4D E-SloMo: 4D Reconstruction for High Speed Scene using a Hybrid RGB-Event Multi-View System

Bo Xu[1], Jun Dai[3], Yutian Chen[3], Linning Xu[3],
Mulin Yu[3], Yujin Wang[3], Shi Guo[3], Xinyi Le[1], Tianfan Xue[1,2,3]
[1]Shanghai Jiao Tong University [2]CUHK MMLab [3]Shanghai AI Lab [4]CPII under InnoHK

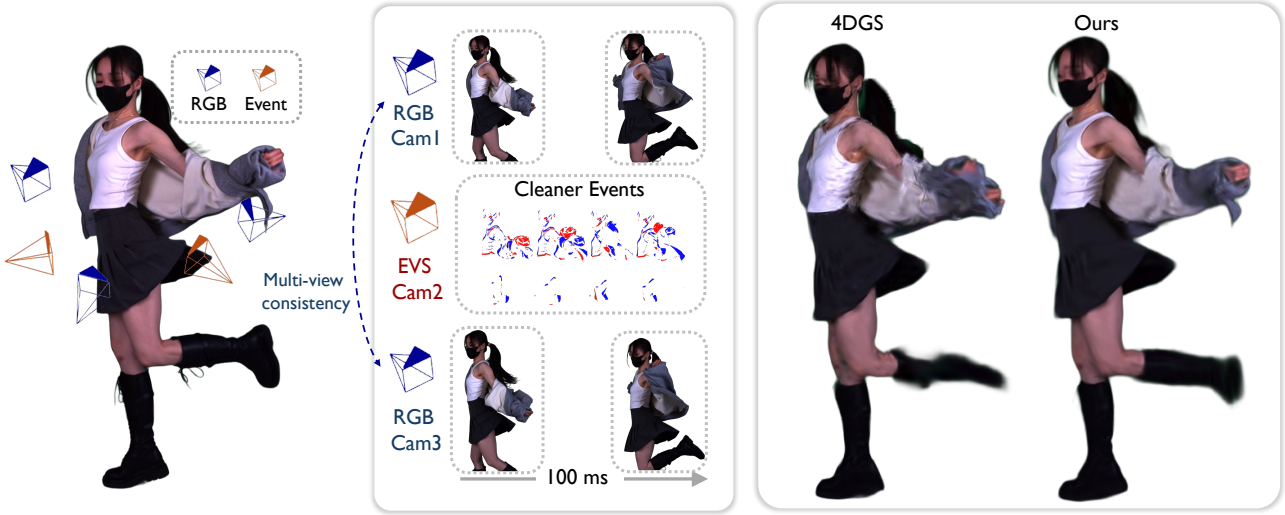

Figure 1. **Our hybrid event–RGB reconstruction framework.** *(a) Camera Setting:* our multi-view capture rig synchronizes high-resolution RGB cameras with event sensors at fixed viewpoints, enabling consistent cross-view observations of fast, large-amplitude motion. *(b) Reconstruction Results:* compared with 4DGS, our method produces sharper geometry and more stable appearance under high-speed motion by leveraging event-driven temporal cues.

Abstract

High-speed dynamic scene reconstruction enables free viewpoint replay in sports, filmmaking, and virtual reality applications. Recent neural rendering approaches have achieved impressive results in general scene reconstruction. However, these approaches struggle with fast motion when the inputs come from low-frame-rate RGB cameras. In such cases, motion blur and inter-frame gaps result in the loss of fine temporal details, compromising the effectiveness of the system. Existing event-based methods can capture high-speed dynamics with microsecond precision. Yet, they are limited by the low resolution and narrow field of single-sensor setups, restricting them to object-level or synthetic scenes. To overcome these limitations, we propose a high-resolution hybrid capture system in which synchronized RGB cameras are paired with co-located event sensors to record room-scale dynamic scenes. This design preserves fine spatial detail while maintaining a microsecond temporal response. In addition to the hardware, we introduce two algorithmic components to improve reconstruction quality. First, an event-guided optimization adaptively selects key frames and emphasizes regions with fast changes. Second, a flow-based loss further aligns event and image cues to enhance structural consistency and rendering fidelity. We also introduce a dataset of synchronized, high-resolution RGB and event streams across diverse, high-speed indoor scenes. Our system can reconstruct motions of up to 5 m/s and produces sharper, more stable renderings than existing neural or event-based baselines.

1. Introduction

4D reconstruction is a critical task in the field of 3D vision. It involves creating a renderable, time-varying 3D representation from multiple video inputs. One of its main challenges is reconstructing fast motion, which has broad applications in sports analysis, action movies, VR/AR, etc.

In recent years, neural rendering–based methods [17, 22, 41, 42, 44, 46, 50] have achieved remarkable results in 4D reconstruction. These approaches typically model a scene by learning implicit radiance fields or explicit Gaussian set, by render them differentiably to match multi-view images across time. Despite their impressive fidelity, they inherently rely on temporally consistent and photometrically sharp image supervision. As shown in Fig. 1, when dealing with fast, non-linear motions, low-frame-rate RGB cameras lose important information between frames.This causes these methods to fundamentally fail. Using high-speed camera arrays can solve this problem to a large extent. However, such setups are expensive and require a lot of bandwidth, which limits their practical use.

