AsyncEvGS: Asynchronous Event-Assisted Gaussian Splatting for Handheld Motion-Blurred Scenes

Jun Dai[1*], Renbiao Jin[2*], Bo Xu[2], Yutian Chen[3], Linning Xu[3], Mulin Yu[1], Tianfan Xue[3,1,4†], and Shi Guo[1†]

[1] Shanghai AI Laboratory
[2] Shanghai Jiao Tong University
[3] CUHK MMLab
[4] CPII under InnoHK

[*]Equal contribution. [†]Corresponding author.
jundai332@gmail.com, guoshi@pjlab.org.cn

Abstract. 3D reconstruction methods such as 3D Gaussian Splatting (3DGS) and Neural Radiance Fields (NeRF) achieve impressive photorealism but fail when input images suffer from severe motion blur. While event cameras provide high-temporal-resolution motion cues, existing event-assisted approaches rely on low-resolution sensors and strict synchronization, limiting their practicality for handheld 3D capture on common devices, such as smartphones. We introduce a flexible, high-resolution **asynchronous** RGB–Event dual-camera system and a corresponding reconstruction framework. Our approach first reconstructs sharp images from the event data and then employs a cross-domain pose estimation module based on the Visual Geometry Transformer (VGGT) to obtain robust initialization for 3DGS. During optimization, we employ a structure-driven event loss and view-specific consistency regularizers to mitigate the ill-posed behavior of traditional event losses and deblurring losses, ensuring both stable and high-fidelity reconstruction. We further contribute AsyncEv-Deblur, a new high-resolution RGB–Event dataset captured with our asynchronous system. Experiments demonstrate that our method achieves state-of-the-art performance on both our challenging dataset and existing benchmarks, substantially improving reconstruction robustness under severe motion blur. Project page: https://openimaginglab.github.io/AsyncEvGS/.

Keywords: 3D Gaussian Splatting · Event Camera · Motion Deblurring · 3D Reconstruction

1 Introduction

Neural Radiance Fields (NeRF) [26] and 3D Gaussian Splatting (3DGS) [12] have recently achieved unprecedented photorealism in novel view synthesis. Their success, however, hinges on a collection of high-quality, sharp input images—an

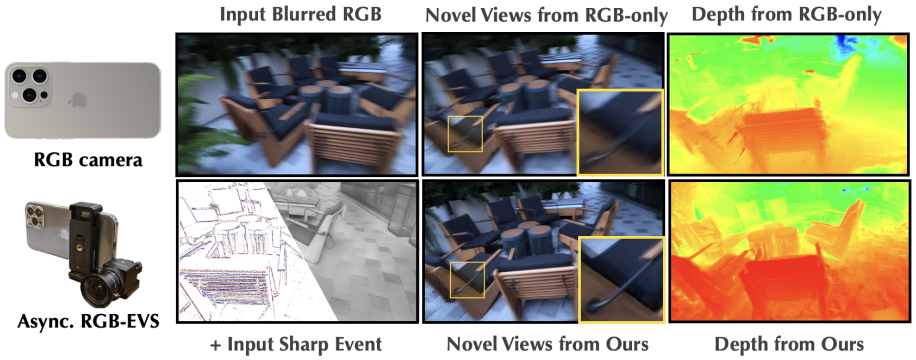


Fig. 1: High-quality 3D reconstruction from severely blurred inputs captured during rapid handheld motion. (Top) Reconstructing from blurred RGB images alone is ill-posed: RGB-only methods (BAGS [29]) fail to resolve motion ambiguity, producing blurry novel views with noticeable artifacts and distorted geometry. (Bottom) We propose a high-resolution asynchronous RGB–EVS system that pairs a handheld RGB camera with an event sensor. Leveraging the sharp, high-frequency cues from events, our method recovers accurate geometry and produces sharp novel views. This shows that *our system can effectively use event signals to boost 3D reconstruction quality for commonly used handheld RGB cameras such as smartphones.*

assumption frequently violated in real-world 3D scanning. Handheld capture, in particular, is often plagued by severe motion blur [16, 23, 47], as also shown in Fig 1. Given the prevalence of handheld devices in 3D capture and robotics, it is critical to enhance reconstruction robustness under motion blur. While computational deblurring 3DGS techniques [23, 29, 41, 47] exist, they grapple with the ill-posed nature of deblurring. Event cameras, in contrast, offer a powerful alternative [1, 10]. With their high temporal resolution and asynchronous measurement of intensity changes, they provide robust motion cues even in the presence of severe blur [9, 11, 25, 30, 33, 38].

However, existing event-assisted 3D reconstruction methods face two critical limitations that preclude their use in common handheld scenarios. (1) **Low Sensor Resolution.** Most methods [11, 14, 24, 30, 31, 33, 34] are built upon low-resolution event sensors (*e.g.*, DAVIS 346 × 260). This resolution, substantially lower than modern multi-megapixel RGB cameras, fundamentally limits the achievable reconstruction fidelity [39]. (2) **Temporal Synchronization Requirement.** While high-resolution stereo setups like LSENeRF [38] address the resolution bottleneck, they mandate rigid, hardware-level temporal synchronization. This reliance on external triggering restricts their use to specialized industrial cameras and is incompatible with prevalent handheld devices (*e.g.*, smartphones, RealSense [37]) where such synchronization is unavailable. This disparity raises a critical question: *Can event cameras be leveraged to deblur 3D reconstructions from common, unsynchronized handheld devices?*

To address these challenges, we introduce a simple but novel, high-resolution **asynchronous** RGB-Event dual-camera rig (Fig. 1). Without hardware synchronization, the rig can be readily paired with a commodity RGB camera

for in-the-wild capture. However, asynchrony breaks a key assumption in prior RGB–Event systems [38]: event-camera poses can no longer be obtained by directly transferring RGB poses via a fixed extrinsic calibration. As a practical solution, we build a two-stage pipeline that couples event-to-intensity reconstruction with cross-modal pose estimation to obtain reliable poses for both RGB and Event. We first reconstruct intensity images from the event stream using E2VID [32]. Although E2VID provides sharp grayscale cues (Fig. 1), COLMAP [35] remains brittle when jointly registering motion-blurred RGB frames and event reconstructions, often yielding fragmented trajectories. Motivated by recent progress in feed-forward 3D foundation models, we leverage VGGT [40] to obtain robust cross-modal pose initialization [6]; using the E2VID reconstructions as an event-derived bridge, this initialization enables stable 3D Gaussian Splatting reconstruction in our asynchronous setting.

Beyond pose initialization, effective data fusion during optimization is also critical. Prior event-assisted methods [30, 33] often rely on *cross-view* supervision. Without direct per-view constraints, this approach renders the appearance estimation problem inherently ill-posed, especially in our asynchronous setting, where each view contains only a single modality, *either* an RGB image *or* an event-derived grayscale observation. We address this by introducing an optimization framework that augments standard event losses [14, 38, 45] with a structure-based loss. Guided by an event-confidence map computed via multi-scale high-frequency consistency, this loss selectively emphasizes reliable event structures while suppressing noisy or uninformative regions, enabling effective exploitation of high-frequency cues from the event. Furthermore, existing deblur modules [23, 31, 47] constrain a blurred observation using an averaging of renderings from neighboring views, which can admit degenerate "compensation" solutions: errors in individual views may cancel out after aggregation and still match the blurred image. To prevent this, we introduce a consistency regularizer for RGB views that encourages neighboring latent appearances to be consistent, reducing artifacts and stabilizing optimization under pose noise.

Since there are no asynchronous event-RGB data for 3D reconstruction, we also have collected **AsyncEv-Deblur**, a new RGB-EVS dataset that we will release. This dataset features diverse scenes captured with our high-resolution setup. Our experiments demonstrate that our method not only excels on this new, challenging dataset but also achieves state-of-the-art performance on public benchmarks. In summary, our contributions are threefold:

- **A Novel High-Resolution Asynchronous System:** We propose the first practical pipeline for high-fidelity 3D reconstruction using a flexible 1280×720 RGB-Event rig, overcoming the low-resolution (*e.g.*, 346×260) limits of prior work.
- **A Robust Cross-Domain Algorithmic Framework:** We introduce a robust initialization pipeline using VGGT for cross-domain pose estimation, and a tailored optimization framework featuring a novel event structure loss and consistency regularizers.

