannici and Davide Scaramuzza. 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 (CVPR)*.

[3] Ang Cao and Justin Johnson. 2023. HexPlane: A Fast Representation for Dynamic Scenes. In *Computer Vision and Pattern Recognition (CVPR)*. https://caoang327.github.io/HexPlane/

[4] Kanghao Chen, Zeyu Wang, and Lin Wang. 2024. ExFMan: Rendering 3D Dynamic Humans with Hybrid Monocular Blurry Frames and Events. *arXiv preprint arXiv:2409.14103* (2024).

[5] Kang Chen, Jiyuan Zhang, Zecheng Hao, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. 2024. USP-Gaussian: Unifying Spike-based Image Reconstruction, Pose Correction and Gaussian Splatting. *arXiv preprint arXiv:2411.10504* (2024).

[6] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7020–7030.

[7] Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam. http://www.blender.org

[8] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems* 36 (2024).

[9] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 2024. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

[10] Burak Ercan, Onur Eker, Canberk Saglam, Aykut Erdem, and Erkut Erdem. 2024. Hypere2vid: Improving event-based video reconstruction via hypernetworks. *IEEE Transactions on Image Processing* (2024).

[11] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. 2024. InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds. arXiv:2403.20309 [cs.CV]

[12] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. 2023. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *arXiv preprint arXiv:2311.17245* (2023).

[13] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. 2020. Event-based vision: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)* (2020).

[14] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. 2018. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3867–3876.

[15] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. 2021. DSEC: A Stereo Event Camera Dataset for Driving Scenarios. *IEEE Robotics and Automation Letters* (2021). doi:10.1109/LRA.2021.3068942

[16] Gaurvi Goyal, Franco Di Pietro, Nicolo Carissimi, Arren Glover, and Chiara Bartolozzi. 2023. Moveenet: online high-frequency human pose estimation with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4024–4033.

[17] Shuang Guo and Guillermo Gallego. 2024. CMax-SLAM: Event-based rotational-motion bundle adjustment and SLAM system using contrast maximization. *IEEE Transactions on Robotics* (2024).

[18] Yijia Guo, Liwen Hu, Yuanxi Bai, Jiawei Yao, Lei Ma, and Tiejun Huang. 2024. Spikegs: Reconstruct 3d scene via fast-moving bio-inspired sensors. *arXiv preprint arXiv:2407.03771* (2024).

[19] Haiqian Han, Jianing Li, Henglu Wei, and Xiangyang Ji. 2024. Event-3DGS: Event-based 3D Reconstruction Using 3D Gaussian Splatting. *Advances in Neural Information Processing Systems* 37 (2024), 128139–128159.

[20] Haiqian Han, Jiacheng Lyu, Jianing Li, Henglu Wei, Cheng Li, Yajing Wei, Shu Chen, and Xiangyang Ji. 2024. Physical-Based Event Camera Simulator. In *European Conference on Computer Vision*. Springer, 19–35.

[21] Bing He, Yunuo Chen, Guo Lu, Qi Wang, Qunshan Gu, Rong Xie, Li Song, and Wenjun Zhang. 2024. S4D: Streaming 4D Real-World Reconstruction with Gaussians and 3D Control Points. arXiv:2408.13036 [cs.CV] https://arxiv.org/abs/2408.13036

[22] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. 2022. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17804–17813.

[23] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. 2022. Event-aided direct sparse odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5781–5790.

[24] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. 2021. v2e: From video frames to realistic DVS events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1312–1321.

[25] Jian Huang, Chengrui Dong, and Peidong Liu. 2024. IncEventGS: Pose-Free Gaussian Splatting from a Single Event Camera. *arXiv preprint arXiv:2410.08107* (2024).

[26] Inwoo Hwang, Junho Kim, and Young Min Kim. 2023. Ev-nerf: Event based neural radiance field. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 837–847.

[27] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*. 13700–13710.

[28] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5148–5157.

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)* (2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[30] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. 2023. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18032–18042.

[31] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. 2023. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18032–18042.

[32] Simon Klenk, Jason Chui, Nikolaus Demmel, and Daniel Cremers. 2021. TUM-VIE: The TUM stereo visual-inertial event dataset. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8601–8608.

[33] Simon Klenk, Lukas Koestler, Davide Scaramuzza, and Daniel Cremers. 2023. E-nerf: Neural radiance fields from a moving event camera. *IEEE Robotics and Automation Letters* 8, 3 (2023), 1587–1594.

[34] Jae Yong Lee, Yuqun Wu, Chuhang Zou, Derek Hoiem, and Shenlong Wang. 2024. Plenoptic PNG: Real-Time Neural Radiance Fields in 150 KB. *arXiv preprint arXiv:2409.15689* (2024).

[35] Seungjun Lee and Gim Hee Lee. 2025. DiET-GS: Diffusion Prior and Event Stream-Assisted Motion Deblurring 3D Gaussian Splatting. arXiv:2503.24210 [cs.CV] https://arxiv.org/abs/2503.24210

[36] Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. 2023. Weakly-supervised 3d spatial reasoning for text-based visual question answering. *IEEE Transactions on Image Processing* 32 (2023), 3367–3382.

[37] Hao Li, Curise Jia, Peng Jin, Zesen Cheng, Kehan Li, Jialu Sui, Chang Liu, and Li Yuan. 2023. Freestyleret: Retrieving images from style-diversified queries. *arXiv preprint arXiv:2312.02428* (2023).