Inspired by the human visual system, event sensors [2, 38] asynchronously record per-pixel brightness changes with microsecond latency, producing temporally dense yet spatially sparse signals. Event camera captures rapid motion without blur and provide fine-grained temporal cues complementary to RGB images. Recent works [8, 24, 36, 45, 54] show that integrating event and RGB modalities notably enhances 4D Gaussian Splatting, enabling accurate motion recovery in high-speed scenes.

However, existing event-assisted 4D reconstruction methods suffer from two key limitations that restrict high-quality and high-resolution results. (1) *Sparse monocular observations*. Most approaches [24, 45, 54] rely on a single DAVIS camera combining RGB and event sensors, where the camera moves around the subject to collect multi-view data. For large-scale or deformable targets such as humans, this dynamic monocular setup yields spatially and temporally sparse signals, making it difficult to constrain complex shapes, while camera motion under non-uniform lighting often introduces noisy events. (2) *Low sensor resolution*. Multi-camera systems [8, 36] partially alleviate sparsity but remain limited by the low resolution of DAVIS sensors (346×260). Although such setups can capture correct motion in high-speed scenes, their low fidelity and resolution still prevent deployment in real-world applications.

To advance high-quality and high-resolution 4D reconstruction in real-world high-speed scenes using event camera, we build a *Hybrid RGB-Event Multi-View System*. Compared with previous setups based on DAVIS sensors, our system employs three high-quality 4K RGB cameras and two high-resolution event cameras (*i.e.*, Prophesee sensor with 720×1280 resolution) to capture dynamic motions. As shown in Fig. 1, the high-resolution imaging enables us to capture complex and challenging human motion such as dancing with fine spatial details. However, unlike DAVIS-based systems where each view provides temporally and spatially aligned RGB-event pairs, our hybrid setup captures either RGB or event signals per view. This design offers higher spatial resolution but also poses challenges for cross-modal alignment and fusion during reconstruction.

To achieve high-quality 4D reconstruction with our hybrid RGB-Event multi-view system, we introduce two key algorithmic improvements. (1) Cross-modal motion representation via optical flow. We were inspired by how optical flow captures motion continuity across time, which is a natural bridge between event streams and RGB frames. Instead of directly enforcing event-based intensity consistency, we reimagine events as a flow field that describes how pixels move. By warping high-quality RGB frames through this flow, the event signals become a dynamic cue guiding motion recovery across time. This cross-modal representation connects sparse event dynamics with dense visual appearance, allowing motion to emerge in a more coherent and physically intuitive way. (2) Event-aware scheduling. We draw inspiration from how human perception naturally attends to motion: our eyes instinctively focus on regions of rapid change. Similarly, we let event activity guide where the model should pay more attention: when the scene moves fast, dense event bursts signal important motion cues. By letting supervision adapt to these dynamics, the model learns to emphasize high-velocity regions, capturing motion with sharper temporal fidelity.

We also release a dataset pairing 4K RGB videos with 1K event streams across 6 indoor scenes of five to twelve seconds. On this real-world dataset, our approach reconstructs photorealistic appearance under fast, large-amplitude motion, yielding sharper boundaries and fewer floating artifacts than existing neural rendering baselines.

In summary, we make the following contributions:

- We propose the first hybrid Event–RGB multi-view system for high-speed 4D reconstruction and view synthesis.
- We develop a key-frame partitioning scheme and a flow-based loss that enables deformable 3D Gaussians to incorporate microsecond-level motion cues.
- We provide a public dataset that pairs 4K RGB videos with real event streams for challenging, room-scale dynamic scenes.
- Extensive experiments demonstrate sharper and more coherent renderings of fast motion than existing neural rendering approaches.

2. Related work

3D Scene Reconstruction. Traditional 3D scene representations can be categorized into matching-, point-, voxel-, and mesh-based approaches. Matching-based methods [27, 39] extract keypoints from adjacent frames and triangulate

them into sparse point clouds. Point-based methods [1, 14, 34] leverage multi-view stereo to estimate dense depth, but independent points often lead to local holes and incomplete geometry. Voxel-based representations [10, 12, 18] divide the scene into fixed-shape volumetric grids to maintain structural integrity, yet suffer from feature mixing across neighboring voxels. Mesh-based approaches [3, 6, 9, 40] refine voxels into adaptive triangular surfaces, enabling more detailed and topology-aware modeling. Recent neural rendering techniques shift to continuous representations. Neural Radiance Fields (NeRF) [25, 29] optimize MLPs with volumetric ray marching for high-quality novel view synthesis. 3D Gaussian Splatting (3DGS) [12] introduces an explicit formulation using a set of anisotropic 3D Gaussians that jointly model geometry and appearance. While these methods perform well for static scenes, extending them to dynamic scenes remains challenging due to fast motion, occlusion changes, and temporal consistency issues.

Dynamic Novel-View Synthesis. Dynamic 3D reconstruction and temporally coherent novel view synthesis remain difficult problems, particularly when scenes exhibit complex or fast motion. A common strategy in recent literature is to capture the scene using multiple synchronized cameras, providing per-frame multi-view observations for supervision [7, 19, 22, 24, 31, 35, 41, 44, 47, 50, 53]. These systems typically build upon a static 3D representation—such as NeRF [25] or 3D Gaussian Splatting (3DGS) [12]—and introduce an explicit motion model to account for temporal changes. The motion component can take various forms, including neural deformation fields [23, 30, 49], factorized space–time planes [37, 44], polynomial parameterizations [20], or Fourier-based motion bases [11]. Despite their differences, these approaches generally treat each time step independently, making it difficult to capture smooth temporal evolution or reconstruct high-frequency dynamics, which are essential for accurate modeling of rapid motion.