– **A New Benchmark Dataset and SOTA Performance:** We present AsyncEv-Deblur, a new high-resolution RGB-Event dataset, and demonstrate that our method achieves state-of-the-art performance, significantly outperforming existing approaches.

2 Related work

2.1 Deblurring 3D reconstruction

Recent research has actively explored deblurring neural rendering to recover sharp and geometrically consistent 3D scenes from motion-blurred inputs. Conventional neural rendering frameworks such as NeRF [26] and 3D Gaussian Splatting (3DGS) [12] assume static scenes and sharp multi-view images, leading to severe reconstruction artifacts when exposed to motion blur. To address this, several trajectory-based deblurring pipelines have been introduced.

Trajectory-based approaches explicitly model the motion trajectory of the camera or scene during exposure. BAD-NeRF [41], ExBluRF [17], DyBluRF [36], and BAD-Gaussians [47] jointly optimize the latent 3D representation and exposure trajectory by synthesizing blurred renderings through temporal integration of multiple sharp latent images. Later methods further extend this paradigm within the Gaussian Splatting framework. CRiM-GS [18] and CoMoGaussian [19] adopt continuous-time neural ODEs to parameterize camera trajectories, achieving smoother and more flexible motion modeling. BARD-GS [22], MoBGS [2], and MoBluRF [3] enhance dynamic scene modeling by disentangling static and moving regions, enabling temporally coherent novel view synthesis under severe motion blur.

Other methods jointly optimize geometric and radiance attributes to better handle spatially varying blur. Deblur-NeRF [23], PDRF [28], and DP-NeRF [16] incorporate differentiable blur kernels and depth-dependent transformations into the rendering process to simulate the blur formation model. Within explicit Gaussian representations, BAGS [29] and Deblurring 3DGS [15] refine per-Gaussian anisotropy to adaptively encode the blur field, while DeepDeblurRF [8] integrates pretrained 2D deblurring priors into the 3D radiance field optimization. However, due to the ill-posed nature of the blur formation, these methods struggle to handle large motion blurs effectively.

2.2 Event-based deblurring 3D reconstruction

Event-based sensors offer microsecond-level temporal resolution and high dynamic range, making them ideal for mitigating motion blur and lighting saturation. Recent methods leverage event streams to guide 3D reconstruction from degraded inputs. E-NeRF [14] formulates NeRF training using event generation-based supervision, comparing predicted brightness changes with real event streams. E2NeRF [30] and Ev-DeblurNeRF [4] employ the Event-based Double Integral (EDI) model [27] to reconstruct sharp latent frames for pose initialization

Table 1: Comparison of key hardware configurations. We compare against DeblurGS [15], E2NeRF [30], and LSENeRF [38]. *Both* denotes the use of RGB and Event cameras. *Temp Sync.* denotes if strict temporal synchronization is required between the event camera and the RGB camera.

	DeblurGS	E2NeRF	LSENeRF	**Ours**
Cam Type.	RGB	Both	Both	Both
Resolution	600×400	346×260	1280×720	1280×720
Temp Sync.	-	Yes	Yes	No

and consistent radiance learning. Moving to explicit representations, EvaGaussians [45], EaDeblur-GS [43], and DiET-GS [20] integrate event-based temporal priors and EDI-guided supervision into 3DGS optimization, jointly refining motion trajectory, event consistency, and Gaussian attributes for high-fidelity reconstruction. These advances demonstrate that integrating event signals into neural rendering provides physically grounded constraints that effectively mitigate motion blur, ensuring temporally precise and geometrically stable 3D reconstruction in dynamic real-world environments. The capture hardware differences are summarized in Tab. 1. Existing event-assisted systems are limited by low sensor resolution and strict temporal synchronization, restricting their use in practical, high-resolution scenarios such as mobile capture. To address this, we propose an asynchronous RGB-Event solution for high-quality 3D reconstruction.

3 Method

We propose an asynchronous RGB-Event system for high-fidelity 3D reconstruction from motion-blurred RGB images and sharp event data. Our pipeline, illustrated in Fig. 2, is structured around two core stages: 1) a robust cross-domain initialization and pose estimation framework that bypasses traditional SfM tools (e.g., COLMAP), and 2) a specialized optimization framework for 3D Gaussian Splatting. For optimization, we introduce a multi-objective loss function. This includes a deblurring loss for the RGB data, which we augment with a novel **event structure loss** to enforce high-frequency details. Concurrently, we employ a **consistency regularization term** to prevent color degradation. We will detail each of these components in the subsequent sections.

3.1 RGB-Event dual-camera system

Our 3D reconstruction pipeline is fed by a high-resolution, dual-camera capture system (shown in Fig. 1). This design addresses the significant resolution gap between common event sensors (*e.g.*, DAVIS346) and the high-definition images required by NeRF or 3DGS. Our system comprises: (1) a Prophesee EVK-3 HD event camera to capture high-resolution (1280×720) event streams, and (2) a separate RGB camera (*i.e.*, an iPhone 13, set to 1280×720) to provide the information for colorful reconstruction. Critically, our system operates in a

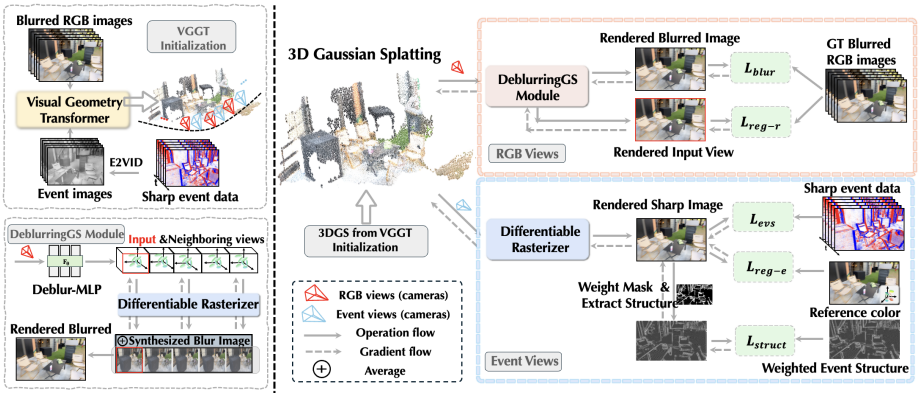


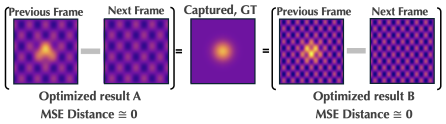
Fig. 2: An overview of our proposed reconstruction pipeline. Our method takes blurred RGB images and sharp event streams as input. We first employ VGGT [40] to process both RGB and event images, providing robust initial camera poses and 3DGS points. The 3DGS representation is then jointly optimized using five key losses, broadly categorized into three groups: **(1) Deblurring Losses:** The blur synthesis loss ($\mathcal{L}_{\text{blur}}$) matches the synthesized blur to the input, while an RGB consistency regularizer ($\mathcal{L}_{\text{reg-r}}$) prevents degradation of the sharp neighboring views. **(2) Event-Guided Losses:** We augment the traditional photometric loss ($\mathcal{L}_{\text{evs}}$), with our novel structure loss ($\mathcal{L}_{\text{struct}}$) to robustly leverage high-frequency event details. **(3) Consistency Loss ($\mathcal{L}_{\text{reg-e}}$):** A color distillation loss ensures that event views match the colors learned from a coarse (Stage 1) 3DGS copy.

much more flexible way; it requires no hardware-synchronization and can use high-resolution RGB sensor (Tab. 1). This is enabled by our novel initialization method (Sec. 3.2), which allows the cameras to be flexibly co-mounted on a simple rigid bracket.

