GEOMETRY BEFORE FUSION: EVENT-FIRST STATE-SPACE MODELING FOR EVENT-FRAME STEREO

Anonymous authors
Paper under double-blind review

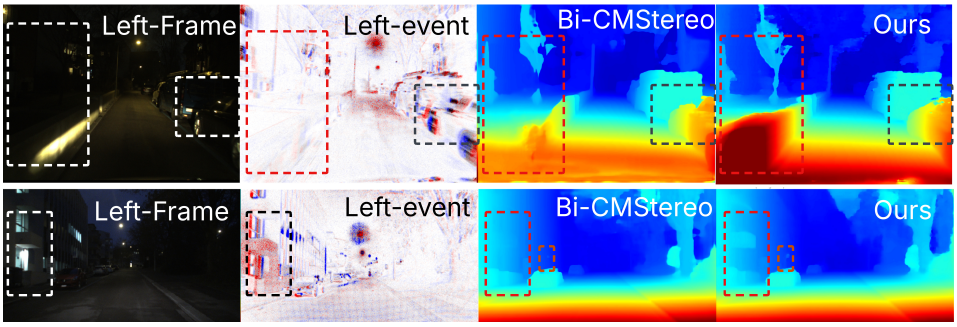

Figure 1: **Qualitative comparison on event–frame stereo.** Compared to Bi-CMPStereo Xu et al. (2026), our method achieves sharper structural details.

ABSTRACT

Event cameras remain informative under fast motion and extreme illumination but are sparse, whereas frame cameras provide dense appearance that degrades under the same conditions. Existing event–frame stereo methods fuse the two modalities before or during matching, so unreliable appearance can bias correspondence before any geometric reference exists. We propose to *establish geometry before multimodal fusion*: disparity is first estimated solely from stereo events with a state-space architecture, then refined by a single frame weighted by a reliability measure derived from event stereo consistency. On DSEC, our method is the most accurate among methods using at most one frame camera and approaches symmetric methods using two, while its event-only variant outperforms prior event-only methods. It also generalizes zero-shot to MVSEC and M3ED. Ablations show that fusion timing is critical: introducing the frame before matching degrades accuracy below the event-only model.

1 INTRODUCTION

Stereo matching Scharstein & Szeliski (2002); Tosi et al. (2025) recovers dense depth from two calibrated views and is widely used for 3D perception in autonomous driving and robotics. Its reliability, however, depends on both the scene and the sensor: correspondence becomes ambiguous in low-texture regions, and conventional frame cameras, while providing dense appearance information, deteriorate under fast motion and limited dynamic range Gallego et al. (2020); Mostafavi et al. (2021). Event cameras provide a complementary sensing mechanism: they asynchronously record brightness changes with microsecond-scale temporal resolution and high dynamic range, offering advantages over conventional frame cameras in challenging scenarios such as high-speed motion and extreme lighting conditions Gallego et al. (2020); Mostafavi et al. (2021). However, events are spatially sparse and their generation depends on temporal brightness changes, making dense correspondence difficult in static or weakly activated regions Gallego et al. (2020); Mostafavi et al. (2021)

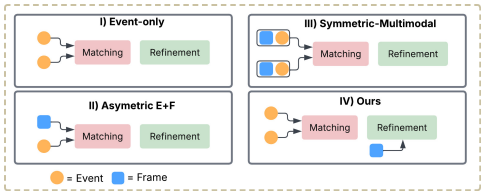

Figure 2: **Event–frame stereo paradigms.** Event-only (I), asymmetric E+F (II), symmetric multimodal (III), and ours (IV).

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.

Nam et al. (2022); Zhang et al. (2022). These complementary sensing characteristics have motivated event–frame stereo methods that combine dense frame appearance with temporally responsive event measurements Mostafavi et al. (2021); Cho & Yoon (2022a;b); Zhao et al. (2024).

A central question is not only *how* the two modalities should interact, but *when*. Existing event–frame stereo methods align, translate, or fuse the two modalities before or during correspondence estimation Mostafavi et al. (2021); Cho & Yoon (2022a;b); Zhao et al. (2024); Lou et al. (2024); Zhuang et al. (2025); Xu et al. (2026). This exploits their complementarity, but lets frame appearance shape the matching evidence before any event-derived geometric hypothesis exists: when frames degrade, their features bias the initial correspondence, and no appearance-independent reference is available to check consistency. A second question is *where* each modality should participate. Prior work adopts three dominant configurations, that we survey in Figure 2. Compared to event stereo (I), asymmetric stereo (II) pairs one RGB view with one event view Xu et al. (2026); Ding et al. (2024), requiring correspondence across heterogeneous feature spaces; the resulting cross-modality matching problem can remain challenging under complex motion and varying depth Ding et al. (2024). In contrast, symmetric multimodal stereo (III) provides both views with both modalities, allowing complementary event and appearance cues to contribute jointly to correspondence estimation. This richer evidence helps when frames are reliable, but lets appearance directly shape the geometric correspondence.

We instead ask a different question: not how much modality information to include, but *when* and *in what role* each modality should participate. To answer this question, we propose a geometry-before-fusion design, shown in Figure 2 (IV), where two event views participate in correspondence matching and a single RGB view in refinement. This sacrifices one frame relative to the symmetric setting (III), reducing deployment cost and simplifying synchronization and calibration. This comes with an accuracy trade-off: on the DSEC benchmark, symmetric two-frame methods remain more accurate when both frames are reliable. Our design is therefore not a claim that additional appearance information is undesirable, but rather that *the role appearance plays matters as much as its quantity*. By restricting matching to the event domain, we obtain an appearance-independent geometric hypothesis before any frame feature is consulted. Frame evidence then enters only to refine a disparity that already exists, evaluated against a geometric reference that it did not influence. This decoupling isolates the initial geometric estimate from frame quality, allowing the refinement stage to down-weight unreliable appearance rather than propagate its errors into the initial estimate.

Obtaining accurate event-only geometry is non-trivial: sparse evidence must support both local discrimination and long-range contextual reasoning. We build on selective state-space models Gu & Dao (2023), adapting them to preserve local event geometry while enabling epipolar matching. Multimodal fusion then enters guided by geometry, where the initial disparity serves two roles: it warps right-view event features to the left view, producing an event stereo-consistency residual that drives a reliability gate for selective integration of the left frame; and it conditions a multiscale refinement network that deformably resamples the fused representation and predicts a residual correction.