[38] Hao Li, Da Long, Li Yuan, Yu Wang, Yonghong Tian, Xinchang Wang, and Fanyang Mo. 2025. Decoupled peak property learning for efficient and interpretable electronic circular dichroism spectrum prediction. *Nature Computational Science* (2025), 1–11.

[39] Jinwei Lin. 2024. Dynamic NeRF: A Review. *arXiv preprint arXiv:2405.08609* (2024).

[40] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. 2022. Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *European Conference on Computer Vision*. Springer, 578–593.

[41] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. 2024. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21136–21145.

[42] Weng Fei Low and Gim Hee Lee. 2023. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[43] Jiahao Lu, Jiacheng Deng, Ruijie Zhu, Yanzhe Liang, Wenfei Yang, Xu Zhou, and Tianzhu Zhang. 2025. Dn-4dgs: Denoised deformable network with temporal-spatial aggregation for dynamic scene rendering. *Advances in Neural Information Processing Systems* 37 (2025), 84114–84138.

[44] Qi Ma, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. 2023. Deformable neural radiance fields using rgb and event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3590–3600.

[45] Branislav Micusik and Tomáš Pajdla. 2006. Structure from motion with wide circular field of view cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 7 (2006), 1135–1149.

[46] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. arXiv:1905.00889 [cs.CV] https://arxiv.org/abs/1905.00889

[47] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*. https://www.matthewtancik.com/nerf

[48] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. 2019. Bringing a blurry frame alive at high frame-rate with an event camera. In *Computer Vision and Pattern Recognition (CVPR)*.

[49] Yatian Pang, Tanghui Jia, Yujun Shi, Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Xing Zhou, Francis EH Tay, and Li Yuan. 2024. Envision3D: One Image to 3D with Anchor Views Interpolation. *arXiv preprint arXiv:2403.08902* (2024).

[50] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. 2022. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*. Springer, 604–621.

[51] Yatian Pang, Bin Zhu, Bin Lin, Mingzhe Zheng, Francis EH Tay, Ser-Nam Lim, Harry Yang, and Li Yuan. 2024. DreamDance: Animating Human Images by Enriching 3D Geometry Cues from 2D Poses. *arXiv preprint arXiv:2412.00397* (2024).

[52] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5865–5874.

[53] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. 2021. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228* (2021).

[54] Shihan Peng, Hanyu Zhou, Hao Dong, Zhiwei Shi, Haoyue Liu, Yuxing Duan, Yi Chang, and Luxin Yan. 2024. CoSEC: A coaxial stereo event camera dataset for autonomous driving. *arXiv preprint arXiv:2408.08500* (2024).

[55] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.

[56] Yunshan Qi, Jia Li, Yifan Zhao, Yu Zhang, and Lin Zhu. 2024. E3 NeRF: Efficient Event-Enhanced Neural Radiance Fields from Blurry Images. *arXiv preprint arXiv:2408.01840* (2024).

[57] Yunshan Qi, Lin Zhu, Yu Zhang, and Jia Li. 2023. E2NeRF: Event Enhanced Neural Radiance Fields from Blurry Images. In *International Conference on Computer Vision (ICCV)*.

[58] Maxime Raafat and contributors. 2024. BlenderNeRF: Easy NeRF synthetic dataset creation within Blender. https://github.com/maximeraafat/BlenderNeRF. Accessed: 2025-04-07.

[59] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. 2019. High Speed and High Dynamic Range Video with an Event Camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)* (2019). http://rpg.ifi.uzh.ch/docs/TPAMI19_Rebecq.pdf

[60] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. 2023. EventNeRF: Neural Radiance Fields from a Single Colour Event Camera. In *Computer Vision and Pattern Recognition (CVPR)*.

[61] Viktor Rudnev, Gereon Fox, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. 2024. Dynamic EventNeRF: Reconstructing General Dynamic Scenes from Multi-view Event Cameras. *arXiv preprint arXiv:2412.06770* (2024).

[62] Arman Savran, Raffaele Tavarone, Bertrand Higy, Leonardo Badino, and Chiara Bartolozzi. 2018. Energy and computation efficient audio-visual voice activity detection driven by event-cameras. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 333–340.

[63] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion Revisited. In *Computer Vision and Pattern Recognition (CVPR)*. https://github.com/colmap/colmap

[64] Jiwei Shan, Zeyu Cai, Cheng-Tai Hsieh, Shing Shin Cheng, and Hesheng Wang. 2025. Deformable Gaussian Splatting for Efficient and High-Fidelity Reconstruction of Surgical Scenes. *arXiv preprint arXiv:2501.01101* (2025).

[65] Zihang Shao, Xuanye Fang, Yaxin Li, Chaoran Feng, Jiangrong Shen, and Qi Xu. 2023. EICIL: joint excitatory inhibitory cycle iteration learning for deep spiking neural networks. *Advances in Neural Information Processing Systems* 36 (2023), 32117–32128.

[66] Noah Snavely, Steven M Seitz, and Richard Szeliski. 2006. Photo tourism: exploring photo collections in 3D. In *ACM siggraph 2006 papers*. 835–846.

[67] Ganchao Tan, Yang Wang, Han Han, Yang Cao, Feng Wu, and Zheng-Jun Zha. 2022. Multi-grained spatio-temporal features perceived network for event-based lip-reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20094–20103.

[68] Wei Zhi Tang, Daniel Rebain, Kostantinos G Derpanis, and Kwang Moo Yi. 2024. LSE-NeRF: Learning Sensor Modeling Errors for Deblured Neural Radiance Fields with RGB-Event Stereo. *arXiv preprint arXiv:2409.06104* (2024).

