EVENT4D: DENSE 4D RECONSTRUCTION FROM RGB AND NEUROMORPHIC EVENT STREAMS

Anonymous authors
Paper under double-blind review

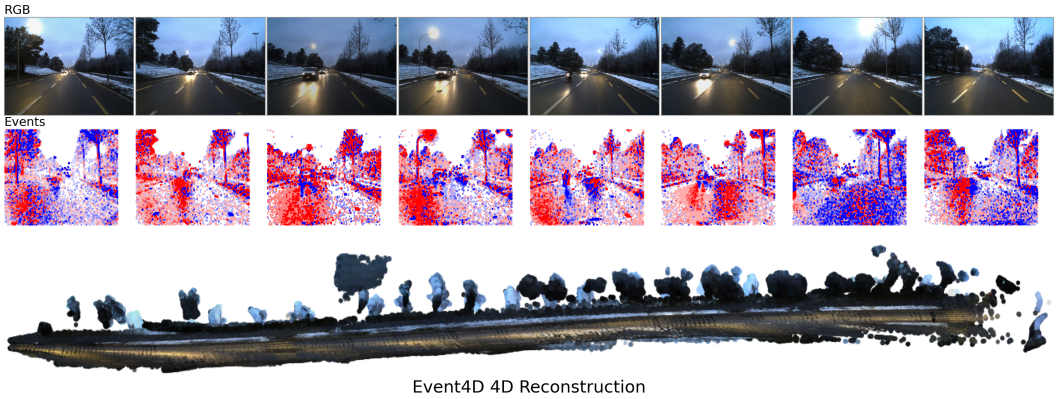


Figure 1: **Dense 4D reconstruction from sparse RGB and events.** Event4D combines temporally sparse RGB observations (top) with event measurements, represented as voxel grids (middle), to reconstruct scene geometry and camera motion at every reconstruction time step. Events provide motion information between RGB captures, enabling higher-fidelity 3D reconstruction at RGB frame rates. Bottom: the resulting 3D reconstruction accumulated over the sequence.

ABSTRACT

Reconstructing dynamic 3D scenes over time (4D reconstruction) from a single moving camera is important for robotics and AR/VR. However, existing RGB-only feed-forward methods are vulnerable to motion blur and temporal aliasing under rapid camera or object motion. Event cameras asynchronously capture per-pixel brightness changes at microsecond resolution, providing complementary temporal information between RGB frames. We introduce Event4D, the first single-stage feed-forward 4D reconstruction pipeline that fuses events and RGB images. Event4D pairs event voxels at fine-grained reconstruction time steps with the most recently captured RGB frame and fuses their features through a learned RGB-staleness conditioning. Using five benchmark datasets, we evaluate Event4D under three challenging conditions for RGB-based 4D reconstruction: fast object motion, low illumination, and temporal aliasing. Event4D improves threshold accuracy over the state-of-the-art by 10.11% on fast-motion sequences, 59.29% under temporal aliasing, and 11.61–24.77% across low-light sequences.

1 INTRODUCTION

Recovering the time-varying 3D geometry of a dynamic scene together with the camera trajectory from monocular video—commonly referred to as *4D reconstruction*—is an important capability for autonomous driving, robotics, and mixed reality applications. The problem is especially challenging because camera ego-motion is entangled with the independent motion of objects in the scene, violating the static-world assumption that underlies classical structure-from-motion, SLAM, and multi-view stereo pipelines Cai et al. (2026); Zhou et al. (2026). A major shift in this direction came with DUSt3R Wang et al. (2024a), which directly regresses dense pointmaps from uncalibrated and unposed image pairs, allowing camera parameters, depth, and correspondences to be recovered from a shared representation. MonST3R Zhang et al. (2025) extended this formulation to dynamic scenes

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.

through fine-tuning on dynamic data, motivating a rapidly growing line of work. Recent methods combine per-frame geometry with motion cues such as optical flow, point tracks, and dynamic masks Wang et al. (2025c); Park & Shin (2026), while others extend the formulation to multi-frame and streaming settings using 3D foundation models Wang et al. (2025a;b); Lan et al. (2026); Zhou et al. (2026). As a result, temporally consistent video depth has become one of the key outputs of modern 4D reconstruction systems. However, these systems still rely almost entirely on a single sensing modality.

This reliance on RGB frames means that existing methods inherit the limitations of conventional cameras. Fast motion can cause severe motion blur, low-light and high-contrast conditions can remove useful visual information, and the fixed frame rate leaves scene dynamics between consecutive frames unobserved. Event cameras provide a natural complement to frame-based sensing. Rather than capturing complete images at fixed intervals, they asynchronously report per-pixel brightness changes with microsecond temporal resolution, offer a higher dynamic range than conventional cameras, and are largely free from motion blur Hidalgo-Carrió et al. (2020); Gehrig et al. (2021a). These properties allow event cameras to preserve useful structural information precisely in conditions where frame-based pipelines tend to degrade Shi et al. (2023). More importantly, events capture scene motion during the temporal gaps between RGB frames, where conventional 4D reconstruction systems receive no measurements. Events alone, however, encode only changes in brightness and contain little absolute appearance information; event cameras also produce few or no measurements in static or weakly textured regions Gehrig et al. (2021a); Devulapally et al. (2024); Bartolomei et al. (2025b). RGB images provide this missing contextual information. The two sensing modalities are therefore highly complementary, suggesting that their fusion could make 4D reconstruction more robust to rapid motion and challenging illumination while also enabling geometry to be recovered at a finer temporal resolution than the RGB frame rate alone allows.

Event-based geometry estimation, however, has largely developed separately from this RGB-centric progress in 4D reconstruction. Most event-based work has instead focused on monocular depth estimation, progressing from recurrent encoder–decoder architectures Hidalgo-Carrió et al. (2020) and RGB–event fusion networks Gehrig et al. (2021a); Devulapally et al. (2024) to more recent state-space model architectures Wang et al. (2026). Several recent approaches also use cross-modal distillation from image-based depth foundation models to address the limited availability of dense ground-truth depth for event cameras Bartolomei et al. (2025b); Lee & Lee (2026). While these methods achieve strong depth estimation performance, they typically predict geometry independently in the coordinate frame of each input view. They do not jointly recover camera pose, maintain a consistent geometric scale over time, or disentangle camera ego-motion from independently moving objects—all of which are central requirements for 4D reconstruction. EAG3R Wu et al. (2026b) remains the only prior work to achieve 4D reconstruction through event and RGB fusion. However, like MonST3R, it relies on a multi-stage pipeline that requires post-processing to align independently predicted pointmaps. RGB-based reconstruction methods have increasingly moved away from such pipelines, as separating depth estimation and alignment can limit overall reconstruction accuracy. Furthermore, EAG3R is evaluated primarily under low-light conditions, leaving other challenging settings in which events can provide complementary information, such as fast motion and temporal aliasing, unexplored.

To this end, we propose Event4D, a feed-forward model for 4D reconstruction that combines RGB images with event measurements. Given a sequence of N reconstruction time steps, Event4D predicts a 3D pointmap at each time step using the corresponding event representation and the most recently captured RGB image. This formulation enables reconstruction at a higher temporal resolution than the RGB capture rate by exploiting event measurements between successive RGB frames.

Training such a model is challenging because datasets containing synchronized RGB images, events, and ground-truth geometry are scarce. We address this limitation by adding synthetically generated events to existing RGB datasets that provide depth supervision. Combining these augmented datasets with real RGB–event datasets, we train a single network to jointly predict video depth, temporally indexed 3D pointmaps, and camera poses.

In summary, we make the following **key contributions**:

1. **Event4D for Event- and RGB-Based 4D Reconstruction:** We present Event4D, the first feed-forward event- and RGB-based 4D reconstruction model designed to accurately estimate scene

geometry and motion under fast motion, temporal aliasing, and low-light conditions. EAG3R, the only prior work on event–RGB 4D reconstruction, requires a separate post-processing stage to temporally align pointmaps. In contrast, Event4D performs 4D reconstruction through a single feed-forward architecture, achieving more accurate reconstruction.