On the official DSEC online benchmark, our model outperforms all event-only and asymmetric event–frame approaches, and comes close to symmetric methods that use one more frame camera. Ablations show the importance of establishing event-based geometry before frame fusion, while zero-shot evaluation on unseen data demonstrates generalization beyond the training distribution. Figure 1 illustrates this qualitatively. Our main contributions are summarized as follows:

- A *geometry-before-fusion* formulation for event–frame stereo, where disparity is estimated from stereo events alone before frame evidence is used for refinement.
- A state-space event stereo architecture combining local state-space modeling, cross-view propagation, and explicit epipolar matching for frame-independent geometry; its event-only variant alone outperforms prior event-only methods on the DSEC online benchmark.
- A reliability-aware refinement stage that uses the event-derived disparity to measure consistency, gate frame evidence, and condition feature sampling.
- Evidence that fusion timing matters: early frame fusion performs worse than no frame at all, while fusing after geometry is established yields the best results among methods using at most one frame camera.[1]

[1] *Code will be made publicly available upon acceptance.*

2 Related Work

Event-Based Stereo Matching. Learning-based event stereo has advanced along three axes. Early work focused on representations: DDES Tulyakov et al. (2019) introduced learnable dense event representations, DTC Zhang et al. (2022) developed discrete-time convolutions encoding spatio-temporal structure at multiple scales, and SE-CFF Nam et al. (2022) addressed non-uniform event density through concentrated representations.

A second line formulates event stereo as dense correspondence. StereoFlow-Net Cho et al. (2024) exploits temporal motion, while EMatch Zhang et al. (2025b) unifies optical flow and stereo matching through shared representations and iterative refinement. A third line targets training strategies: cross-sensor distillation Zhang et al. (2025a) transfers frame knowledge while retaining event-only inference, and EventHub Bartolomei et al. (2026) investigates data generation and generalization.

Despite these advances, event-only stereo remains constrained by measurement sparsity: events fire only at brightness changes, leaving regions with weak motion or texture under-constrained. This motivates using frame observations as dense appearance evidence for geometric refinement.

Event-Frame Stereo Matching. Early symmetric event-image stereo methods use both modalities across both views. EIS Mostafavi et al. (2021) jointly exploits event and frame information, SMC-Net Cho & Yoon (2022a) improves interaction through cross-modality feature propagation, SCS-Net Cho & Yoon (2022b) introduces differentiable event selection and cross-similarity features, and EGEI-Net Zhao et al. (2024) injects image-derived edge guidance into recurrent disparity updates.

A second line considers asymmetric sensing, where the two views originate from different modalities. ZEST Lou et al. (2024) constructs physically motivated event–intensity representations compatible with pretrained stereo networks. Temporal feature gating Zhuang et al. (2025) enhances frame representations with motion cues before cross-modal matching. Bi-CMPStereo Xu et al. (2026) learns complementary canonical representations from each modality, while ARES Lee & Lee (2026) constructs a shared event-RGB representation through multimodal attention for joint disparity and motion reasoning.

Across all these approaches, multimodal interaction occurs during correspondence construction, via feature fusion, alignment, or cross-modal attention. This leaves one question unanswered: can correspondence be estimated from a single modality, with the other entering only for refinement? Our design answers yes.

State-Space Models for Stereo and Multimodal Vision. State-space models have emerged as efficient alternatives to attention for long-range visual representation. Mamba Gu & Dao (2023) introduces input-dependent selective dynamics, while Vision Mamba Zhu et al. (2024) and VMamba Liu et al. (2024) adapt this to visual data through bidirectional and multidirectional scanning. Several works specialize SSMs for stereo: StereoMamba Wang et al. (2025) introduces intra- and cross-view state-space interaction before correlation, RRESM Ding et al. (2025) applies bidirectional Mamba to coordinate-aware cost aggregation, and further architectures explore state-space regularization and multi-branch aggregation Suo et al. (2026); Lu et al. (2025). Beyond stereo, AIMDepth Jing et al. (2026) demonstrates asymmetric image–event state-space interaction for monocular depth.

These works establish SSMs as effective for long-range stereo and multimodal vision, but none addresses the combination required here: preserving local event geometry during state formation, propagating context across two event views without performing matching, and conditioning frame refinement on an event-only disparity. Prior stereo SSMs either interleave matching with propagation Wang et al. (2025) or apply state-space modeling only during cost aggregation Ding et al. (2025); Suo et al. (2026); Lu et al. (2025). Neither separates geometry estimation from multimodal fusion, which is the design principle we pursue.

3 Method

3.1 Problem Formulation and Overview

Given synchronized stereo event streams $\mathcal{E}_l, \mathcal{E}_r$ and a single conventional frame I_l, we estimate the dense left-view disparity $D \in \mathbb{R}^{H \times W}$. Events are represented as voxel grids with $B_e = 15$ temporal

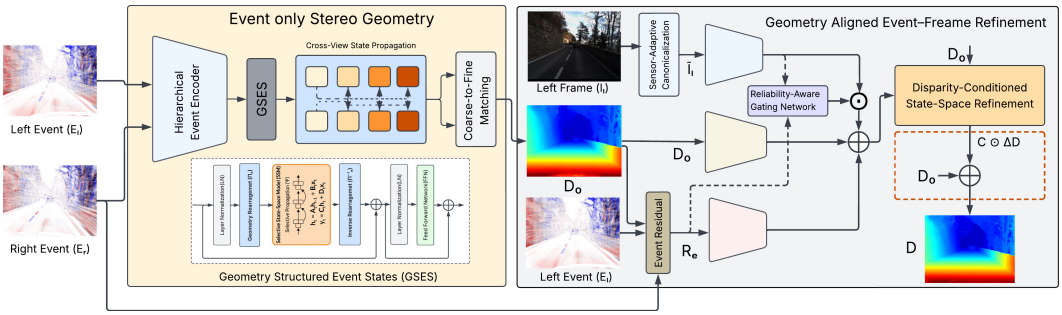

Figure 3: **Geometry-before-fusion framework.** Stage I estimates frame-independent stereo geometry from events. Stage II uses this geometry to align modality-specific evidence, selectively integrates canonicalized frame information according to reliability, and applies a confidence-gated residual correction through disparity-conditioned state-space refinement.

bins, following standard practice Zhu et al. (2019). Our design follows a simple principle: *establish geometry before multimodal fusion*. As illustrated in Fig. 3, Stage I predicts an initial disparity as

$$D_0 = \mathcal{G}_e(E_l, E_r), \tag{1}$$

making the initial correspondence estimate independent of frame appearance. Stage II then uses D_0 as a geometric reference to assess whether event and frame features are consistent with the current stereo hypothesis before any fusion occurs. It predicts a residual correction whose influence is controlled by local reliability,

$$D = D_0 + C \odot \Delta D_{\text{cand}}, \tag{2}$$

where $C \in [0,1]^{H\times W}$ is a reliability-modulated confidence map derived from event-domain geometric consistency and a learned refinement confidence (Sec. 3.3). This decomposition makes geometry both a prediction and the interface through which heterogeneous modalities interact: D_0 organizes cross-view evidence, controls multimodal interaction, and conditions the spatial computation used for refinement.

3.2 Event-Only Stereo Geometry (Stage I)

Stage I must establish sufficiently accurate correspondence *without* relying on conventional-frame appearance. Three components make this possible: geometry-structured event states that preserve local spatial layout within a state-space recurrence, bidirectional cross-view propagation that exchanges long-range context between the two event streams, and explicit coarse-to-fine epipolar matching that enforces physical correspondence.