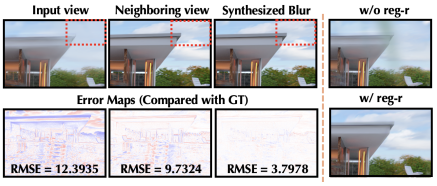
3.2 Camera poses and 3DGS initialization

Estimating camera poses for our dual-camera system is non-trivial. A conventional approach might derive event camera poses from the RGB camera's COLMAP estimates via a pre-calibrated relative extrinsic between the RGB and Event cameras [38]. However, this not only requires meticulous pre-calibration of both the relative extrinsics and COLMAP's global scale but also mandates temporal synchronization. An alternative is to convert event streams to grayscale frames via models like E2VID [32] for subsequent COLMAP processing. However, the joint calibration struggles due to severe motion blur in the RGB frames and the inherent domain gap between the RGB and reconstructed grayscale images.

To ensure flexibility and achieve robust, efficient pose estimation, we leverage VGGT [40] for joint calibration. Benefiting from its strong data priors, VGGT can process challenging, motion-blurred RGB inputs while producing a denser and more accurate initialization for 3D Gaussian Splatting compared to COLMAP. Specifically, we first convert the raw event stream into a sequence



(a) 2D toy example of the ill-posed classical event loss. It renders two consecutive frames and minimizes the difference against acquired event data, which is prone to ambiguity and limits reconstruction quality.



(b) The blur synthesis loss is ill-posed. The model can achieve a low loss by rendering accurate *Neighboring views* while producing artifacts in the (training) *Input view* ("w/o reg."). Our regularization recovers a sharper result ("w/ reg.").

Fig. 3: Illustration of ill-posed problems in our optimization. **(a)** The classical event loss only constrains the intensity *difference* between adjacent frames, providing no absolute supervision and resulting in limited reconstruction quality. **(b)** Synthesizing motion blur by averaging neighboring views can also converge to a degenerate solution; our consistency regularizer effectively mitigates this.

of sharp gray-scale images using E2VID. These images are then post-processed with bilateral denoising and multi-frame brightness equalization. Finally, we feed both the blurred RGB frames and the sharp gray-scale frames into VGGT. This process yields a dense point cloud, which serves as the 3DGS initialization, along with the corresponding camera poses for all input images. Compared to COLMAP, our initialization method is significantly more robust to severe motion blur and provides a more accurate initial geometry as shown in Fig. 7a.

3.3 Event Structure Loss

Conventional event-based 3D reconstructions [30, 33, 38] employ event-based losses, such as the photometric consistency loss, to supervise the change of log-intensity between two adjacent timestamps t_s and t_e:

$$\mathcal{L}_{\text{evs}} = \mathbb{E}_{t_s,t_e}\left[\left\| \log I(t_e) - \log I(t_s) - c\sum E(t_s,t_e)\right\|_2^2\right], \tag{1}$$

where c represents the contrast threshold that triggers events, and E denotes the event signal. This formulation is inherently ill-posed, as it only constrains the *difference* and provides no supervision for the absolute intensity. Such ill-posedness limits the recovery of fine textures, as illustrated in Fig. 3a. While using an event-to-video network (*e.g.*, E2VID [32]) to generate grayscale "ground truth" images I_{GT} for direct supervision can alleviate this ambiguity, the E2VID outputs suffer from significant brightness inconsistencies—both across frames and within weakly textured regions. Naively supervising on these images would bake these artifacts directly into the 3DGS, leading to severe degradation.

To resolve this, we propose an **Event Structure Loss**, $\mathcal{L}_{\text{struct}}$, designed to be robust to both brightness artifacts and potential pose inaccuracies. First, motivated by the fact that events are most reliable at edges, we isolate high-frequency structural information using a structure extractor $\mathrm{S}(\cdot)$. We then derive a confidence map W from the cross-scale consistency of S, where structurally consistent regions receive high confidence and textureless areas are down-weighted. Second,

we must account for small pose inaccuracies from our estimator, which can cause minor view shifts. Therefore, we compute our loss using the Structural Similarity (SSIM) index [42], which is inherently robust to small translations and focuses on structural correctness rather than unstable pixel-wise alignment.

Our final event structure loss is defined as a weighted SSIM, computed as the expectation over all pixels p:

$$\mathcal{L}_{\text{struct}} = 1 - \mathbb{E}_p \left[W(p) \cdot \text{SSIM}\left(\text{S}(I_{\text{ren}}), \text{S}(I_{\text{evs}})\right)(p) \right], \tag{2}$$

where I_{ren} denotes the Gaussian-rendered image, I_{evs} is the E2VID-reconstructed target, $\text{S}(\cdot)$ is the structure extractor (following [7]), $W(p)$ is the confidence weight at pixel p, and $\text{SSIM}(\cdot)(p)$ computes the pixel-wise SSIM value. Note that the luminance term is removed from SSIM to eliminate brightness inconsistencies between RGB and event-based grayscale images.

3.4 Consistency regularization

To handle motion blur in the RGB inputs, we adopt the deblurring strategy from Deblurring 3DGS [15]. This method uses an MLP to estimate per-view offsets and render N sharp neighboring images $I_{i=1}^{N}$ for each RGB view. These images are averaged to synthesize a blurred image $I_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} I_i$, which is supervised by comparing with the captured blurred image I',

$$\mathcal{L}_{\text{blur}} = (1-\alpha) \cdot \|I_{\text{avg}} - I'\|_1 + \alpha \cdot \left(1 - \text{SSIM}(I_{\text{avg}}, I')\right), \tag{3}$$

where α denotes the loss weighting factor.

However, we find this deblurring loss is ill-posed on its own, as it only enforces cross-view constraints. As shown in Fig. 3b, 3DGS may converge to a local optimum, causing the rendered input training views to exhibit severe degradation. The incorporation of event signals further amplifies this instability. To mitigate this ill-posedness, we introduce consistency regularizers on both the RGB and event views, denoted as $\mathcal{L}_{\text{reg-r}}$ and $\mathcal{L}_{\text{reg-e}}$, respectively.

For the RGB views, we impose additional constraints to stabilize the rendered image. We further enforce that the rendered image itself should remain close to both the neighboring sharp frames and the observed blurred input:

$$\mathcal{L}_{\text{reg-r}} = \frac{1}{N}\sum_{i=1}^{N} \left\|g(I_i) - g(I)\right\|_2^2 + \left\|g(I) - I'\right\|_2^2, \tag{4}$$

where $g(\cdot)$ denotes Gaussian blurring, I_i is the i-th neighboring sharp frame estimated by the deblurring MLP module, I' is the blurry observation.

For the event views, we aim to introduce a color constraint to prevent the rendered appearance from drifting. To obtain such a reference color prior, we first train a coarse Gaussian model $\mathcal{G}_{\text{ref}}$ using only the RGB images. Although this model provides poor structural quality, it still provides a rough but reliable color reference. Thus, for an event-view pose P, the corresponding regularizer $\mathcal{L}_{\text{reg-e}}$ is defined as:

$$\mathcal{L}_{\text{reg-e}} = \mathbb{E}_{P \in \mathcal{R}_{\text{evt}}} \left[\left\| I(P) - \mathcal{G}_{\text{ref}}(P) \right\|_2^2 \right], \tag{5}$$

where P denotes a pose from the set of event-view poses $\mathcal{R}_{\text{evt}}$, and $I(P)$ is the rendered image.

4 Experiments

4.1 Implementation Details

Training Details. We implement our method based on the Gsplat framework [44] and adopt the RGB deblurring module from Deblurring 3DGS [15]. In Stage 1, we train a 3DGS using only RGB images, initialized from VGGT-reconstructed 3D points filtered by a 50% confidence threshold. We render $N = 5$ views to synthesize the final blurred image. The model is optimized for $10k$ iterations using Adam [13] to minimize the blur loss $\mathcal{L}_{\text{blur}}$. The learning rates are default setting from Gsplat. Following this, in Stage 2, the 3DGS from Stage 1 is copied to serve as a fixed reference for color supervision. We then train a new 3DGS ($30k$ iters), again initialized from VGGT points, using both RGB and event data. This model is optimized using Adam with our multi-objective loss. We set the loss weights to $\lambda_{\text{blur}} = 1.0$, $\lambda_{\text{struct}} = 0.2$, $\lambda_{\text{evs}} = 0.002$, $\lambda_{\text{reg-r}} = 0.2$, and $\lambda_{\text{reg-e}} = 1.0$. The learning rates are identical to those in Stage 1. Further details regarding the structure extractor and the deblurring MLP are provided in the supplementary material.