2. **Synthetic Event, RGB, and Depth Datasets:** To address the scarcity of paired event, RGB, and depth datasets for training a foundational model such as Event4D, we propose an automated augmentation approach that synthesizes event representations from RGB videos using the v2e simulator Hu et al. (2021). Through this approach, we introduce five synthetic event, RGB, and depth datasets: N-PointOdyssey, N-Spring, N-TartanAir, N-Waymo, and N-Sintel, extending PointOdyssey Zheng et al. (2023), Spring Mehl et al. (2023); Oei et al. (2026), TartanAir Wang et al. (2020), Waymo Sun et al. (2020), and Sintel Butler et al. (2012), respectively. We use N-Sintel for evaluation and the remaining four datasets to train Event4D, consistent with prior RGB-based 4D reconstruction methods Wang et al. (2024a) that primarily use Sintel as an evaluation dataset. These datasets will be publicly released.
3. **Performance of Event4D:** We evaluate Event4D on three complementary 4D reconstruction tasks: (a) monocular depth estimation, (b) video depth estimation, and (c) camera pose estimation. Across these tasks, we primarily examine three known failure modes of conventional RGB cameras: (a) fast object motion, (b) temporal aliasing, and (c) low-light conditions. Event4D achieves superior reconstruction accuracy under these conditions, improving threshold accuracy by 10.11%, 59.29%, and 11.61–24.77% compared with prior state-of-the-art models under fast object motion, temporal aliasing, and low-light conditions, respectively.

2 Related Work

Prior work relevant to Event4D falls into two categories: monocular depth estimation and video depth estimation. Monocular depth estimation predicts per-pixel depth but generally does not recover camera pose or a globally consistent 3D representation. Even video-based variants typically reason over individual frames or short temporal windows without explicit cross-view geometric reasoning. Video depth methods instead exploit correspondences across multiple views to recover geometry consistent in scale and coordinate frames, enabling camera ego-motion to be distinguished from independently moving scene elements.

Monocular Depth Estimation. These works predict a dense depth map from a single RGB image without explicitly reasoning about camera geometry or motion. Applied independently to video frames, such models often produce temporal flickering and scale drift Ke et al. (2025), motivating temporally consistent video extensions that still predict depth as the primary output.

Recent efforts to extend monocular depth estimation to video follow three broad strategies. The first repurposes video diffusion priors to produce temporally consistent depth through fine-tuned image-to-video diffusion, cross-frame context sharing, or training-free guidance that mitigates sliding-window drift Hu et al. (2025b); Dong et al. (2025). The second adapts feed-forward single-image depth backbones with lightweight temporal components, such as spatio-temporal heads, flow-based feature propagation, and temporal embeddings, tackling the computational cost of diffusion Chen et al. (2025); Feiden et al. (2026). A third treats temporal consistency as a model-agnostic add-on, with plug-and-play stabilizers such as NVDS+ Wang et al. (2024b) and PRIDEV Xu et al. (2025).

Event cameras provide an alternative temporal signal, either replacing or complementing RGB. E2Depth Hidalgo-Carrió et al. (2020) predicts dense depth directly from events using a recurrent architecture. Because events lack absolute appearance and are sparse in static or weakly textured regions, several methods fuse them with RGB: RAM Gehrig et al. (2021a) performs asynchronous recurrent fusion, ER-F2D Devulapally et al. (2024) jointly attends over event and RGB tokens, EVEN Shi et al. (2023) targets adverse illumination, and EF-SAT Arigela et al. (2026) learns fusion without dense ground-truth depth.

The scarcity of dense event-depth supervision has also motivated cross-modal distillation. Depth AnyEvent Bartolomei et al. (2025b) transfers dense proxy supervision from a vision foundation model, while Distil-E2D Lee & Lee (2026) combines this with sparse LiDAR. Other methods tailor architectures to sparse asynchronous events, including state-space models such as ESDepth Wang et al. (2026), EDE-Distill Hu et al. (2025a), and GLU-Net Wu et al. (2026a). Despite these advances,

both RGB-only and event-based monocular depth methods predict depth without jointly recovering the camera trajectory or maintaining a globally consistent 3D coordinate system.

Video Depth Estimation. Video depth estimation, the setting considered in our work, jointly recovers scene geometry and camera motion, typically as per-frame depth or pointmaps in a consistent 3D coordinate system. DUSt3R Wang et al. (2024a) directly regresses dense pointmaps from uncalibrated, unposed image pairs, estimating camera parameters, depth, and correspondences from a shared representation. MASt3R Leroy et al. (2024) adds a dense matching head, Spann3R Wang & Agapito (2025) replaces optimization-based multi-view alignment with spatial memory, and VGGT Wang et al. (2025a) uses a single feed-forward transformer to infer cameras, depth, pointmaps, and 3D tracks from many views.

This formulation has also been extended to dynamic scenes. MonST3R Zhang et al. (2025) adapts DUSt3R through dynamic-scene fine-tuning. C4D Wang et al. (2025c) adds optical flow and long-term tracking. Track3R Park & Shin (2026) predicts pointmaps together with separate object- and camera-motion trajectories. For streaming settings or larger sets of views, CUT3R Wang et al. (2025b) and STream3R Lan et al. (2026) use persistent state and causal attention, respectively. Mem4D Cai et al. (2026) combines short- and long-term memory. Fast3R Yang et al. (2025) processes many views in one forward pass. PAGE-4D Zhou et al. (2026) extends VGGT with dynamics-aware masking for pose and geometry.

A key distinction is whether pointmap prediction and multi-view alignment are solved jointly or separately. DUSt3R Wang et al. (2024a), MonST3R Zhang et al. (2025), and MASt3R Leroy et al. (2024) operate pairwise and rely on a separate optimization-based alignment stage. EAG3R Wu et al. (2026b), the only prior event-based pointmap reconstruction method, follows the same MonST3R-based two-stage design and is evaluated primarily under low-light conditions. In contrast, VGGT Wang et al. (2025a), Fast3R Yang et al. (2025), STream3R Lan et al. (2026), and PAGE-4D Zhou et al. (2026) are feed-forward and single-stage.

Event4D differs along two axes. First, unlike DUSt3R, MonST3R, MASt3R, and EAG3R, Event4D performs 4D reconstruction in a single feed-forward architecture without a separate global alignment stage. Second, unlike VGGT, Fast3R, STream3R, PAGE-4D, and other RGB-only methods, Event4D incorporates events alongside RGB frames. At each time step, it predicts a pointmap from an event representation and the most recent RGB image, enabling 4D reconstruction with higher fidelity. We evaluate this capability under rapid motion, temporal aliasing, and low-light illumination.

3 METHODOLOGY

Problem Formulation: We illustrate the proposed Event4D architecture in Figure 2. We consider a sequence of N reconstruction time steps. At time step i, the model receives the most recently captured RGB image $\bar{\mathbf{I}}_i \in \mathbb{R}^{3\times H\times W}$ and an event voxel $\mathbf{V}_i \in \mathbb{R}^{5\times H\times W}$. Each voxel aggregates signed brightness-change measurements into five temporal bins over the event window associated with timestamp j as follows. An asynchronous event stream over the time interval $(t, t+\Delta)$ is defined as $\mathcal{E}_{(t,t+\Delta)} = \{e_j = (p_j, x_j, y_j, t_j)\}_{j=0}^{M}$, where $p_j \in \{-1, 1\}$ denotes the event polarity corresponding to an increase or decrease in brightness, $(x_j, y_j) \in \mathbb{R}^{H\times W}$ represents the pixel coordinates, and t_j denotes the event timestamp.