[69] Zhenyu Tang, Chaoran Feng, Xinhua Cheng, Wangbo Yu, Junwu Zhang, Yuan Liu, Xiaoxiao Long, Wenping Wang, and Li Yuan. 2025. NeuralGS: Bridging Neural Fields and 3D Gaussian Splatting for Compact 3D Representations. *arXiv preprint arXiv:2503.23162* (2025).

[70] Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Wangbo Yu, Chaoran Feng, Yatian Pang, Bin Lin, and Li Yuan. 2024. Cycle3D: High-quality and Consistent Image-to-3D Generation via Generation-Reconstruction Cycle. *arXiv preprint arXiv:2407.19548* (2024).

[71] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. 2021. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16155–16164.

[72] Diwen Wan, Yuxiang Wang, Ruijie Lu, and Gang Zeng. 2024. Template-free Articulated Gaussian Splatting for Real-time Reposable Dynamic View Synthesis. *arXiv preprint arXiv:2412.05570* (2024).

[73] Haoyang Wang, Ruishan Guo, Pengtao Ma, Ciyu Ruan, Xinyu Luo, Wenhua Ding, Tianyang Zhong, Jingao Xu, Yunhao Liu, and Xinlei Chen. 2025. Towards Mobile Sensing with Event Cameras on High-mobility Resource-constrained Devices: A Survey. *arXiv preprint arXiv:2503.22943* (2025).

[74] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2555–2563.

[75] Jiaxu Wang, Junhao He, Ziyi Zhang, Mingyuan Sun, Jingkai Sun, and Renjing Xu. 2024. EvGGS: A Collaborative Learning Framework for Event-based Generalizable Gaussian Splatting. *arXiv preprint arXiv:2405.14959* (2024).

[76] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. 2024. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764* (2024).

[77] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[78] Jiahao Wu, Rui Peng, Zhiyan Wang, Lu Xiao, Luyang Tang, Jinbo Yan, Kaiqiang Xiong, and Ronggang Wang. 2025. Swift4D: Adaptive divide-and-conquer Gaussian Splatting for compact and efficient reconstruction of dynamic scene. *arXiv preprint arXiv:2503.12307* (2025).

[79] Jingqian Wu, Shuo Zhu, Chutian Wang, and Edmund Y Lam. 2024. Ev-GS: Event-based gaussian splatting for efficient and accurate radiance field rendering. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[80] Jingqian Wu, Shuo Zhu, Chutian Wang, Boxin Shi, and Edmund Y Lam. 2024. SweepEvGS: Event-Based 3D Gaussian Splatting for Macro and Micro Radiance

Field Rendering from a Single Sweep. *arXiv preprint arXiv:2412.11579* (2024).

[81] Renlong Wu, Zhilu Zhang, Mingyang Chen, Xiaopeng Fan, Zifei Yan, and Wangmeng Zuo. 2024. Deblur4DGS: 4D Gaussian Splatting from Blurry Monocular Video. *arXiv preprint arXiv:2412.06424* (2024).

[82] Ziyi Wu, Mathias Gehrig, Qing Lyu, Xudong Liu, and Igor Gilitschenski. 2024. Leod: Label-efficient object detection for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 16933–16943.

[83] Tianyi Xiong, Jiayi Wu, Botao He, Cornelia Fermuller, Yiannis Aloimonos, Heng Huang, and Christopher A Metzler. 2024. Event3DGS: Event-based 3D Gaussian Splatting for Fast Egomotion. *arXiv preprint arXiv:2406.02972* (2024).

[84] Chuanzhi Xu, Haoxian Zhou, Haodong Chen, Vera Chung, and Qiang Qu. 2025. A Survey on Event-driven 3D Reconstruction: Development under Different Categories. *arXiv preprint arXiv:2503.19753* (2025).

[85] Wenhao Xu, Wenming Weng, Yueyi Zhang, Ruikang Xu, and Zhiwei Xiong. 2024. Event-boosted Deformable 3D Gaussians for Fast Dynamic Scene Reconstruction. *arXiv preprint arXiv:2411.16180* (2024).

[86] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 2023. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 1179–1189.

[87] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. 2024. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[88] Jinbo Yan, Rui Peng, Luyang Tang, and Ronggang Wang. 2024. 4D Gaussian Splatting with Scale-aware Residual Field and Adaptive Optimization for Real-time rendering of temporally complex dynamic scenes. In *Proceedings of the 32nd ACM International Conference on Multimedia.* 7871–7880.

[89] Zhiwen Yan, Chen Li, and Gim Hee Lee. 2023. Nerf-ds: Neural radiance fields for dynamic specular objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8285–8295.

[90] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. 2024. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 20331–20341.

[91] Zeyu Yang, Zijie Pan, Xiatian Zhu, Li Zhang, Yu-Gang Jiang, and Philip HS Torr. 2024. 4D Gaussian Splatting: Modeling Dynamic Scenes with Native 4D Primitives. *arXiv preprint arXiv:2412.20720* (2024).

[92] Xiaoting Yin, Hao Shi, Yuhan Bao, Zhenshan Bing, Yiyi Liao, Kailun Yang, and Kaiwei Wang. 2024. E-3DGS: Gaussian Splatting with Exposure and Motion Events. *arXiv preprint arXiv:2410.16995* (2024).

[93] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025. TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models. *arXiv preprint arXiv:2503.05638* (2025).

