TRIPLEX4DGS: HIGH-SPEED 4D GAUSSIAN SPLATTING WITH COMPLEMENTARY VISION SENSORS

Anonymous authors
Paper under double-blind review

ABSTRACT

High-speed dynamic scene 4D reconstruction is crucial for real-world applications. Conventional RGB-based reconstruction suffers from limited temporal resolution to record fast-moving objects, while high-speed RGB cameras incur prohibitive bandwidth and data overhead. As an alternative, the complementary vision sensor (CVS) simultaneously captures synchronized 30-fps RGB frames alongside high-frequency, sparse and multi-bit Temporal Difference (TD) and Spatial Difference (SD) signals, enabling high-speed imaging with much lower bandwidth than traditional sensors. However, directly reconstructing RGB frames from TD/SD may introduce temporal inconsistencies that propagate into subsequent 4D reconstruction. Moreover, TD and SD are monochromatic and follow photometric characteristics distinct from RGB pixels, making joint optimization challenging. Meanwhile, existing 4D Gaussian Splatting frameworks lack differentiable rendering models for TD and SD supervision, so such data still fails to yield improvements for high-speed 4D reconstruction. To tackle the above challenges, we propose Triplex4DGS, a two-stage framework that fully exploits CVS' three complementary data modalities, achieving the capability of ultra-high-speed 4D scene reconstruction with extremely low hardware overhead. In the first stage of Triplex4DGS, a continuous dynamic Gaussian representation is initialized from sparse RGB keyframes. The second stage builds unified supervision within the TD–SD response space via a differentiable spatio-temporal difference simulator, enabling hybrid observation-guided optimization, where SD refines fine-grained spatial geometry, TD regularizes temporal motion trajectories, and RGB frames maintain consistent visual appearance. Furthermore, we build a multi-CVS data acquisition system and construct the first simulated and real-world CVS 4D reconstruction benchmark. Extensive experiments on both datasets demonstrate state-of-the-art performance, validating the effectiveness of Triplex4DGS with CVS for high-speed 4D reconstruction.

1 INTRODUCTION

In recent years, dynamic scene reconstruction has achieved remarkable progress, enabling the recovery of time-varying 3D scenes from multi-view observations and the synthesis of high-quality novel views. Accurate modeling of dynamic motions not only facilitates emerging applications such as slow-motion playback and continuous-time novel view synthesis, but also provides reliable dynamic 3D representations for motion analysis, intelligent perception, and digital twin systems. Building upon Neural Radiance Fields (NeRF) Mildenhall et al. (2021) and 3D Gaussian Splatting (3DGS) Kerbl et al. (2023), a series of dynamic scene representation methods have been developed by incorporating temporal modeling mechanisms, deformation fields, and time-varying Gaussian representations. These approaches effectively extend static scene representations into the spatio-temporal domain, achieving significant improvements in reconstruction fidelity and rendering efficiency for dynamic environments. In particular, 4D Gaussian Splatting (4DGS)-based methods inherit the explicit and efficient representation advantages of 3DGS, enabling high-quality and real-time novel view synthesis for dynamic scenes. Large vision foundation models including VGGT, VGGT-Ω Wang et al. (2025a; 2026a) and DynamicVGGT He et al. (2026) realize robust geometric perception via scalable, temporally consistent point modeling. Yet these models focus on geometry estimation rather than photorealistic rendering, making them unsuitable for high-fidelity dynamic novel view

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

synthesis. Traditional 4DGS methods, meanwhile, only adopt RGB supervision and deformation optimization, which restricts their performance on complex high-speed motion capture.

Despite these advances, achieving accurate and robust 4D reconstruction remains challenging, especially for high-speed dynamic scenes. Existing methods primarily rely on RGB image sequences, where limited frame rates may result in insufficient temporal sampling of fast motions and degraded reconstruction quality Wang et al. (2026b). Although high-speed cameras provide higher temporal resolution, their high acquisition costs and large data overhead hinder practical deployment Chen et al. (2025a). Recently, event cameras have been explored to address this limitation due to their microsecond-level temporal resolution Xu et al. (2026b). However, event cameras only capture brightness changes without complete appearance information, requiring additional RGB cameras and multi-sensor calibration for joint reconstruction He et al. (2025); Xu et al. (2025). Moreover, rapid camera motion can further complicate pose estimation and geometric reconstruction from event observations Huang et al. (2025). A recently proposed complementary vision sensor (CVS),

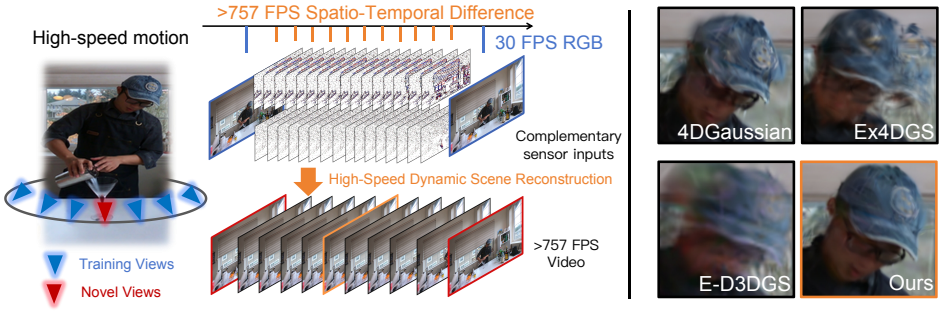

Figure 1: High-speed dynamic scene reconstruction with complementary vision sensing. Tianmouc simultaneously provides RGB observations and high-frequency TD/SD signals, enabling the reconstruction of high-temporal-resolution dynamic videos.

Tianmouc Yang et al. (2024b), provides a promising alternative for high-speed dynamic scene reconstruction. Inspired by biological vision, Tianmouc simultaneously captures RGB images through the Cognition-Oriented Pathway (COP) and high-frequency Temporal Difference (TD) and Spatial Difference (SD) signals through the Action-Oriented Pathway (AOP). By combining stable appearance information with high-temporal-resolution motion cues, Tianmouc enables comprehensive observation of both scene appearance and rapid dynamics, offering a new sensing paradigm for dynamic scene reconstruction. Existing studies have verified that RGB paired with TD and SD signals enables high-quality, blur-free video reconstructionMeng et al. (2025; 2026); Lin et al. (2026). However, such reconstructions are often plagued by minor artifacts and temporal fluctuations, which prove detrimental for precise 4D reconstruction. These works nevertheless reveal that RGB, TD and SD collectively contain sufficient information to faithfully characterize the observed scene, motivating an end-to-end, dedicated 4D reconstruction paradigm as a superior alternative, which has great potential to achieve more accurate and robust 4D reconstruction without introducing additional computing costs.

However, effectively integrating RGB observations with TD/SD complementary signals for accurate high-speed dynamic reconstruction while preserving color fidelity and geometric consistency remains a challenging problem. Existing 4D reconstruction frameworks are primarily designed for dense RGB inputs and lack dedicated differentiable rendering mechanisms for sparse spatio-temporal differential signals. Direct integration of raw TD and SD supervision into standard Gaussian optimization pipelines poses substantial challenges. Stemming from the intrinsic modality gap between intensity RGB frames and monochromatic differential measurements, such supervision also exhibits poor native compatibility with 4D Gaussian primitive formulations. Unlike dense RGB or event stream, TD and SD signals are with high precision, sparse and monochromatic. This fundamental difference introduces cross-modal correspondence ambiguity and inconsistent optimization constraints. Furthermore, high-frequency differential measurements are inevitably affected by sensor noise, quantization errors, and environmental illumination variations. These challenges collectively hinder stable Gaussian optimization, often resulting in color inconsistency, inaccurate motion estimation, and geometric distortions.