The event stream is converted into a voxel grid representation Zhu et al. (2019) by accumulating events within each temporal bin as

$$V(b, x, y) = \sum_{e_j \in \mathcal{E}_{(t,t+\Delta)}} p_j \mathbf{1}\left\{b = \left\lfloor B \cdot \frac{t_j - t}{\Delta} \right\rfloor, x = x_j, y = y_j\right\}, \tag{1}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and $B = 5$ is the number of temporal bins.

Event windows remain available at every reconstruction step, independently of the RGB capture rate. We assume that preprocessing places the two modalities on a common spatial grid.

Let $\tau(i)$ denote the index of the latest RGB capture available at step i. We define a stale RGB input as $\bar{\mathbf{I}}_i = \mathbf{I}_{\tau(i)}$ and its *staleness* at step i as $s_i = i - \tau(i)$.

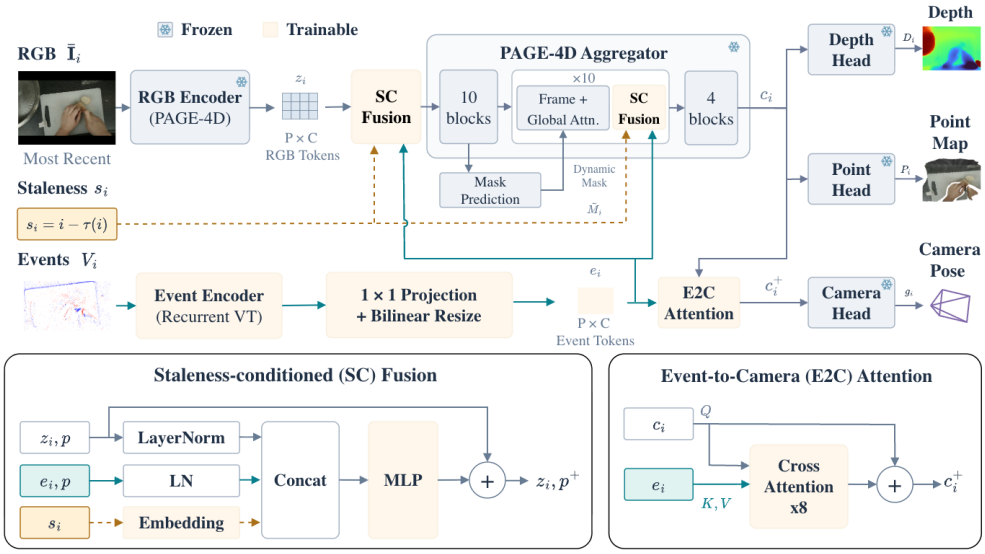


Figure 2: **Overview of Event4D.** (a) A frozen PAGE-4D RGB encoder processes the latest available image, while a recurrent event encoder provides features at every reconstruction time step. Event features are projected and resized to the RGB token grid, then injected through eleven staleness-conditioned residual adapters: one before aggregation and one in each of the ten middle blocks. (b) Each adapter concatenates normalized RGB and event features with a learned embedding of RGB staleness and predicts a tokenwise residual using an MLP. (c) Eight-head cross-attention updates the final camera token using spatial event features before pose prediction. The pretrained aggregator, dynamics masking, and prediction heads are retained.

Thus, $s_i = 0$ denotes a fresh RGB observation, while $s_i > 0$ denotes a repeated image. Given $\{(\bar{\mathbf{I}}_i, \mathbf{V}_i, s_i)\}_{i=1}^N$, our objective is to predict camera pose $\mathbf{g}_i$, depth $\mathbf{D}_i$, and a 3D point map $\mathbf{P}_i$ at every reconstruction time step, including those without a new RGB capture.

RGB Feature Backbone. We use the pretrained RGB encoder from PAGE-4D without modification and keep its parameters frozen. The encoder maps each input RGB image to a grid of P spatial tokens with feature dimension C, which are then fused with the corresponding event features.

Event Feature Backbone. We encode event voxels using a compact recurrent vision transformer (RVT) Gehrig & Scaramuzza (2023), initialized from its pretrained weights. The encoder combines spatial feature extraction with recurrent state updates to capture motion across successive event windows. We extract a 192-channel feature map at $1/16$ spatial resolution and project it to the RGB token dimension C using a 1×1 convolution. We then bilinearly resample the features to the RGB token grid and flatten them to obtain event tokens $\mathbf{e}_i \in \mathbb{R}^{P \times C}$. The resulting event features match the RGB tokens in spatial grid size and feature dimension, enabling tokenwise fusion.

Staleness-Conditioned Residual Fusion. To capture the gradual divergence of a scene from a captured RGB image, we condition event fusion explicitly on RGB staleness. At fusion site ℓ, a learned lookup table embeds the capped staleness into a 64-dimensional vector, $\mathbf{d}_i^\ell = \mathrm{Embed}_\ell\big(\min(s_i, 16)\big) \in \mathbb{R}^{64}$.

For RGB token $\mathbf{z}_{i,p}^\ell$ at spatial position p, the adapter predicts

$$\boldsymbol{\Delta}_{i,p}^\ell = \mathrm{MLP}_\ell\big([\mathrm{LN}(\mathbf{z}_{i,p}^\ell) \,\|\, \mathrm{LN}(\mathbf{e}_{i,p}) \,\|\, \mathbf{d}_i^\ell]\big), \qquad \mathbf{z}_{i,p}^\ell \leftarrow \mathbf{z}_{i,p}^\ell + \boldsymbol{\Delta}_{i,p}^\ell. \tag{2}$$

Here, LN denotes layer normalization, $\|$ denotes channel concatenation, and the MLP consists of two linear layers separated by GELU. The staleness embedding is shared across spatial positions within a frame. Fusion operates on corresponding grid positions, without additional cross-attention over spatial tokens.

We apply this adapter once before the first aggregation block and once after global attention in each of the ten middle blocks, giving eleven independently parameterized fusion modules. The input adapter supplies event-conditioned features to the initial frozen blocks, while the intermediate

adapters update the representation as cross-frame information is aggregated. Each adapter has its own staleness embedding and MLP.

Event-to-Camera Attention. Spatial event fusion contributes to camera pose estimation through the frozen aggregator. To additionally provide the camera head with direct access to event features, we update each frame's camera token after the final aggregation block using residual cross-attention:

$$\mathbf{c}_i^+ = \mathbf{c}_i + \mathrm{MHA}_\phi\big(Q = \mathbf{c}_i,\ K = \mathbf{e}_i,\ V = \mathbf{e}_i\big), \tag{3}$$

where the camera token $\mathbf{c}_i$ serves as the query and the event tokens $\mathbf{e}_i$ serve as keys and values. The eight-head attention module aggregates spatial event features to refine the camera representation before pose prediction.

The updated token $\mathbf{c}_i^+$ replaces $\mathbf{c}_i$ in the output token sequence. We retain PAGE-4D's pretrained camera, depth, and point-map heads without modification and keep their parameters frozen. The camera head predicts pose from the updated camera token, while the depth and point-map heads operate on patch tokens enriched by the spatial fusion modules described earlier.

Training Datasets. To address the scarcity of paired event, RGB, and depth datasets for training Event4D, we use a combination of synthetic and real event datasets and train on seven datasets with geometry and camera-pose supervision. These comprise four synthetic event datasets—N-PointOdyssey, N-Spring, N-TartanAir, and N-Waymo—extending PointOdyssey Zheng et al. (2023), Spring Mehl et al. (2023); Oei et al. (2026), TartanAir Wang et al. (2020), and Waymo Sun et al. (2020), respectively, and three real event datasets: EVIMO Mitrokhin et al. (2019), DSEC Gehrig et al. (2021b), and EventKitchen Feng et al. (2026). Together, these datasets cover diverse camera and object motion patterns: trajectories derived from real videos in N-PointOdyssey, cinematic camera motion in N-Spring, simulated navigation trajectories in N-TartanAir, vehicle-mounted motion in N-Waymo and DSEC, indoor camera motion alongside independently moving objects in EVIMO, and head-mounted egocentric motion during cooking in EventKitchen.