[94] Wangbo Yu, Chaoran Feng, Jiye Tang, Xu Jia, Li Yuan, and Yonghong Tian. 2024. EvaGaussians: Event Stream Assisted Gaussian Splatting from Blurry Images. *arXiv preprint arXiv:2405.20224* (2024).

[95] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).

[96] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. 2024. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 19447–19456.

[97] Sohaib Zahid, Viktor Rudnev, Eddy Ilg, and Vladislav Golyanik. 2025. E-3DGS: Event-based Novel View Rendering of Large-scale Scenes Using 3D Gaussian Splatting. *3DV* (2025).

[98] Yifan Zhan, Zhuoxiao Li, Muyao Niu, Zhihang Zhong, Shohei Nobuhara, Ko Nishino, and Yinqiang Zheng. 2024. KFD-NeRF: Rethinking Dynamic NeRF with Kalman Filter. *arXiv preprint arXiv:2407.13185* 3 (2024).

[99] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. 2024. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825* (2024).

[100] Junwu Zhang, Zhenyu Tang, Yatian Pang, Xinhua Cheng, Peng Jin, Yida Wei, Wangbo Yu, Munan Ning, and Li Yuan. 2023. Repaint123: Fast and high-quality one image to 3d generation with progressive controllable 2d repainting. *arXiv preprint arXiv:2312.13271* (2023).

[101] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

[102] Zixin Zhang, Kanghao Chen, and Lin Wang. 2024. Elite-EvGS: Learning Event-based 3D Gaussian Splatting by Distilling Event-to-Video Priors. *arXiv preprint arXiv:2409.13392* (2024).

[103] Zixin Zhang, Kanghao Chen, and Lin Wang. 2024. Elite-evgs: Learning event-based 3d gaussian splatting by distilling event-to-video priors. *arXiv preprint arXiv:2409.13392* (2024).

[104] Ziran Zhang, Xiaohui Li, Yihao Liu, Yujin Wang, Yueting Chen, Tianfan Xue, and Shi Guo. 2025. Event-Guided Video Diffusion Model for Physically Realistic Large-Motion Frame Interpolation. *arXiv preprint arXiv:2503.20268* (2025).

[105] Chunhui Zhao, Yakun Li, and Yang Lyu. 2023. Event-based real-time moving object detection based on imu ego-motion compensation. In *2023 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 690–696.

[106] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. 2021. Eventhpe: Event-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10996–11005.

[107] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. 2024. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5819–5828.

Detailed descriptions of dataset construction and training configurations are provided in Section A of the appendix. Section B presents the implementation of our proposed initialization strategy and compares it with existing methods. Further experimental results and ablation studies are reported in Section C.

A Dataset Preparations

A.1 Synthetic datasets

We manually create eight synthetic scenes with six viewpoints arranged in a 360-degree configuration around the object or scene. Each scene is designed as a center-focus setup, with an object placed at the center. For these scenes, we render six dynamic scenarios at a resolution of 346 × 260 in Blender [7] at 3000 FPS with the BlenderNeRF addon [58]. Six moving viewpoints are uniformly distributed around the object in a spherical spiral motion at a constant height. Event streams are generated using the v2e framework [24]. Additionally, leveraging the camera trajectory data, we simulate blurry images by integrating RGB frames over the exposure time, with varying degrees of motion *blur—mild*, *medium*, and *strong*.

For training and evaluation, we use six viewpoints for training and set the llffhold value to 8 for testing. For event-only dynamic reconstruction, RGB frames are converted to grayscale for evaluation, with event streams used exclusively as input. In the event-RGB fusion dynamic reconstruction, full-resolution color images are used in conjunction with event slices as input modalities.

Data Composition The proposed synthetic dataset consists of five dynamic objects, three dynamic indoor scenes, as follows:

- **Dynamic objects.** We design five object models in Blender, including *Lego*, *Rubik's Cube*, *MC Toy*, *Hinge*, and *Cubes*. The dynamic *Lego* model is derived from the static *Lego* in the NeRF dataset [47], to which we add animation.
- **Dynamic indoor scenes.** We design three indoor models with dynamic objects in Blender, including *Capsule*, *Restroom* and *Garage*.

All models are licensed under *CC-BY 4.0* and will be open-source.

Data Limitations. The synthetic data in this work is generated using the v2e framework [24], which simulates events based on images. However, this approach is inherently limited in handling extreme lighting conditions, such as overexposure or very low light. In these scenarios, the images themselves lack crucial information due to the nature of the lighting, which restricts the ability to accurately simulate event data for such conditions.The left is the pointcloud of Gaussian initilization and the right is the novel view of the *Restroom* scene.

A.2 Real-scene datasets

We adopt the DSEC dataset [15], a large-scale real-world dataset designed for driving scenarios, to evaluate our method under realistic and dynamic conditions. The dataset was captured using a synchronized sensor rig mounted on a vehicle, consisting of a Prophesee Gen3.1 event camera, a global shutter RGB camera, and a Velodyne LiDAR. The event camera records asynchronous brightness changes at a spatial resolution of 640×480 and provides high temporal resolution (down to microseconds), enabling the capture of fast motion and high dynamic range scenes. The RGB camera outputs global shutter images at 1024×768 resolution with fixed frame intervals. Calibration files are provided to align coordinate systems of the sensors.