Evaluation datasets. We evaluate our method on our newly captured real-world **AsyncEv-Deblur** dataset and a modified **Ev-DeblurBlender** dataset [4]. Our AsyncEv-Deblur dataset contains 7 scenes: *Patio, Bin, Lounge, Bench, Stair, Bus,* and *Wall*. For each scene, we perform a rapid handheld sweep using the RGB camera and the event sensor to capture the blurred inputs at the same time, followed by a slow, stable pass with the RGB camera alone to record high-quality ground-truth images. For the Ev-DeblurBlender dataset, we utilize all four scenes: *factory, pool, tanabata,* and *trolley*. Crucially, as VGGT initialization is vital for reconstruction, we use our VGGT pipeline to re-calibrate all camera poses and 3DGS initializations for both datasets, ensuring an unbiased comparison. Further details are provided in the supplementary material.

Baselines. We evaluate our method against two categories of baselines: RGB-only and RGB-Event fusion methods. For the RGB-only category, we select the original 3DGS [12], BAGS [29], and DeblurringGS [15] as recent state-of-the-art deblurring and reconstruction methods. For the RGB-Event category, finding a directly comparable baseline is challenging due to our use of an asynchronous dual-camera system. Most existing work employs low-resolution, single-camera setups (*e.g.*, DAVIS346). To the best of our knowledge, LSENeRF [38] is the only other work utilizing a dual-camera system. However, its requirement for strict camera synchronization, which our setup lacks, results in incompatible data formats. Therefore, we adapt its event loss configuration to our data format and re-implement it in our codebase. To ensure a fair comparison, all baselines uniformly utilize our camera poses and 3DGS initialization calibrated by VGGT.

Fig. 4: Qualitative comparison on synthetic data, *factory* (top) and *trollet* (bottom). Our method recovers sharp details, such as the stair in the first example, as well as accurate colors, outperforming other event-based and RGB-only methods.

4.2 Experimental Validations

Evaluation results. Tab. 2 and Fig. 4 present the quantitative and qualitative evaluations on the synthetic Ev-DeblurBlender dataset. As evidenced by the poor performance of original 3DGS, reconstruction quality is significantly degraded by motion blur. Baselines relying solely on RGB images (*e.g.*, DeblurringGS) offer limited mitigation for this ill-posed problem. In contrast, incorporating motion cues from event cameras yields higher-quality reconstructions. Our method achieves the best performance across all metrics.

Table 2: Quantitative results on the Ev-DeblurBlender dataset. We color code the best **PSNR**, **SSIM**, and **LPIPS** performances.

Scene	Original 3D GS			BAGS			DeblurringGS			LSENeRF*			Ours		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Factory	18.01	0.514	0.446	19.12	0.634	0.312	21.26	0.620	0.232	20.47	0.707	0.219	**23.18**	**0.830**	**0.165**
Pool	17.24	0.439	0.599	23.84	0.640	0.339	15.20	0.015	0.739	24.22	0.639	0.280	**25.16**	**0.696**	**0.252**
Tanabata	17.51	0.517	0.487	18.31	0.596	0.381	19.58	0.556	0.256	19.60	0.659	0.257	**20.11**	**0.727**	**0.221**
Trolley	18.52	0.609	0.413	20.25	0.722	0.287	21.08	0.659	0.204	20.75	0.753	0.184	**23.49**	**0.854**	**0.135**
Average	17.82	0.520	0.486	20.38	0.648	0.330	19.28	0.463	0.358	21.26	0.689	0.235	**22.99**	**0.777**	**0.193**

Tab. 3 and Fig. 5 further validate our method on the real-world AsyncEv-Deblur dataset. The performance trends are consistent: baselines relying solely on RGB images suffer from blur artifacts, while event-based supervision substantially improves reconstruction quality. A key benefit is that event-based

Table 3: Quantitative results on our AsyncEv-Deblur dataset. We color code the best **PSNR**, **SSIM**, and **LPIPS** performances.

Scene	Original 3D GS			BAGS			DeblurringGS			LSENeRF*			Ours		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Patio	22.59	0.744	0.382	23.41	0.779	0.302	23.07	0.628	0.289	23.47	0.779	0.273	**24.45**	**0.835**	**0.223**
Bin	22.26	0.804	0.412	25.01	0.819	0.316	23.69	0.633	0.281	**25.88**	0.827	**0.258**	25.67	**0.829**	0.265
Lounge	22.20	0.778	0.434	24.14	0.834	0.347	**26.34**	0.764	**0.182**	25.47	0.852	0.256	25.91	**0.881**	0.199
Bench	22.95	0.815	0.373	23.03	0.824	0.326	24.19	0.684	0.192	24.48	0.846	0.217	**27.77**	**0.896**	**0.176**
Stair	24.89	0.833	0.376	26.51	0.863	0.294	27.95	0.792	0.173	26.25	0.868	0.202	**28.43**	**0.900**	**0.169**
Bus	20.53	0.737	0.466	22.06	0.762	0.403	23.43	0.615	0.255	22.14	0.760	0.295	**23.94**	**0.808**	**0.251**
Wall	23.42	0.707	0.466	24.12	0.704	0.242	18.01	0.312	0.513	25.14	0.749	0.225	**25.82**	**0.785**	**0.224**
Average	22.69	0.774	0.416	24.04	0.798	0.319	23.81	0.633	0.269	24.69	0.812	0.247	**26.00**	**0.847**	**0.215**

supervision (shared by LSENeRF [38] and our method) effectively mitigates the edge artifacts introduced by the deblurring module. However, our proposed event structure loss provides more direct supervision than classical event losses. It avoids the ill-posedness associated with calculations between adjacent frames and enables a more comprehensive utilization of the event signals, leading to superior reconstruction quality.

Complementary strengths of RGB and Event modalities. A core design principle of our method is to exploit the complementary strengths inherent to each sensor modality. As illustrated in Fig. 6, reconstructing from RGB images alone preserves color fidelity but yields severely blurred results, since the RGB camera inevitably suffers from motion blur during rapid handheld capture. Conversely, the event camera is inherently blur-free and thus captures sharp, high-frequency textures, yet its reconstructions lack color information entirely. Our method effectively fuses both modalities: the event structure loss transfers fine-grained details from the event stream, while the color consistency regularization preserves the rich chromatic information from the RGB frames. The result is a high-quality 3D reconstruction that simultaneously achieves color-accurate appearance and sharp structural detail.

VGGT Initialization. Obtaining robust camera poses from our cross-domain data, consisting of blurred RGB and event-reconstructed grayscale images, is a key bottleneck for traditional SfM pipelines like COLMAP. As shown in Tab. 4, which details the percentage of images each method successfully registered, COLMAP exhibits low registration rates on the blurred RGB frames, failing to provide a complete set of camera poses. In contrast, our VGGT-based initialization demonstrates strong robustness, successfully registering all RGB and event-reconstructed frames and demonstrating its generalization to these challenging, heterogeneous sources. VGGT's advantage also extends to the quality of the initial point cloud. As illustrated in Fig. 7a, the COLMAP initialization leads to incorrect camera pose estimations and a sparse and noisy point cloud. This provides a poor initialization that is insufficient to guide the downstream optimization. In contrast, VGGT directly produces a dense and spatially coherent point cloud even from the blurred inputs. This dense initialization is crucial for

Fig. 5: Qualitative comparison on real-world camera motion blur, *Patio* (top) and *Bus* (bottom). Our method recovers high frequency details, such as the text in *Patio* and the logo in *Bus*.