We generate the four synthetic event datasets using v2e Hu et al. (2021), processing each sequence once at its native image resolution with a shared configuration across all four datasets. We first apply SuperSloMo Jiang et al. (2018) video interpolation to increase the temporal sampling rate by a factor of five, providing the event emulator with intermediate intensity observations between consecutive source frames. We then use the standard v2e configuration to generate raw timestamped event streams, which are converted into a five-bin voxel grid representation, as defined in Eq. 1. The same voxel grid representation is used for both synthetic and real events.

Combining synthetic and real event datasets provides Event4D with sufficient training samples to learn a foundational representation of 4D geometry from events and RGB and adapt to real-world sensor artifacts and errors.

Training Objective. We retain the depth, point-map, and absolute camera pose losses from VGGT/PAGE-4D without modification, denoted by $\mathcal{L}_{\mathrm{depth}}$, $\mathcal{L}_{\mathrm{point}}$, and $\mathcal{L}_{\mathrm{camera}}$, respectively. We additionally introduce relative-pose supervision over multiple temporal strides and a translation-direction term to explicitly constrain camera motion between frames.

To complement absolute camera pose supervision, we penalize relative-motion errors over temporal strides $\mathcal{K} = \{1, 2, 4, 8\}$. For world-to-camera extrinsics $[\mathbf{R}_i \mid \mathbf{t}_i]$, the relative transformation from frame $i - k$ to frame i is

$$\mathbf{R}_{i,\mathrm{rel}}^{(k)} = \mathbf{R}_i \mathbf{R}_{i-k}^\top, \qquad \mathbf{t}_{i,\mathrm{rel}}^{(k)} = \mathbf{t}_i - \mathbf{R}_{i,\mathrm{rel}}^{(k)} \mathbf{t}_{i-k}. \tag{4}$$

We compare these transformations with the ground-truth transformations using element-wise L_1 penalties:

$$\mathcal{L}_{\mathrm{rel}} = \left\langle \left| \mathbf{t}_{i,\mathrm{rel}}^{(k)} - \hat{\mathbf{t}}_{i,\mathrm{rel}}^{(k)} \right| \right\rangle_{i,k} + \left\langle \left| \mathbf{R}_{i,\mathrm{rel}}^{(k)} - \hat{\mathbf{R}}_{i,\mathrm{rel}}^{(k)} \right| \right\rangle_{i,k} \tag{5}$$

where the brackets denote averaging over valid frame pairs, temporal strides, and vector or matrix entries. Using multiple strides constrains camera motion across both short and long temporal intervals.

We also introduce a cosine loss on consecutive-frame translations to supervise direction independently of magnitude:

$$\mathcal{L}_{\mathrm{dir}} = \left\langle 1 - \frac{\mathbf{t}_{i,\mathrm{rel}}^{(1)} \cdot \hat{\mathbf{t}}_{i,\mathrm{rel}}^{(1)}}{\|\mathbf{t}_{i,\mathrm{rel}}^{(1)}\|_2 \|\hat{\mathbf{t}}_{i,\mathrm{rel}}^{(1)}\|_2} \right\rangle_{i \in \mathcal{M}}. \tag{6}$$

Table 1: Video and monocular depth estimation on five datasets. Lower is better for AbsRel. Higher is better for $\delta < 1.25$. The frame rate is set to the original rate F. Best results are in bold.

	DSEC		EventKitchen		N-Sintel		MVSEC		EVIMO	
Model	AbsRel ↓	$\delta < 1.25$ ↑	AbsRel ↓	$\delta < 1.25$ ↑	AbsRel ↓	$\delta < 1.25$ ↑	AbsRel ↓	$\delta < 1.25$ ↑	AbsRel ↓	$\delta < 1.25$ ↑
Video Depth Estimation										
CUT3R Wang et al. (2025b)	0.244	0.535	0.214	0.638	0.534	0.558	0.526	0.306	0.309	0.523
VGGT Wang et al. (2025a)	0.151	0.793	0.173	0.726	0.261	0.639	0.495	0.253	0.472	0.470
MonST3R Zhang et al. (2025)	0.211	0.710	0.198	0.676	0.360	0.489	**0.344**	**0.433**	0.397	0.443
MASt3R Leroy et al. (2024)	0.208	0.721	0.210	0.696	0.480	0.517	0.404	0.342	0.360	0.433
PAGE-4D Zhou et al. (2026)	0.078	0.940	0.135	0.807	0.212	**0.763**	0.486	0.250	0.229	0.768
Event4D (Ours)	**0.047**	**0.977**	**0.117**	**0.867**	**0.197**	0.735	0.451	0.251	**0.125**	**0.828**
Monocular Depth Estimation										
CUT3R Wang et al. (2025b)	0.182	0.692	0.163	0.755	0.458	0.594	0.367	0.363	0.120	0.841
VGGT Wang et al. (2025a)	0.063	0.966	0.122	0.847	0.269	0.697	0.481	0.257	0.131	0.836
MonST3R Zhang et al. (2025)	0.171	0.715	0.153	0.771	0.328	0.571	0.267	0.523	0.154	0.800
MASt3R Leroy et al. (2024)	0.170	0.717	0.144	0.788	0.349	0.602	0.301	0.451	0.169	0.808
Depth AnyEvent Bartolomei et al. (2025a)	0.442	0.374	0.289	0.560	0.550	0.354	**0.241**	**0.602**	0.332	0.470
PAGE-4D Zhou et al. (2026)	0.064	0.958	0.132	0.825	0.256	**0.754**	0.437	0.280	0.119	0.858
Event4D (Ours)	**0.045**	**0.979**	**0.117**	**0.869**	**0.231**	0.737	0.453	0.251	**0.048**	**0.927**

Here, $\mathcal{M}$ contains frame pairs whose ground-truth displacement exceeds a small fraction of the batch's median displacement, excluding pairs with negligible motion. We optimize the following weighted sum: $\mathcal{L} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{point}} + 5\mathcal{L}_{\text{camera}} + 25\mathcal{L}_{\text{rel}} + 0.0475\mathcal{L}_{\text{dir}}$.

We retain the original PAGE-4D/VGGT weights for the depth, point, and camera losses, and select the weights of the two added terms by matching their contributions to the existing objective. Specifically, we set the relative-pose weight to 25 to avoid the negligible contribution observed at 5 and trajectory degradation observed at 100, and scale the larger-magnitude direction loss by 0.0475 to provide meaningful supervision without dominating the objective.

Training Setup. We optimize only the recurrent event encoder and its projection, the eleven residual fusion modules, and the event-to-camera attention module, keeping all remaining parameters frozen at their pretrained PAGE-4D values. We train Event4D for eight epochs using AdamW with weight decay 0.01 and gradient clipping at 1. The learning rate increases linearly from 10^{-6} to 2×10^{-4} over the first 5% of training, then follows a cosine decay to 10^{-6} over the remaining 95%.

4 RESULTS

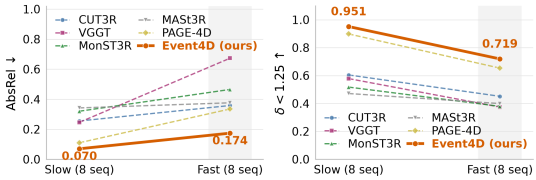
Evaluation Datasets. We evaluate Event4D on five datasets providing image sequences and reference depth: DSEC Gehrig et al. (2021b), EventKitchen Feng et al. (2026), N-Sintel (synthetically generated from Sintel Butler et al. (2012)), MVSEC Zhu et al. (2018), and EVIMO Mitrokhin et al. (2019), with recorded event streams available for all except N-Sintel. DSEC and MVSEC include outdoor driving sequences with LiDAR-derived depth and variations in camera motion, illumination, and scene content. EVIMO contains indoor recordings with independently moving objects for evaluating slow and fast object motion, while EventKitchen captures egocentric cooking activities with camera motion and close-range object interactions. Sintel comprises synthetic animated-film sequences with ground-truth depth, complex camera trajectories, and non-rigid scene motion. For N-Sintel, we synthesize events from Sintel's RGB sequences using the procedure in Section 3 and aggregate them into event voxels temporally aligned with the corresponding image frames.