Each sequence in DSEC contains temporally synchronized event streams, RGB frames, LiDAR point clouds, camera intrinsics/extrinsics, and time-stamped poses obtained via visual-inertial odometry. For our experiments, we select three representative sequences: *interlaken_00_c*, *interlaken_00_d*, and *zurich_city_00_a*, which cover diverse urban and suburban environments.

Since the dataset is not originally designed for Novel View Synthesis (NVS), we perform several processing steps to construct suitable input-output pairs:

- **Image-Event Alignment:** For each RGB frame, we extract a corresponding event stream by accumulating events within a fixed temporal window around the image timestamp. Events outside the desired range are discarded to reduce background noise.
- **View Subsampling:** We uniformly sample camera viewpoints along the driving trajectory. Following the standard LLFF [46] protocol, we use every 8 consecutive views for training and hold out the next view for evaluation.
- **Modality Handling:** For event-only models, RGB frames are converted to grayscale as evaluation and only event streams are used as input. For event-RGB fusion settings, the full-resolution color images are used jointly with the event slices as input modalities.
- **Frame Curation:** Frames suffering from severe motion blur or under-/over-exposure are excluded to ensure a clean evaluation set. We also ignore frames with poor localization confidence based on pose metadata.

Although the dataset offers event streams, RGB images, and LiDAR data, its forward-facing setup with narrow baseline viewpoints makes it inherently unsuitable for tasks requiring diverse multi-view observations, such as high-fidelity 3D reconstruction and novel view synthesis. As a result, we use these sequences only for qualitative visualization.

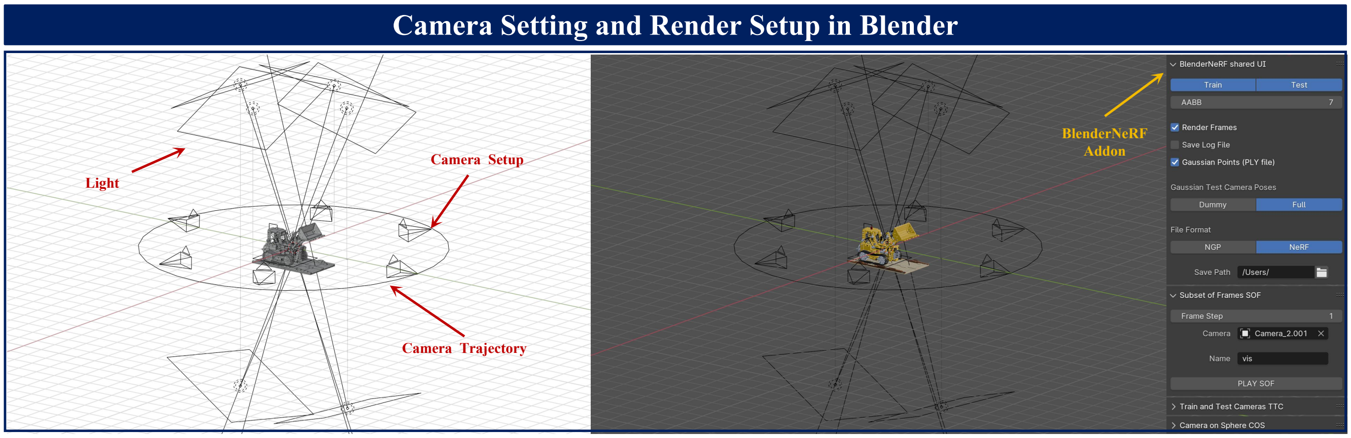
B More details of Pointcloud Initialization

Table 5: Ablation Study on Different Initialization Method.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Time/h
Random Init.	21.56	0.785	0.233	0.9
E2VID+SfM	24.87	0.866	0.170	2.5
Ours	24.21(-0.66)	0.854(-0.012)	0.176(+0.006)	1.1(-1.4)

In this section, we explore the impact of different point cloud initialization methods on the rendering performance of *E-4DGS* in three proposed indoor scenes. Compared to the random initialization, commonly used in methods such as [25, 79], using the sparse point clouds from Structure-from-Motion (SfM) [45] significantly improves rendering accuracy when only motion events are utilized, with the PSNR metric increasing from 24.21 dB to 24.87 dB.

To further demonstrate the trade-off between efficiency and performance achieved by our proposed method, we compare the



AABB 7

☑ Render Frames

☐ Save Log File

☑ Gaussian Points (PLY file)

Save Path /Users/

Frame Step 1

Camera Camera_2.001

Name vis

Figure 7: Virtual camera setup in Blender for synthetic dataset generation. The six simulated DAVIS 346C event cameras are positioned to match the layout of our real-world multi-view recording environment.

