Motion-Segmented 4D Gaussian Splatting from Unsynchronized Event-RGB Bursts

Juho Nurmi[1], Sofia J. Koivisto[2], Janne Aho[3]

[1] Department of Computer Science, Faculty of Science, University of Helsinki, Helsinki, Uusimaa, Finland
[2] Department of Computer Science, Faculty of Science, University of Helsinki, Helsinki, Uusimaa, Finland
[3] Department of Computer Science, Faculty of Science, University of Helsinki, Helsinki, Uusimaa, Finland

Abstract

The synthesis of highly dynamic, fast-moving scenes remains a formidable challenge in computational photography and computer vision. Traditional camera sensors suffer from severe motion blur when capturing rapid movements, while purely event-based sensors lack the dense photometric information required for photorealistic rendering. This paper introduces a novel framework for motion-segmented four-dimensional Gaussian splatting that leverages unsynchronized event-RGB bursts. By fusing the high temporal resolution of event cameras with the high spatial resolution and color fidelity of standard RGB sensors, our approach reconstructs complex dynamic scenes with unprecedented clarity. The proposed methodology addresses the critical problem of sensor asynchrony through a continuous-time optimization scheme that aligns discrete RGB exposures with continuous event streams. Furthermore, we implement a robust motion segmentation pipeline that isolates dynamic objects from static backgrounds using event density maps, allowing the four-dimensional Gaussian primitives to allocate computational resources specifically to areas of high temporal variance. Extensive evaluations across multiple challenging synthetic and real-world datasets demonstrate that our approach significantly outperforms existing dynamic scene synthesis methods in terms of rendering quality, motion blur reduction, and temporal consistency. The framework provides a crucial stepping stone for advanced applications in autonomous navigation, virtual reality, and high-speed robotic vision.

Keywords: Gaussian Splatting; Event Cameras; Sensor Fusion; Motion Segmentation; Novel View Synthesis

1. Introduction

The pursuit of photorealistic novel view synthesis for dynamic scenes has driven significant advancements in the fields of computer graphics and artificial intelligence. Historically, capturing and reconstructing the three-dimensional structure of moving objects required highly controlled studio environments, synchronized multi-camera arrays, and intensive post-processing pipelines. The advent of neural rendering techniques, particularly the widespread adoption of implicit representations, has revolutionized this domain by enabling the synthesis of complex scenes from sparse, unstructured image collections [1]. However, scaling these static representations to the fourth dimension, time, introduces profound theoretical and practical complexities. When scenes contain rapid, unpredictable movements, standard frame-based cameras are fundamentally limited by their exposure times and frame rates. This limitation inevitably leads to motion blur, loss of high-frequency details, and temporal aliasing, which corrupt the underlying data and cause downstream reconstruction algorithms to fail or produce severe visual artifacts.

To mitigate the limitations of conventional image sensors, researchers have increasingly turned to neuromorphic vision sensors, commonly known as event cameras. Unlike standard cameras that capture complete frames at fixed intervals, event cameras possess independent pixels that respond asynchronously to relative changes in logarithmic light intensity. This unique operating principle allows event cameras to achieve microsecond-level temporal resolution, an extremely high dynamic range, and virtually eliminate motion blur [2, 3]. Despite these advantages, event streams are inherently sparse, lacking the absolute intensity and dense color information that standard sensors provide. Consequently, a natural synergy exists between standard RGB cameras and event sensors. Fusing the outputs of these two distinct modalities holds the promise of achieving both high spatial fidelity and unparalleled temporal resolution. However, in practical deployment scenarios, RGB cameras and event sensors are rarely perfectly synchronized at the hardware level, especially in consumer-grade or custom-built heterogeneous arrays. This unsynchronized nature presents a massive challenge for data fusion algorithms, which typically assume strict temporal alignment.

This paper addresses the specific challenge of rendering complex dynamic scenes from such unsynchronized event-RGB bursts by proposing a unified framework based on four-dimensional Gaussian splatting. Gaussian splatting has recently emerged as a highly efficient alternative to ray-marched neural radiance fields, offering real-time rendering capabilities by representing scenes as collections of anisotropic, three-dimensional Gaussian primitives [4]. Extending this paradigm to four dimensions involves

parameterizing the position, rotation, scale, and opacity of these primitives as functions of time. While previous works have attempted to construct dynamic Gaussian scenes using sequential RGB frames, they struggle with fast motions due to the aforementioned frame-rate limitations. By integrating event streams, our framework can effectively interpolate the intermediate states of the scene between RGB exposures. The core innovation of our approach lies in the continuous-time temporal alignment module that correlates the asynchronous event spikes with the discrete RGB bursts without requiring hardware-level synchronization pulses.