Evaluation Tasks and Metrics. Following prior 4D reconstruction methods such as PAGE-4D, we evaluate Event4D on three tasks: monocular depth estimation, video depth estimation, and camera pose estimation. For both depth estimation tasks, we report absolute relative error (AbsRel) and threshold accuracy ($\delta < 1.25$). For camera pose estimation, we report absolute trajectory error (ATE) and the translational and rotational relative pose errors ($\text{RPE}_{\text{trans}}$ and RPE_{rot}, respectively).

Video Depth Estimation. Following prior work Zhang et al. (2025); Zhou et al. (2026), we first evaluate Event4D for video depth estimation across all five evaluation datasets. As summarized in Table 1, Event4D establishes a new state of the art in video depth estimation when compared with current RGB-based 4D reconstruction methods. For example, on EVIMO, which includes sequences with both slow and fast object motion, Event4D reduces AbsRel by 45.41% relative to PAGE-4D, the strongest RGB-based baseline on this benchmark.

Table 2: Camera pose estimation on N-Sintel, MVSEC and EVIMO. The frame rate is set to the original rate F. Lower is better for all metrics. Best results among the reported methods are in bold.

	N-Sintel			MVSEC			EVIMO		
Model	ATE ↓	$\text{RPE}_{\text{trans}}$ ↓	RPE_{rot} ↓	ATE ↓	$\text{RPE}_{\text{trans}}$ ↓	RPE_{rot} ↓	ATE ↓	$\text{RPE}_{\text{trans}}$ ↓	RPE_{rot} ↓
CUT3R	0.198	0.071	0.639	0.370	0.042	0.835	0.113	0.035	4.844
VGGT	0.175	**0.061**	**0.482**	0.486	0.097	5.999	0.055	0.035	4.878
MonST3R	**0.088**	0.079	2.208	0.074	0.065	2.530	0.036	0.037	7.804
MASt3R	0.089	0.075	7.279	**0.068**	0.066	3.551	0.042	0.037	8.962
PAGE-4D	0.179	0.070	0.547	0.298	0.098	3.165	0.049	0.036	5.003
Event4D (Ours)	0.225	0.094	0.844	0.552	**0.040**	**0.523**	**0.034**	**0.031**	**2.366**



(a) AbsRel ↓. (b) $\delta < 1.25$ ↑.

Figure 3: Reconstruction accuracy on the EVIMO slow- and fast-motion subsets (8 sequences each).

Table 3: Video depth estimation on MVSEC nighttime sequences. Best results are in bold.

	Night1		Night2		Night3	
Method	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
DUSt3R	0.407	0.393	0.415	0.384	0.463	0.335
MonST3R	0.370	0.373	0.309	0.469	0.317	0.453
EAG3R	**0.353**	0.491	0.307	0.518	0.288	0.533
Event4D (Ours)	0.441	**0.548**	**0.279**	**0.612**	**0.279**	**0.665**

Table 4: Video depth estimation ablations on DSEC and EventKitchen.

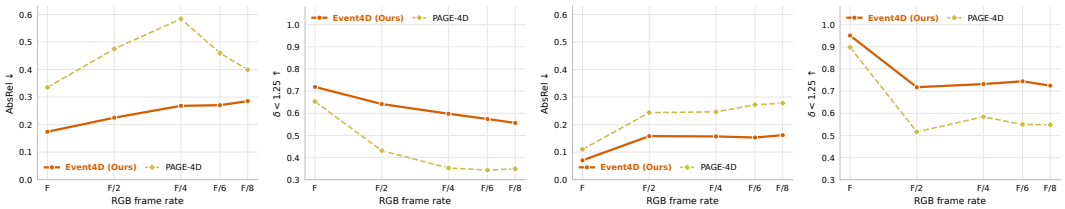
	DSEC		EventKitchen	
Condition	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
Input modalities				
RGB only	0.078	0.940	0.135	0.807
Ev. only	0.394	0.381	0.247	0.615
Ev. + E2VID	0.500	0.330	0.184	0.722
Ours (Ev. + RGB)	**0.047**	**0.977**	**0.117**	**0.867**
Input noise				
RGB noise	0.101	0.901	0.133	0.836
Ev. noise	0.051	0.973	0.120	0.863
w/o noise	**0.047**	**0.977**	**0.117**	**0.867**
Model & loss variants				
w/o RVT	0.095	0.913	0.162	0.760
w/o $\mathcal{L}_{rel}$ w/o $\mathcal{L}_{dir}$	0.096	0.906	0.139	0.794
w/o $\mathcal{L}_{rel}$	0.101	0.893	0.136	0.803
w/o $\mathcal{L}_{dir}$	0.098	0.908	0.137	0.799
Ours (w. RVT)	**0.047**	**0.977**	**0.117**	**0.867**
Staleness conditioning @ F/4				
w/o stale.	0.092	0.920	0.162	0.785
wrong stale.	0.097	0.917	0.162	0.780
Ours	**0.077**	**0.945**	**0.154**	**0.795**

Monocular Depth Estimation. Following the evaluation protocol of prior work Zhang et al. (2025); Zhou et al. (2026), we also evaluate Event4D on monocular depth estimation. As shown in Table 1, Event4D outperforms both the compared 4D reconstruction methods and the state-of-the-art event-based monocular depth estimation method, Depth AnyEvent Bartolomei et al. (2025b), consistent with its video depth estimation results. On DSEC, which includes driving sequences with varying vehicle speeds, traffic conditions, and illumination, Event4D reduces AbsRel by 28.57% relative to VGGT, the strongest RGB-based baseline on this benchmark.

Camera Pose Estimation. Table 2 presents the camera pose estimation results for Event4D. Across most evaluation metrics and benchmark datasets, Event4D achieves state-of-the-art accuracy on majority of the metrics. On N-Sintel, VGGT and PAGE-4D outperform Event4D in camera pose estimation, which may be attributable to Event4D relying on synthetic events generated by v2e that could degrade pose accuracy. However, on real-event datasets such as EVIMO, which includes both slow and fast object motion, Event4D reduces ATE, RPEtrans, and RPErot by 5.56%, 11.43%, and 51.16%, respectively.

Fast vs. Slow Motion 4D Reconstruction. The finite frame rate of RGB cameras limits their ability to capture rapid object motion, constraining RGB-based 4D reconstruction methods' ability to recover scene geometry over time. We evaluate Event4D on video depth estimation separately for the fast- and slow-motion subsets of EVIMO, as shown in Figure 3. While all methods exhibit higher depth estimation error under fast motion, Event4D achieves the lowest AbsRel in both subsets. Furthermore, its accuracy degrades less as motion speed increases: AbsRel rises by only 0.104 from slow to fast motion, compared with an increase of 0.225 for PAGE-4D (Figure 3a).

4D Reconstruction under Temporal Aliasing. The finite frame rate of RGB cameras can lead to temporal aliasing when scene motion is undersampled. Event4D addresses this limitation by formulating reconstruction over N time steps, at each of which the model receives an event voxel and the most recently captured RGB frame, with RGB staleness explicitly encoded in the fusion



(a) Fast: AbsRel. (b) Fast: $\delta < 1.25$. (c) Slow: AbsRel. (d) Slow: $\delta < 1.25$.