To address the challenges of modal inconsistency, rendering incompatibility, sparse supervision, and optimization instability in CVS-based high-speed dynamic reconstruction, we propose **Triplex4DGS**, a two-stage 4D Gaussian reconstruction framework tailored to CVS (see **Fig. 1**). The core difficulty lies not only in the heterogeneous signal forms of RGB, TD, and SD observations, but also in the fact that directly injecting sparse differential measurements into 4D Gaussian optimization can easily destabilize training when geometry and motion remain inaccurate. To bridge this gap, we introduce a differentiable Tianmouc sensor simulation module that maps rendered RGB images into the differential response domain while modeling sensor-specific characteristics, including pixel response, noise, and quantization. Built upon this bridge, Triplex4DGS adopts a progressive two-stage optimization strategy: the first stage establishes stable geometry, appearance, and coarse deformation priors from sparse RGB keyframes, while the second stage progressively incorporates TD and SD observations for cross-modal refinement. This design explicitly exploits the complementary properties of the three modalities: RGB preserves global appearance, SD provides spatial variation constraints, and TD regularizes temporal dynamics. As a result, our framework reduces modal gap during supervision and improves the stability of 4D Gaussian optimization for high-speed dynamic scene reconstruction. Our main contributions are summarized as follows: (1) We propose Triplex4DGS, a two-stage 4D Gaussian reconstruction framework tailored for complementary vision sensors. A differentiable Tianmouc sensor simulation module is introduced to model sensor-specific characteristics, including pixel response, noise, and quantization, enabling effective optimization with sparse differential measurements. (2) We develop a hybrid RGB-TD-SD supervision strategy that progressively integrates appearance and high-frequency motion cues. RGB observations preserve global appearance, SD signals provide spatial variation constraints, and TD signals regularize temporal dynamics, improving reconstruction consistency for high-speed scenes. (3) We establish a Tianmouc-based high-speed dynamic scene dataset with synchronized RGB, TD, and SD observations, and demonstrate that Triplex4DGS achieves superior performance in high-speed dynamic novel view synthesis and 4D reconstruction.

2 Related Works

2.1 Dynamic Scene Rendering on Traditional Cameras

Dynamic scene reconstruction aims to recover time-varying scene geometry and appearance from multi-view observations. The emergence of 3DGS Kerbl et al. (2023) has significantly improved reconstruction efficiency. Existing 3DGS-based dynamic scene reconstruction methods can be broadly divided into two categories. The first models scene motion through deformation fields in a canonical space, where time-dependent deformation fields drive Gaussian motion, such as Deformable-3DGS Yang et al. (2024c), Gaufre Liang et al. (2025) and 4DGaussians Wu et al. (2024). The second directly incorporates temporal information into Gaussian representations by introducing time-dependent Gaussian attributes or motion trajectories, such as 4DGS Yang et al. (2024a), STG Li et al. (2024), 4DRotorGS Duan et al. (2024) and Ex4DGS Lee et al. (2024). Despite their success, these methods remain limited in high-speed dynamic scenes due to the frame-rate constraints of RGB cameras. To alleviate this issue, 4DSlomo Chen et al. (2025a) improves the effective temporal sampling rate through asynchronous multi-view RGB capture and employs a video generation model to recover unobserved views. RetimeGS Wang et al. (2026b) introduces optical flow priors to provide additional motion constraints for Gaussian motion. However, 4DSlomo relies on an additional asynchronous capture system and a video generation model, while RetimeGS depends on external optical flow priors, limiting their applicability.

2.2 4D Reconstruction with Novel Vision Sensors

Early studies demonstrated the effectiveness of event cameras for static scene reconstruction, leveraging their high dynamic range and high temporal resolution for HDR imaging Chen et al. (2025b), motion deblurring Li et al. (2026); Yu et al. (2025); Lee & Lee (2025), and static 3D reconstruction Yura et al. (2025). More recently, event information has been introduced into dynamic Gaussian Splatting. Existing methods mainly exploit event information in two ways. The first utilizes the high-temporal-resolution motion cues provided by events to facilitate dynamic deformation modeling, such as DEGS He et al. (2025), STD-GS Zhou et al. (2025), and EDyGS Lu et al. (2025). The second explores the use of event cameras for high-speed dynamic scene reconstruction. FastEvent-

DGS Dai et al. (2026) investigates monocular event-driven dynamic scene reconstruction, while E-D3DGS Xu et al. (2025), E-4DGS Feng et al. (2025), Sensor Fusion Splatting Zou et al. (2025), and 4D E-SloMo Xu et al. (2026a) further combine RGB images with event streams to improve reconstruction performance in high-speed dynamic scenes. However, most existing methods require joint 2D video reconstruction with RGB images together with cross-modal calibration. Moreover, event data generally rely on continuous camera motion to recover scene geometry, limiting their applicability to fixed-camera scenarios while making accurate calibration of continuous camera poses difficult in practice. By contrast, we leverage data merits of Tianmouc to directly adopt raw TD and SD data for calibration and reconstruction.

3 PRELIMINARY

3.1 4D GAUSSIANS PRIMITIVES

4DGaussians Wu et al. (2024) extends static 3D Gaussian Splatting Kerbl et al. (2023) to dynamic scene reconstruction by introducing a continuous deformation field. Specifically, the canonical Gaussian set is defined as $\mathcal{G} = \{\mathbf{x}, \mathbf{r}, \mathbf{s}, \alpha, \mathbf{c}\}$, where $\mathbf{x}$, $\mathbf{r}$, $\mathbf{s}$, α, and $\mathbf{c}$ denote the Gaussian center, rotation, scale, opacity, and color, respectively. Given a Gaussian $\mathcal{G}$ and timestamp t, a spatial-temporal encoder extracts the feature $\mathbf{f} = H(\mathcal{G}, t)$, which is fed into a deformation network to predict the Gaussian deformation:

$$(\Delta\mathbf{x}, \Delta\mathbf{r}, \Delta\mathbf{s}) = D(\mathbf{f}), \tag{1}$$

where $H(\cdot)$ and $D(\cdot)$ denote the spatial-temporal encoder and deformation decoder, respectively. The deformed Gaussian is then computed as

$$\mathcal{G}_t = \{\mathbf{x} + \Delta\mathbf{x}, \mathbf{r} + \Delta\mathbf{r}, \mathbf{s} + \Delta\mathbf{s}, \alpha, \mathbf{c}\}, \tag{2}$$

which is rendered using differentiable Gaussian splatting to synthesize images at arbitrary viewpoints and timestamps.

3.2 COMPLEMENTARY VISION SENSOR

Tianmouc is a kind of CVS simultaneously acquires scene appearance and high-temporal-resolution spatio-temporal information through two complementary pathways: The COP outputs 10-bit RGB images at approximately 30 FPS, which provide appearance information such as texture and color. The AOP outputs TD and SD at frame rates ranging from 757 to 10,000 FPS with a precision from ±1 to ±7 bits, capturing high-frequency spatio-temporal variations in dynamic scenes. The TD and SD signals are defined as:

$$\begin{cases} TD = \nabla_t I, \\ SD_L = I(x, y, t) - I(x-1, y-1, t), \\ SD_R = I(x, y, t) - I(x-1, y+1, t), \end{cases} \tag{3}$$

where $I \in \mathbb{R}^{H \times W}$ denotes the image intensity and ∇_t is the temporal gradient operator. The TD signal characterizes temporal intensity variations between consecutive observations, whereas the two SD channels capture local spatial gradients along the diagonal directions, providing complementary structural information. Since RGB, TD, and SD are naturally synchronized in space and time during acquisition, they jointly provide appearance, motion, and structural cues for high-speed dynamic scene reconstruction. In this work, Tianmouc operates in the single-gain mode, with the COP running at 30 FPS and the AOP outputting TD/SD signals at 757 FPS with a precision of ±7 bits.

4 METHOD

An overview of the proposed framework is shown in **Fig. 2**. The core of our framework consists of a differentiable simulator tailored to the characteristics of the Tianmouc and a two-stage progressive optimization pipeline.