Geometry-Structured Event States (GSES). A shared hierarchical encoder extracts multi-scale event features $\{F_l^s, F_r^s\}_{s=1}^3$ with channel dimensions 48, 96, and 192 at progressively coarser resolutions. Standard vision SSMs flatten a two-dimensional feature map into a one-dimensional sequence before applying the recurrence, which breaks the spatial adjacency of neighboring pixels and discards local geometric structure. This is particularly harmful for event data, where informative content is defined by the relative arrangement of sparse events within small local neighborhoods. We therefore preserve local geometry explicitly instead of flattening.

Given $X \in \mathbb{R}^{B\times C\times H\times W}$ and a neighborhood size $a = 3$, the reversible rearrangement operator Π_a maps X to $\tilde{X} \in \mathbb{R}^{B\times C'\times L}$ with $C' = Ca^2$ and $L = \lfloor H/a \rfloor \lfloor W/a \rfloor$. The a^2 spatial offsets within each $a \times a$ neighborhood are placed in the channel dimension and ordered lexicographically (top-left to bottom-right), producing an ordered channel–offset sequence. Because Π_a is a fixed permutation, its inverse Π_a^{-1} restores the original spatial layout and channel dimension without information loss. A local state-space block is then applied as

$$Y = X + \Pi_a^{-1}[\Psi(\Pi_a(\mathrm{LN}(X)))], \qquad X' = Y + \phi_{\mathrm{FFN}}(\mathrm{LN}(Y)), \tag{3}$$

where Ψ denotes a Mamba-style selective state-space operator with input-dependent parameters, applied along the ordered channel–offset sequence. Because the recurrence operates on geometrically ordered local states, the operator can learn offset-dependent behavior while retaining linear

complexity in the sequence length. The residual structure in Eqs. equation 3 stabilizes training and preserves the encoder's original feature statistics.

Cross-view State Propagation. At the matching scales, the left and right features are flattened into spatial sequences $X_l, X_r \in \mathbb{R}^{B\times L\times C}$ and processed by a shared intra-view state-space block,

$$\bar{X}_v = X_v + \mathcal{M}_{\text{intra}}(\text{LN}(X_v)), \qquad v \in \{l, r\}, \tag{4}$$

where $\mathcal{M}_{\text{intra}}$ shares parameters across views. The processed features $\bar{X}_l, \bar{X}_r$ are then interleaved at the token level,

$$S = [\bar{x}_{l,1}, \bar{x}_{r,1}, \ldots, \bar{x}_{l,L}, \bar{x}_{r,L}], \tag{5}$$

so that adjacent tokens correspond to the same spatial location in the two views. Bidirectional propagation is applied as

$$U = \mathcal{M}_{\rightarrow}(\text{LN}(S)) + \mathcal{R}[\mathcal{M}_{\leftarrow}(\mathcal{R}[\text{LN}(S)])], \tag{6}$$

where $\mathcal{M}_{\rightarrow}$ and $\mathcal{M}_{\leftarrow}$ are forward and backward selective state-space operators and $\mathcal{R}$ reverses the sequence. Interleaving ensures that the forward pass attends to left-then-right pairs, while the backward pass attends to right-then-left pairs, so long-range context is exchanged symmetrically between the two views. The outputs are separated into $\bar{U}_l, \bar{U}_r$ and added to their corresponding residual features. This step exchanges contextual information but does not itself define stereo correspondence: physical matching remains explicitly constrained along epipolar lines, as described next.

Coarse-to-Fine Epipolar Matching. At the coarsest scale, we compute correspondence along the epipolar line using feature correlation and soft-argmax:

$$C_c(y, x, d) = \frac{\langle F_l^c(y, x), F_r^c(y, x-d)\rangle}{\sqrt{C}}, \qquad D_c(y, x) = \sum_d d, \text{softmax}\, d(C_c(y, x, d)). \tag{7}$$

At finer scales, the upsampled disparity $\hat{D}_{s-1}$ defines a local search window,

$$d \in [\hat{D}_{s-1}(y, x) - r_s, \hat{D}_{s-1}(y, x) + r_s], \tag{8}$$

within which correspondence is refined by the same correlation and soft-argmax operation. The resulting event-only disparity D_0 provides the geometric reference for the subsequent frame refinement stage.

3.3 Geometry-Aligned Event–Frame Refinement (Stage II)

Once D_0 has been established, multimodal interaction can be conditioned on a frame-independent geometric hypothesis. Stage II therefore asks not merely what information a modality contains, but whether that information is consistent with the current stereo geometry.

Geometry-Aligned Modality Evidence. We extract shallow event features using a view-shared encoder, $F_l^e = \psi_e(E_l)$ and $F_r^e = \psi_e(E_r)$. The available left frame is first mapped to a canonical structural representation before feature extraction, $\widetilde{I}_l = \mathcal{C}(I_l)$ and $F_l^f = \psi_f(\widetilde{I}_l)$, where $\mathcal{C}$ produces three structural channels: normalized intensity, local contrast, and gradient magnitude. The same construction is used for RGB and grayscale inputs, reducing sensitivity to sensor-specific appearance while preserving spatial structure. The event-derived disparity D_0 provides the geometric reference for aligning stereo event features. We warp the right event feature toward the left view, $\widehat{F}_r^e(x, y) = F_r^e(x - D_0(x, y),\, y)$, using bilinear interpolation. The resulting event stereo-consistency residual is

$$R^e = \left(\widehat{F}_r^e - F_l^e\right) \odot V_e, \tag{9}$$

where $V_e \in [0, 1]^{H\times W}$ is a warp-validity map that is 1 at pixels whose bilinear interpolation support lies entirely within the image bounds and decays smoothly to 0 at the boundary. The frame branch provides unary appearance evidence rather than frame–frame stereo correspondence. The modality evidence is therefore projected separately as

$$X_e = \phi_e([F_l^e, R^e]), \qquad X_f = \phi_f\left(F_l^f\right), \tag{10}$$

The event-derived disparity is encoded through a separate geometry branch, $\bar{D}_0 = D_0/\max(W - 1, 1)$ and $X_d = \phi_d(\bar{D}_0)$. The normalized disparity $\bar{D}_0$ is used only for neural conditioning; all geometric warps operate on disparity in pixel units.