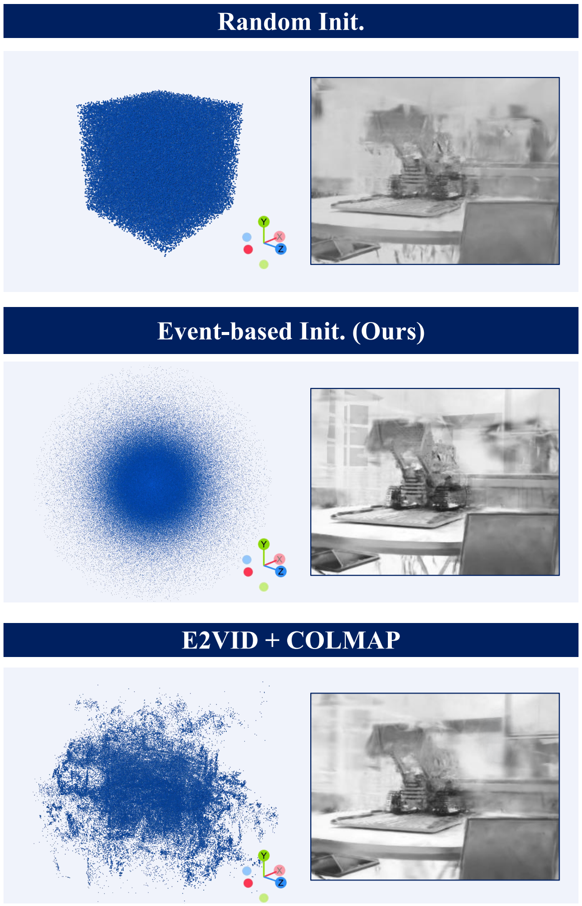


Figure 8: Qualitative comparison of different initialization methods and our method achieves a trade-off between efficiency and performace.

effect of point cloud initialization using event-to-video approaches in Table 5. Using E2VID[59] to convert event data into images and generating point clouds through SfM yields further accuracy improvements. However, this process introduces additional computational costs and time due to reliance on learning-based methods.

As shown in Figure 8, we visualize the impact of different initialization methods on event-based 4DGS rendering. When random initialization is used, the 4DGS reconstruction based on motion events suffers from noticeable artifacts and a lack of detail. The E2VID + COLMAP-based SfM method improves scene reconstruction, but at the cost of significantly lower runtime efficiency. In contrast, our method employs a radial initialization after considering a center-focus environment, yielding comparable rendering results to the two-stage initialization approach, despite slightly lower quantitative metrics. This validates the key role of our approach in improving the efficiency of event-driven explicit dynamic reconstruction.

C Additional experiments

C.1 Performance of adaptive event threshold

To bridge the gap between dense image rendering and sparse event streams, our E-4DGS framework incorporates a learnable event contrast threshold $\hat{C}$. This parameter governs the sensitivity of event triggering, and is jointly optimized with other model parameters. Rather than relying on a fixed threshold, we allow $\hat{C}$ to dynamically evolve to better accommodate diverse temporal changes in intensity. As shown in Fig. 9, the synthetic dataset demonstrates a relatively stable threshold behavior, aligning with its lower noise and controlled motion. In contrast, real-scene data produces more frequent and stronger burst patterns, requiring a more adaptive threshold to handle high-frequency voltage changes effectively. This adaptiveness ensures accurate contrast modeling for event supervision, contributing to the photometric alignment between rendered and observed event data.

C.2 Qualitative comprisons of the motion deblurring

In the main paper, we have already presented qualitative results under varying levels of motion blur. In this section, we further provide additional visual comparisons on the synthetic dataset to