Figure 4: Video depth accuracy as RGB frame rate is reduced on the fast- and slow-motion subsets of EVIMO. Event4D outperforms PAGE-4D across all evaluated frame rates in both motion subsets.

modules. We evaluate this setting on EVIMO by progressively reducing the RGB frame rate while keeping the event update and reconstruction rates fixed (Figure 4). At an RGB frame rate of F/k, each image is held for k reconstruction steps, and its staleness ranges from 0 to $k-1$. Thus, $F/2$ and $F/4$ correspond to maximum staleness values of 1 and 3, respectively. Across all evaluated frame rates, Event4D achieves lower AbsRel and higher threshold accuracy than PAGE-4D in both motion subsets. Even at $F/8$, Event4D achieves an AbsRel of 0.2852 on fast-motion sequences, compared with 0.3992 for PAGE-4D, corresponding to a 28.56% reduction.

4D Reconstruction under Low-Light Conditions. Low-light conditions present another challenge for RGB-based 4D reconstruction. EAG3R Wu et al. (2026b) addresses this setting by combining event and RGB observations within a multi-stage reconstruction pipeline. Following its evaluation protocol, we evaluate Event4D on the three nighttime sequences of MVSEC (Table 3). Event4D achieves the best results in five of the six sequence–metric comparisons. In particular, it achieves the highest threshold accuracy on all three sequences, exceeding EAG3R by 5.7%, 9.4%, and 13.2% on Night1, Night2, and Night3, respectively. It also achieves the lowest AbsRel on Night2 and Night3, reducing the error from 0.307 to 0.279 and from 0.288 to 0.279, respectively. These results demonstrate strong low-light reconstruction performance within a single feed-forward architecture.

Ablation Studies. Table 4 presents ablations under four conditions. (a) *Input modalities:* We evaluate RGB-only and events-only inputs, as well as replacing RGB frames with images reconstructed from events using E2VID Rebecq et al. (2019a;b). Fusing RGB and events with Event4D achieves the lowest video depth estimation error. (b) *Input noise:* We add Gaussian noise to RGB images ($\sigma = 0.2$) or event voxels ($\sigma = 0.5$). Both perturbations increase reconstruction error, with a larger degradation under the tested RGB-noise setting. (c) *Model and loss variants:* Replacing the RVT event encoder with a three-layer CNN increases DSEC AbsRel from 0.047 to 0.095, a relative increase of 102.13%. Removing either or both of the additional loss terms, $\mathcal{L}_{\text{rel}}$ and $\mathcal{L}_{\text{dir}}$, also increases reconstruction error, supporting their contribution to reconstruction accuracy. (d) *Staleness conditioning:* We evaluate all variants at an RGB frame rate of $F/4$, where F denotes the original frame rate. In *w/o stale.*, staleness is always set to zero despite repeated RGB inputs. In *wrong stale.*, repeated RGB frames are assigned randomly corrupted staleness values. Both variants increase video depth estimation error, with incorrect staleness producing the lowest threshold accuracy on both datasets. These results support the importance of accurately encoding the age of the available RGB observation.

5 Conclusion

In this work, we present Event4D, an event–RGB fusion approach for reconstructing dynamic 3D scenes over time, a task known as 4D reconstruction. Event4D is the first single-stage feed-forward architecture for event–RGB 4D reconstruction that jointly predicts depth, camera poses, and dynamic pointmaps. Through the proposed fusion mechanism, Event4D addresses three key limitations of traditional RGB-based 4D reconstruction: fast object motion, temporal aliasing, and low-light conditions. Across these settings, Event4D establishes a new state of the art, improving threshold accuracy by 10.11%, 59.29%, and 11.61–24.77% under fast motion, temporal aliasing, and low-light conditions, respectively.

AI USE STATEMENT

In this work, we used generative AI tools for help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools for generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, clean and reformat dataset and assist with translation is not applicable to this work. Additionally, we used generative AI tools for create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, creation of artifacts, draft parts of a research paper, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. For example, we checked LLM-generated research ideas for potential plagiarism through a manual literature survey, LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the full training and evaluation code, with all default parameters used to produce our reported results, as supplementary material. The synthetic dataset generation scripts are also included. A README file accompanying the code gives clear, step-by-step instructions for running the experiments end to end, covering synthetic dataset generation, training, and inference. Scripts for generating N-PointOdyssey, N-Spring, N-TartanAir, N-Waymo, and N-Sintel are also included. Due to their large size, the generated datasets and trained model checkpoint are not included in the supplementary material; both will be publicly released upon acceptance of the paper

REFERENCES

Arun Kumar Arigela, D David Neels Ponkumar, Martin Margala, K Jayasri, A Leela Sravanthi, and Prasun Chakrabarti. Self-supervised cross frame-rate event-frame depth learning without ground truth. *Scientific Reports*, 2026.

Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19669–19678. IEEE, 2025b.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.) (ed.), *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012.

Xudong Cai, Shuo Wang, Peng Wang, Yongcai Wang, Zhaoxin Fan, Wanting Li, Tianbao Zhang, Jianrong Tao, Yeying Jin, and Deying Li. Mem4d: Decoupling static and dynamic memory for dynamic scene reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2552–2560, 2026.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22831–22840. IEEE, 2025.

Anusha Devulapally, Md Fahim Faysal Khan, Siddharth Advani, and Vijaykrishnan Narayanan. Multi-modal fusion of event and rgb for monocular depth estimation using a unified transformer-based architecture. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2081–2089. IEEE, 2024.

Yue-Jiang Dong, Wang Zhao, Jiale Xu, Ying Shan, and Song-Hai Zhang. Depthsync: Diffusion guidance-based depth synchronization for scale-and geometry-consistent video depth estimation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5415–5425. IEEE, 2025.

Johann-Friedrich Feiden, Tim Küchler, Denis Zavadski, Bogdan Savchynskyy, and Carsten Rother. Online video depth anything: Temporally-consistent depth prediction with low memory consumption. In *2026 International Conference on 3D Vision (3DV)*, pp. 1760–1770. IEEE, 2026.

Chengming Feng, Hesam Araghi, Liming Zheng, Julien Dupeyroux, Xucong Zhang, Jan van Gemert, and Nergis Tömen. Cooking beyond frames: A stereo event camera dataset in the kitchen. In *European Conference on Computer Vision*, pp. 641–660. Springer, 2026.

Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021a.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021b.

Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, pp. 534–542. IEEE, 2020.

Chenming Hu, Junjie Jiang, Yaohui Li, Mingyuan Sun, and Zheng Fang. Ede-distill: Boosting event-based monocular depth estimation performance via knowledge distillation. *IEEE Robotics and Automation Letters*, 10(8):8252–8259, 2025a.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2005–2015. IEEE, 2025b.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1312–1321. IEEE, 2021.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9000–9008, 2018.

Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7233–7243. IEEE, 2025.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. In *International Conference on Learning Representations*, volume 2026, pp. 42746–42768, 2026.

Jie Long Lee and Gim Hee Lee. Distil-e2d: Distilling image-to-depth priors for event-based monocular depth estimation. *Advances in Neural Information Processing Systems*, 38:125764–125785, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4991, 2023.

Anton Mitrokhin, Chengxi Ye, Cornelia Fermüller, Yiannis Aloimonos, and Tobi Delbruck. Evimo: Motion segmentation dataset and learning pipeline for event cameras. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6105–6112. IEEE, 2019.

Victor Oei, Jenny Schmalfuss, Lukas Mehl, Madlen Bartsch, Shashank Agnihotri, Margret Keuper, Andreas Bulling, and Andres Bruhn. Robustspring: Benchmarking robustness to image corruptions for optical flow, scene flow and stereo. In *Proc. International Conference on Learning Representations (ICLR)*, 2026.

Seong Hyeon Park and Jinwoo Shin. Track3r: Joint point map and trajectory prior for spatiotemporal 3d understanding. *Advances in Neural Information Processing Systems*, 38:166097–166116, 2026.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, 2019a. URL `http://rpg.ifi.uzh.ch/docs/TPAMI19_Rebecq.pdf`.