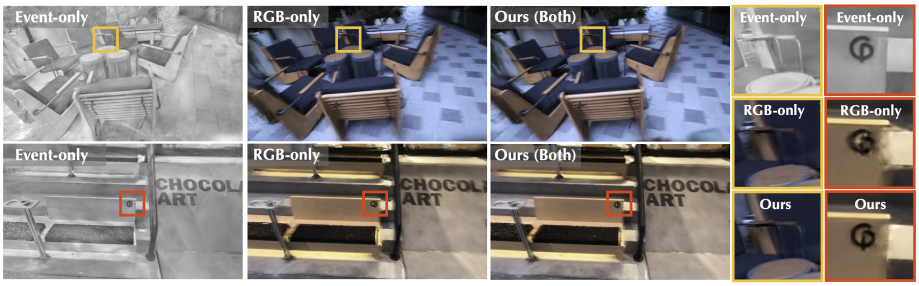


Fig. 6: Qualitative ablation on input modalities. **Event-only** reconstruction captures fine-grained structural details but lacks color information. **RGB-only** reconstruction preserves color fidelity but suffers from severe blur artifacts. **Ours (Both)** combines both modalities, achieving sharp details with faithful color reproduction. Zoom-in patches (right) highlight the complementary strengths of each modality.

preserving geometric continuity and constraining the early 3DGS optimization. Although these initial poses are sufficiently accurate to bootstrap the reconstruction, they are not perfect, so we jointly refine all poses during the 3DGS optimization stage.

Table 4: Ablation studies and comparisons. **(a)** Pose estimation success rates (%) for COLMAP and VGGT. **(b)** Loss component ablation on the synthetic dataset (Average). **(c)** Comparison with 2D deblurring + 3DGS pipelines (Average). We color code the best **PSNR↑**, **SSIM↑**, and **LPIPS↓**.

(a) Pose Estimation Success Rates

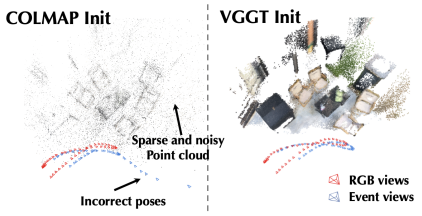
Method	Cam	Patio	Bin	Lounge	Bench	Stair	Bus	Wall
COLMAP	RGB	96	84	44	36	64	48	92
	Event	85	100	100	100	100	95	100
VGGT	RGB	100	100	100	100	100	100	100
	Event	100	100	100	100	100	100	100

(b) Loss Component Ablation

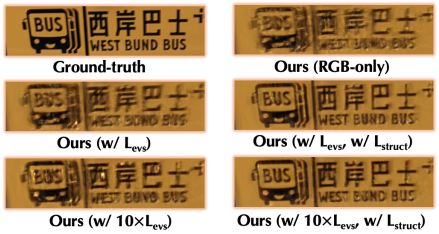
Configuration	PSNR↑	SSIM↑	LPIPS↓
w/o pose optim.	21.26	0.691	0.217
w/o $\mathcal{L}_{\text{reg-r}}$	22.33	0.773	0.199
w/o $\mathcal{L}_{\text{struct}}$	22.26	0.751	0.219
w/o $\mathcal{L}_{\text{reg-e}}$	22.78	0.777	0.198
LSENeRF* [38]	21.26	0.689	0.235
Ours (Full)	**22.99**	**0.777**	**0.193**

(c) 2D Deblurring + 3DGS

Method	PSNR↑	SSIM↑	LPIPS↓
NAFNet [5] + 3DGS	22.72	0.778	0.399
Restormer [46] + 3DGS	22.79	0.778	0.403
ShiftNet [21] + 3DGS	21.84	0.750	0.415
Ours	**26.00**	**0.847**	**0.215**



(a) Comparison of initialization methods. COLMAP (left) fails under motion blur, producing sparse, noisy points. Our VGGT-based method (right) robustly estimates accurate poses and generates a dense point cloud.



(b) Qualitative ablation of our event loss. RGB-only or traditional $\mathcal{L}_{\text{evs}}$ fails to recover sharp details; increasing its weight (10×) does not help. Only our $\mathcal{L}_{\text{struct}}$ recovers fine, sharp text.

Fig. 7: Qualitative analysis of key components. **(a)** Our VGGT-based initialization is robust to severe motion blur and cross-domain inputs, providing dense, high-quality 3DGS initialization compared to COLMAP. **(b)** Our proposed event structure loss $\mathcal{L}_{\text{struct}}$ successfully incorporates high-frequency event details, outperforming the classical event loss.

Impact of loss functions. We conduct a comprehensive ablation study to validate the design of our multi-objective loss function, with quantitative results in Tab. 4 and qualitative insights in Fig. 7b. Our analysis confirms that our full model significantly outperforms all ablated variants. We validate our two primary algorithmic contributions: the two-part color consistency regularizer ($\mathcal{L}_{\text{reg-r}}, \mathcal{L}_{\text{reg-e}}$) and the event structure loss ($\mathcal{L}_{\text{struct}}$). Quantitatively, removing either of the color regularization components degrades performance, validating our claim in Sec. 3.4 that they are crucial for stabilizing the deblurring module and maintaining global color consistency. Disabling our $\mathcal{L}_{\text{struct}}$ results in an even more significant performance drop, confirming its vital role in integrating high-frequency event details. Qualitatively, Fig. 7b reinforces our motivation for $\mathcal{L}_{\text{struct}}$ (from Sec. 3.3). Reconstructing with only RGB data fails to recover sharp

details from the motion blur. Adding the traditional event loss provides some structural guidance, but the result remains ill-defined. In contrast, our proposed $\mathcal{L}_{\text{struct}}$ successfully harnesses the high-frequency event data to restore sharp, fine-grained text, proving its superiority over traditional event-based supervision. We also conduct ablations on the pose optimization, as it is important for the reconstruction. The results confirm that both components indeed provide substantial improvements to the final reconstruction quality.

Comparison with 2D deblurring pipelines. An alternative to our end-to-end approach is to first apply a 2D deblurring network per frame and then reconstruct with standard 3DGS. As shown in Tab. 4, this pipeline consistently underperforms our method by a large margin, despite using state-of-the-art deblurring models [5, 21, 46]. The gap arises because (1) 2D deblurring cannot fully recover sharp details under severe motion blur, and (2) per-frame processing provides no multi-view consistency guarantee, leading to degraded 3D reconstruction.

Resolution flexibility. Our asynchronous dual-camera setup does not require the two sensors to share the same resolution. To verify this, we halve the RGB resolution to 640×360 while keeping the event camera at 1280×720. As shown in Fig. 8, the pipeline still produces reasonable reconstructions, confirming its flexibility. Nevertheless, the quality degradation compared to the full-resolution setting demonstrates the advantage of our high-resolution system in achieving superior reconstruction quality.

Fig. 8: Reconstruction with mismatched resolutions.

5 Conclusions

We introduce a novel, flexible, high-resolution asynchronous RGB-Event dual-camera system that effectively leverages blurry RGB images and high-frame-rate event signals for high-quality 3D reconstruction. Our approach addresses the critical initialization bottleneck—where standard SfM (*e.g.*, COLMAP) fails due to motion blur—by leveraging VGGT for robust cross-domain pose estimation. To optimize the 3DGS representation, we augment traditional event-based losses with a novel event structure loss to robustly harness high-frequency motion details. Furthermore, we introduce a crucial two-part consistency regularizer to prevent deblurring artifacts and distill color to event-only views. This system design facilitates more efficient scene acquisition. Extensive evaluations on both synthetic and real-world datasets demonstrate that our method achieves state-of-the-art reconstruction quality, significantly outperforming existing baselines.