Existing multi-camera capture rigs for 4D scenes are also constrained by their low frame rates (*e.g.*, 15–30 FPS [5, 17, 21]), making it difficult to reconstruct high-speed motions. To achieve higher temporal resolution without increasing system complexity, recent works have begun exploring event cameras, whose microsecond-level temporal density makes them a promising direction for capturing fast, dynamic 4D scenes.

Event-based Novel-View Synthesis.

Event cameras offer microsecond temporal resolution and extremely high dynamic range, making them effective for mitigating motion blur and saturation in frame-based capture. Recent methods use event streams to enhance static 3D reconstruction: E-NeRF [13] imposes event-generation consistency on NeRF, while E2NeRF [31] and Ev-DeblurNeRF [4] leverage the EDI model [28] to recover sharp latent frames for pose and radiance optimization. In explicit representations, EvaGaussians [52], EaDeblur-GS [43], and DiET-GS [16] inject event-driven temporal priors into 3D Gaussian Splatting to refine motion, event consistency, and appearance. Overall, these studies demonstrate that event signals provide strong physical constraints that substantially improve temporal sharpness and geometric stability under challenging conditions.

For dynamic scenes, event signals have recently been incorporated into 4D reconstruction pipelines. Several works [8, 24, 36, 45, 54] combine RGB and event modalities within 4D Gaussian Splatting frameworks to better capture fast motion, typically by using monocular [24, 45, 54] or multi-camera [8, 36] DAVIS setups to provide temporally dense constraints. However, these systems still rely on low-resolution event sensors and constrained capture configurations, which limit the achievable reconstruction quality and practical applicability. In contrast, we introduce a high-resolution, hybird RGB–Event multi-view system together with a tailored reconstruction framework that fully leverages event cues for motion-aware Gaussian modeling.

3. Method

3.1. Preliminaries

Event Camera Model. Event camera monitors per-pixel changes in the logarithmic image intensity. Instead of recording full image frames, each pixel triggers an event whenever the change in log-brightness exceeds a contrast threshold C. For a pixel located at $\mathbf{x}$, an event $e = (\mathbf{x}, t, p)$ is fired at time t with polarity $p \in \{-1, +1\}$ when

$$L(\mathbf{x}, t) - L(\mathbf{x}, t - \Delta t) = p\,C, \tag{1}$$

where Δt denotes the time since the previous event at the same pixel. The sensor thus produces a sparse but highly temporally resolved stream of events, each indicating the sign and timing of local brightness changes rather than absolute intensities. For clarity, we denote by $e_{\mathbf{x},t}$ the event observed at position $\mathbf{x}$ and time t.

4D Gaussian Splatting. We follow the 4D Gaussians framework [44] to represent dynamic scenes. Instead of defining Gaussians directly in a 4D space, 4D Gaussians maintains a canonical set of 3D Gaussians $\mathcal{G}$ and models motion via a learned deformation field. Given a timestamp t, a spatio–temporal encoder $\mathcal{H}(G, t)$ extracts multi-resolution 4D features for each Gaussian $G \in \mathcal{G}$, and a Gaussian deformation decoder $\mathcal{D}$ predicts its time-dependent deformation: $\Delta G = \mathcal{D}(\mathcal{H}(G, t))$. The posed Gaussian at time t is then computed as: $G' = G + \Delta G$, where ΔG includes offsets of the Gaussian attributes (position, scaling, rotation, and SH coefficients).

Finally, the deformed Gaussians $\{G'\}$ are rendered via standard differentiable Gaussian splatting. Given a pixel