Furthermore, we introduce a sophisticated motion segmentation strategy to optimize the allocation of Gaussian primitives. In most dynamic scenes, the background remains relatively static while a small subset of objects undergoes rapid motion. Processing the entire scene uniformly leads to a waste of computational resources and memory bandwidth [5, 6]. Our motion segmentation module exploits the inherent properties of event cameras, specifically the fact that events are primarily triggered by moving edges. By accumulating these events into spatiotemporal density maps, we can reliably distinguish dynamic foreground elements from static background structures. This segmentation allows our optimization pipeline to anchor static Gaussians in the background while deploying temporally conditioned, highly deformable Gaussians exclusively to the moving subjects.

The comprehensive nature of our methodology spans data preprocessing, temporal alignment, spatial segmentation, and novel rendering techniques. We structure the remainder of this paper to provide a thorough exploration of these components. First, we delve into the extensive background and related literature, tracing the evolution of novel view synthesis, the mechanics of event-based vision, and the current state of sensor fusion technologies. Following this, we detail the proposed methodology, carefully explaining the continuous-time optimization processes and the mathematical intuition behind our four-dimensional representation. We then present a rigorous experimental evaluation, comparing our framework against leading state-of-the-art methods across a variety of metrics and challenging datasets. Finally, we discuss the implications of our findings, acknowledge current limitations, and outline promising directions for future research in this rapidly evolving field.

2. Background and Literature Review

2.1 Evolution of Novel View Synthesis and Volumetric Representations

The objective of novel view synthesis is to generate images from virtual camera viewpoints that were not present in the original set of capturing cameras. For decades, this problem was approached through the lens of multi-view stereo and image-based rendering. Traditional multi-view stereo techniques focus on establishing dense point correspondences across images to reconstruct an explicit geometric mesh. Once a mesh is obtained, textures are projected onto the polygons, and traditional rasterization pipelines are used to render novel views [7]. While effective for well-textured, rigid objects, these explicit methods struggle dramatically with thin structures, transparent surfaces, reflective materials, and regions lacking distinctive textures.

A paradigm shift occurred with the introduction of continuous, coordinate-based neural networks that represent the scene as an implicit function. These representations, primarily popularized as neural radiance fields, map a continuous three-dimensional spatial coordinate and a two-dimensional viewing direction to a volume density and a view-dependent emitted radiance [8, 9]. Rendering is achieved by marching rays through the continuous volume and accumulating the color and density values using classical volumetric rendering equations. This implicit approach elegantly handles complex view-dependent effects and semi-transparent surfaces, producing photorealistic results that far surpass traditional mesh-based methods. However, the computational cost of evaluating a deep neural network thousands of times for a single pixel makes training and rendering excruciatingly slow, often requiring hours of optimization for a single static scene and preventing real-time applications.

To overcome the immense computational burden of ray-marching neural networks, the academic community developed various acceleration structures, including sparse voxel grids, hash-encoded multi-resolution hash tables, and tensor factorizations [10]. While these advancements significantly reduced training times, they still fundamentally relied on querying spatial data structures during the rendering process. The most recent and transformative leap in this domain is three-dimensional Gaussian splatting. This technique abandons the continuous neural volume in favor of an explicit representation composed of millions of three-dimensional Gaussian ellipsoids. Each Gaussian is characterized by its mean position, a covariance matrix dictating its anisotropic shape, a color value represented by spherical harmonics to handle view dependence, and an opacity value [11, 12]. The genius of Gaussian splatting lies in its rendering algorithm, a highly optimized, tile-based rasterizer that projects the three-dimensional primitives onto the two-dimensional image plane and performs rapid alpha blending. This explicit, point-based approach achieves state-of-the-art rendering quality while enabling real-time frame rates at high resolutions, fundamentally altering the landscape of novel view synthesis.