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

Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank Po Wen Lo, and Benny Lo. Even: An event-based framework for monocular depth estimation at adverse night conditions. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1–7. IEEE, 2023.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jian Wang, Liuyi Li, Wenze Shao, and Liqian Wang. Event camera monocular depth estimation: An event-tailored state-space model-based approach. *IEEE Transactions on Instrumentation and Measurement*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Shizun Wang, Zhenxiang Jiang, Xingyi Yang, and Xinchao Wang. C4d: 4d made from 3d through dual correspondences. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7570–7580. IEEE, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024a.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Yiran Wang, Min Shi, Jiaqi Li, Chaoyi Hong, Zihao Huang, Juewen Peng, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Nvds+: Towards efficient and versatile neural stabilizer for video depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 47(1): 583–600, 2024b.

Shangjie Wu, Jihua Zhu, Zhikuan Zhou, Siqi Li, Shaoyi Du, and Yue Gao. Glu-net: Global–local fusion network for event-based monocular depth estimation via uncertainty optimization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(6):1–20, 2026a.

Xiaoshan Wu, Yifei Yu, Xiaoyang Lyu, Yihua Huang, Bo Wang, Baoheng Zhang, Zhongrui Wang, and Xiaojuan Qi. Eag3r: Event-augmented 3d geometry estimation for dynamic and extreme-lighting scenes. *Advances in Neural Information Processing Systems*, 38:73314–73342, 2026b.

Jing Xu, Hong Liu, Jianbing Wu, and Xinhua Xu. Pridev: A plug-and-play refinement for improved depth estimation in videos. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10929–10935. IEEE, 2025.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Forrester Cole, Deqing Sun, Ming-Hsuan Yang, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, volume 2025, pp. 82863–82886, 2025.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023.

Kaichen Zhou, Yuhan Wang, Grace Chen, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. Page-4d: Disentangled pose and geometry estimation for vggt-4d perception. In *International Conference on Learning Representations (ICLR)*, pp. 36401–36414, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3d55170799265c03b37993e02b71b2cc-Abstract-Conference.html.

Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 989–997. IEEE, 2019.

A Appendix

B Additional Qualitative Results

We show additional qualitative reconstructions from our model on few selected sequences. For each sequence we show, at every time step: the input RGB frame, the corresponding event representation, our predicted depth map, and the resulting point map.

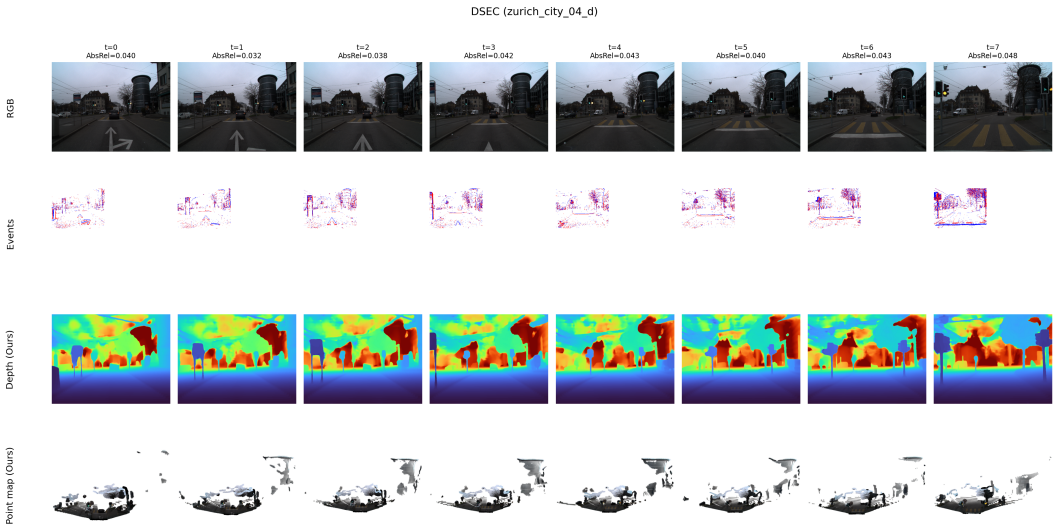


Figure 5: DSEC, `zurich_city_04_d`. Mean AbsRel 0.041 across the shown window.

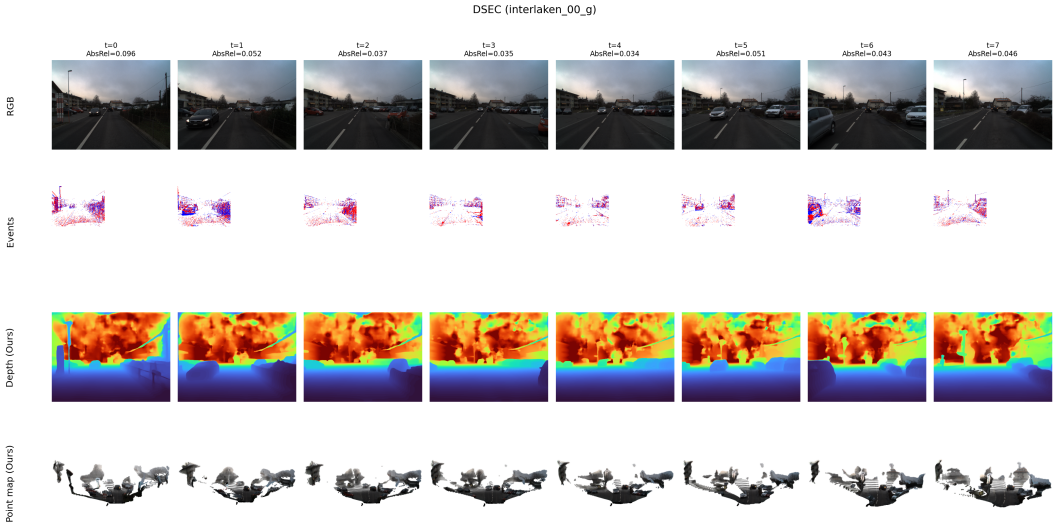


Figure 6: DSEC, `interlaken_00_g`. Mean AbsRel 0.049 across the shown window.

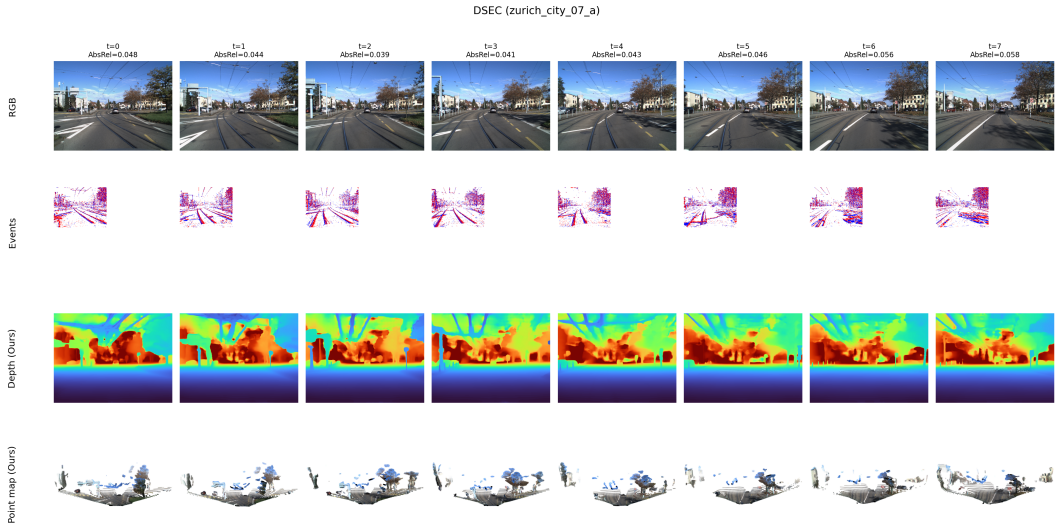


Figure 7: DSEC, `zurich_city_07_a`. Mean AbsRel 0.047 across the shown window.