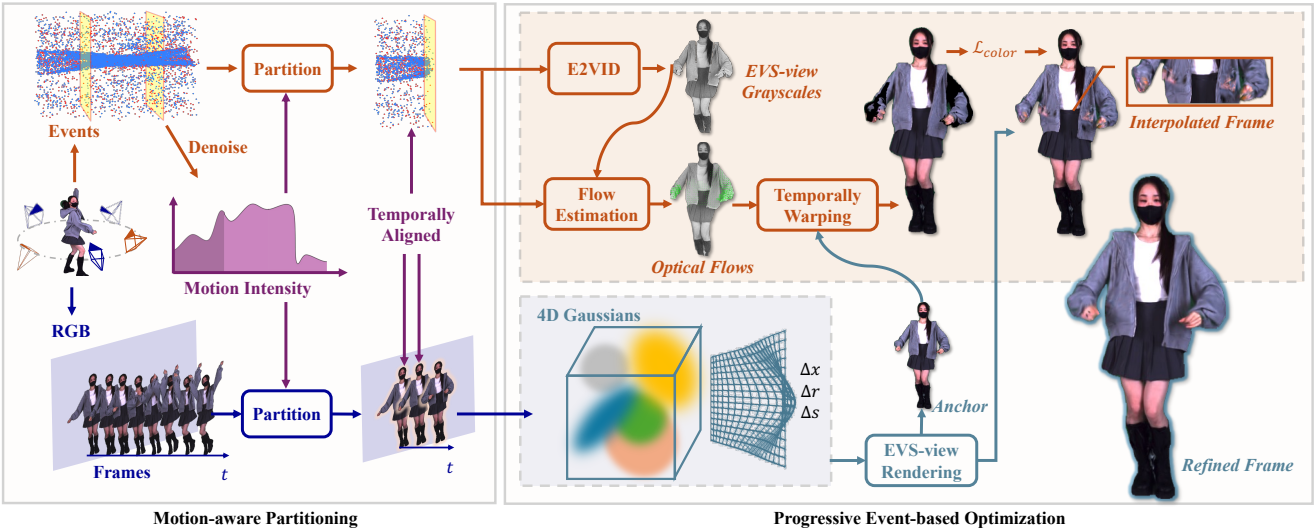

Figure 2. **Overview of 4D E-SloMo.** *Left:*To achieve better reconstruction quality, 4D E-SloMo first performs adaptive segmentation of the input mixed-stream data based on the characteristics of the event signals, while preserving temporal alignment. *Right:* We first train a deformable set of 4D Gaussians using only RGB data, and then guide the color space by converting events into EVS-view optical flow, enabling high-frame-rate and sharp reconstruction of fast motion.

(u, v) at time t, the rendered color is

$$C(u,v,t) = \sum_{i=1}^{N} \alpha_i \, w_i(u,v,t) \, c_i(t) \prod_{j<i} \big(1 - \alpha_j \, w_j(u,v,t)\big), \tag{2}$$

where $w_i(u, v, t)$ is the projected Gaussian weight at pixel (u, v), α_i is its opacity, and $c_i(t)$ is its time-dependent color predicted by 4D spherical harmonics.

3.2. Overview

The input to our system consists of hardware-synchronized multi-view RGB images $\{C^k\}_{k=1}^{K_1}$ and event streams $\{E^k\}_{k=1}^{K_2}$, captured by K_1 RGB cameras and K_2 event cameras. For each timestamp t, the k-th observations are denoted as $C^k(t)$ and $E^k(t)$. Unlike DAVIS-style setups, our hybrid system provides *either* RGB or event data at each view rather than spatially aligned RGB–event pairs. This design removes the need for per-pixel cross-modal alignment and allows us to employ high-resolution RGB and event sensors, enabling higher imaging quality.

Standard RGB cameras operate at low frame rates (*e.g.*, 30 FPS), causing rapid motion to exceed their temporal sampling capacity and leading to severe loss of fine geometric details. In contrast, event cameras offer microsecond-level motion cues, while RGB cameras provide high-quality appearance. Our goal is to reconstruct a temporally dense 4D representation that recovers these missing details by jointly leveraging the complementary strengths of RGB and event modalities in our hybrid system.

The overall pipeline is illustrated in Fig. 2. Our method integrates RGB and event observations through three complementary components, each addressing a different challenge in reconstructing fast 4D dynamics:

- **Motion-Aware Partitioning (Sec. 3.3)** Event streams encode instantaneous motion magnitude. We exploit this property to partition the input sequence into motion-consistent segments, where each segment can be reliably represented by a single 4D deformation field. This prevents a model from being forced to explain overly large or highly non-linear motion within one deformation space, avoiding tearing, drift, and excessive smoothing artifacts.
- **Flow-based Event Loss (Sec. 3.4)** RGB and event cameras capture complementary cues: RGB provides high-quality appearance, while events offer microsecond-level temporal information. To leverage this, we convert short event windows into dense optical flow using a two-stage pipeline: (1) a recurrent E2VID module that reconstructs latent event-conditioned intensity frames, and (2) a lightweight video optical-flow estimator computing flow between consecutive latent frames. The resulting event-derived flow enables backwarping sharp event-driven frames to RGB timestamps, providing direct photometric supervision for blurred RGB views without relying solely on log-intensity or structural event losses that are inherently ill-posed.
- **Progressive Event-based Optimization (Sec. 3.5)** Training a full spatio-temporal 4D Gaussian model from scratch is unstable under rapid motion. We therefore adopt a coarse-to-fine optimization strategy: we first learn a stable static appearance prior using RGB supervision, and then gradually introduce temporal parameters and event-based losses to refine motion trajectories and re-

cover high-frequency dynamic details. This progressive schedule avoids early overfitting to noisy event reconstructions and ensures stable 4D convergence.

Together, these components form an end-to-end reconstruction framework capable of producing photorealistic and temporally coherent 4D reconstructions of extremely fast motions.

3.3. Motion-Aware Partitioning

A single 4D deformation field can model only a limited range of motion. When motion within a temporal window becomes too large or highly non-linear, the canonical scene must stretch excessively, leading to tearing or over-smoothed geometry [48]. To ensure that each 4D volume remains within its expressive range, we partition the sequence into motion-consistent segments.

Event-driven segmentation. Event activity directly reflects motion magnitude: rapid motion produces dense event bursts, while slow motion results in sparse activity. For each candidate segment, we track an aggregated event representation $\mathbf{E}(t)$ and evaluate both the instantaneous motion (event count in a short window) and the accumulated motion span across the segment. When a weighted combination of these quantities exceeds a threshold T, the segment is terminated and a new one begins.

Canonical keyframe selection. 4D Gaussian Splatting models motion relative to a canonical timestamp. For a segment containing frames $\{\tau_i\}$, we identify the canonical time as the one whose event representation is closest to the segment-wide "event mean." For each τ_i, we compute a total event deviation

$$V_i = \sum_{j \neq i} \left\| \mathbf{E}(\tau_i) - \mathbf{E}(\tau_j) \right\|_1 , \tag{3}$$

where $\mathbf{E}(\tau)$ denotes the accumulated event map at timestamp τ. The canonical timestamp is chosen as $\tau^\star = \arg\min_{\tau_i} V_i$, which minimizes the deformation magnitude required to explain the segment.