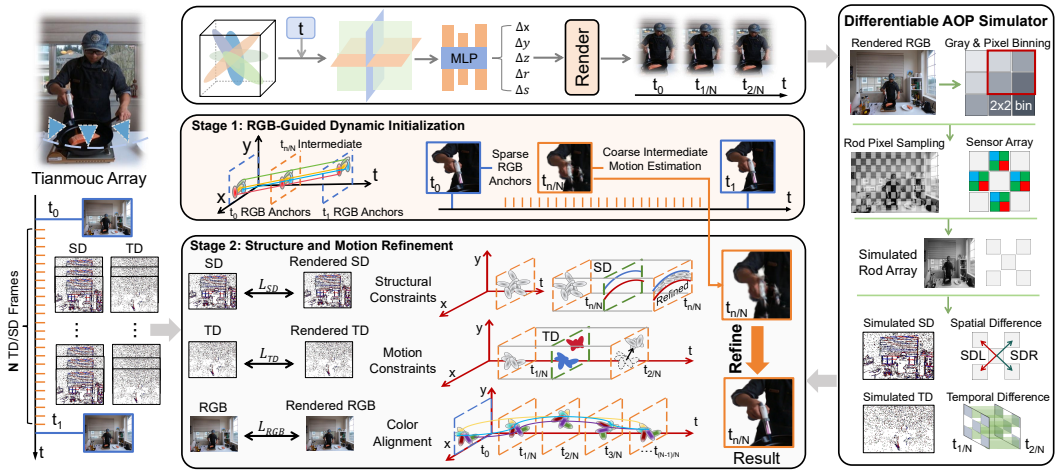

Figure 2: Overall pipeline of the proposed Triplex4DGS. Our method adopts a two-stage optimization paradigm for high-speed dynamic 4D reconstruction based on Tianmouc CVS. In the first stage, sparse RGB keyframes are utilized to initialize the dynamic Gaussian representation with reliable scene geometry and appearance priors. In the second stage, a differentiable spatio-temporal difference simulator is used to convert rendered RGB outputs into supervisable TD and SD responses, which are compared with measured Tianmouc signals to refine spatial structures and temporal dynamics while RGB supervision preserves appearance consistency.

4.1 RGB-Guided Dynamic Initialization

We first optimize the dynamic Gaussian representation using only sparse RGB supervision to establish a reliable initialization. Given an RGB observation captured by camera c at timestamp t, the deformed Gaussian representation $\mathcal{G}_t$ is rendered to produce the corresponding image:

$$\widehat{\mathbf{I}}_{c,t} = \mathcal{R}\left(\mathcal{G}_t, \mathbf{C}_c\right), \tag{4}$$

where $\mathcal{R}(\cdot)$ denotes the differentiable Gaussian rasterizer and $\mathbf{C}_c$ represents the camera parameters. The rendered image is supervised against the corresponding RGB observation $\mathbf{I}_{c,t}$ using a combination of the ℓ_1 loss and structural similarity:

$$\mathcal{L}_{\text{RGB}} = \mathbb{E}_{(c,t)}\left[\|\widehat{\mathbf{I}}_{c,t} - \mathbf{I}_{c,t}\|_1 + \lambda_{\text{s}}\left(1 - \text{SSIM}(\widehat{\mathbf{I}}_{c,t}, \mathbf{I}_{c,t})\right)\right], \tag{5}$$

where λ_{s} balances the two loss terms. This stage mainly serves to provide an appearance-consistent canonical representation across time and to establish a stable geometric and deformation prior before multi-modal supervision is introduced. Because RGB observations are temporally sparse, they are sufficient for coarse dynamic initialization but insufficient for accurately recovering high-speed intermediate states. The finer spatial details and temporal variations are therefore deferred to the second stage.

4.2 Differentiable TD/SD Simulator

The TD and SD signals are multi-bit differential measurements rather than conventional intensity images, and therefore cannot be directly supervised by rendered RGB outputs. To reduce this modal gap, we introduce a differentiable TD/SD simulator that converts rendered RGB images into the AOP response domain, enabling supervision from TD and SD observations without modifying the underlying 4D Gaussian renderer. Given the rendered RGB image $\widehat{\mathbf{I}}_{c,t}$ produced by the dynamic Gaussian representation at camera c and timestamp t, we first convert it into the corresponding AOP pixel output voltage $\widehat{\mathbf{V}}_{c,t}$ through a luminance conversion operator $\mathcal{Y}(\cdot)$ and an AOP sampling operator $\mathcal{P}(\cdot)$:

$$\widehat{\mathbf{V}}_{c,t} = \mathcal{P}\left(\mathcal{Y}\left(\widehat{\mathbf{I}}_{c,t}\right)\right), \tag{6}$$

where $\mathcal{Y}(\cdot)$ converts RGB images into grayscale intensity to estimate the full-spectrum response of large AOP pixels, and $\mathcal{P}(\cdot)$ approximates the Tianmouc sensing process through pixel binning, downsampling, and spatially interleaved sampling. The predicted differential responses are computed from the simulated sensor response. Specifically, the predicted TD signal is obtained as the temporal difference between two consecutive responses:

$$\widehat{\mathbf{TD}}_{c,t} = \widehat{\mathbf{V}}_{c,t} - \widehat{\mathbf{V}}_{c,t-1}, \tag{7}$$

Similarly, two SD channels correspond to spatial gradients along two orthogonal directions:

$$\widehat{\mathbf{SD}}_{L,c,t} = \widehat{\mathbf{V}}(x,y,t) - \widehat{\mathbf{V}}(x-1,y-1,t), \tag{8}$$

$$\widehat{\mathbf{SD}}_{R,c,t} = \widehat{\mathbf{V}}(x,y,t) - \widehat{\mathbf{V}}(x-1,y+1,t), \tag{9}$$

respectively. Since all operations in the simulator are differentiable, TD and SD losses can be backpropagated to optimize the dynamic Gaussian representation. In our framework, the simulator mainly serves as the bridge that makes cross-modal supervision feasible during refinement.

4.3 Progressive Optimization Strategy

Direct joint optimization of 4DGaussians with RGB, TD and SD from scratch suffers from severe instability: heterogeneous signal forms create inherent modality gaps, and unreliable initial geometry and deformation cause simulated differential responses to deviate from ground truth, driving the model to fit modal discrepancies rather than true scene dynamics. We therefore propose a two-stage progressive optimization strategy, which functions not merely as a training schedule but as a core mechanism to mitigate the modality gap and stabilize convergence.

Stage 1: RGB-Guided Dynamic Initialization. The first stage optimizes dynamic Gaussians using only sparse RGB keyframes to establish reliable anchors of scene geometry, appearance and coarse deformation. RGB is selected for initialization as it provides direct, semantically complete intensity supervision that stabilizes early optimization, though its low temporal sampling rate is insufficient to recover high-speed intermediate states. The stage objective is:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{RGB}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{10}$$

where $\mathcal{L}_{\text{RGB}}$ denotes the RGB reconstruction loss and $\mathcal{L}_{\text{reg}}$ is the regularization term inherited from the base 4DGS formulation, which encourages the deformation field to vary smoothly across space and time and stabilizes model optimization.

Stage 2: Structure and Motion Refinement. Built on the stable RGB prior, we introduce TD and SD supervision for joint refinement, with each modality serving a dedicated complementary role: RGB preserves global appearance consistency, SD refines fine-grained spatial structures, and TD regularizes temporal motion continuity. SD captures diagonal spatial intensity gradients and is highly sensitive to object boundaries and local geometries. Compared with sparse RGB keyframes, it provides denser structural constraints for intermediate timestamps, effectively correcting spatial artifacts in fast-motion regions:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{(c,t)}\left[\|\widehat{\mathbf{SD}}_{c,t} - \mathbf{SD}_{c,t}\|_1\right], \tag{11}$$

where $\widehat{\mathbf{SD}}_{c,t}$ is generated by our differentiable simulator. TD measures temporal brightness changes between consecutive frames, delivering high-frequency motion supervision for weakly constrained intermediate states. Beyond temporal regularization, TD prevents SD from degenerating into per-frame structural overfitting by coupling adjacent states and propagating the RGB appearance prior across timestamps:

$$\mathcal{L}_{\text{TD}} = \mathbb{E}_{(c,t)}\left[\left\|\widehat{\mathbf{TD}}_{c,t} - \mathbf{TD}_{c,t}\right\|_1\right]. \tag{12}$$

The overall objective of the second stage is:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{RGB}}\mathcal{L}_{\text{RGB}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_{\text{TD}}\mathcal{L}_{\text{TD}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{13}$$

with weighting coefficients balancing the loss terms. This progressive design avoids premature cross-modal alignment under unreliable representations, assigns each modality to its most suitable optimization target, and reframes multi-modal learning as a prior-guided refinement process, improving both optimization stability and reconstruction quality for high-speed dynamic scenes.