Reliability-Aware Frame Integration. Canonicalization reduces appearance variation but does not guarantee that frame evidence is equally reliable across the image. We therefore estimate its contribution using local geometric consistency and cross-modal structural agreement. We construct a reliability descriptor from five per-pixel cues: normalized event activity $\widetilde{A}_e(x, y)$, normalized canonical-frame contrast $\widetilde{A}_f(x, y)$, event consistency error $\epsilon_e(x, y)$, event warp validity $V_e(x, y)$, and event–frame structural agreement $A_{ef}(x, y)$. The event consistency error is defined as

$$\epsilon_e = \tanh\left(\frac{\mathrm{mean}_c\,|R_c^e|}{\mathrm{sg}\Big(\mathrm{mean}_c\,|F_{l,c}^e|\Big) + \varepsilon}\right), \tag{11}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and ε is a small constant. Normalizing the residual by the corresponding event-feature magnitude reduces sensitivity to feature scale. The resulting descriptor is $Q_g = \Big[\widetilde{A}_e, \widetilde{A}_f, \epsilon_e, V_e, A_{ef}\Big]$, $A_{ef} = \exp\Big(-\Big|\widetilde{A}_e - \widetilde{A}_f\Big|\Big)$, where A_{ef} measures local structural agreement between the two modalities. A lightweight gating network maps Q_g to a spatial-channel frame gate, $G_f = g_{\max}\,\sigma(f_g(Q_g))$, $g_{\max} = 0.75$, where the scalar upper bound $g_{\max}$ prevents frame evidence from dominating the persistent event and geometric representation. The refinement representation is then formed as

$$X_{\mathrm{ref}} = \mathrm{ReLU}(\mathrm{IN}(X_e + X_d + G_f \odot X_f)), \tag{12}$$

so that event and geometric features provide the persistent representation, while frame features are incorporated according to their estimated reliability. Since the frame branch is unary, we additionally attenuate its contribution by $\eta_u = 0.1$ before applying the gate. During training, per-sample frame dropout with probability p_f exposes the refiner to both frame-assisted and event-only inputs, reducing dependence on frame appearance.

Disparity-Conditioned State-Space Refinement. We refine X_{ref} using a three-scale encoder–decoder with bidirectional state-space blocks at spatial resolutions $H/4$, $H/8$, and $H/16$. At each scale s, a feature-dependent predictor estimates a bounded offset $O_{\mathrm{raw}}^{(s)} = \tanh\Big(f_{\mathrm{off}}^{(s)}(X^{(s)})\Big)$, and its magnitude is modulated by the normalized disparity through $A^{(s)} = \sigma\Big(f_{\mathrm{scale}}^{(s)}(\bar{D}_0^{(s)})\Big)$. To respect the predominantly horizontal geometry of rectified stereo, we constrain the resulting offset anisotropically, $O_x^{(s)} = \delta A^{(s)} O_{\mathrm{raw},x}^{(s)}$, $O_y^{(s)} = \eta_v \delta A^{(s)} O_{\mathrm{raw},y}^{(s)}$, where $\delta = 0.025$ bounds the displacement to at most 2.5% of the image width and $\eta_v = 0.5$ suppresses vertical deformation. We resample the feature map as $X_{\mathrm{def}}^{(s)} = \mathcal{W}\big(X^{(s)}, G^{(s)} + O^{(s)}\big)$, where $\mathcal{W}$ denotes the offset-based feature resampling operator and $G^{(s)}$ is the regular sampling grid. After flattening $X_{\mathrm{def}}^{(s)}$ into a spatial sequence $Z^{(s)}$, we apply bidirectional state-space propagation followed by an FFN,

$$\widetilde{Z}^{(s)} = Z^{(s)} + \mathcal{M}_{\leftrightarrow}^{(s)}\Big(\mathrm{LN}(Z^{(s)})\Big), \qquad Y^{(s)} = \widetilde{Z}^{(s)} + \phi_{\mathrm{FFN}}^{(s)}\Big(\mathrm{LN}(\widetilde{Z}^{(s)})\Big). \tag{13}$$

Here, $\mathcal{M}_{\leftrightarrow}^{(s)}$ denotes bidirectional selective state-space propagation at scale s. Thus, disparity controls the spatial support used for feature sampling, while the state-space operator aggregates the sampled evidence over long spatial ranges.

Confidence-gated Conservative Residual. To avoid overwriting a reliable D_0 under uncertain cross-sensor evidence, we predict a bounded residual correction together with a reliability-modulated confidence,

$$\Delta D_{\mathrm{cand}} = \beta W \tanh(f_{\mathrm{disp}}(X_{\mathrm{dec}})), \qquad C = C_\theta \odot R_e, \tag{14}$$

where the reliability term is $R_e = r_{\min} + (1 - r_{\min}) \cdot \exp(-2\epsilon_e) \cdot V_e$, $C_\theta \in [0, 1]^{H \times W}$ is a learned refinement confidence produced by a sigmoid-activated head, X_{dec} is the decoder output at the finest scale, β bounds the residual to a fixed fraction of the maximum disparity, and $r_{\min} \in (0, 1]$ is a minimum reliability floor. We set $r_{\min} = 0.25$ in all experiments. R_e makes the gate interpretable: $R_e = 1$ when event-domain evidence is locally consistent ($\epsilon_e \to 0$) and the warp is valid ($V_e = 1$), and R_e decays smoothly toward $r_{\min}$ as consistency or validity degrades. The floor $r_{\min}$ guarantees a non-zero residual budget even in regions where event evidence is weak, which we found necessary to avoid the network collapsing to $D = D_0$ early in training. The final confidence is clipped to $[0, 1]$. The final disparity is obtained by applying the gated residual to D_0 as in Eq. equation 2. The residual prediction head is zero-initialized, so that Stage II initially implements the identity mapping $D = D_0$ and subsequently learns only evidence-supported corrections to the initial stereo estimate.

Table 1: **Comparison with state-of-the-art stereo methods on DSEC.** Zu, In, and Th: Zurich City, Interlaken, and Thun sequences. Best, second, and third results are highlighted in **1st**, **2nd**, and **3rd**. Params and MACs are measured by us for all methods with the same tool at 480×640 input.

	MAE ↓				1PE ↓				2PE ↓				RMSE ↓				Params ↓	MACs ↓
Method	Zu	In	Th	All	Zu	In	Th	All	Zu	In	Th	All	Zu	In	Th	All	**(M)**	**(G)**
SECFF Nam et al. (2022)	0.627	0.555	0.771	0.612	13.123	10.397	17.745	12.477	3.506	2.536	4.408	3.288	1.462	1.325	1.862	1.445	6.92	223.0
DTC Zhang et al. (2022)	0.631	0.563	0.768	0.621	12.809	10.183	17.569	12.069	3.153	2.481	4.129	3.018	1.501	1.357	1.736	1.508	**2.01**	**158.0**
ZEST Lou et al. (2024)	2.483	1.282	1.339	2.168	48.325	39.415	38.127	46.920	20.195	13.847	13.303	19.288	3.779	2.319	2.608	3.394	340.75	2034.0
ZEST† Lou et al. (2024)	0.785	0.580	0.797	0.763	21.003	12.913	19.922	20.382	4.845	3.303	4.615	4.646	1.474	1.211	1.713	1.438	340.75	2034.0
SEVFI Ding et al. (2024)	0.729	0.623	0.875	0.711	17.374	14.558	24.225	16.932	4.497	3.408	5.970	4.307	1.552	1.309	1.844	1.509	39.69	1332.0
Bi-CMPS Xu et al. (2026)	0.547	0.450	0.715	0.532	11.112	7.992	16.444	10.613	2.606	1.507	3.587	2.415	1.244	1.026	1.651	1.210	55.85	1076.7
Ours (Stage-I)	0.567	0.437	0.780	0.544	11.248	6.228	13.879	10.450	1.777	0.745	2.525	1.615	1.158	0.912	1.572	1.116	7.89	437.0
Ours	**0.510**	**0.396**	**0.695**	**0.491**	**9.596**	**5.529**	**12.302**	**8.839**	**1.512**	**0.696**	**2.519**	**1.368**	**1.149**	**0.908**	**1.557**	**1.109**	10.34	654.4