Boundary smoothing. To avoid visible seams across segments, Gaussians near segment boundaries are duplicated with shared latent features, and temporal smoothness regularization is applied to their deformation parameters.

3.4. Flow-based Event Loss

RGB frames are too sparse in time to describe fast motion: large displacements may occur between two consecutive RGB timestamps, making interpolation ambiguous. Events fill in these temporal gaps, but their measurements differ in contrast and brightness from RGB, making direct intensity matching unreliable. Instead of enforcing event–RGB intensity consistency, we convert event streams into dense optical flow and use this flow to align event-derived "clear" snapshots to RGB timestamps.

Event-to-flow conversion. A short event window is first processed by an E2VID-style recurrent network [32], producing an intermediate intensity snapshot $\hat{I}_{\text{ev}}(t)$ that preserves spatial structure suitable for motion estimation, even though its absolute brightness does not match RGB. To obtain dense motion consistent with RGB appearance, we estimate optical flow using a DOT-inspired formulation [15], modified to operate on consecutive event snapshots. This yields a high-temporal-resolution flow field $\mathbf{F}^k(t_1 \to t_2) \in \mathbb{R}^2$, which represents the pixel displacement observed by event camera k from time t_1 to t_2. This flow captures high-temporal-resolution motion without relying on RGB input.

Flow-guided event warping. Let $C^k(t)$ denote the RGB observation of camera k at an RGB-valid timestamp t. For event views (which do not provide RGB), we obtain their corresponding RGB appearance at timestamp t from our Stage 1&2 reconstruction pipeline (Sec. 3.5), denoted as $\hat{C}^k(t)$. Let $\hat{I}^k_{\text{ev}}(t')$ be the nearest event-derived snapshot in the same view, where t' lies between two RGB timestamps and thus provides additional fine-scale motion information. We estimate an event-only optical flow $\mathbf{F}^k(t \to t')$ that describes the pixel displacement from time t to the event time t'. To propagate the RGB-based appearance to the event timestamp t', we apply a warping operator $W(\cdot)$ based on the predicted flow:

$$\hat{C}^k_{\text{ev}}(t') = W\Big(\hat{C}^k(t),\ \mathbf{F}^k(t \to t')\Big) . \tag{4}$$

The warped image $\hat{C}^k_{\text{ev}}(t')$ represents the scene appearance at time t' according to the microsecond-level motion encoded by the event stream. It therefore provides a temporally dense supervision signal for the rendered color $C^k(u, v, t')$ of the 4D Gaussian model.

Event-based photometric supervision. Given the rendered color $C^k(u, v, t')$ from the 4D Gaussian model (Eq. 2), we supervise it using the event-warped appearance $\hat{C}^k_{\text{ev}}(t')$ obtained in Eq. 4. The flow-based reconstruction loss is defined as: $\mathcal{L}_{\text{flow}} = \left\| C^k(u, v, t') - \hat{C}^k_{\text{ev}}(t') \right\|_1$. This formulation uses events solely as a *motion cue*: events provide dense temporal dynamics, while RGB provides reliable appearance. By supervising the rendered color only through the flow-aligned snapshot, we avoid any reliance on event intensity or contrast assumptions and obtain a clean, temporally aligned photometric signal for 4D reconstruction.

3.5. Progressive Event-based Optimization

Directly optimizing a fully dynamic 4D Gaussian model from scratch is unstable: early iterations suffer from inaccurate color, weak geometry, and noisy temporal gradients that amplify event noise rather than true motion. To stabilize training, we adopt a progressive strategy that first builds a clean RGB-based appearance model, then gradually introduces temporal resolution and event-guided supervision.

Stage 1: Static RGB initialization. We begin by training a purely static 3D Gaussian scene using all RGB frames, with no temporal deformation enabled. The objective is an L_1 photometric loss:

$$\mathcal{L}_{\text{rgb}} = \sum_{k,t,(u,v)} \left\| C^k(u,v,t) - \hat{C}^k(u,v,t) \right\|_1, \quad (5)$$

which establishes reliable geometry and appearance.

Stage 2: RGB-only deformation learning. Next, we activate the deformation field but restrict training to RGB timestamps. This stage allows Gaussians to learn per-frame appearance variations and initializes a reasonable motion space without any event supervision. The loss remains identical to Eq. 5.

Stage 3: Temporal densification and event-guided refinement. We then insert intermediate timestamps obtained from our motion-aware partitioning (Sec. 3.3) and duplicate Gaussians to initialize their parameters. At this stage, RGB supervision alone becomes sparse, so we incorporate two event-driven losses:

(1) Structure-only SSIM loss. To stabilize appearance at inserted timestamps, we use a structure-only SSIM loss:

$$\mathcal{L}_{\text{struct}} = \sum_{k,t,(u,v)} \Big[1 - \text{SSIM}\big(C^k(u,v,t),\ \hat{C}^k(u,v,t)\big) \Big], \quad (6)$$

which encourages consistent local structure while remaining robust to brightness differences between RGB and event-derived images.