Figure 4: Qualitative comparison on the synthetic Tianmouc benchmark. The red boxes highlight fast-motion regions where our refinement recovers sharper boundaries and more accurate local structures than the 4DGaussians and other methods.

5 Experiment

5.1 Synthetic 4D benchmark

Since no public Tianmouc dataset is currently available for dynamic scene reconstruction, we construct a simulated dataset based on the multi-view Neural 3D Video Li et al. (2022) dataset, which was captured at 30 FPS. Specifically, we first downsample the images to 640×320 and the playback frame rate is boosted to 25 times the original, resulting in an equivalent temporal resolution of approximately 750 FPS. We then retain one frame every 25 frames as the 30 FPS RGB of Tianmouc and use the complete sequence to simulate TD/SD signals with threshold response, quantization, and noise, as illustrated in **Fig 3(a)**.

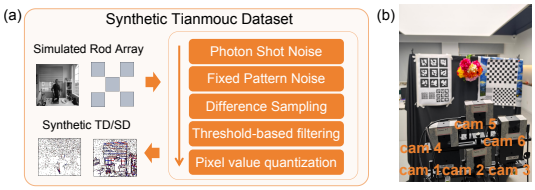

Figure 3: Dataset construction for evaluation. (a) High-fidelity TD/SD generation by simulating the Tianmouc sampling mechanism, sensor noise, and acquisition pipeline. (b) A synchronized six-camera Tianmouc acquisition system.

We compare Triplex4DGS with representative RGB-based, event-based, and two-stage reconstruction methods. RGB-based baselines include 4DGS Yang et al. (2024a), 4D Gaussians Wu et al. (2024), and Ex4DGS Lee et al. (2024), all of which are trained using only the simulated 30 FPS RGB observations. For event-based reconstruction, we adapt E-D3DGS Xu et al. (2025) from a moving-camera to a fixed-view setting and use ESIM-generated events Rebecq et al. (2018). We further construct a two-stage baseline by combining the Tianmouc 2D reconstruction method CBRDM Meng et al. (2025) with 4DGaussians.

5.2 Evaluation Results

We evaluate reconstruction quality using PSNR Hore & Ziou (2010), SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018). Quantitative comparison results are shown in **Table 1**. Our method achieves state-of-the-art PSNR (30.61 dB) and SSIM (0.929), and delivers competitive LPIPS performance. Temporal fluctuations in Tianmouc 2D reconstruction are amplified under multi-view inputs, leading to severe artifacts. Compared with RGB-only baselines, our approach yields consistent gains by using high-frequency TD/SD cues to compensate for motion blur in low-frame-rate

RGB inputs. The event-based E-D3DGS obtains the lowest LPIPS, while our method outperforms it in PSNR and SSIM, showing that RGB-TD/SD fusion better balances texture fidelity and structural accuracy. Qualitative results are shown in **Fig. 4**. In fast-motion regions, our method recovers clearer structures and sharper motion boundaries, validating the effectiveness of the proposed refinement strategy.

5.3 Ablation Study

To quantify the contribution of each supervision modality and validate our triple-modal complementary optimization strategy, we conduct ablation studies by disabling different loss terms on the synthetic benchmark, with results summarized in **Table 2**. The RGB-only baseline achieves moderate reconstruction quality but suffers from inaccurate motion estimation due to the lack of high-frequency temporal constraints. Adding SD supervision to RGB instead degrades performance, because SD provides only monochromatic spatial structural constraints and its response domain is not fully consistent with that of RGB observations. Optimizing spatial structures in the grayscale SD response domain can distort the reconstructed colors, resulting in a noticeable performance degradation. Likewise, using only TD and SD fails to recover consistent appearance without RGB intensity priors.

Table 1: Quantitative comparison on the synthetic benchmark.

Method	PSNR↑	SSIM↑	LPIPS↓
CBRDM+4DGaussians	21.95	0.722	0.257
4DGS	25.93	0.892	0.139
4DGaussians	29.44	<u>0.921</u>	<u>0.102</u>
Ex4DGS	29.40	0.920	0.119
E-D3DGS	<u>29.55</u>	0.920	**0.067**
Ours	**30.61**	**0.929**	<u>0.102</u>

In contrast, combining RGB with TD supervision significantly improves reconstruction quality by introducing temporal motion constraints and propagating appearance information from sparse RGB anchor frames to intermediate time steps. Building upon this baseline, the full model further incorporates SD supervision to refine local geometric structures, achieving the best PSNR and SSIM. These results demonstrate the complementary roles of RGB, TD, and SD in high-speed dynamic scene reconstruction.

As visualized in **Fig. 5**, our progressive pipeline improves reconstruction quality stage by stage, and the full model attains the best spatio-temporal performance in challenging fast-motion regions. These results jointly verify the complementarity of the three modalities and the rationality of our progressive hybrid-observation optimization strategy.



Figure 5: Ablation of the different complementary modalities.

To avoid the increase in model complexity caused by introducing high-frequency TD/SD supervision at an early stage of training, we adopt a two-stage training strategy. Compared with jointly applying RGB and TD/SD supervision from the beginning of training, our strategy reduces the average number of Gaussian primitives from $333{,}069$ to $128{,}676$ and the model size from 78.78 MB to 30.43 MB, corresponding to a reduction of approximately 61.37% in both metrics. Meanwhile, the reconstruction quality remains largely unchanged: PSNR decreases by only 0.046 dB and SSIM by only 0.0010, while LPIPS further improves from 0.1047 to 0.1021. These results demonstrate that the proposed two-stage training strategy significantly reduces redundant Gaussian primitives and storage overhead while preserving reconstruction quality.

5.4 Real-World Implementation

To evaluate the proposed method under practical sparse-view settings, we build a synchronized multi-camera acquisition system consisting of six Tianmouc cameras and construct the Tianmouc dataset, as shown in **Fig 3(b)**. The dataset contains two dynamic sequences, including a freely falling basketball and a rapidly moving flower. Camera poses and the initial sparse point cloud are estimated using COLMAP Schönberger & Frahm (2016). Owing to the limited number of cameras, our real-world experiments focus on evaluating high-speed dynamic reconstruction under sparse input views. Specifically, every other RGB frame in the original sequence is held out as a refer-

ence image, while the remaining RGB frames together with all TD/SD measurements are used for training. Since the reconstructed results of Triplex4DGS are jointly optimized using both RGB and TD/SD observations, their appearance reflects the fused responses of the two modalities. Renderings from synthetic time sometimes outperform raw RGB frames, for instance by eliminating motion blur, so the held-out RGB frames are used only as references.

Fig. 6 presents representative qualitative results on real-world data. RGB-based baselines, including 4DGaussians and Ex4DGS, fail to accurately reconstruct fast-moving objects, exhibiting severe motion blur and structural distortion in dynamic regions. In contrast, Triplex4DGS produces significantly clearer high-speed dynamic reconstructions, with spatial structures that closely match the reference RGB images. Moreover, by jointly leveraging the complementary RGB and TD/SD observations, Triplex4DGS achieves a higher effective dynamic range, enabling faithful recovery of sharp object boundaries and fine structural details even in overexposed regions of the reference RGB images.