3.4 Training Objective

Stage II is trained to refine the event-only disparity estimate D_0 while preserving the frame-independent geometry established by Stage I. Let $e_0 = |D_0 - D_{\mathrm{gt}}|$ denote the error of the event-only estimate, $e_r = |D - D_{\mathrm{gt}}|$ the error after refinement, and $\Delta = D - D_0$ the predicted correction. The training objective combines three terms,

$$\mathcal{L} = \underbrace{\frac{1}{|\Omega|}\sum_{p\in\Omega} \ell_{\mathrm{H}}(D(p), D_{\mathrm{gt}}(p))}_{\mathcal{L}_{\mathrm{fit}}} + \underbrace{\frac{\lambda_h}{|\Omega|}\sum_{p\in\Omega} \max\big(e_r(p) - e_0(p),\, 0\big)}_{\mathcal{L}_{\mathrm{harm}}} + \underbrace{\frac{\lambda_p}{|\Omega_{\mathrm{easy}}|}\sum_{p\in\Omega_{\mathrm{easy}}} \big|\Delta(p)\big|}_{\mathcal{L}_{\mathrm{pres}}}, \quad (15)$$

where Ω is the set of valid ground-truth pixels, $\Omega_{\mathrm{easy}} = \{p \in \Omega : e_0(p) < 1\,\mathrm{px}\}$, ℓ_{H} is the Huber loss with transition point $\beta = 1$, and $\lambda_h = \lambda_p = 1$ in all experiments. The fitting term $\mathcal{L}_{\mathrm{fit}}$ optimizes the final disparity estimate, while $\mathcal{L}_{\mathrm{harm}}$ penalizes regressions relative to the event-only estimate. The preservation term $\mathcal{L}_{\mathrm{pres}}$ discourages unnecessary corrections where Stage I is already accurate, acting as a trust region on Δ. Together, these terms encourage improvements in difficult regions while preserving reliable event-derived geometry.

4 Experiments

4.1 Implementation and Experimental Settings

Implementation Details. We implement our networks in PyTorch and train on a single NVIDIA RTX PRO 6000 Blackwell GPU with batch size 8. Each stage is trained for 200k steps at 288×384 using random crops, followed by 100k steps at full resolution (480×640). Stage-I is trained first, followed by Stage-II with Stage-I frozen. We use AdamW with a learning rate of 1×10^{-4}. The disparity objective uses Huber loss with $\beta = 1$, while Stage-II additionally uses the harm and preservation terms with $\lambda_h = \lambda_p = 1$. We use OneCycleLR with cosine annealing, 1% warm-up, and no cycle momentum, and clip gradients at 0.5. The left frame is randomly dropped with probability 0.2 during training.

Datasets. We train on DSEC Gehrig et al. (2021) and follow the Bi-CMPStereo split Xu et al. (2026): 31 training sequences (18,950 samples) and 10 held-out sequences (7,109 samples). As in Bi-CMPStereo, the left frame is warped to the left event camera using the calibration provided with the dataset. Our method uses both event streams for matching and only the left frame for refinement; although a second frame is available, we exclude it so that appearance cannot influence matching. We also evaluate zero-shot generalization on MVSEC Zhu et al. (2018) and M3ED Chaney et al. (2023) without fine-tuning, using the same subsets as ZEST Lou et al. (2024): MVSEC Indoor Flying S1–S3 and M3ED *car urban day horse*.

Metrics. Following Bi-CMPStereo Xu et al. (2026), we report MAE, RMSE, and the n-pixel error (nPE, percentage of pixels with error above n pixels), with $n \in \{1, 2\}$ on DSEC and $n \in \{2, 3\}$ on MVSEC and M3ED.

Baselines. We compare against ZEST Lou et al. (2024) and its DSEC-trained variant ZEST† Xu et al. (2026), SEVFI Ding et al. (2024), Bi-CMPStereo Xu et al. (2026), SE-CFF Nam et al. (2022), DTC Zhang et al. (2022), and, on the DSEC online benchmark, Zhuang et al. Zhuang et al. (2025) and EMatch Zhang et al. (2025b). All baselines use published results and splits.

Table 2: **Zero-shot cross-dataset generalization from DSEC.** Models are evaluated on unseen MVSEC and M3ED without target-domain adaptation, following the ZEST protocol. S1–S3 denote the MVSEC Indoor Flying subsets; ZEST† is the DSEC-trained variant.

	DSEC → MVSEC																DSEC → M3ED			
	MAE ↓				2PE ↓				3PE ↓				RMSE ↓				MAE ↓	2PE ↓	3PE ↓	RMSE ↓
Method	S1	S2	S3	All	S1	S2	S3	All	S1	S2	S3	All	S1	S2	S3	All				
SE-CFF Nam et al. (2022)	2.428	2.867	2.518	2.614	52.993	53.451	52.837	53.076	30.102	32.864	32.238	31.972	5.729	6.241	5.807	5.934	1.615	17.364	12.182	2.976
DTC Zhang et al. (2022)	2.175	2.502	2.204	2.297	48.767	50.646	50.237	50.047	27.382	29.928	29.530	29.185	5.313	5.571	5.210	5.353	1.902	22.375	15.661	3.068
ZEST Lou et al. (2024)	3.340	7.190	5.130	5.220	33.220	61.380	43.970	46.190	19.060	41.720	28.660	29.820	3.830	7.830	5.600	5.763	2.060	29.020	19.040	3.390
ZEST† Lou et al. (2024)	4.923	6.022	4.267	4.998	44.676	60.123	45.688	50.275	32.255	46.015	31.699	36.594	6.626	8.023	5.799	6.724	2.653	36.011	23.020	5.069
SEVFI Ding et al. (2024)	5.537	4.728	5.779	5.375	63.568	60.108	68.703	64.697	48.482	45.459	53.147	49.548	10.143	8.352	10.440	9.678	3.634	52.107	36.615	5.282
Bi-CMPS Xu et al. (2026)	1.743	2.181	1.674	1.858	29.247	37.041	29.869	32.121	14.924	20.889	15.336	17.095	2.708	3.419	2.466	2.837	1.557	**17.044**	**11.420**	2.877
Ours (Stage-I)	1.613	1.949	1.506	1.678	23.212	26.371	21.465	23.489	12.304	16.121	10.264	12.670	2.470	2.868	2.304	2.529	1.570	24.160	15.080	3.000
Ours	**1.359**	**1.762**	**1.302**	**1.468**	**17.362**	**21.716**	**16.627**	**18.487**	**10.600**	**14.811**	**9.688**	**11.598**	**2.269**	**2.721**	**2.159**	**2.371**	**1.314**	18.063	11.468	**2.691**