(2) Event-based flow loss. Using the flow-guided warping defined in Sec. 3.4, we enforce:

$$\mathcal{L}_{\text{flow}} = \sum_{k,t',(u,v)} \left\| C^k(u,v,t') - \hat{C}^k_{\text{ev}}(u,v,t') \right\|_1, \quad (7)$$

providing microsecond-level temporal alignment unavailable from RGB.

(3) Spatio–temporal TV regularization. To suppress flicker and stabilize the deformation over time, we apply a spatio–temporal total variation term:

$$\mathcal{L}_{\text{tv}} = \sum_{k,t,(u,v)} \sqrt{\|\nabla_{xy} C^k(u,v,t)\|_2^2 + \beta^2 \|\nabla_t C^k(u,v,t)\|_2^2}. \quad (8)$$

Final objective. The total training loss is:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{struct}} \mathcal{L}_{\text{struct}} + \lambda_{\text{flow}} \mathcal{L}_{\text{flow}} + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}}. \quad (9)$$

This progressive schedule ensures that RGB establishes clean geometry and appearance before event signals are introduced, resulting in stable and temporally consistent 4D reconstructions even under extremely fast motion.

4. Experiments

4.1. Datasets

To address the absence of a public multi-view RGB-EVS video dataset, we establish complementary benchmarks in real-world environments for comprehensive evaluation. Our capture rig consists of three frame-based RGB cameras and two event cameras, with the event sensors interleaved among the RGB viewpoints (Fig. 4). The RGB cameras record raw images at 4000×3096 resolution, later resized to 1280×720, while the Prophesee event sensors output high-resolution polarity streams at 1280×720. All devices are synchronized via a global hardware trigger, and a single host PC controls every sensor to ensure tight timing. The RGB cameras operate at 30 FPS. To minimize illumination-induced event noise, we darken the studio and illuminate the scene using two controlled, front-facing LED panels. We capture 6 sequences of roughly 10 s each, focusing on fast human dance motions.

RGB intrinsics and extrinsics are estimated per camera using OpenCV stereo calibration with a classical chessboard pattern. For event cameras, which do not produce standard intensity images, we follow the e2calib procedure [26]: moving-chessboard event streams are first converted into grayscale frames and then calibrated using the same pipeline. We further denoise the event streams by applying a Background Activity Filter (BAF), removing isolated events lacking spatio–temporal support. An event is retained only if it has at least two neighbors of the same polarity within a spatial radius of $r = 3$ pixels and a temporal window of $\Delta t = 30\,\mu$s.

4.2. Implementation Details

Initialization. Under sparse viewpoints, we initialize a 3D Gaussian representation with NoPoSplat [51]. NoPoSplat takes two RGB images and their relative pose as input and directly infers a 3D Gaussian set. We select the pair of temporal keyframes with the largest baseline (maximum relative displacement) among available RGB frames, and feed this pair to NoPoSplat. The resulting dense set of 3D Gaussians is uniformly subsampled, yielding a subset that serves as our initialization for subsequent training.

Training Details. The entire three-stage schedule converges in ~ 4 hours on a single RTX A800 for a scene containing 50 k Gaussians. For experiments that require the contrast threshold of the event camera, we set the value to 0.3 after performing calibration in the brightness domain, which is slightly higher than the hardware setting. The overall loss consists of a RGB reconstruction loss, a flow-based loss, a structure only SSIM loss and a total variation regularization term, weighted by $\lambda_{\text{rgb}} = 1.0$, $\lambda_{\text{flow}} = 1.0$, $\lambda_{\text{struct}} = 0.005$ and $\lambda_{\text{tv}} = 0.005$.

Figure 3. Qualitative results on real dataset.

Table 1. Quantitative comparisons on real dataset. Higher PSNR-ROI/PSNR/SSIM and lower LPIPS are better.

Real sequences												
Methods	**K-pop 0**				**K-pop 0 (faster part)**				**K-pop 1**			
	PSNR-ROI↑	PSNR↑	SSIM↑	LPIPS↓	PSNR-ROI↑	PSNR↑	SSIM↑	LPIPS↓	PSNR-ROI↑	PSNR↑	SSIM↑	LPIPS↓
DynEventNeRF	19.8	23.5	0.60	0.449	18.6	19.4	0.59	0.448	18.6	29.8	0.60	0.464
4D GS	28.2	31.1	0.81	0.272	26.2	30.3	0.81	0.270	26.8	30.2	0.80	0.273
4D Gaussians	31.0	34.3	0.80	0.266	27.9	32.7	0.82	0.267	29.7	33.0	0.82	0.265
4D Gaussians$_{w/\ E}$	30.8	34.1	0.78	0.264	28.3	32.9	0.81	0.266	29.5	33.0	0.83	0.255
Ours	**32.4**	**36.0**	**0.83**	**0.257**	**30.7**	**35.4**	**0.83**	**0.260**	**31.3**	**34.4**	**0.84**	**0.254**
Methods	**K-pop 2**				**Contemporary (blue dress)**				**Contemporary (pink dress)**			
	PSNR-ROI↑	PSNR↑	SSIM↑	LPIPS↓	PSNR-ROI↑	PSNR↑	SSIM↑	LPIPS↓	PSNR-ROI↑	PSNR↑	SSIM↑	LPIPS↓
DynEventNeRF	19.8	22.4	0.61	0.447	12.6	14.4	0.58	0.349	fail			
4D GS	26.2	30.1	0.80	0.271	24.2	26.8	0.80	0.252	24.8	27.8	0.79	0.285
4D Gaussians	28.9	32.7	0.83	0.255	25.4	27.6	0.82	0.240	27.0	29.5	0.82	0.274
4D Gaussians$_{w/\ E}$	29.7	33.4	0.82	0.256	25.6	27.9	**0.83**	0.245	27.0	29.6	0.79	0.268
Ours	**31.1**	**34.6**	**0.88**	**0.248**	**26.1**	**28.2**	0.82	**0.240**	**27.6**	**30.0**	**0.83**	**0.268**