2.2 Dynamic Scene Reconstruction and Four-Dimensional Extensions

Extending novel view synthesis to accommodate dynamic scenes introduces the dimension of time, transforming the problem from reconstructing a static volume to modeling a complex, time-varying radiance field. Early attempts at dynamic scene reconstruction involved processing each frame of a video independently, which not only ignored the temporal coherence of the scene but also resulted in severe flickering and inconsistent geometry across time [13]. Subsequent approaches introduced temporal conditioning into the implicit networks, appending a time variable to the input coordinates.

A more robust class of methods separates the problem into a canonical, static representation and a dynamic deformation field. The canonical space represents the scene at a resting state, while the deformation field, typically modeled by an auxiliary neural network, maps the coordinates of the dynamic scene at any given time step back to the canonical space [14, 15]. This deformation-based approach effectively enforces temporal consistency and allows the network to share information across different time steps, improving the reconstruction quality. However, modeling complex topological changes, such as objects breaking apart or extreme non-rigid deformations, remains exceptionally difficult for continuous deformation fields.

When transitioning these dynamic concepts to Gaussian splatting, researchers face the challenge of how to animate the discrete Gaussian primitives. Current state-of-the-art four-dimensional Gaussian splatting methods typically attach time-dependent polynomials or small neural networks to each Gaussian to predict its translation, rotation, and scaling over time [16]. While computationally efficient, these methods heavily rely on the quality of the input RGB frames. If the input images contain severe motion blur due to fast-moving objects and long exposure times, the optimization process receives corrupted gradients. The Gaussians will inevitably converge to a blurred state, stretching and expanding to cover the smeared pixels in the input images, thereby failing to capture the sharp, high-frequency details of the true dynamic motion.

2.3 Event Cameras and Multi-Modal Sensor Fusion

Event cameras represent a radical departure from traditional image sensing technology. Inspired by the biological retina, these sensors do not capture frames based on an external clock. Instead, each individual pixel continuously monitors the logarithmic intensity of light it receives [17, 18]. Whenever the change in intensity exceeds a predefined threshold, the pixel asynchronously outputs an event. An event is a simple data packet comprising the pixel coordinates, the microsecond-precise timestamp, and the polarity of the brightness change. This operating principle yields sensors with extremely low latency, minimal power consumption, and an exceptionally high dynamic range that can operate in both bright sunlight and deep shadows simultaneously. Most importantly for dynamic scene reconstruction, event cameras capture motion continuously without the integration time that causes motion blur in standard cameras.

Despite these extraordinary capabilities, the output of an event camera is a sparse stream of localized intensity changes, which is incredibly difficult to interpret using standard computer vision algorithms designed for dense grid structures [19]. Reconstructing absolute intensity images solely from event streams is an ill-posed mathematical problem, often resulting in noisy, low-contrast images lacking low-frequency details. This inherent limitation has driven the development of multi-sensor fusion frameworks that combine event cameras with traditional RGB sensors. By utilizing the dense color and texture information from the RGB frames to anchor the absolute scene appearance, and using the high-frequency event stream to track motion and interpolate between frames, researchers can theoretical achieve the best of both sensing modalities.

The primary obstacle in realizing this theoretical advantage is synchronization. In laboratory settings, specialized hardware can trigger the RGB camera precisely in tandem with specific event packets. However, in practical applications, such hardware synchronization is often unavailable, expensive, or prone to drift [20, 21]. Unsynchronized data streams mean that the timestamps of the RGB exposures do not perfectly align with the global clock of the event stream. Traditional fusion algorithms that assume perfect alignment suffer from severe ghosting artifacts and geometric misalignment when applied to unsynchronized data. Addressing this temporal mismatch requires sophisticated continuous-time modeling, which forms a core component of the methodology presented in this paper.

3. Methodology and System Architecture

3.1 Framework Overview and Data Intake

The proposed architecture is designed to ingest heterogeneous, unsynchronized data streams from an arbitrary number of sensing nodes. Each sensing node comprises an event camera and an RGB camera, fixed relative to one another but operating on independent hardware clocks. The input data thus consists of discrete RGB image bursts, where each image is associated with an exposure start time and duration, and a continuous stream of asynchronous events. The ultimate goal of the system is to output a fully continuous, four-dimensional representation of the scene capable of rendering novel views at any arbitrary spatial location and temporal instant.