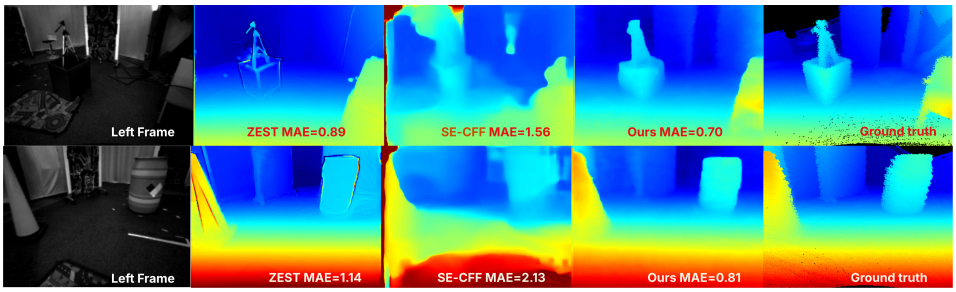

Figure 4: **Zero-shot DSEC → MVSEC.** Left to right: frame, ZEST, SE-CFF, ours, ground truth.

4.2 Results

In-Domain Evaluation. We follow the evaluation protocol of Bi-CMPStereo Xu et al. (2026). Table 1 reports results on the DSEC 31/10 split (together with parameter counts and MACs). Our full model achieves the best accuracy among event-only (E+E) and asymmetric E+F methods that use at most a single frame, ranking first across all reported metrics. Specifically, our model achieves 0.491 MAE, 8.839 1PE, 1.368 2PE, and 1.109 RMSE, improving over Bi-CMPStereo by 7.7%, 16.7%, and 43.3% on MAE, 1PE, and 2PE respectively, and also improving RMSE (1.109 vs. 1.210). At 480×640 input, our full model uses 10.34M parameters and 654.4 GMACs, 5× fewer parameters than Bi-CMPStereo and 33× fewer than ZEST Lou et al. (2024) while achieving the best accuracy. The gains are largest on 1PE and 2PE, indicating that the frame mainly corrects residual errors of an already accurate event-only estimate, sharpening fine structures and depth boundaries (Fig. 1).

Table 3: **Comparison on the DSEC online benchmark.** Input: number of event (E) and frame (F) cameras required at inference. First, second, and third best results over all methods are highlighted, best in each group are bolded.

Method	**Input**	**MAE↓**	**1PE↓**	**2PE↓**	**RMSE↓**
SE-CFF Nam et al. (2022)	2E	0.519	9.583	2.620	1.231
DTC Zhang et al. (2022)	2E	0.527	9.277	2.416	1.291
EMatch Zhang et al. (2025b)	2E	0.467	8.338	2.168	**1.075**
Ours (Stage-I)	2E	0.507	9.583	2.607	1.179
Ours (event-only)	2E	**0.463**	**7.795**	**1.989**	1.099
Zhuang et al. Zhuang et al. (2025)	1E+1F	0.488	8.881	2.335	1.168
Bi-CMPS Xu et al. (2026)	1E+1F	**0.475**	**8.557**	**2.253**	**1.101**
Ours	2E+1F	**0.411**	**5.993**	**1.381**	**1.001**
EI-Stereo Mostafavi et al. (2021)	2E+2F	0.396	5.814	1.055	0.906
SCS-Net Cho & Yoon (2022b)	2E+2F	**0.387**	**5.673**	**0.992**	**0.852**

Cross-Dataset Generalization. Table 2 evaluates zero-shot transfer from DSEC to MVSEC and M3ED without target-domain adaptation. Our full model achieves the best MAE, 2PE, 3PE, and RMSE on MVSEC, and the best MAE and RMSE on M3ED, reducing MVSEC MAE from 1.858 to 1.468 (21.0%) and RMSE from 2.837 to 2.371 (16.4%) relative to Bi-CMPStereo. On MVSEC, the Stage-I estimator alone already outperforms all baselines (1.678 vs. 1.858 MAE), indicating that event-only geometry transfers well across sensor configurations. Frame refinement then reduces MAE by 12.5% on MVSEC, compared to 9.7% on DSEC, showing that its benefit is preserved under domain shift. These gains are consistent across all three MVSEC subsets, where our full model and Stage-I rank first and second on every metric. On M3ED, our model ranks first on MAE and RMSE, while Bi-CMPStereo remains slightly better on 2PE and 3PE. Qualitative examples are shown in Figure 4.

DSEC Online Benchmark. Table 3 reports results on the DSEC online benchmark. Our 2E+1F model achieves 0.411 MAE, 5.993 1PE, 1.381 2PE, and 1.001 RMSE, the best among all methods

Table 4: **Component ablation on DSEC.** We progressively enable each component of our pipeline. Stage I establishes event-only geometry; Stage II refines it with geometry-aligned frame evidence. ✓ denotes an enabled component. First, second, and third best results are highlighted.

	Stage I		Stage II			Performance			
ID	Structured Event State	Cross-View Matching	Reliability Gate	Frame Dropout	D_0-Cond. Sampling	MAE↓	1PE↓	2PE↓	RMSE↓
A	-	-	-	-	-	0.656	14.24	4.27	1.41
B	✓	-	-	-	-	0.601	12.51	3.52	1.33
C	✓	✓	-	-	-	0.544	10.45	1.62	1.116
D	✓	✓	✓	-	-	0.533	8.69	1.55	1.114
E	✓	✓	✓	✓	-	0.515	**8.61**	1.45	1.112
Ours	✓	✓	✓	✓	✓	**0.491**	8.84	**1.37**	**1.109**

using at most one frame camera. Our event-only variant, which uses the full refinement network with the frame removed, achieves the best MAE, 1PE, and 2PE among event-only methods (0.463, 7.795, and 1.989), while EMatch Zhang et al. (2025b) remains better on RMSE (1.075 vs. 1.099). Symmetric 2E+2F methods, which use one more frame camera, remain more accurate across all metrics.

4.3 Ablation Study

Component Ablation. Table 4 isolates each component. Structured event states (B) reduce MAE from 0.656 to 0.601, and cross-view propagation (C) yields the event-only Stage-I estimator (0.544 MAE). In Stage II, the reliability gate (D) provides the largest 1PE reduction (10.45 to 8.69), and frame dropout (E) further improves all metrics. D_0-conditioned sampling (Ours) improves MAE, 2PE, and RMSE, reaching 0.491 MAE, at a slight cost in 1PE.