4.3. Metrics

We evaluate reconstruction quality using PSNR, SSIM, and LPIPS. Given that our primary focus is the fidelity of dynamic motion, we additionally report PSNR-ROI, a region-of-interest variant that measures reconstruction accuracy specifically within motion areas.

Our real-world dataset is captured at 30 FPS. For training, we subsample the sequences to 15 FPS, while the interleaved frames are held out as unseen ground truth. This setup allows us to directly assess the model's ability to reconstruct motion at frames it has never observed, without temporal leakage.

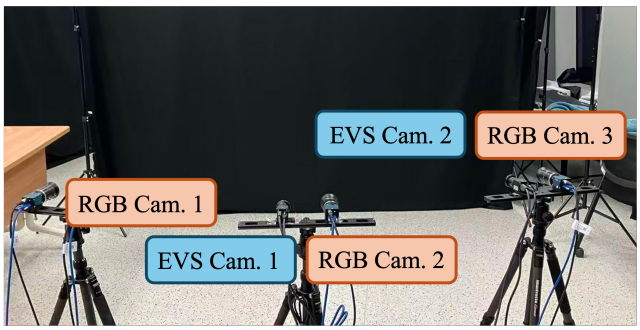

Figure 4. Our hybrid RGB-Event multi-view capture system.

Table 2. **Ablation Study.** We investigate the effectiveness of event-weighted Sampling, flow-based loss, and motion-aware partitioning. Each proposed component contributes to the final performance, and the full model achieves the best results on our real dataset.

Method	PSNR-ROI↑	SSIM↑	LPIPS↓
w/o Flow-based Loss	27.53	0.792	0.296
w/o Motion-aware Partitioning	30.08	0.812	0.267
w/o Event-weighted Sampling	31.45	0.801	0.277
Full	**32.46**	**0.832**	**0.257**

4.4. Baselines

We evaluate two baseline families under identical training and evaluation settings. First, for event-based baseline, we adopt DynamicEventNeRF[36] as the representative method, which is the only open-source approach available for event-based setting. Second, for RGB-only baseline, we use two representative Gaussian-based 4D reconstruction methods: 4D Gaussian Splatting (4D GS) [50] and 4D Gaussians[44], and optionally add a naive event supervision derived directly from the mathematical model of event camera. Since our goal is RGB-consistent reconstruction, we do not employ E2VID[32, 33] or other event-to-intensity pipelines to produce grayscale images for training.

4.5. Comparison

We perform a quantitative comparison in Tab. 1 and a qualitative one in Fig. 3, evaluating our method against several state-of-the-art baselines on our capture dataset. In the K-pop 0 (Faster part) scenario, which features the most rapid motion, our method achieves a PSNR-ROI of 30.7, outperforming all baselines. The closest competitor, 4D Gaussians with Event Loss, reaches only 28.3 PSNR-ROI, showing our clear advantage in reconstructing fast motion. Our event-weighted sampling and motion-aware partitioning help focus on rapidly moving regions, boosting reconstruction quality in high-speed dynamics. Qualitative results in Fig. 3 confirm these findings: 4D E-SloMo produces sharper reconstructions with reduced blur, particularly around fast-moving limbs and clothing, whereas baselines like 4D GS and 4D Gaussians with Event Loss exhibit noticeable blur and artifacts.

4.6. Ablation study

We ablate our method on event-weighted sampling, flow-based loss, and motion-aware partitioning. The quantitative results are presented in Tab. 2. Without the flow-based loss, the model loses motion consistency supervision, causing blurred boundaries and temporal misalignment in fast-moving regions.

Finally, disabling motion-aware partitioning (i.e., adopting a unified 4D representation for the entire sequence) causes the model to underfit scenes with large motion amplitude, since a single representation cannot effectively capture spatially varying motion dynamics once deformation exceeds its representational capacity. Overall, these results confirm that both our event-driven supervision and decomposition strategies are crucial for translating event signals into coherent 4D motion representations, and all proposed components contribute meaningfully to the final performance.

5. Conclusion

In this work, we introduce a novel hybrid capture system for high-speed dynamic scene reconstruction, featuring a flow-based event loss, a motion-aware representation partitioning strategy, and dedicated training refinements. Our system is capable of capturing high-fidelity 3D reconstructions of complex, fast-moving scenes, achieving sub-millisecond temporal resolution and maintaining manageable bandwidth and hardware cost. It achieves superior results in our real datasets, demonstrating great improvements in motion detail recovery compared to existing methods.