Table 2: Ablation study of hybrid supervision on the synthetic Tianmouc benchmark.

Method	PSNR↑	SSIM↑	LPIPS↓
$\mathcal{L}_{\mathrm{RGB}}$	29.44	0.921	<u>0.102</u>
$\mathcal{L}_{\mathrm{RGB}} + \mathcal{L}_{\mathrm{SD}}$	24.61	0.884	0.212
$\mathcal{L}_{\mathrm{TD}} + \mathcal{L}_{\mathrm{SD}}$	24.78	0.886	0.204
$\mathcal{L}_{\mathrm{RGB}} + \mathcal{L}_{\mathrm{TD}}$	<u>30.35</u>	<u>0.928</u>	**0.101**
$\mathcal{L}_{\mathrm{RGB}} + \mathcal{L}_{\mathrm{TD}} + \mathcal{L}_{\mathrm{SD}}$	**30.61**	**0.929**	<u>0.102</u>

We further compare our method with CBRDM's 2D reconstruction results. Although CBRDM demonstrates strong reconstruction performance for single-view high-speed 2D reconstruction, its iterative diffusion inference incurs substantial computational overhead. Consider a multi-view scene with 20 views and 300 TD/SD frames per view, resulting in 6,000 TD/SD frames in total. CBRDM requires approximately 6 min 38 s to reconstruct every 25 frames, leading to about 26 h 32 min for 2D preprocessing alone, excluding the subsequent 4D reconstruction. In contrast, our method directly optimizes the dynamic Gaussian representation from RGB images and differential measurements without reconstructing intermediate 2D intensity images, completing the entire 4D reconstruction in only 1 h 49 min. Even compared with CBRDM's 2D preprocessing time alone, our method reduces the runtime by approximately 93.2%, achieving at least a $14.6\times$ speedup.

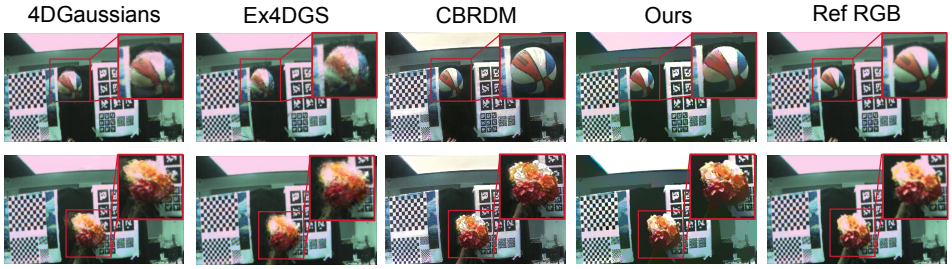

Figure 6: Real-world sparse-view Tianmouc results. Real-world examples rendered at an intermediate timestamp: the RGB-only 4DGaussians and Ex4dgs baseline exhibits blur and detail loss in fast-motion regions, while Triplex4DGS preserves clearer structures with TD/SD-guided refinement (see zoom-in boxes).

6 CONCLUSION

In this work, we propose Triplex4DGS, a two-stage multi-modal 4D reconstruction framework tailored for Tianmouc CVS for high-fidelity high-speed dynamic scene reconstruction. We design a differentiable spatio-temporal difference simulator, bridging the domain gap between standard RGB rendering and sparse TD/SD signals. Our progressive triple-modal optimization leverages RGB to preserve global appearance, SD to refine fine-grained geometry, and TD to regularize high-speed motion trajectories, mitigating color drift, motion distortion and optimization divergence. We also build synthetic and real-world Tianmouc datasets to fill the vacancy of public CVS-specific benchmarks. Extensive experiments demonstrate our method outperforms state-of-the-art RGB-based and event-based baselines. We hope this work offers a solid baseline and insights for future CVS-driven high-speed 4D reconstruction research.

AI USE STATEMENT

All scientific aspects of this work, including the formulation of the research problem, development of the methodology, implementation of the proposed approach, experimental design, data analysis, and interpretation of the results, were carried out by the authors. Generative AI tools were introduced only during manuscript preparation, after the technical content had been established. They were used as language-editing aids to improve English grammar, academic phrasing, readability, and sentence structure. In particular, these tools were not relied upon as sources of scientific evidence or for establishing the technical claims and conclusions of this work. All AI-assisted edits were subsequently inspected by the authors and revised when necessary to ensure consistency with the original technical meaning and the actual experimental findings. The final wording, scientific statements, and conclusions were approved by the authors, who assume full responsibility for the entire content of the manuscript.

ETHICS STATEMENT

Our research investigates high-speed dynamic scene reconstruction using complementary visual sensing, with experiments conducted on both simulated data and visual data captured in controlled environments. The study does not involve human or animal subjects, and the collected data are intended solely for evaluating visual reconstruction and sensing algorithms. We avoid collecting personally identifiable or sensitive information during data acquisition, and all external datasets and algorithmic resources are used in accordance with their respective licenses and terms of use. While high-speed visual reconstruction technologies may potentially be extended to broader sensing and monitoring applications, such applications are beyond the scope of this work; we encourage responsible use of the proposed techniques with appropriate consideration of privacy and applicable regulations. We report our methodology, experimental settings, and evaluation procedures as transparently as possible to support verifiable research and reproducibility. The research has been conducted in accordance with the ICLR Code of Ethics and established standards of research integrity.

REPRODUCIBILITY STATEMENT

We aim to support the reproducibility of this work by providing detailed descriptions of the proposed framework and experimental procedures. The paper specifies the main components of our method, the training and optimization process, data processing procedures, and experimental settings. We also describe the construction of the simulated benchmark and the configuration of our real-world data acquisition system to facilitate an accurate understanding and reproduction of our experimental pipeline. To further support future research and independent verification of our results, we plan to publicly release the source code, data processing tools, and datasets associated with this work. Upon release, these resources will provide the necessary implementations and data for reproducing the experiments reported in the paper.

REFERENCES

Yutian Chen, Shi Guo, Tianshuo Yang, Lihe Ding, Xiuyuan Yu, Jinwei Gu, and Tianfan Xue. 4DSloMo: 4D Reconstruction for High Speed Scene with Asynchronous Capture. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a.

Zehao Chen, Zhan Lu, De Ma, Huajin Tang, Xudong Jiang, Qian Zheng, and Gang Pan. EvHDR-GS: Event-Guided HDR Video Reconstruction with 3D Gaussian Splatting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2367–2375, 2025b.

Zijia Dai, Nico Messikommer, Rong Zou, Nikola Zubic, Davide Scaramuzza, and Laurent Kneip. FastEventDGS: Deformable Gaussian Splatting for Fast Dynamic Scenes from a Single Event Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29537–29546, 2026.

Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4D-Rotor Gaussian Splatting: Towards Efficient Novel View Synthesis for Dynamic Scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Chaoran Feng, Zhenyu Tang, Wangbo Yu, Yatian Pang, Yian Zhao, Jianbin Zhao, Li Yuan, and Yonghong Tian. E-4DGS: High-Fidelity Dynamic Reconstruction from the Multi-View Event Cameras. In *Proceedings of the ACM International Conference on Multimedia*, pp. 7356–7365, 2025.

Junhao He, Jiaxu Wang, Jia Li, Mingyuan Sun, Qiang Zhang, Jiahang Cao, Ziyi Zhang, Yi Gu, Jingkai Sun, and Renjing Xu. DEGS: Deformable Event-Based 3D Gaussian Splatting from RGB and Event Stream. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Zhuolin He, Jing Li, Guanghao Li, Xiaolei Chen, Jiacheng Tang, Siyang Zhang, Zhounan Jin, Feipeng Cai, Bin Li, Jian Pu, et al. DynamicVGGT: Learning Dynamic Point Maps for 4D Scene Reconstruction in Autonomous Driving. *arXiv preprint arXiv:2603.08254*, 2026.