Table 5: **Fusion timing ablation on DSEC.**

Fusion Strategy	MAE↓	1PE↓	2PE↓	RMSE↓
Early fusion	0.561	11.02	1.78	1.142
Post-D_0 fusion	0.518	9.31	1.49	1.118
Post-D_0 + reliability (Ours)	**0.491**	**8.839**	**1.37**	**1.109**

Fusion Timing. Table 5 isolates where frame evidence enters the pipeline, holding the event-only backbone and refinement capacity fixed. Performance improves as fusion is delayed: early fusion yields 0.561 MAE, post-D_0 fusion 0.518, and our full scheme (post-D_0 with reliability gating) 0.491. Notably, concatenating frame and event features before matching performs worse than the event-only Stage I (0.561 vs. 0.544 MAE): introducing the frame before matching not only fails to help but degrades the estimate. Delaying fusion reduces MAE by 7.7%, while reliability gating provides a further 5.2% reduction over unconditional post-D_0 fusion. The largest relative gain from gating is in 2PE, which decreases from 1.49 to 1.36 (8.7%).

Efficiency. At 480×640 input, our full model uses 10.34M parameters and 654.4 GMACs, $5\times$ fewer parameters than Bi-CMPStereo and $33\times$ fewer than ZEST Lou et al. (2024), while achieving the best accuracy among methods using at most one frame camera. The event-only Stage-I estimator uses 7.89M parameters and 437.0 GMACs.

5 Conclusion

We introduced an event-first stereo framework that establishes geometry before frame evidence is consulted. On DSEC, our 2E+1F model is the most accurate among methods using at most one frame camera, its event-only variant outperforms prior event-only methods, and the model generalizes zero-shot to MVSEC and M3ED. Ablations show that fusion timing is a key design choice for event–frame stereo: introducing the frame before matching performs worse than using no frame at all. A limitation is that symmetric two-frame methods remain more accurate when both frames are reliable, reflecting the price of appearance-independent geometry.

AI Assistance Statement

LLM were used only to polish the writing of this manuscript, not to generate ideas, implement methods, produce results, or write technical content. All scientific contributions are the authors' own.

References

Luca Bartolomei, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Guillermo Gallego. Eventhub: Data factory for generalizable event-based stereo networks without active sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37063–37074, 2026.

Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *2023 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 4016–4023. IEEE, 2023.

Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 454–462, 2022a.

Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pp. 470–486. Springer, 2022b.

Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *European Conference on Computer Vision*, pp. 294–314. Springer, 2024.

Chao Ding, Mingyuan Lin, Haijian Zhang, Jianzhuang Liu, and Lei Yu. Video frame interpolation with stereo event and intensity cameras. *IEEE Transactions on Multimedia*, 26:9187–9202, 2024.

Yang Ding, Can Han, Sijia Du, Yaqi Wang, and Dahong Qian. Robust real-time endoscopic stereo matching under fuzzy tissue boundaries. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 392–406. Springer, 2025.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. doi: 10.1109/LRA.2021.3068942.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Luoxi Jing, Dianxi Shi, Yushe Cao, Yuanze Wang, Junze Zhang, Yuning Cui, and Mengzhu Wang. Aimdepth: Asymmetric image-event mamba for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8033–8044, 2026.

Jie Long Lee and Gim Hee Lee. Ares: Unifying asymmetric rgb-event stereo for probabilistic scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37022–37031, 2026.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Hanyue Lou, Jinxiu Liang, Minggui Teng, Bin Fan, Yong Xu, and Boxin Shi. Zero-shot event-intensity asymmetric stereo via visual prompting from image domain. *Advances in Neural Information Processing Systems*, 37:13274–13301, 2024.

Xingyuan Lu, Yanbing Xue, Leida Li, Shiyin Li, and Zan Gao. Mamba-based multi-branch cost aggregation for stereo matching. *Applied Soft Computing*, pp. 113973, 2025.

Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4258–4267, 2021.

Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6104–6113. IEEE, 2022.

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

Shaoxuan Suo, Jinxin Liu, and Huihui Miao. Mamba-stereo: Mamba regularization for stereo matching. *Pattern Recognition*, 170:112120, 2026.

Fabio Tosi, Luca Bartolomei, and Matteo Poggi. A survey on deep stereo matching in the twenties. *International Journal of Computer Vision*, 133(7):4245–4276, 2025.

Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1527–1537. IEEE, 2019.

Xu Wang, Jialang Xu, Shuai Zhang, Baoru Huang, Danail Stoyanov, and Evangelos B Mazomenos. Stereomamba: Real-time and robust intraoperative stereo disparity estimation via long-range spatial dependencies. *IEEE Robotics and Automation Letters*, 2025.

Ninghui Xu, Fabio Tosi, Lihui Wang, Jiawei Han, Luca Bartolomei, Zhiting Yao, Matteo Poggi, and Stefano Mattoccia. Bidirectional cross-modal prompting for event-frame asymmetric stereo. *arXiv preprint arXiv:2604.15312*, 2026.

Haihao Zhang, Yunjian Zhang, Jianing Li, Lin Zhu, Meng Lv, Yao Zhu, Yanwei Liu, and Xiangyang Ji. Enhanced event-based dense stereo via cross-sensor knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5437–5447. IEEE, 2025a.

Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, and Luziwei Leng. Discrete time convolution for fast event-based stereo. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8666–8676. IEEE, 2022.

Pengjie Zhang, Lin Zhu, Xiao Wang, Lizhi Wang, and Hua Huang. Ematch: A unified framework for event-based optical flow and stereo matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5845–5855. IEEE, 2025b.

Fengan Zhao, Qianang Zhou, and Junlin Xiong. Edge-guided fusion and motion augmentation for event-image stereo. In *European Conference on Computer Vision*, pp. 190–205. Springer, 2024.

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.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Hao Zhuang, Yan Yang, and Liyuan Pan. Asymmetric event-image stereo with temporal feature gating and iterative structure-detail refinement. In *36th British Machine Vision Conference*, pp. 24–27, 2025.

A Appendix

A.1 Event Representation

Stage-I operates exclusively on events, encoded as a voxel grid. Given the set of events $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$ falling within a temporal window of $\Delta t = 100\,\text{ms}$ preceding the reference timestamp, with polarity $p_i \in \{-1, +1\}$, we build a tensor $V \in \mathbb{R}^{B\times H\times W}$ with $B = 15$ bins at the sensor resolution $H \times W = 480 \times 640$.

Temporal Normalization. Timestamps are mapped to the bin axis using the empirical extent of the window rather than its nominal bounds,

$$\tau_i = (B-1)\,\frac{t_i - t_{\min}}{t_{\max} - t_{\min}}, \tag{16}$$

where $t_{\min}$ and $t_{\max}$ are the first and last event timestamps in the window. This makes the representation invariant to the absolute event rate: a window containing few events still spans the full temporal axis, so the encoder sees a consistent temporal scale in both sparse and dense regions.