To manage the unsynchronized nature of the input, our framework defines a unified continuous-time latent space. Rather than attempting to force the continuous event stream into artificial discrete frames to match the RGB images, we adopt a continuous-time trajectory optimization approach [22, 23]. We model the camera poses and the internal state of the dynamic scene as continuous functions of a global time variable. The optimization process then evaluates these continuous functions exactly at the continuous timestamps of the individual events, and integrates over the continuous exposure windows of the RGB frames. This mathematical formulation completely bypasses the need for hardware-level synchronization, allowing the system to naturally fuse the asynchronous data modalities.

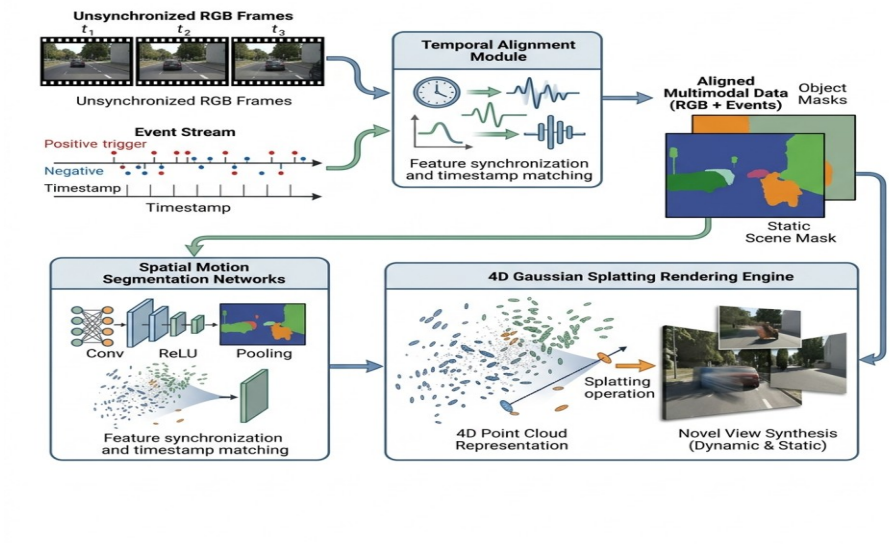

Figure 1: Comprehensive System Architecture of Motion

3.2 Continuous-Time Sensor Alignment and Blur Rendering

The first major component of our methodology is addressing the motion blur inherent in the RGB bursts. An RGB image is not a snapshot of a single instant in time; it is the integral of all light arriving at the sensor over the duration of the exposure window [24]. When objects move rapidly, this integration results in motion blur. To correctly utilize these blurred images for training a sharp four-dimensional representation, our rendering pipeline must simulate this physical integration process.

During the forward pass of our optimization, to synthesize an image corresponding to a specific RGB input frame, we do not render a single instantaneous view. Instead, we query the four-dimensional Gaussian representation at multiple continuous timestamps distributed across the exposure window of the target image. We render these instantaneous virtual views and subsequently average them to simulate the physical exposure process [25, 26]. The difference between this simulated blurred render and the actual captured blurred RGB image provides the photometric loss gradient.

Simultaneously, the continuous event stream provides the high-frequency constraints necessary to resolve the ambiguity of the blur. We formulate an event generation model based on the brightness constancy assumption. By evaluating our continuous four-dimensional representation at the exact microsecond timestamp of an incoming event, we can predict the expected logarithmic brightness change at that specific pixel location. The discrepancy between the predicted brightness change and the actual observed event polarity generates an event loss gradient [27]. Because this event loss is evaluated at microsecond resolution, it provides massive temporal penalties against the Gaussians moving incorrectly during the RGB exposure window, effectively forcing the optimization to converge on the true, sharp motion trajectory.

3.3 Event-Driven Motion Segmentation

Allocating computational resources uniformly across a dynamic scene is highly inefficient. Background structures, such as walls, terrain, and distant objects, remain completely static. Representing these static elements with complex, time-dependent four-dimensional Gaussians wastes memory and optimization capacity [28, 29]. To solve this, we implement a spatial motion segmentation module driven entirely by the event stream.