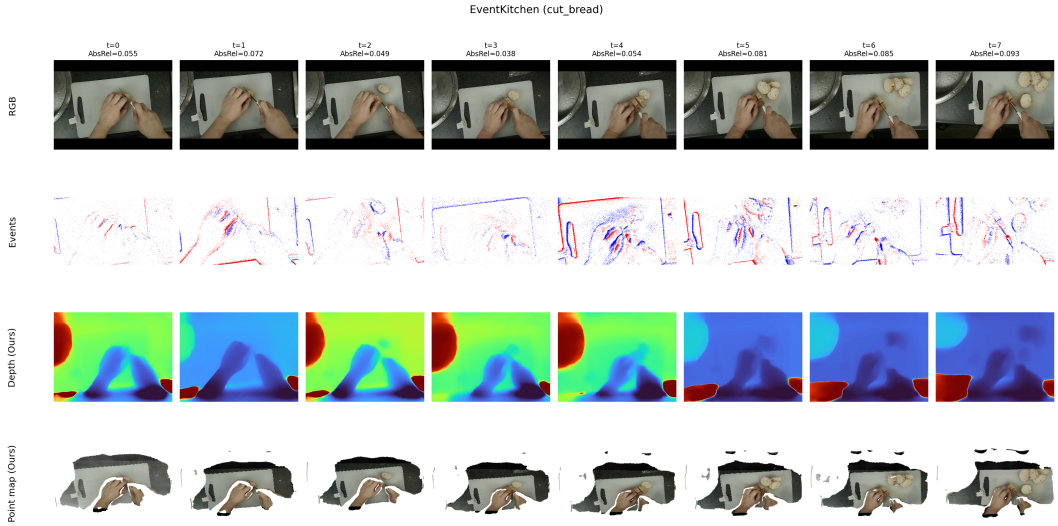


Figure 8: EventKitchen, `cut_bread`. Mean AbsRel 0.066 across the shown window.

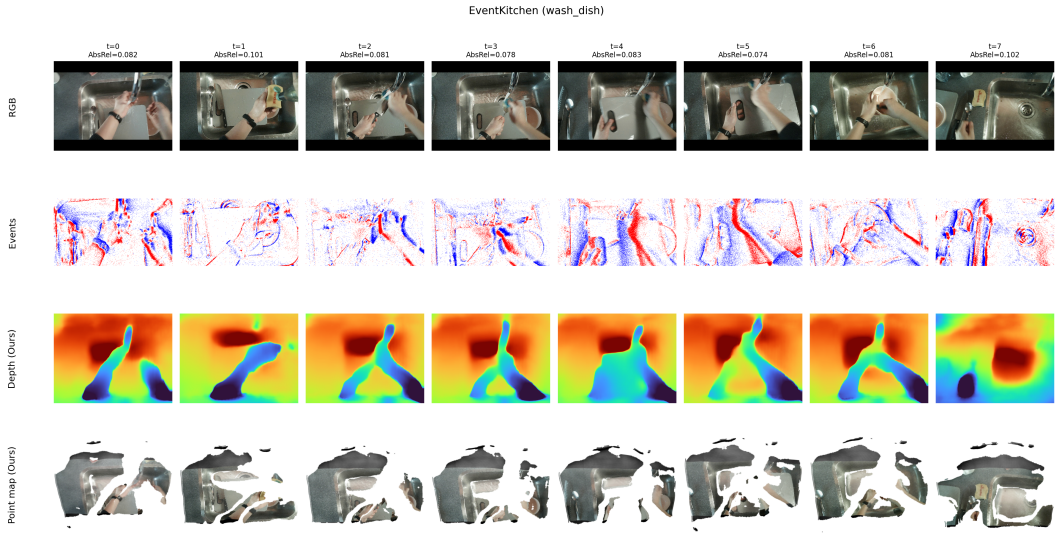


Figure 9: EventKitchen, `wash_dish`. Mean AbsRel 0.085 across the shown window.

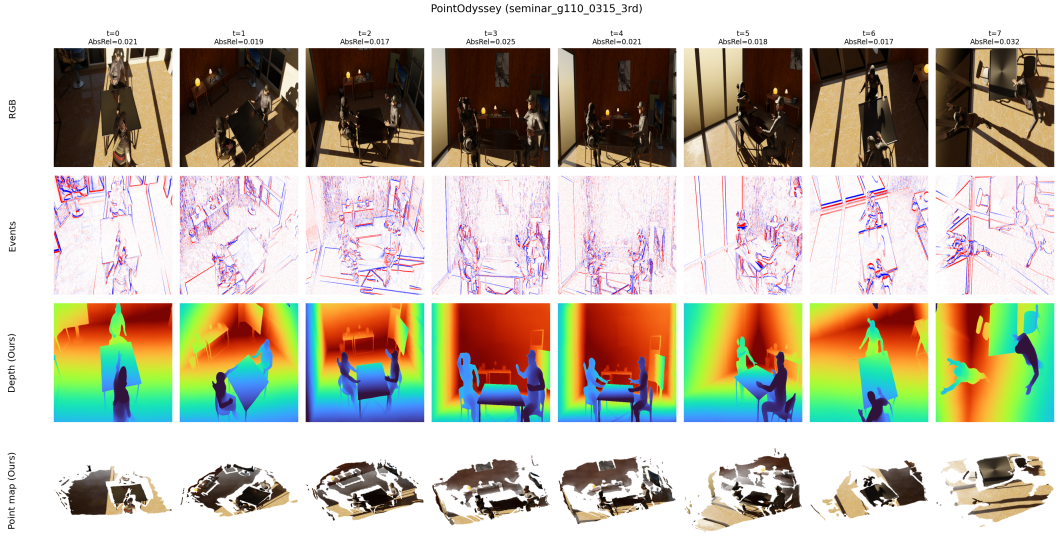


Figure 10: PointOdyssey, `seminar_g110_0315_3rd`. Mean AbsRel 0.021 across the shown window.

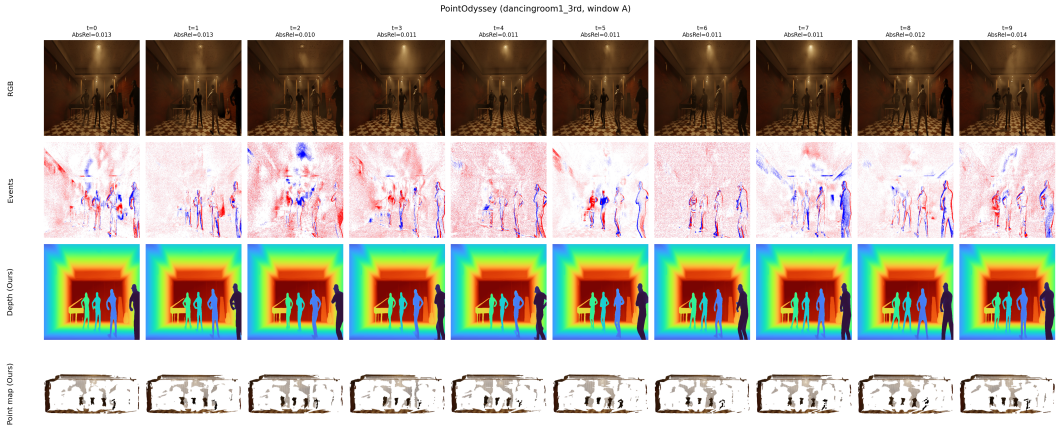


Figure 11: PointOdyssey, `dancingroom1_3rd`. Mean AbsRel 0.012 across the shown window.