However, our system still faces several limitations. While we demonstrate the effectiveness of progressive training, the scalability of our approach to even larger scenes and more complex motions remains an open challenge. Moreover, the accuracy of reconstruction is still dependent on the alignment of the event and RGB cameras, requiring further optimization for extreme scene deformations and more diverse environmental conditions. Future work will focus on improving the robustness of our model to diverse lighting conditions and enhancing its capability for real-time applications.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory. This study was supported in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd.,

a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. The work is supported by the National Key R&D Program of China (No. 2025YFE0201300)

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *European conference on computer vision*, pages 696–712. Springer, 2020. 3

[2] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 us latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 2

[3] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020. 3

[4] Marco Cannici and Davide Scaramuzza. Mitigating motion blur in neural radiance fields with events and frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9286–9296, 2024. 3

[5] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023. 3

[6] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 3

[7] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 3

[8] Chaoran Feng, Zhenyu Tang, Wangbo Yu, Yatian Pang, Yian Zhao, Jianbin Zhao, Li Yuan, and Yonghong Tian. E-4dgs: High-fidelity dynamic reconstruction from the multi-view event cameras. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7356–7365, 2025. 2, 3

[9] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3083–3091, 2015. 3

[10] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 3

[11] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama. A compact dynamic 3d gaussian representation for real-time dynamic view synthesis. In *European Conference on Computer Vision*, pages 394–412. Springer, 2024. 3

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 3

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

[14] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. In *Computer Graphics Forum*, pages 29–43. Wiley Online Library, 2021. 3

[15] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: Connecting the dots. In *CVPR*, 2024. 5

[16] Seungjun Lee and Gim Hee Lee. Diet-gs: Diffusion prior and event stream-assisted motion deblurring 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21739–21749, 2025. 3

[17] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5521–5531, 2022. 2, 3

[18] Zhong Li, Yu Ji, Wei Yang, Jinwei Ye, and Jingyi Yu. Robust 3d human motion reconstruction via dynamic template construction. In *2017 International Conference on 3D Vision (3DV)*, pages 496–505. IEEE, 2017. 3

[19] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8508–8520, 2024. 3

[20] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8508–8520, 2024. 3

[21] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[22] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 2, 3

[23] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Min Yang, Xiao Tang, Feng Zhu, and Yuchao Dai. 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8900–8910, 2024. 3

[24] Qi Ma, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Deformable neural radiance fields using rgb and

event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3590–3600, 2023. 2, 3

[25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[26] Manasi Muglikar, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. How to calibrate your event camera. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2021. 6

[27] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[28] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019. 3

[29] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 3

[30] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 3

[31] Yunshan Qi, Lin Zhu, Yu Zhang, and Jia Li. E2nerf: Event enhanced neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13254–13264, 2023. 3

[32] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 5, 8

[33] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, 2019. 8

[34] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics (ToG)*, 41(4):1–14, 2022. 3

[35] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. Eventnerf: Neural radiance fields from a single colour event camera. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[36] Viktor Rudnev, Gereon Fox, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. Dynamic eventnerf: Reconstructing general dynamic scenes from multi-view rgb and event streams. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4866–4876, 2025. 2, 3, 8

[37] Sara Fridovich-Keil and Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 3

[38] Teresa Serrano-Gotarredona and Bernabé Linares-Barranco. A 128×128 1.5% contrast sensitivity 0.9% fpn 3us latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. *IEEE Journal of Solid-State Circuits*, 48(3):827–838, 2013. 2

[39] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2

[40] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *European Conference on Computer Vision*, pages 246–264. Springer, 2020. 3

[41] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. 2024. 2, 3

[42] Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels. *Advances in Neural Information Processing Systems*, 37:131316–131343, 2025. 2

[43] Yuchen Weng, Zhengwen Shen, Ruofan Chen, Qi Wang, and Jun Wang. Eadeblur-gs: Event assisted 3d deblur reconstruction with gaussian splatting. *arXiv preprint arXiv:2407.13520*, 2024. 3

[44] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 2, 3, 8

[45] Wenhao Xu, Wenming Weng, Yueyi Zhang, Ruikang Xu, and Zhiwei Xiong. Event-boosted deformable 3d gaussians for dynamic scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28334–28343, 2025. 2, 3

[46] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20029–20040, 2024. 2

[47] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. EmerneRF: Emergent spatial-temporal scene decomposition via self-supervision. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[48] Ling Yang, Kaixin Zhu, Juanxi Tian, Bohan Zeng, Mingbao Lin, Hongjuan Pei, Wentao Zhang, and Shuicheng Yan. Widerange4d: Enabling high-quality 4d reconstruction with wide-range movements and scenes. *arXiv preprint arXiv:2503.13435*, 2025. 5

[49] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 20331–20341, 2024. 3

[50] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 8

[51] Botao Ye, Sifei Liu, Haofei Xu, Li Xueting, Marc Pollefeys, Ming-Hsuan Yang, and Peng Songyou. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 6

[52] Wangbo Yu, Chaoran Feng, Jiye Tang, Jiashu Yang, Zhenyu Tang, Xu Jia, Yuchao Yang, Li Yuan, and Yonghong Tian. Evagaussians: Event stream assisted gaussian splatting from blurry images. *arXiv preprint arXiv:2405.20224*, 2024. 3

[53] Ruijie Zhu, Yanzhe Liang, Hanzhi Chang, Jiacheng Deng, Jiahao Lu, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:101790–101817, 2024. 3

[54] Zihao Zou, Ziyuan Qu, Xi Peng, Vivek Boominathan, Adithya Pediredla, and Praneeth Chakravarthula. High-speed dynamic 3d imaging with sensor fusion splatting. *arXiv preprint arXiv:2502.04630*, 2025. 2, 3