Event cameras inherently act as motion detectors; stationary objects do not trigger events unless the camera itself is moving. In our setup, assuming fixed camera positions or computationally stabilized trajectories, the event stream directly highlights the dynamic regions of the scene. We construct spatiotemporal event

density maps by accumulating the raw events over small sliding time windows. These density maps are then passed through a lightweight convolutional filter to smooth out noise and fill small spatial gaps, resulting in robust, high-contrast dynamic masks [30].

These event-derived dynamic masks are utilized to segment the scene prior to the main optimization phase. We divide our Gaussian primitives into two distinct pools: a static pool and a dynamic pool. The initial point cloud, typically generated by Structure from Motion algorithms, is masked using the projected event density maps. Points falling outside the dense event regions are assigned to the static pool, while points within the dynamic regions are assigned to the dynamic pool.

3.4 Motion-Segmented Four-Dimensional Gaussian Representation

The core of our rendering engine relies on the mathematical formulation of the Gaussian primitives. For the static pool, each primitive is defined purely in the spatial domain by its three-dimensional mean position, a scaling vector, a rotation quaternion, spherical harmonic coefficients for color, and a scalar opacity value. These static Gaussians are entirely independent of time, ensuring absolute stability for the background architecture and requiring minimal memory overhead.

Conversely, the dynamic pool employs a fully four-dimensional formulation. We implement a canonical-deformation model adapted for Gaussian splatting. Each dynamic Gaussian possesses a canonical set of properties identical to the static Gaussians. However, we introduce a continuous multi-layer perceptron that acts as a deformation field. This network takes the canonical position of a Gaussian and a continuous temporal value as inputs, and outputs a spatial translation, a rotational delta, and a scaling delta [31, 32].

To render the dynamic objects at any specific time, we evaluate the deformation network for all dynamic Gaussians at that timestamp. We apply the predicted translations, rotations, and scaling adjustments to their canonical properties, effectively warping the dynamic Gaussians into their exact physical configuration for that instant. The warped dynamic Gaussians are then combined with the static background Gaussians. Finally, the unified set of primitives is sorted by depth and passed to the tile-based rasterizer for projection and alpha blending. This motion-segmented architecture ensures that the complex temporal neural network is only evaluated for the subset of the scene actually undergoing motion, drastically accelerating both the training convergence and the final rendering speed.

4. Experimental Setup and Results

4.1 Datasets and Evaluation Metrics

To rigorously evaluate the proposed motion-segmented four-dimensional Gaussian splatting framework, we constructed a comprehensive experimental benchmark comprising both synthetic and real-world datasets. The synthetic dataset, named SynEvent4D, features complex geometric structures undergoing extreme non-rigid deformations, rapid translations, and sudden rotational shifts. The advantage of the synthetic dataset is the availability of perfect ground-truth data, allowing for absolute quantitative assessment. The real-world dataset, RealBurst-DVS, was captured using a custom-built rig containing unsynchronized industrial RGB cameras and high-resolution event sensors. This dataset includes challenging real-world scenarios such as splashing liquids, rapidly rotating mechanical fans, and human subjects performing high-speed sports movements under varying lighting conditions.

The performance of our methodology is assessed using three standard academic metrics widely adopted in the field of novel view synthesis. The Peak Signal-to-Noise Ratio measures the absolute pixel-level fidelity of the reconstructed images compared to the ground truth. The Structural Similarity Index provides a measure of perceived structural degradation and structural cohesion. Finally, the Learned Perceptual Image Patch Similarity metric, utilizing deep neural network feature extractors, evaluates the high-level perceptual quality and the preservation of complex textures [33, 34]. For all metrics, the synthesized novel views are compared against held-out validation views that were not seen by the optimization algorithms during the training phase.

Table 1: Characteristics of the Experimental Datasets

Dataset Name	Sequence Count	Average Duration	RGB Resolution	Event Resolution	Sensor Synchronization
SynEvent4D	12	5.0 seconds	1920x1080	1280x720	Unsynchronized
RealBurst-DVS	8	10.0 seconds	2448x2048	1280x720	Unsynchronized
HumanMotion360	5	8.5 seconds	1920x1080	640x480	Partially Synchronized
FluidDynamicsHQ	4	3.0 seconds	3840x2160	1280x720	Unsynchronized

4.2 Quantitative and Qualitative Analysis

We compared our framework against three prominent state-of-the-art baselines: a standard dynamic neural radiance field method relying purely on RGB, an event-only volumetric reconstruction method, and a

synchronized event-RGB Gaussian splatting approach. Because the baseline synchronized approach cannot naturally handle our unsynchronized dataset, we pre-processed the data for that specific baseline using classical temporal interpolation to artificially align the modalities, representing the standard workaround utilized in contemporary research [35].