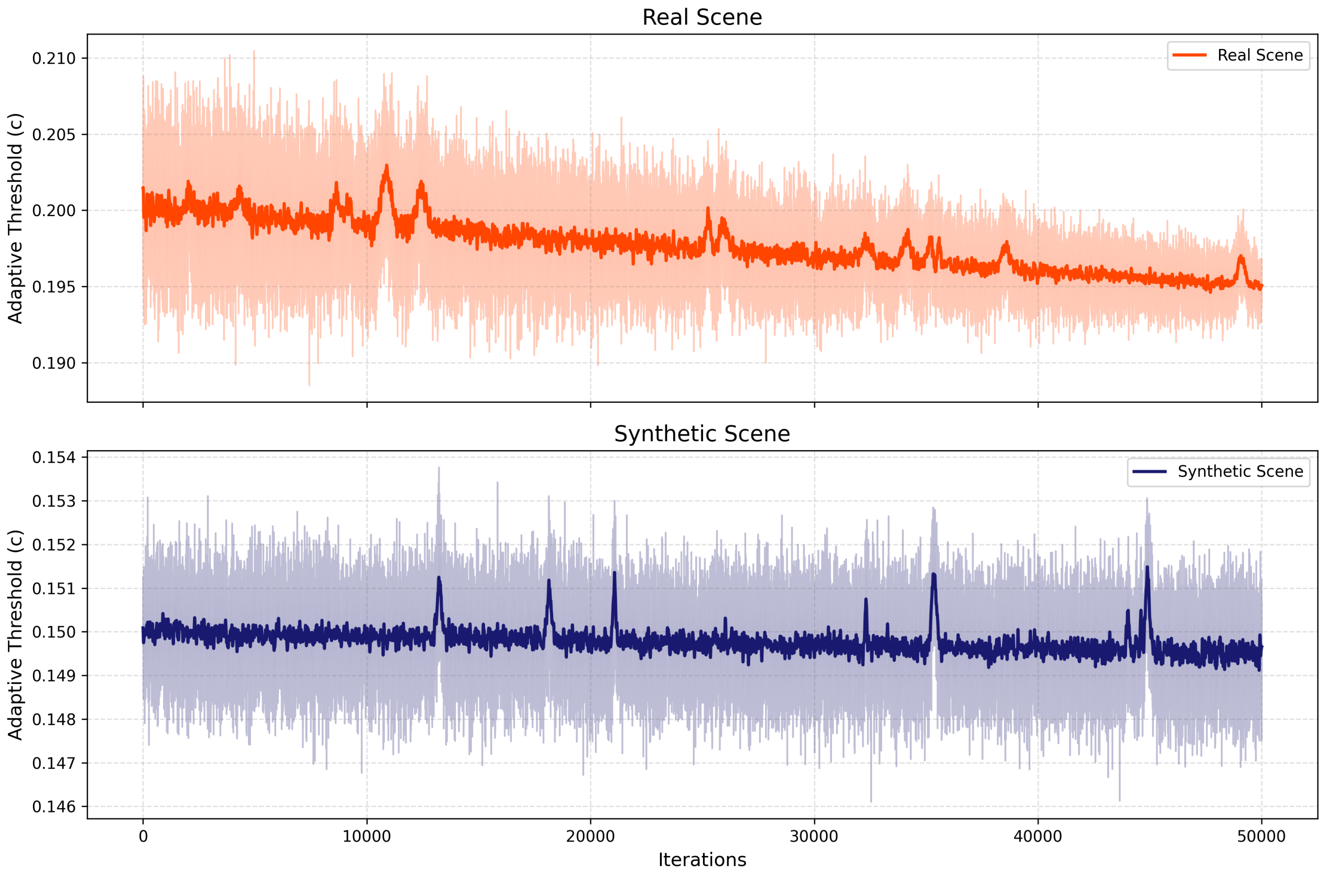


Figure 9: Visualization of the adaptive event threshold $\hat{C}$ during training on both synthetic and real-scene datasets. For the synthetic dataset (bottom), $\hat{C}$ is initialized to 0.15 and remains relatively stable with occasional spikes. For the real-scene dataset (top), $\hat{C}$ is initialized to 0.2 and exhibits more pronounced temporal fluctuations due to sensor noise and real-world intensity transitions. These burst-like perturbations reflect dynamic changes in photovoltage, which are used to compute the contrast between adjacent frames. A properly adjusted $\hat{C}$ is critical for robustly converting such contrast into events during training.

evaluate the robustness of different methods across mild, medium, and severe blur conditions. As shown in Figure 10, increasing blur levels degrade the reconstruction quality of baseline methods to varying degrees. Compared to D3DGS, which struggles to recover sharp structures under heavy blur, and E2VID+D3DGS, which introduces artifacts from video reconstruction, our method E-4DGS consistently produces sharper and more temporally coherent results. Although Deblur4DGS mitigates some blur-related degradation, it lacks the geometric consistency offered by our event-guided framework. Overall, **E-4DGS** achieves high-fidelity reconstructions across all blur settings, demonstrating its robustness and effectiveness under challenging motion scenarios.

C.3 Per-Scene Breakdown

Table 6 presents the quantitative results of all methods for each of the eight synthetic scene sequences, simulated with default settings that are optimal for all methods. The per-scene results are generally consistent with the aggregate metrics, as discussed in Section 5.1.2. Our method outperforms the baselines in most scenes and shows comparable performance in others.

D Broader Impact and Limitations

Broader Impact. The proposed *E-4DGS* framework opens up new possibilities for high-fidelity 4D reconstruction in domains where traditional cameras fall short due to motion blur or limited dynamic range. By leveraging the high temporal resolution of event cameras, our method enables temporally coherent scene modeling under rapid motion, which is beneficial for a variety of real-world applications including autonomous robotics, high-speed inspection, sports analytics, and scientific visualization in challenging illumination conditions. Furthermore, the ability to reconstruct dynamic scenes using purely event-based supervision contributes to the development of low-latency, power-efficient visual systems, which are particularly relevant for resource-constrained or edge computing scenarios.

Table 6: Per-synthetic scene breakdown under the default setting.

Metric	Method	Synthetic Scene								Average
		Lego	Rubik’s Cube	MC-Toy	Hinge	Cubes	Capsule	Restroom	Garage	
PSNR ↑	D3DGS$_{w/o\ blur}$	26.47	20.30	31.23	28.05	21.75	21.64	20.67	20.36	23.81
	D3DGS$_{w/\ blur}$	23.62	18.12	27.51	26.46	19.67	20.08	19.32	19.05	21.73
	E2VID+D3DGS	20.57	16.16	26.06	24.87	17.98	18.49	17.17	17.79	19.88
	Deblur4DGS	23.17	17.68	28.06	26.35	19.27	20.10	19.39	19.23	21.66
	E-4DGS$_{event\text{-}only}$	26.85	20.97	31.85	28.83	22.36	24.23	23.17	24.81	25.38
	E-4DGS$_{event\&\ RGB}$	**27.23**	**21.23**	**32.41**	**29.02**	**22.42**	**24.39**	**23.30**	**24.95**	**25.62**
LPIPS ↓	D3DGS$_{w/o\ blur}$	0.099	0.207	0.077	0.074	0.129	0.271	0.278	0.251	0.173
	D3DGS$_{w/\ blur}$	0.250	0.351	0.181	0.160	0.175	0.436	0.406	0.409	0.296
	E2VID+D3DGS	0.346	0.404	0.267	0.247	0.298	0.595	0.527	0.493	0.397
	Deblur4DGS	0.265	0.375	0.176	0.162	0.181	0.402	0.386	0.385	0.291
	E-4DGS$_{event\text{-}only}$	0.084	0.185	0.071	0.069	0.120	0.183	0.189	0.172	0.134
	E-4DGS$_{event\&\ RGB}$	**0.078**	**0.172**	**0.068**	**0.067**	**0.119**	**0.178**	**0.184**	**0.165**	**0.129**
SSIM ↑	D3DGS$_{w/o\ blur}$	0.910	0.868	0.956	0.936	0.924	0.770	0.765	0.757	0.861
	D3DGS$_{w/\ blur}$	0.821	0.804	0.905	0.908	0.905	0.730	0.686	0.620	0.797
	E2VID+D3DGS	0.765	0.752	0.851	0.856	0.852	0.655	0.547	0.549	0.728
	Deblur4DGS	0.813	0.786	0.908	0.900	0.898	0.736	0.695	0.643	0.797
	E-4DGS$_{event\text{-}only}$	0.912	0.882	0.959	0.942	0.931	0.842	0.829	0.874	0.896
	E-4DGS$_{event\&\ RGB}$	**0.925**	**0.895**	**0.963**	**0.949**	**0.933**	**0.848**	**0.835**	**0.879**	**0.903**