Alain Hore and Djemel Ziou. Image Quality Metrics: PSNR vs. SSIM. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 2366–2369. IEEE, 2010.

Jian Huang, Chengrui Dong, Xuanhua Chen, and Peidong Liu. IncEventGS: Pose-Free Gaussian Splatting from a Single Event Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26933–26942, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023.

Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully Explicit Dynamic Gaussian Splatting. *Advances in Neural Information Processing Systems*, 37:5384–5409, 2024.

Seungjun Lee and Gim Hee Lee. DiET-GS: Diffusion Prior and Event Stream-Assisted Motion Deblurring 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21739–21749, 2025.

Pengteng Li, Pinhao Song, Weiyu Guo, Huizai Yao, Yunfan Lu, F. Richard Yu, and Hui Xiong. DeblurSplat: Traditional SfM-Free 3D Gaussian Splatting with Event Camera for Robust Deblurring. *IEEE Transactions on Multimedia*, 2026.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3D Video Synthesis from Multi-View Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5521–5531, 2022.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8508–8520, 2024.

Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao. GauFRe: Gaussian Deformation Fields for Real-Time Dynamic Novel View Synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2642–2652. IEEE, 2025.

Yihan Lin, Yuguo Chen, Yapeng Meng, Xiangru Chen, Taoyi Wang, Rong Zhao, and Luping Shi. Brain-Inspired Visual Sensing–Perception for Open-World Environments. *Nature Sensors*, pp. 1–17, 2026.

Mengxu Lu, Zehao Chen, Yan Liu, De Ma, Huajin Tang, Qian Zheng, and Gang Pan. EDyGS: Event Enhanced Dynamic 3D Radiance Fields from Blurry Monocular Video. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1684–1692, 2025.

Yapeng Meng, Yihan Lin, Taoyi Wang, Yuguo Chen, Lijian Wang, and Rong Zhao. Diffusion-Based Extreme High-Speed Scenes Reconstruction with the Complementary Vision Sensor. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5701–5710, 2025.

Yapeng Meng, Lin Yang, Yuguo Chen, Xiangru Chen, Taoyi Wang, Lijian Wang, Zheyu Yang, Yihan Lin, and Rong Zhao. Spatio-Temporal Difference Guided Motion Deblurring with the Complementary Vision Sensor. *arXiv preprint arXiv:2604.10554*, 2026.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An Open Event Camera Simulator. In *Conference on Robot Learning (CoRL)*, pp. 969–982. PMLR, 2018.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Taoyi Wang, Yuxiang Li, Yihan Lin, Yapeng Meng, Yuguo Chen, and Rong Zhao. An adaptive filter for denoising brain-inspired complementary vision sensor. *Neuromorphic Computing and Engineering*, 2025b.

Xuezhen Wang, Li Ma, Yulin Shen, Zeyu Wang, and Pedro V. Sander. RetimeGS: Continuous-Time Reconstruction of 4D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7340–7350, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, 2024.

Bo Xu, Jun Dai, Yutian Chen, Linning Xu, Mulin Yu, Yujin Wang, Shi Guo, Xinyi Le, and Tianfan Xue. 4D E-SloMo: 4D Reconstruction for High Speed Scene Using a Hybrid RGB-Event Multi-View System. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43–53, 2026a.

Chuanzhi Xu, Haoxian Zhou, Langyi Chen, Haodong Chen, Zeke Zexi Hu, Zhicheng Lu, Ying Zhou, Vera Chung, Qiang Qu, and Weidong Cai. A Survey of 3D Reconstruction with Event Cameras. *Computational Visual Media*, 2026b.

Wenhao Xu, Wenming Weng, Yueyi Zhang, Ruikang Xu, and Zhiwei Xiong. Event-Boosted Deformable 3D Gaussians for Dynamic Scene Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28334–28343, 2025.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-Time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 9142–9159, 2024a.

Zheyu Yang, Taoyi Wang, Yihan Lin, Yuguo Chen, Hui Zeng, Jing Pei, Jiazheng Wang, Xue Liu, Yichun Zhou, Jianqiang Zhang, et al. A Vision Chip with Complementary Pathways for Open-World Sensing. *Nature*, 629(8014):1027–1033, 2024b.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20331–20341, 2024c.

Wangbo Yu, Chaoran Feng, Jianing Li, Jiye Tang, Jiashu Yang, Zhenyu Tang, Meng Cao, Xu Jia, Yuchao Yang, Li Yuan, et al. EvaGaussians: Event Stream Assisted Gaussian Splatting from Blurry Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24780–24790, 2025.

Toshiya Yura, Ashkan Mirzaei, and Igor Gilitschenski. EventSplat: 3D Gaussian Splatting from Moving Event Cameras for Real-Time Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26876–26886, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Hanyu Zhou, Haonan Wang, Haoyue Liu, Yuxing Duan, Luxin Yan, and Gim Hee Lee. STD-GS: Exploring Frame-Event Interaction for SpatioTemporal-Disentangled Gaussian Splatting to Reconstruct High-Dynamic Scene. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24801–24810, 2025.

Zihao Zou, Ziyuan Qu, Xi Peng, Vivek Boominathan, Adithya Pediredla, and Praneeth Chakravarthula. High-Speed Dynamic 3D Imaging with Sensor Fusion Splatting. *arXiv preprint arXiv:2502.04630*, 2025.

A Appendix

A.1 Implementation Details

All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. We adopt a progressive training strategy to optimize the dynamic Gaussian representation. Specifically, we first perform 3,000 coarse-stage iterations, followed by 20,000 fine-stage iterations. During the first 10,000 iterations of the fine stage, the model is optimized using only sparse RGB keyframes to establish stable scene geometry, appearance, and coarse deformation. TD and SD supervision is introduced from the 10,001st iteration, and the remaining 10,000 iterations jointly refine scene appearance, spatial structure, and temporal motion. The weights of the RGB, SDL, SDR, and TD loss terms are set to $\lambda_{\mathrm{RGB}} = 1.0$, $\lambda_{\mathrm{SDL}} = 0.5$, $\lambda_{\mathrm{SDR}} = 0.5$, and $\lambda_{\mathrm{TD}} = 1$, respectively.

A.2 Physical Noise Model

When generating the simulated dataset, we introduce fixed-pattern noise and temporally random noise according to the noise characteristics of the Tianmouc sensor. Wang et al. (2025b) demonstrated Tianmouc noise distribution model. We perform noise simulation based on this noise model. For each channel $c \in \{\mathrm{TD}, \mathrm{SDL}, \mathrm{SDR}\}$, the simulated sensor measurement is formulated as

$$Y_c(i,j,t) = D_c(i,j,t) + N_c^{\mathrm{FPN}}(i,j) + N_c^{\mathrm{rand}}(i,j,t), \tag{14}$$

where D_c denotes the ideal sensor response, N_c^{FPN} denotes the fixed-pattern noise, and N_c^{rand} denotes the temporally random noise.

Fixed-pattern noise mainly originates from pixel-to-pixel variations in photoelectric conversion gain, device threshold, and readout offset. Based on dark-frame statistics collected from the real sensor, it is modeled as

$$N_c^{\mathrm{FPN}}(i,j) \sim \mathcal{N}\left(\mu_c, \sigma_c^2\right), \tag{15}$$

where μ_c and σ_c are the calibrated mean and standard deviation for channel c, respectively. The sampled fixed-pattern noise map remains unchanged within the same simulated sequence, thereby reproducing the spatially structured, pixel-dependent offsets of the physical sensor.