References

1. Bauersfeld, L., Scaramuzza, D.: A monocular event-camera motion capture system (2025), https://arxiv.org/abs/2502.12113
2. Bui, M.Q.V., Park, J., Bello, J.L.G., Moon, J., Oh, J., Kim, M.: Mobgs: Motion deblurring dynamic 3d gaussian splatting for blurry monocular video. arXiv preprint arXiv:2504.15122 (2025)
3. Bui, M.Q.V., Park, J., Oh, J., Kim, M.: Moblurf: Motion deblurring neural radiance fields for blurry monocular video. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
4. Cannici, M., Scaramuzza, D.: Mitigating motion blur in neural radiance fields with events and frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9286–9296 (2024)
5. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV (2022)
6. Chen, Y., Potamias, R.A., Ververas, E., Song, J., Deng, J., Lee, G.H.: Deep gaussian from motion: Exploring 3d geometric foundation models for gaussian splatting. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
7. Chen, Z., Wang, Y., Cai, X., You, Z., Lu, Z., Zhang, F., Guo, S., Xue, T.: Ultrafusion: Ultra high dynamic imaging using exposure fusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 16111–16121 (June 2025)
8. Choi, H., Yang, H., Han, J., Cho, S.: Exploiting deblurring networks for radiance fields. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6012–6021 (2025)
9. Deguchi, H., Masuda, M., Nakabayashi, T., Saito, H.: E2gs: Event enhanced gaussian splatting (2024), https://arxiv.org/abs/2406.14978
10. Gehrig, D., Scaramuzza, D.: Low latency automotive vision with event cameras (2024)
11. Huang, J., Dong, C., Chen, X., Liu, P.: Inceventgs: Pose-free gaussian splatting from a single event camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
12. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015), http://arxiv.org/abs/1412.6980
14. Klenk, S., Koestler, L., Scaramuzza, D., Cremers, D.: E-nerf: Neural radiance fields from a moving event camera. IEEE Robotics and Automation Letters **8**(3), 1587–1594 (2023)
15. Lee, B., Lee, H., Sun, X., Ali, U., Park, E.: Deblurring 3d gaussian splatting. In: European Conference on Computer Vision. pp. 127–143. Springer (2024)
16. Lee, D., Lee, M., Shin, C., Lee, S.: Dp-nerf: Deblurred neural radiance field with physical scene priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12386–12396 (2023)
17. Lee, D., Oh, J., Rim, J., Cho, S., Lee, K.M.: Exblurf: Efficient radiance fields for extreme motion blurred images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17639–17648 (2023)

18. Lee, J., Kim, D., Lee, D., Cho, S., Lee, M., Lee, S.: Crim-gs: Continuous rigid motion-aware gaussian splatting from motion-blurred images. arXiv preprint arXiv:2407.03923 (2024)
19. Lee, J., Kim, D., Lee, D., Cho, S., Lee, M., Lee, W., Kim, T., Wee, D., Lee, S.: Comogaussian: Continuous motion-aware gaussian splatting from motion-blurred images. arXiv preprint arXiv:2503.05332 (2025)
20. Lee, S., Lee, G.H.: Diet-gs: Diffusion prior and event stream-assisted motion deblurring 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21739–21749 (2025)
21. Li, D., Shi, X., Zhang, Y., Cheung, K.C., See, S., Wang, X., Qin, H., Li, H.: A simple baseline for video restoration with grouped spatial-temporal shift. In: CVPR. pp. 9822–9832 (2023)
22. Lu, Y., Zhou, Y., Liu, D., Liang, T., Yin, Y.: Bard-gs: Blur-aware reconstruction of dynamic scenes via gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16532–16542 (2025)
23. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblur-nerf: Neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12861–12870 (2022)
24. Ma, Q., Paudel, D.P., Chhatkuli, A., Gool, L.V.: Deformable neural radiance fields using rgb and event cameras (2023)
25. Matta, G.R., Reddypalli, T., Mitra, K.: Besplat: Gaussian splatting from a single blurry image and event stream. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 917–927 (February 2025)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
27. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: 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. pp. 6820–6829 (2019)
28. Peng, C., Chellappa, R.: Pdrf: Progressively deblurring radiance field for fast and robust scene reconstruction from blurry images. arXiv preprint arXiv:2208.08049 (2022)
29. Peng, C., Tang, Y., Zhou, Y., Wang, N., Liu, X., Li, D., Chellappa, R.: Bags: Blur agnostic gaussian splatting through multi-scale kernel modeling. In: European Conference on Computer Vision. pp. 293–310. Springer (2024)
30. Qi, Y., Zhu, L., Zhang, Y., Li, J.: E2nerf: Event enhanced neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13254–13264 (2023)
31. Qi, Y., Zhu, L., Zhao, Y., Bao, N., Li, J.: Deblurring neural radiance fields with event-driven bundle adjustment. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 9262–9270 (2024)
32. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2019)
33. Rudnev, V., Elgharib, M., Theobalt, C., Golyanik, V.: Eventnerf: Neural radiance fields from a single colour event camera. In: Computer Vision and Pattern Recognition (CVPR) (2023)
34. Rudnev, V., Fox, G., Elgharib, M., Theobalt, C., Golyanik, V.: Dynamic eventnerf: Reconstructing general dynamic scenes from multi-view rgb and event streams. CVPR Workshop on Event-based Vision (2025)

35. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
36. Sun, H., Li, X., Shen, L., Ye, X., Xian, K., Cao, Z.: Dyblurf: Dynamic neural radiance fields from blurry monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7517–7527 (2024)
37. Tadic, V., Odry, A., Kecskes, I., Burkus, E., Kiraly, Z., Odry, P.: Application of intel realsense cameras for depth image generation in robotics. WSEAS Transac. Comput **18**, 2224–2872 (2019)
38. Tang, W.Z., Rebain, D., Derpanis, K.G., Yi, K.M.: Lse-nerf: Learning sensor modeling errors for deblured neural radiance fields with rgb-event stereo. In: 2025 International Conference on 3D Vision (3DV). pp. 534–543. IEEE (2025)
39. Wang, C., Wu, X., Guo, Y.C., Zhang, S.H., Tai, Y.W., Hu, S.M.: Nerf-sr: High-quality neural radiance fields using super-sampling. arXiv (2021)
40. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
41. Wang, P., Zhao, L., Ma, R., Liu, P.: Bad-nerf: Bundle adjusted deblur neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4170–4179 (2023)
42. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
43. Weng, Y., Shen, Z., Chen, R., Wang, Q., Wang, J.: Eadeblur-gs: Event assisted 3d deblur reconstruction with gaussian splatting. arXiv preprint arXiv:2407.13520 (2024)
44. Ye, V., Li, R., Kerr, J., Turkulainen, M., Yi, B., Pan, Z., Seiskari, O., Ye, J., Hu, J., Tancik, M., Kanazawa, A.: gsplat: An open-source library for gaussian splatting. Journal of Machine Learning Research **26**(34), 1–17 (2025)
45. Yu, W., Feng, C., Tang, J., Yang, J., Tang, Z., Jia, X., Yang, Y., Yuan, L., Tian, Y.: Evagaussians: Event stream assisted gaussian splatting from blurry images. arXiv preprint arXiv:2405.20224 (2024)
46. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022)
47. Zhao, L., Wang, P., Liu, P.: Bad-gaussians: Bundle adjusted deblur gaussian splatting. In: European Conference on Computer Vision. pp. 233–250. Springer (2024)

A Working principle of Event camera

Unlike conventional cameras that capture full frames at a fixed rate, an event camera, also known as a Dynamic Vision Sensor (DVS), is a bio-inspired visual sensor that operates asynchronously. It independently monitors the change in logarithmic intensity for each pixel $\mathbf{x} = (x, y)$.

When the change in log intensity $\mathcal{I}(\mathbf{x}, t)$ at time t exceeds a predefined contrast threshold Θ compared to the value at the last event $t - \Delta t$ for that pixel, it asynchronously triggers an event $e = (\mathbf{x}, t, p)$. The polarity $p \in \{+1, -1\}$ indicates the direction of the brightness change (increase or decrease). This triggering mechanism is shown in Equation (1):

$$p = \begin{cases} +1, & \text{if } \log(\mathcal{I}(\mathbf{x}, t)) - \log(\mathcal{I}(\mathbf{x}, t - \Delta t)) > \Theta \\ -1, & \text{if } \log(\mathcal{I}(\mathbf{x}, t)) - \log(\mathcal{I}(\mathbf{x}, t - \Delta t)) < -\Theta \end{cases} \tag{1}$$

Therefore, an event camera outputs a spatially sparse but temporally dense (microsecond resolution) stream of events, recording only the dynamic information in the scene, which effectively avoids motion blur and offers a high dynamic range.