Limitations. While *E-4DGS* demonstrates promising results in dynamic 3D scene reconstruction, certain scenarios, such as those involving extreme motion or significant occlusions, may present challenges for the method. The performance is highly dependent on the availability of synchronized multi-view event data and precise camera calibration. These aspects are areas for further exploration to enhance robustness and generalizability in more complex environments.

Project Release. We implemented *E-4DGS* based on the official code of Deformable3DGS [90], Gaussianflow [41], E-NeRF [33] and Event3DGS [19] with Pytorch Upon the publication of the paper, we will release the project materials.

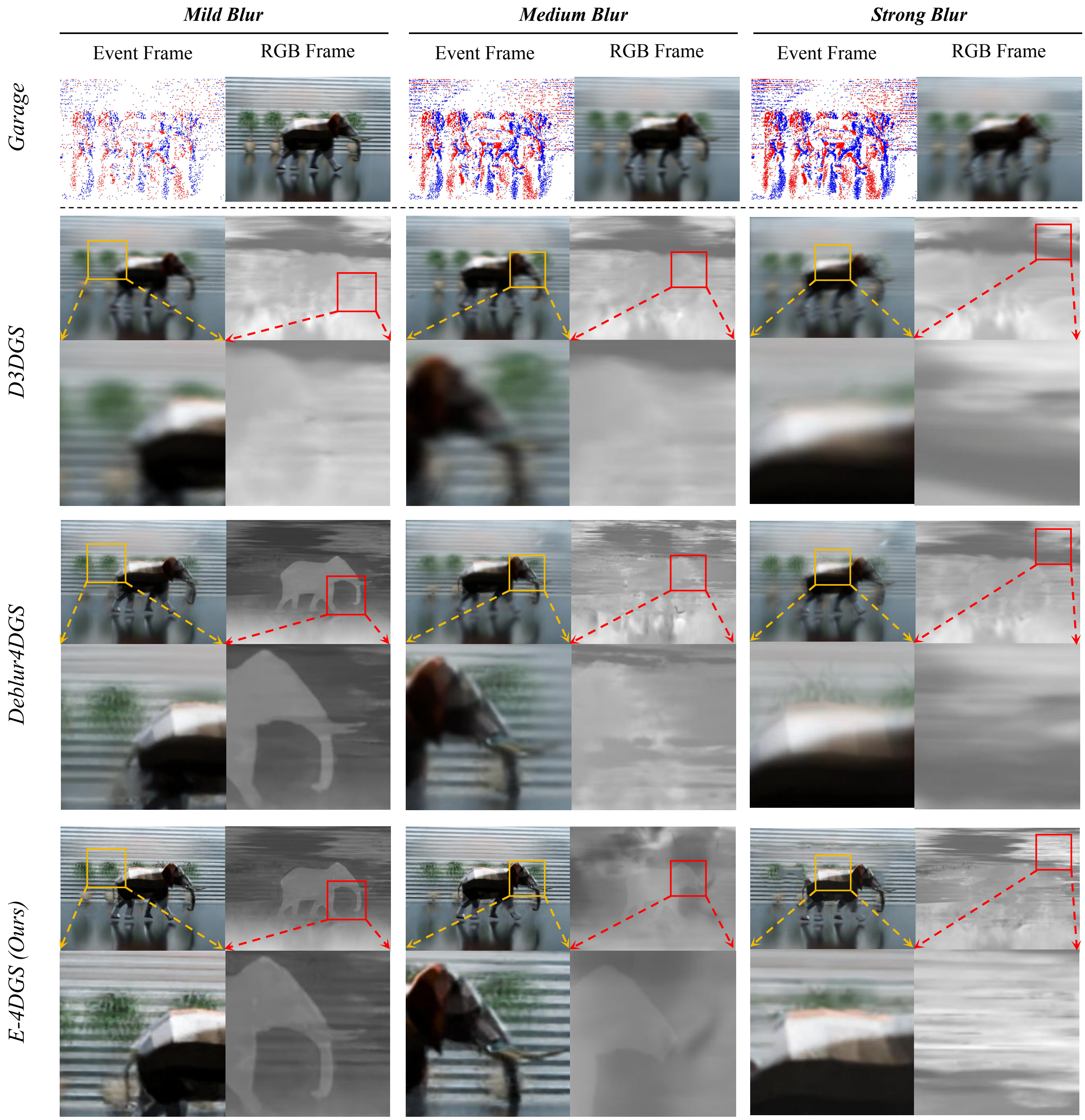


Figure 10: Qualitative comparison under varying motion blur levels on synthetic scenes. As blur severity increases, baseline methods (*D3DGS* and *Deblur4DGS*) suffer from degraded reconstructions with noticeable artifacts or loss of geometric consistency. In contrast, our method (E-4DGS) produces high-fidelity renderings with sharper details and improved temporal coherence across all blur levels, demonstrating its robustness and effectiveness under fast motion.