To model photon shot noise and random fluctuations introduced during differential readout, we use the difference between two independent Poisson random variables:

$$P_1, P_2 \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \qquad N_c^{\text{rand}} = \alpha_c(P_1 - P_2). \tag{16}$$

The difference $P_1 - P_2$ follows a Skellam distribution. Consequently,

$$\mathbb{E}\big[N_c^{\text{rand}}\big] = 0, \qquad \text{Var}\big(N_c^{\text{rand}}\big) = 2\alpha_c^2\lambda. \tag{17}$$

Given a calibrated target standard deviation σ_c^{rand}, we set

$$\alpha_c = \frac{\sigma_c^{\text{rand}}}{\sqrt{2\lambda}}, \tag{18}$$

such that

$$\text{Std}\big[N_c^{\text{rand}}\big] = \sigma_c^{\text{rand}}. \tag{19}$$

This formulation preserves the discrete statistics of photon counting and is consistent with the physical process in which the sensor forms a differential signal from two photoelectric measurements. Based on the measured channel-dependent noise characteristics, separate parameters are assigned to the TD, SDL, and SDR channels.

Importantly, the above noise terms are injected only during the offline generation of the simulated dataset. During model training, the differentiable simulator retains the deterministic grayscale response, spatial sampling, and TD/SD differential readout processes, but does not additionally inject fixed-pattern or random noise.

A.3 Details of the AOP Sampling Operator

In the main paper, the mapping from a rendered RGB image to the simulated AOP pixel output voltage is formulated as

$$\widehat{\mathbf{V}}_{c,t} = \mathcal{P}\left(\mathcal{Y}\left(\widehat{\mathbf{I}}_{c,t}\right)\right), \tag{20}$$

where $\mathcal{Y}(\cdot)$ denotes the luminance conversion and $\mathcal{P}(\cdot)$ denotes the AOP sampling operator. Here, we provide additional implementation details of $\mathcal{P}(\cdot)$ to clarify how the rendered image is spatially mapped to the AOP sensing layout.

Luminance conversion. Given the rendered RGB image $\widehat{\mathbf{I}}_{c,t}$, we first resize it to the sensor resolution $H_s \times W_s$ using bilinear interpolation. Following the implementation of our simulator, the luminance conversion $\mathcal{Y}(\cdot)$ is defined as

$$\widehat{\mathbf{I}}^Y_{c,t}(i,j) = 0.299\,\widehat{\mathbf{I}}^R_{c,t}(i,j) + 0.587\,\widehat{\mathbf{I}}^G_{c,t}(i,j) + 0.114\,\widehat{\mathbf{I}}^B_{c,t}(i,j), \tag{21}$$

where $\widehat{\mathbf{I}}^Y_{c,t}$ denotes the monochromatic intensity response used as the input to the AOP sampling operator.

Pixel binning. The first step of $\mathcal{P}(\cdot)$ performs non-overlapping 2×2 pixel binning. We denote the intermediate binned response as $\widehat{\mathbf{V}}^{\text{bin}}_{c,t}$. For each spatial location (m, n), it is computed as

$$\begin{aligned}\widehat{\mathbf{V}}^{\text{bin}}_{c,t}(m,n) = \frac{1}{4}\big[&\widehat{\mathbf{I}}^Y_{c,t}(2m,2n) + \widehat{\mathbf{I}}^Y_{c,t}(2m+1,2n) \\ &+ \widehat{\mathbf{I}}^Y_{c,t}(2m,2n+1) + \widehat{\mathbf{I}}^Y_{c,t}(2m+1,2n+1)\big].\end{aligned} \tag{22}$$

Thus, each element of $\widehat{\mathbf{V}}^{\text{bin}}_{c,t}$ represents the average response over a 2×2 region of the resized luminance image.

Spatially interleaved sampling. After pixel binning, we reproduce the spatially interleaved sampling pattern of the AOP by selecting alternating columns according to the parity of each row. The final simulated AOP pixel output voltage $\widehat{\mathbf{V}}_{c,t}$ is obtained as

$$\widehat{\mathbf{V}}_{c,t}(r,k) = \begin{cases} \widehat{\mathbf{V}}^{\text{bin}}_{c,t}(r,2k), & r \bmod 2 = 0, \\ \widehat{\mathbf{V}}^{\text{bin}}_{c,t}(r,2k+1), & r \bmod 2 = 1. \end{cases} \tag{23}$$

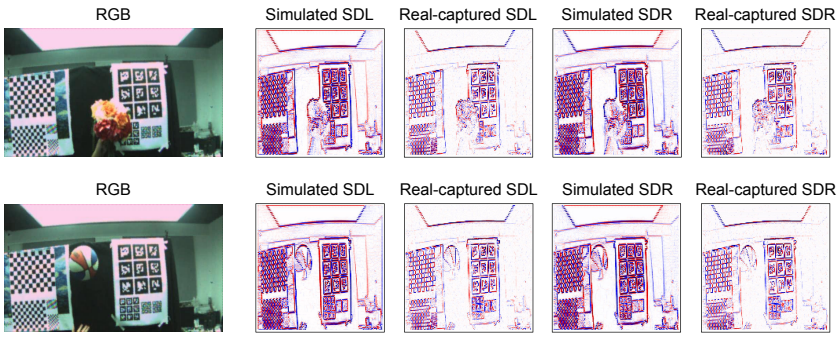

Figure 7: Comparison between Simulated and Real SD Responses.

That is, even rows retain samples from even columns, whereas odd rows retain samples from odd columns. This row-dependent sampling produces the spatially staggered arrangement used to approximate the physical AOP sensing layout.

For the sensor resolution used in our implementation, $H_s = 320$ and $W_s = 640$. The 2×2 pixel-binning operation first transforms the 320×640 luminance image into a 160×320 intermediate response. The subsequent spatially interleaved sampling retains half of the samples along the horizontal dimension, resulting in the final 160×160 AOP response. The resulting $\widehat{\mathbf{V}}_{c,t}$ is then directly used to construct the predicted temporal and spatial difference responses $\widehat{\mathbf{TD}}_{c,t}$, $\widehat{\mathbf{SD}}_{L,c,t}$, and $\widehat{\mathbf{SD}}_{R,c,t}$ following Eqs. (7)–(9) in the main paper. We therefore omit the repeated TD/SD formulations here. All operations involved in the above mapping, including bilinear resizing, luminance conversion, pixel averaging, and indexed sampling, are implemented with differentiable tensor operations, allowing the TD/SD supervision to backpropagate through $\mathcal{P}(\cdot)$ to the rendered Gaussian representation.

A.4 Validation of the Difference Response Simulator

To validate the effectiveness of the proposed difference response simulator, we use real captured RGB images to generate simulated SD responses and compare them with SD measurements captured by the sensor at nearby timestamps. As shown in Fig. 7, the simulated SD responses successfully reproduce the major spatial structures and response characteristics observed in the real SD measurements, demonstrating the simulator's ability to effectively model the SD sensor response. Based on this simulator, we further generate TD/SD signals for constructing the simulated dataset and providing supervision during model training. It should be noted that TD responses depend on high-frequency temporal variations and therefore require visual observations with sufficiently high temporal resolution for direct validation. Since our acquisition system cannot simultaneously capture high-frame-rate RGB sequences with a temporal resolution comparable to that of the TD measurements, we are unable to perform a direct comparison between RGB-derived simulated TD responses and real TD measurements in the same manner.

A.5 More Results on Synthetic Datasets

In Table 3, we provide the quantitative results of our method and the compared approaches on each scene in the synthetic dataset. Our method achieves the best PSNR on five out of the six scenes and the best SSIM on four scenes, demonstrating its effectiveness in recovering high-fidelity appearance and spatial structures. Although E-D3DGS Xu et al. (2025) obtains the lowest LPIPS values, our method achieves the second-best LPIPS on four scenes while consistently delivering superior PSNR and SSIM results. To further evaluate the reconstruction performance under diverse dynamic conditions, we provide additional visual comparisons on the remaining three scenes in fig. 8.

Table 3: Quantitative comparison on dynamic scenes. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. The best and second-best results are shown in **bold** and underlined, respectively.