B More details about the Event Structure Loss

B.1 More details about structure extractor

Our method extracts a high-frequency *structure* component from an input image via local contrast normalization, separating it from low-frequency *color* information.

For an RGB input image, we first convert it to the YUV colorspace. The luminance channel (Y) is isolated for structure extraction, while the chrominance channels (U, V) are preserved as the color component. If the input is already grayscale, it is processed directly as the luminance channel.

The structure is extracted by standardizing the luminance channel Y. We compute local mean μ and local standard deviation σ for each pixel with a Gaussian blur operator (G_{k,σ_X}) to approximate the local statistics. The structure component S is defined as:

$$S(\mathbf{x}) = \frac{Y(\mathbf{x}) - \mu(\mathbf{x})}{\sigma(\mathbf{x}) + c}$$

where $\mathbf{x}$ denotes the pixel coordinates and c is a small constant for stability in low-variance regions.

Finally, this structure component S is normalized to the range $[0, 1]$ to produce the final structure map, S_{norm}. This normalized map is returned along with the color component (if applicable).

B.2 More details about weight mask

To generate a weight map that highlights salient and stable features from the event-reconstructed grayscale image $I \in [0, 1]$, we implement a multi-scale gradient analysis pipeline. The goal is to identify structures that are not only strong at a coarse level but also persistent across scales (i.e., not just fine-scale noise).

First, we compute image gradients at two distinct scales. The input image I is convolved with two Gaussian kernels, G_s (with standard deviation σ_{small}) and G_l (with σ_{large}), to produce a fine-scale version I_s and a coarse-scale version I_l. Denoted as: $I_s = G_s * I, I_l = G_l * I$ We then apply Sobel operators (∇_x, ∇_y) to both blurred images to obtain their respective gradient magnitudes, M_s and M_l:

$$M_s = \sqrt{(\nabla_x I_s)^2 + (\nabla_y I_s)^2 + \epsilon}$$

$$M_l = \sqrt{(\nabla_x I_l)^2 + (\nabla_y I_l)^2 + \epsilon}$$

where ϵ is a small constant (e.g., 10^{-12}) for numerical stability.

To ensure the detected structures are stable and not just fine-scale artifacts, we compute a *cross-scale persistence* score P. We first robustly normalize M_s and M_l to range $[0, 1]$ using a normalization function $\mathcal{N}_R(\cdot)$ (which clips outliers and performs min-max scaling), yielding M'_s and M'_l. The persistence P is then calculated as ratio of coarse-to-fine magnitude, clamped at 1.0 and modulated by an exponent γ:

$$P = \left(\min \left(\frac{M'_l}{M'_s + \epsilon}, 1.0 \right) \right)^{\gamma}$$

This term assigns high scores to structures present at both scales ($M'_l \approx M'_s$) and suppresses features that are strong at fine scale but absent at coarse scale ($M'_l \ll M'_s$).

We further refine the mask by gating out weak structures at the coarse level. A soft gate G is computed using the normalized coarse magnitude M'_l: $G = \text{sigmoid}(s \cdot (M'_l - \tau))$, where s is a sharpness parameter (e.g., 25.0) and τ is a dynamic threshold, typically set as the 85th percentile of M'_l (i.e., $\tau = Q_{0.85}(M'_l)$). This gate G effectively binarizes the coarse map, retaining only the most salient features.

The unnormalized weight map $\hat{W}$ is defined as the product of the coarse-scale salience, the cross-scale persistence, and the coarse-scale gate: $\hat{W} = M'_l \cdot P \cdot G$. This map $\hat{W}$ is then robustly normalized to produce the final weight map $W = \mathcal{N}_R(\hat{W})$.

Optionally, a morphological dilation (implemented as a 2D max-pooling operation with stride 1) with a kernel k_d is applied to W to slightly thicken the resulting structural mask for downstream tasks. We visualize the examples of extracted event structures and created weight mask in Fig. S1.

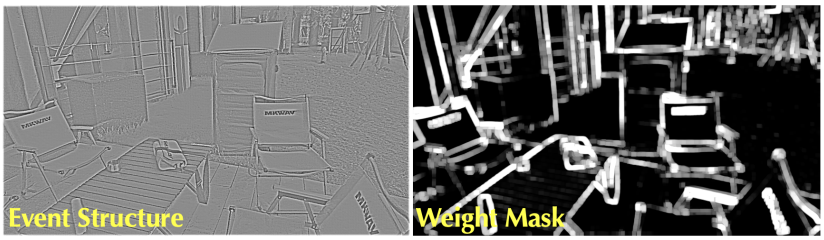


Fig. S1: Extracted Event Structure and Weight Mask.

C More details about Deblurring MLP

To simulate motion blur within the 3D Gaussian Splatting framework, we utilize a multi-layer perceptron (MLP), adopted from Deblurring 3DGS [15], which predicts a series of deformations for each 3D Gaussian to represent its state at multiple discrete moments during exposure. The final blurred rendering is achieved by averaging these deformed states. The network input is an 85-dimensional feature vector, which concatenates positional embeddings of the 3D position ($L = 3$) and 3D view direction ($L = 10$), along with the raw 3D scale (3 channels) and 4D rotation quaternion (4 channels). The core MLP (using default parameters `num_hidden=3`, `width=64`) consists of 3 linear layers with ReLU activations, transforming the input into a 64-dimensional feature vector. This feature is then passed to three separate linear output heads to predict the deltas for position (`self.p`, $3 \times 4 = 12$ channels), scale (`self.s`, $3 \times 5 = 15$ channels), and rotation (`self.r`, $4 \times 5 = 20$ channels). The learning rate for the MLP is constantly 0.001.

D Camera Pose Calibration Pipeline Comparison

Fig. S2 compares the camera pose calibration pipelines of three representative setups. DAVIS-based methods benefit from co-located sensors but are limited to low resolution (346×260). LSE-NeRF achieves high resolution via a synchronized dual-camera rig, but requires a complex multi-step calibration (intrinsics, stereo extrinsics, SfM, and scale alignment), where errors accumulate across stages. In contrast, our method feeds both event-reconstructed frames and RGB images directly into VGGT for end-to-end joint pose estimation, eliminating manual calibration entirely. This simple pipeline supports any event camera at arbitrary resolution, offering both flexibility and robustness.

E More Details about AsyncEv-Deblur Dataset.

We introduce the AsyncEv-Deblur Dataset, which offers two significant advantages over prior work. First, it features a substantially higher resolution. Both our RGB and EVS sensors provide a resolution of 1280×720, which is significantly higher than the 346×260 resolution of typical DAVIS cameras. Second, we employ a more flexible asynchronous dual-camera setup that does not require temporal synchronization between the sensors.

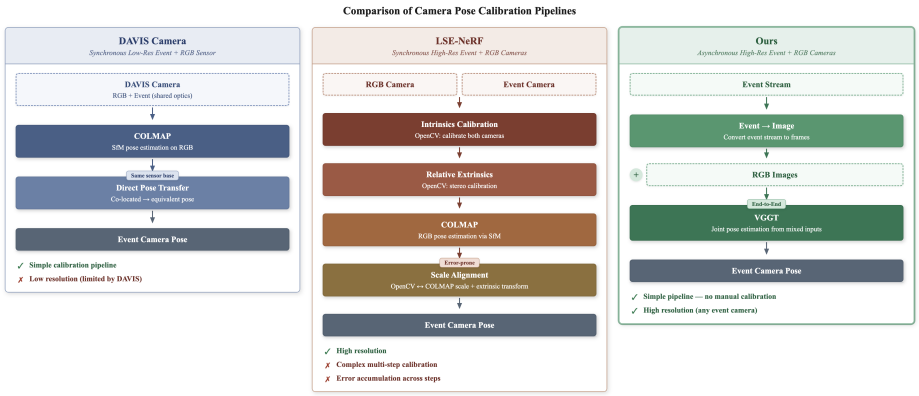


Fig. S2: Comparison of camera pose calibration pipelines. Our approach replaces complex multi-step calibration with a single end-to-end VGGT estimation, achieving simplicity, high resolution, and flexibility.