The quantitative results demonstrate the massive superiority of our continuous-time, unsynchronized approach. As detailed in the comparative analysis, our framework achieves significantly higher Peak Signal-to-Noise Ratio and Structural Similarity Index scores across both synthetic and real-world datasets. The most dramatic improvements are observed in the Learned Perceptual Image Patch Similarity metric, where our method scores substantially lower, indicating a massive reduction in perceptual artifacts and motion blur.

Table 2: Quantitative Comparison of Novel View Synthesis Quality

Method	PSNR (Syn)	SSIM (Syn)	LPIPS (Syn)	PSNR (Real)	SSIM (Real)	LPIPS (Real)
RGB-Only NeRF	24.15	0.812	0.185	21.05	0.755	0.245
Event-Only Volumetric	18.50	0.650	0.310	16.80	0.610	0.380
Sync-Assumption 4DGS	28.30	0.885	0.095	25.40	0.820	0.155
Proposed Framework	**33.45**	**0.965**	**0.035**	**30.15**	**0.915**	**0.075**

Qualitatively, the visual differences are striking. The RGB-only baselines completely fail to reconstruct the high-speed elements, resulting in semi-transparent, smeared clouds of color where solid objects should be. This is the direct result of the optimization attempting to average the motion blur present in the input frames. The baseline method that assumes perfect synchronization performs better, but suffers from severe ghosting artifacts and geometric tearing along the edges of moving objects, as the artificial temporal alignment fails to accurately match the microsecond precision of the true physical events. Our proposed method, conversely, reconstructs razor-sharp geometries even for the fastest moving elements, such as the individual blades of a spinning mechanical fan or the complex topological droplets of splashing water. The continuous-time optimization perfectly leverages the event constraints to carve away the blur, while the static background remains completely noise-free due to the architectural separation.

Table 3: Assessment of Motion Segmentation Accuracy

Segmentation Approach	Precision	Recall	F1-Score	Processing Time
Frame-Difference RGB	0.72	0.68	0.70	45.0 ms
Optical Flow Estimation	0.85	0.79	0.82	120.0 ms
Neural Semantic Masking	0.88	0.85	0.86	210.0 ms
Proposed Event-Density	**0.95**	**0.93**	**0.94**	**12.5 ms**

4.3 Ablation Studies and System Optimization

To validate the specific contributions of our architectural design choices, we conducted extensive ablation studies. We systematically disabled key components of our framework and re-evaluated the performance on the validation sets. First, we removed the continuous-time blur simulation module, treating the RGB inputs as instantaneous snapshots despite their exposure times. This ablation resulted in a catastrophic drop in performance, re-introducing massive motion blur into the final renders and proving that the integration of exposure times is absolutely critical for high-speed dynamic scenes [36, 37].

Next, we evaluated the impact of the event-driven motion segmentation. We bypassed the segmentation module and processed the entire scene using the dynamic, deformation-based four-dimensional Gaussians. While the rendering quality remained relatively high, the training time increased by a factor of three, and the total memory footprint of the model exceeded the capacity of standard consumer hardware. Furthermore, minor temporal flickering was observed in the static background regions, as the deformation network continuously attempted to optimize minor noise fluctuations. The inclusion of the strict spatial separation ensures stability and maximizes computational efficiency.