Scene	Metric	CBRDM+ 4DGaussians	4DGS	4D Gaussians	Ex4DGS	E-D3DGS	Ours
Coffee	PSNR ↑	21.19	24.94	26.30	**27.86**	26.19	26.58
	SSIM ↑	0.679	0.876	0.887	**0.902**	0.884	0.892
	LPIPS ↓	0.294	0.158	0.145	0.137	**0.077**	0.148
Spinach	PSNR ↑	22.43	26.79	30.60	31.02	31.03	**32.22**
	SSIM ↑	0.739	0.901	0.931	0.932	0.927	**0.940**
	LPIPS ↓	0.221	0.121	0.091	0.103	**0.070**	0.089
Cut	PSNR ↑	22.63	26.64	30.54	27.86	31.05	**32.36**
	SSIM ↑	0.759	0.911	0.935	0.902	0.929	**0.947**
	LPIPS ↓	0.224	0.109	0.083	0.137	**0.068**	0.078
Salmon	PSNR ↑	21.24	22.87	26.26	27.01	26.74	**27.32**
	SSIM ↑	0.691	0.830	0.893	**0.901**	**0.901**	0.900
	LPIPS ↓	0.285	0.221	0.140	0.141	**0.065**	0.139
Flame	PSNR ↑	21.07	26.68	30.81	30.57	30.68	**32.52**
	SSIM ↑	0.688	0.915	0.942	0.939	0.941	**0.949**
	LPIPS ↓	0.328	0.121	0.078	0.097	**0.067**	0.077
Sear	PSNR ↑	23.18	27.70	32.15	32.10	31.60	**32.67**
	SSIM ↑	0.778	0.922	0.942	0.941	0.943	**0.944**
	LPIPS ↓	0.188	0.103	0.077	0.097	**0.055**	0.080

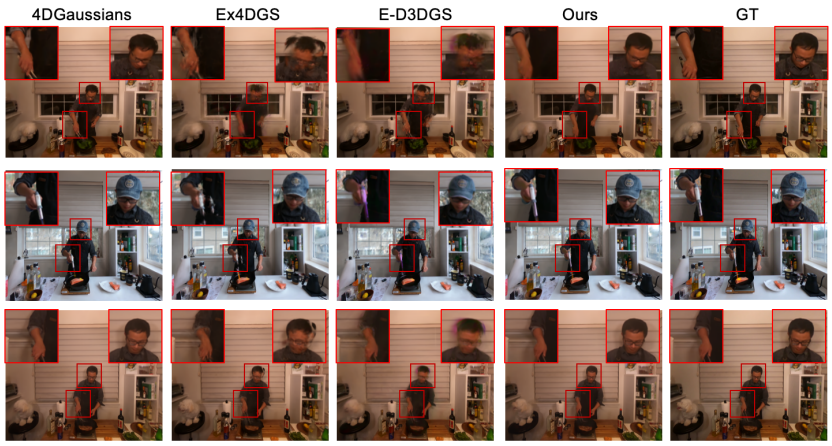

Figure 8: Additional qualitative comparisons on the remaining three dynamic scenes, complementing the three scenes presented in the paper.

A.6 Reconstruction Efficiency Comparison

We further compare our method with a preprocessing-based reconstruction pipeline. This pipeline first employs CBRDM Meng et al. (2025) to reconstruct conventional 2D intensity images from the Tianmouc differential measurements and then uses the reconstructed images as preprocessing results for subsequent 4D scene reconstruction. The evaluated sequence, which is a scene from the synthetic dataset, contains 20 views, with 12 RGB frames and 300 TD/SD frames per view, resulting in a total of 6,000 TD/SD frames. CBRDM requires approximately 6 minutes and 38 seconds to reconstruct

Preprocessing-based Pipeline (CBRDM)	
Completed stage	2D preprocessing only
Preprocessing time	26 h 32 min
Total runtime	$>$ 26 h 32 min
Ours	
Completed stage	Complete 4D reconstruction
Preprocessing time	Not required
Total runtime	1 h 49 min

Table 4: Runtime comparison on a multi-view sequence containing 20 views and 300 TD/SD frames per view.

every 25 frames. Therefore, preprocessing all 6,000 frames requires approximately 26 hours and 32 minutes.

It should be emphasized that this runtime covers only the CBRDM-based 2D image preprocessing stage and does not include the subsequent 4D scene reconstruction. Therefore, the total runtime of the complete preprocessing-based reconstruction pipeline would be substantially longer than 26 hours and 32 minutes. In contrast, our method directly optimizes the dynamic Gaussian representation using RGB images and differential measurements, eliminating the need to reconstruct dense 2D intensity images in advance. Our method requires only approximately 1 hour and 49 minutes to complete the entire 4D dynamic scene reconstruction.

As shown in Table 4, the time required by our method to complete the entire 4D reconstruction is approximately 93.2% shorter than that required by the preprocessing-based method to perform only 2D image preprocessing, corresponding to a speedup of at least $14.6\times$. Since the additional time required by the preprocessing-based method for subsequent 4D scene reconstruction is not included, the reported $14.6\times$ speedup represents a conservative lower bound on the actual acceleration achieved by our method.

A.7 Effectiveness of the Two-Stage Training Strategy

Directly applying joint RGB and TD/SD supervision from the beginning of training imposes a considerable representational burden on the Gaussian model. TD and SD signals predominantly encode high-frequency temporal and spatial information, with strong responses concentrated around motion boundaries, fine textures, and local intensity discontinuities. If these high-frequency constraints are introduced before the scene geometry and low-frequency appearance have sufficiently converged, the model tends to reduce local differential errors by aggressively splitting and duplicating Gaussian primitives. This premature densification produces redundant primitives, substantially increases the model size, and may interfere with early-stage geometry optimization.

To address this issue, we adopt a coarse-to-fine two-stage training strategy. In the first stage, RGB supervision is used to establish stable scene geometry and appearance. TD/SD supervision is subsequently introduced in the second stage to refine temporal variations, motion boundaries, and high-frequency spatial details. As shown in Table 5, our strategy substantially suppresses excessive Gaussian densification while preserving reconstruction quality.

Our two-stage strategy reduces the average number of Gaussian primitives from 333,069 to 128,676 and the model size from 78.78 MB to 30.43 MB, corresponding to reductions of approximately 61.37% in both model complexity and storage. Equivalently, the direct joint-training baseline requires approximately $2.59\times$ as many Gaussian primitives and as much storage as our method.

This substantial complexity reduction is achieved with negligible changes in pixel-level reconstruction quality: PSNR decreases by only 0.046 dB and SSIM by only 0.0010. Meanwhile, LPIPS improves from 0.1047 to 0.1021. These results demonstrate that the proposed two-stage strategy provides a considerably better quality–complexity trade-off, effectively suppressing the redundant Gaussian proliferation induced by prematurely introducing high-frequency TD/SD supervision while preserving both pixel-level fidelity and perceptual quality.

Metric	w/o Two-Stage	Two-Stage
Image Quality		
PSNR ↑	**30.657**	30.611
SSIM ↑	**0.9296**	0.9286
LPIPS ↓	0.1047	**0.1021**
Storage Efficiency		
# Gaussians ↓	333,069	**128,676**
Model Size ↓	78.78 MB	**30.43 MB**

Table 5: Comparison of reconstruction quality and storage efficiency with and without the proposed two-stage training strategy.

A.8 LIMITATIONS

Despite the promising performance of Triplex4DGS in high-speed dynamic scene reconstruction, several limitations remain. First, the refinement stage relies on reliable geometry and deformation initialization provided by sparse RGB keyframes. Consequently, severe motion blur, low-texture regions, or inaccurate camera calibration may introduce errors during initialization that cannot be fully corrected using differential supervision alone. Second, although the differentiable simulator models the primary TD/SD sensing process, discrepancies between simulated and real sensor responses may persist under complex illumination changes. Finally, our current real-world experiments are conducted using a six-camera Tianmouc system and cover a relatively limited number of scenes. Due to the lack of accurate high-speed 4D ground truth, reconstruction quality on real-world scenes cannot yet be comprehensively evaluated quantitatively. Future work will investigate more robust sensor modeling and calibration-aware optimization, as well as evaluation on larger and more diverse real-world datasets.