Trilinear Accumulation. Each event is distributed over its eight neighbouring cells in (x, y, τ) with weights given by the product of per-axis linear kernels,

$$V[b, v, u] \;=\; \sum_i p_i\, \max(0, 1 - |u - x_i|)\, \max(0, 1 - |v - y_i|)\, \max(0, 1 - |b - \tau_i|). \tag{17}$$

Interpolating in space as well as time avoids the quantisation artefacts that a nearest-neighbour assignment introduces at high event rates, and keeps the representation differentiable with respect to sub-pixel event coordinates.

Signed Polarity. Polarity enters as the sign of the accumulated contribution rather than as separate channels. A cell therefore records the *net* brightness change over its temporal support, and opposite-polarity events at the same location cancel. This halves the channel count relative to a polarity-split encoding, at the cost of discarding the distinction between a quiet region and one with balanced activity. We found the compact encoding preferable because the stereo matching cost depends on the structure of brightness change rather than on event volume.

Standardization. The grid is standardized over its non-zero entries only,

$$V[b, v, u] \leftarrow \frac{V[b, v, u] - \mu_{\neq 0}}{\sigma_{\neq 0}}, \qquad \forall\, V[b, v, u] \neq 0, \tag{18}$$

leaving empty cells at exactly zero. Restricting the statistics to occupied cells prevents the large fraction of empty space in a typical window from dominating μ and σ, which would otherwise compress the dynamic range of the informative cells and make normalisation depend on scene sparsity. Empty cells remain identically zero across sequences and sensors, giving the encoder an unambiguous "no evidence" value a property the adaptive-resolution pipeline of Section 3.3 relies on when letterboxing sensors of differing resolution onto a common canvas.

Both Views. The left and right event streams are converted independently with identical parameters, yielding $V_L, V_R \in \mathbb{R}^{15\times 480\times 640}$, which form the two inputs to the shared Stage-I encoder. No frame information enters at this stage.

A.2 Additional Qualitative Results

Figure 5 shows four additional zero-shot DSEC $\rightarrow$ MVSEC samples, one per row. The same qualitative pattern as Figure 4 emerges: ZEST recovers the overall structure but lacks sharp boundaries, while SE-CFF over-smooths the disparity map and yields the highest MAE on every sample. Our method preserves thin structures and depth discontinuities, achieving the lowest MAE on all four samples (0.80, 0.78, 0.87, and 0.91 vs. 1.12, 0.90, 1.05, 1.04 for ZEST and 2.43, 2.41, 2.16, 2.55 for SE-CFF).

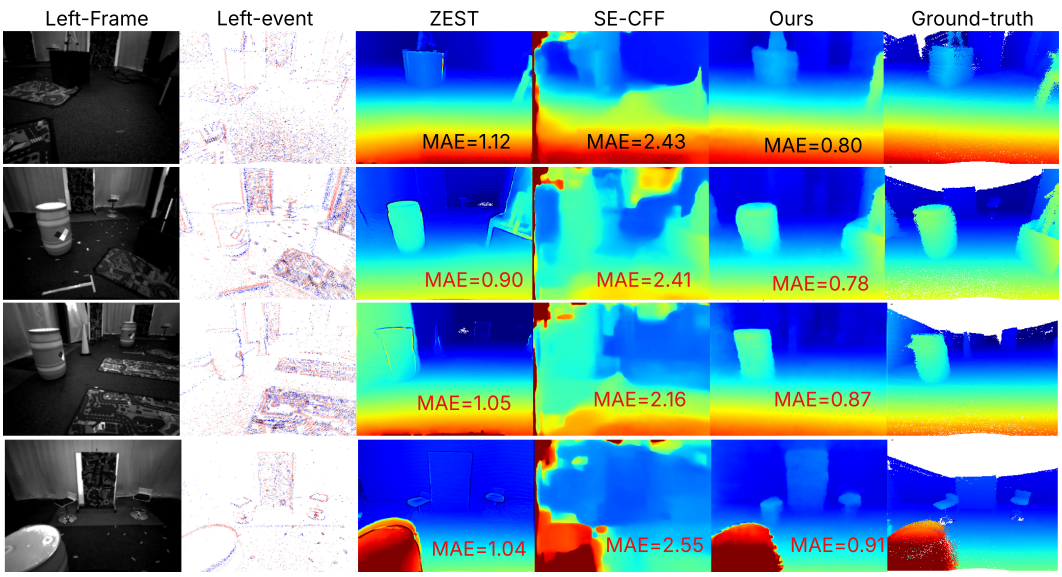

Figure 5: **Additional zero-shot DSEC → MVSEC qualitative results.** Left to right: left frame, left events, ZEST, SE-CFF, ours, and ground truth. MAE values are shown on each prediction panel. Our method consistently produces sharper depth boundaries and lower error than both baselines across all four sequences.

Table 6: **Effect of frame corruption on reliability gating.** MAE on a subset of 1000 DSEC validation samples under clean and blurred frame inputs, reported separately for day and night sequences. Stage-1 uses events only; 2E is an event-only control at the same refinement capacity. Gated refinement remains bounded below the event-only baseline under every condition, while ungated refinement degrades sharply on blurred frames.

	Clean		**Blurred**	
Setting	Day	Night	Day	Night
Stage-1 (no frame)	0.507	0.463	0.507	0.463
2E (control)	0.458	0.414	0.458	0.414
2E+F, gated	**0.379**	**0.361**	**0.425**	**0.412**
2E+F, no gate	0.383	0.372	0.530	0.560

A.3 Additional Discussion of Corruption Study

Table 6 evaluates the reliability gate under controlled frame corruption. We compare four configurations on a subset of 1000 DSEC validation samples, separately for day and night sequences, and under clean or blurred frame inputs. Stage-1 uses events only; 2E is an event-only control that receives no frame input but has access to the same refinement network, isolating the effect of frame fusion from the effect of additional refinement capacity.

Four findings emerge. First, frame refinement helps substantially when frames are reliable: gated refinement reduces MAE from 0.458 to 0.379 on clean day sequences (-17%) and from 0.414 to 0.361 on clean night sequences (-13%). Second, the gate costs almost nothing when frames are clean: gated and ungated refinement differ by only 1% on day and 3% on night, so gating does not sacrifice accuracy in the well-behaved regime. Third, the gate is essential under corruption: with blurred frames, ungated refinement degrades to 0.530 (day) and 0.560 (night), while gated refinement remains at 0.425 and 0.412, yielding a 20% and 26% improvement respectively. Fourth, gated refinement is bounded: under every condition, including blurred inputs, it remains below both the event-only 2E control and the Stage-1 baseline, so adding frame evidence never harms the model. Together, these results confirm that the reliability gate selectively incorporates frame evidence when it is informative and suppresses it when it is not.