Notably, unlike the capture conditions of many previous datasets which focus on low-light, long-exposure scenarios, our data was collected under normal outdoor illumination with standard exposure times. This makes our dataset more representative of real-world use cases and provides a more practical benchmark for evaluation.

The dataset comprises seven distinct scenes, we demonstrate more scenes in Fig. S6. For each scene, we provide 25-50 RGB images, which include blurry training views and sharp testing views. Concurrently, we offer a comparable number of event-reconstructed images generated via E2VID. All views were calibrated to obtain camera intrinsics and extrinsics, with initialization performed using VGGT. Furthermore, while our asynchronous setup does not necessitate camera synchronization or timestamp recording, we provide the event camera timestamps and raw event signals for the convenience of future work and comparative analysis.

F Event images pre-procession

The grayscale event images reconstructed via E2VID [32] exhibit two primary artifacts that degrade downstream performance. First, they suffer from significant noise, as illustrated in Fig. S3, which is a byproduct of unavoidable event noise during real-world capture. Second, the reconstructed sequence displays pronounced inter-frame brightness inconsistencies, particularly in non-edge regions, as shown in Fig. S5.

To mitigate the adverse effects of these issues on subsequent VGGT initialization and Gaussian Splatting reconstruction, we apply a two-stage preprocessing pipeline to all event-reconstructed images. Initially, we apply Bilateral Denoising to all frames. This step markedly reduces the noise level, as shown in Fig. S3. Subsequently, to enhance temporal photometric consistency, we perform a brightness balancing procedure across the image sequence. As demonstrated in Fig. S5

and Fig. S4, this step significantly alleviates inter-frame brightness inconsistency, thereby reducing potential artifacts in the final reconstruction.

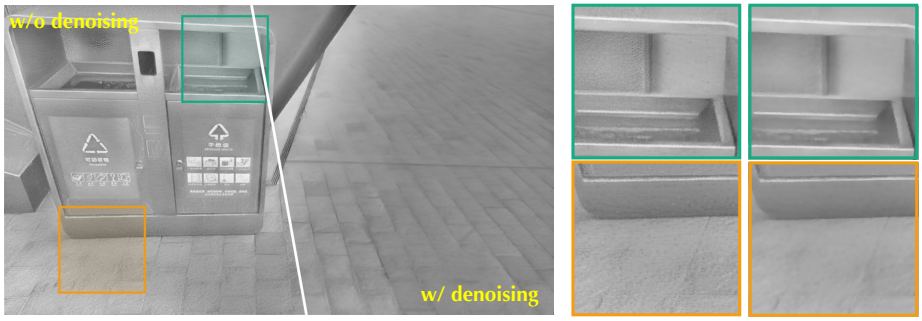


Fig. S3: Comparison of bilateral denoising pre-processing.

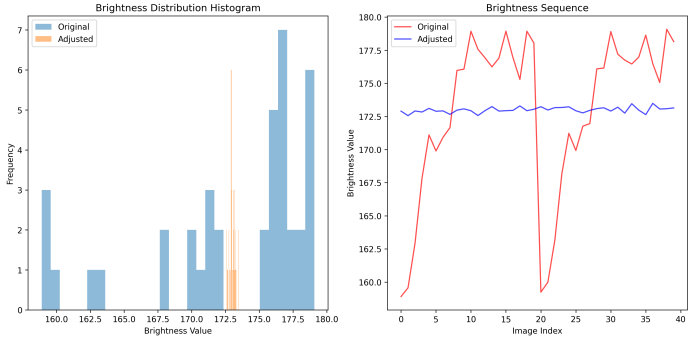


Fig. S4: A representative brightness balance pre-processing results.

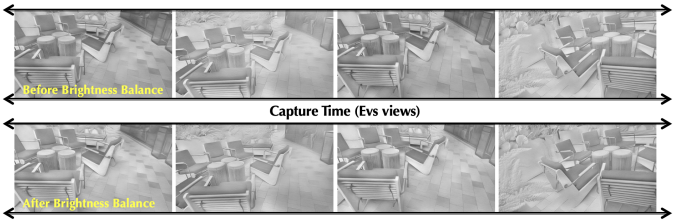


Fig. S5: Comparison of brightness balance.

Fig. S6: More details about our AsynEv-Deblur datasets. AsynEv-Deblur dataset contains diverse scenes, and captured by our dual-camera system with high capturing speed.

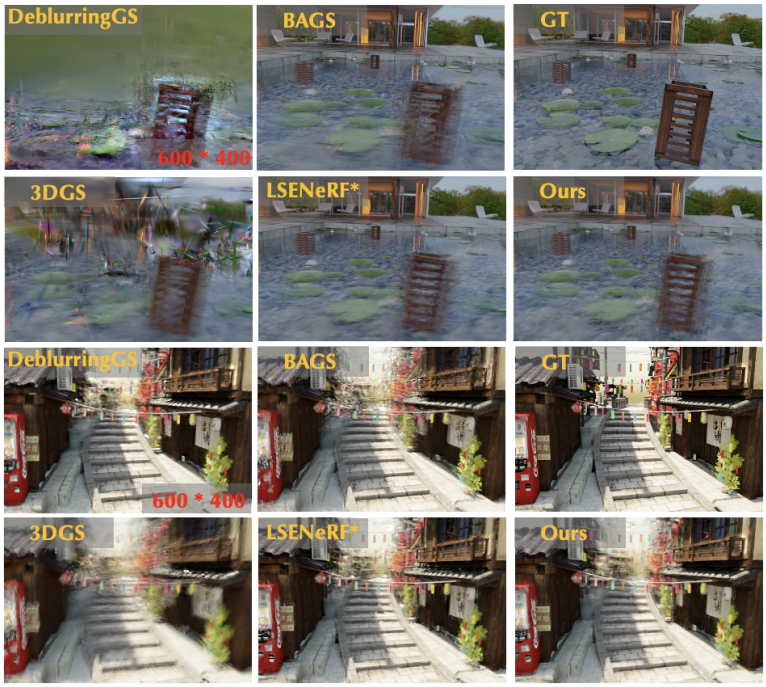


Fig. S7: More qualitative results on Ev-Deblu Blender dataset.

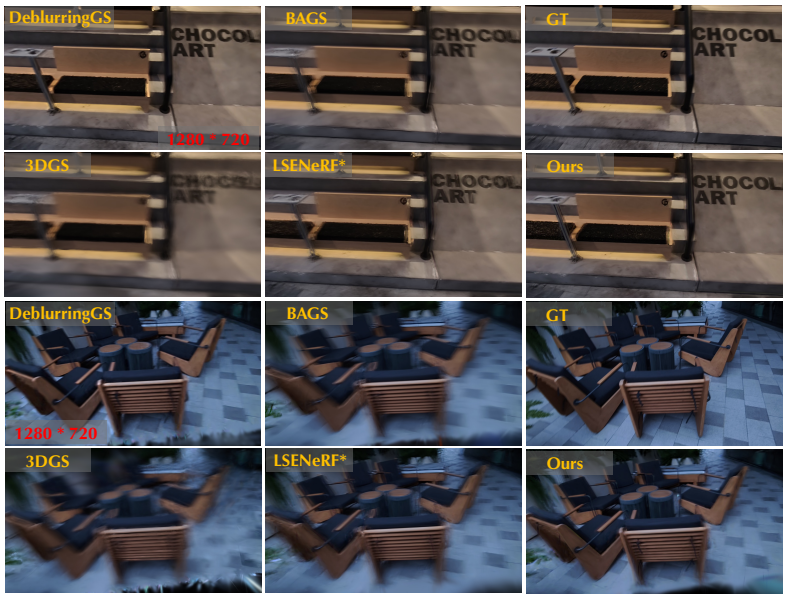


Fig. S8: More qualitative results on our AsyncEv-Deblur dataset.