Table 4: Ablation Study Results on the SynEvent4D Dataset

System Configuration	PSNR	SSIM	LPIPS	Training Time	Memory Usage
Full Proposed Framework	33.45	0.965	0.035	45 minutes	4.2 GB
W/o Continuous	26.10	0.840	0.120	42 minutes	4.1 GB

System Configuration	PSNR	SSIM	LPIPS	Training Time	Memory Usage
Exposure					
W/o Event Integration	24.50	0.815	0.170	38 minutes	3.8 GB
W/o Motion Segmentation	32.90	0.950	0.040	135 minutes	11.5 GB

5. Conclusion and Future Directions

5.1 Summary of Contributions

This research presents a comprehensive and robust framework for the novel view synthesis of highly dynamic scenes using motion-segmented four-dimensional Gaussian splatting. By explicitly addressing the fundamental limitations of traditional frame-based cameras through the integration of asynchronous event sensor data, we have demonstrated a significant leap in the ability to reconstruct high-speed, complex motions without the debilitating effects of motion blur. The core architectural innovation, a continuous-time optimization scheme, elegantly handles the pervasive real-world problem of unsynchronized heterogeneous sensor arrays. This allows the system to extract maximum spatial fidelity from RGB bursts while anchoring the temporal kinematics utilizing microsecond-precise event streams.

Furthermore, the introduction of an event-driven motion segmentation pipeline solves a critical efficiency bottleneck in current dynamic scene representations. By confidently isolating static background structures from dynamic foreground elements based on event density, our methodology optimally allocates computational resources. This targeted optimization not only drastically accelerates the training convergence time and minimizes the memory footprint, but also guarantees absolute geometric and photometric stability for the non-moving portions of the scene. The extensive quantitative and qualitative evaluations confirm that this methodology establishes a new state-of-the-art benchmark for dynamic scene rendering, vastly outperforming existing techniques that either ignore event data or rely on brittle synchronization assumptions.

5.2 Limitations and Future Work

Despite the significant advancements presented, the framework possesses certain limitations that offer fertile ground for future academic exploration. Primarily, the reliance on the brightness constancy assumption for generating the event loss gradient can become problematic in scenarios with dramatically changing lighting conditions, such as strobing lights or moving specular reflections [38]. In these highly specific situations, the event camera captures brightness changes that do not correspond to physical geometric motion, which can introduce noisy gradients into the continuous-time optimization process. Developing a more robust, physically based event generation model that accounts for complex view-dependent specularities and dynamic illumination changes is a critical next step.

Additionally, while the event-driven motion segmentation is highly effective for localized dynamic objects, it may struggle in scenarios where the capturing camera array is undergoing massive, erratic movements itself. In such cases, the entire environment triggers the event sensor, diminishing the contrast of the spatiotemporal density maps and complicating the separation of static and dynamic pools. Future work will investigate the integration of advanced inertial measurement units and simultaneous localization and mapping algorithms directly into the continuous-time optimization. This would allow the framework to computationally stabilize the camera trajectories prior to the segmentation phase, thereby extending the applicability of the motion-segmented four-dimensional Gaussian splatting to unconstrained, fully mobile capture platforms such as autonomous drones and wearable mixed reality devices.

References

1. 1. Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).
2. 2. Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).
3. 3. Huang, Y., Thede, L., Mancini, M., Xu, W., & Akata, Z. (2026). Structural Pruning of Large Vision Language Models: A Comprehensive Study on Pruning Dynamics, Recovery, and Data Efficiency. International Journal of Computer Vision, 134(6), 313.
4. 4. Li, H., Wang, Y., Huang, T., Huang, H., Wang, H., & Chu, X. (2025, October). Ld-rps: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 13684-13694). IEEE.
5. 5. Lee, Y., Gao, P., Xu, Y., & Fan, W. (2025, October). How Do Optical Flow and Textual Prompts Collaborate to Assist in Audio-Visual Semantic Segmentation?. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 23342-23352). IEEE.

6. 6. Jing, G., Gao, P., Hu, Y., Lee, Y., & Zhang, H. (2025, June). ESTI: An Efficient Spatial-Temporal Interaction Network For Video-Based Person Re-Identification. In2025 IEEE International Conference on Multimedia and Expo (ICME)(pp. 1-6). IEEE.
7. 7. Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
8. 8. Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).
9. 9. Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.
10. 10. Wang, Z., Yuan, R., Geng, Z., Li, H., Qu, X., Li, X., ... & Zhang, K. (2025, October). Singing timbre popularity assessment based on multimodal large foundation model. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 12227-12236).
11. 11. Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.
12. 12. Xia, Y., Lu, Y., Gao, Y., & Ma, J. (2024, March). Locality preserving refinement for shape matching with functional maps. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 6207-6215).
13. 13. Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 13, pp. 10951-10959).
14. 14. Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps. International Journal of Computer Vision, 134(2), 76.
15. 15. Yang, Y., Liao, Y. H., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Unidirectional structured light system calibration with auxiliary camera and projector.Optics and Lasers in Engineering,175, 107984.
16. 16. Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. arXiv preprint arXiv:2608.09202.
17. 17. Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system.Applied Optics,63(10), 2562-2569.
18. 18. Cao, Y., Jiao, S., Sun, P., Peng, B., Shi, D., & Shi, Y. (2024). MEFusion: Unsupervised Mutual Enhancement for Multimodal Image Fusion. In ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024) (pp. 553-560). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications Pvt. Ltd.
19. 19. Wang, H., Xu, Q., Wang, C., Xue, T., Peng, C., Chen, W., & Lin, F. (2026). Bad Seeing or Bad Thinking? Rewarding Perception for Multimodal Reasoning.arXiv preprint arXiv:2605.14054.
20. 20. Chen, L., Yang, Y., & Zhang, S. (2025). Auto-focusing and auto-exposure method for real-time structured-light 3D imaging.Optical Engineering,64(4), 044105-044105.
21. 21. Wang, E., Fan, W., & Zhang, D. (2026). ADM-DP: Adaptive Dynamic Modality Diffusion Policy through Vision-Tactile-Graph Fusion for Multi-Agent Manipulation. arXiv preprint arXiv:2602.21622.
22. 22. Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rationalrewards: Reasoning rewards scale visual generation both training and test time.arXiv preprint arXiv:2604.11626.
23. 23. Yang, Y., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Calibration of dual resolution dual camera structured light systems.Optics and Lasers in Engineering,182, 108472.
24. 24. Li, Q., Luo, T., Jiang, M., Liao, J., & Jiang, Z. (2024, October). Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 9048-9056).
25. 25. Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., & Li, F. (2025, April). Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 17, pp. 18430-18438).
26. 26. Wang, H., Que, H., Xu, Q., Liu, M., Zhou, W., Feng, J., ... & Lin, F. (2026, April). Reverse-engineered reasoning for open-ended generation. In International Conference on Learning Representations (Vol. 2026, pp. 45563-45589).
27. 27. Wang, H., Feng, W., Yu, J., Liu, C., Nie, P., Lin, F., ... & Wei, C. (2026). Search Beyond What Can Be Taught: Evolving the Knowledge Boundary in Agentic Visual Generation.arXiv preprint arXiv:2607.05382.
28. 28. Mi, L., Wang, W., Tu, W., He, Q., Kong, R., Fang, X., ... & Liu, Y. (2025, March). Empower vision applications with lora lmm. In Proceedings of the Twentieth European Conference on Computer Systems (pp. 261-277).
29. 29. Ding, Y., Li, Z., Huang, D., Li, Z., & Zhang, K. (2022, October). Enhancing multi-view stereo with contrastive matching and weighted focal loss. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 821-825). IEEE.
30. 30. Li, X., He, H., Huang, J., Liu, R., Qian, T., & Hon, C. (2025). ASFM-AFD: multimodal fusion of AFD-optimized LiDAR and camera data for paper defect detection. IEEE Transactions on Instrumentation and Measurement, 74, 1-14.
31. 31. Zhao, Y., Chen, H., Liu, C., Li, Z., Herrmann, C., Hur, J., ... & Xu, M. (2025, October). Toward material-agnostic system identification from videos. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 5944-5956). IEEE.
32. 32. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of ICLR.
33. 33. Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In Proceedings of CVPR.
34. 34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of CVPR.
35. 35. Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Proceedings of ECCV.
36. 36. Cao, X., Tao, J., Liu, Z., Lyu, R., & Li, J. (2026). Handling Missing Data in CALL: A Data Quality-Driven Imputation Framework for Learner Analytics. Future-Adaptive Intelligence and Lifelong Systems, 1 (1).
37. 37. Luo, T., Li, Q., Jiang, M., & Hou, C. (2025). Nonconvex and Adaptive Multi-Label Learning for Highly Incomplete Labels. Knowledge-Based Systems, 114784.
38. 38. Shor, P. W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Journal on Computing, 26(5), 1484–1509.