Bidirectional Cross-Modal Prompting for Event-Frame Asymmetric Stereo
Supplementary Material

Ninghui Xu[1,2,†] Fabio Tosi[2] Lihui Wang[1,*] Jiawei Han[2,4] Luca Bartolomei[2]
Zhiting Yao[3] Matteo Poggi[2] Stefano Mattoccia[2]

[1]School of Instrument Science and Engineering, Southeast University,
State Key Lab of Comprehensive PNT Network and Equipment Technology,
Key Lab of Micro-Inertial Instrument and Advanced Navigation Technology, MOE
[2]Department of Computer Science and Engineering, University of Bologna
[3]College of Computer Science and Software Engineering, Hohai University
[4]Beijing Institute of Technology

This supplementary document provides additional information for the paper "Bidirectional Cross-Modal Prompting for Event-Frame Asymmetric Stereo", including an extended description of the event representations used in our networks (sec. 1), details of the train–test split of the DSEC [3] dataset (sec. 2), and extensive qualitative visual comparisons on DSEC (sec. 3).

1. Event Representation

Different from conventional cameras that capture intensity frames at a fixed rate, event cameras asynchronously report the per-pixel illumination variation (i.e., event) with microsecond resolution. Letting $L(x, y, t)$ define the logarithmic brightness on pixel (x, y) at time t, an event $e = (x, y, t, p)$ is generated whenever the log-intensity change within the pixel (x, y) reaches a threshold C_{th}:

$$L(x, y, t) - L(x, y, t - \Delta t) = p \cdot C_{th} \tag{1}$$

where Δt is the time elapsed since the previous event at the same pixel, and the polarity $p \in \{+1, -1\}$ indicates whether the brightness has increased or decreased.

Owing to the asynchronous nature of events, we need to convert the event stream into a tensor-like representation that can be interpreted by neural networks. To better leverage the complementary cues from image and event domains in our bidirectional framework, imgCMPStereo and evCMPStereo adopt event representations tailored to their respective target domains.

Voxel grid [7] for imgCMPStereo. Voxel grid takes a spatial-temporal quantization approach. Given a set of N events $\{e_i\}_{i=1}^{N}$, by discretizing its duration $\Delta T = t_N - t_1$ into $B = 5$ uniform bins, each event $e_i = (x_i, y_i, t_i, p_i)$ distributes its polarity to the two nearest voxels using bilinear sampling kernel as follows:

$$V(x, y, t) = \sum_{x_i=x y_i=y} p_i \cdot \max\left(0, 1 - \left|t - \widetilde{t_i}\right|\right) \tag{2}$$

where $\widetilde{t_i} := \frac{B-1}{\Delta T}(t_i - t_1)$ is the normalized timestamp.

Given the proven effectiveness of voxel grid representation in event-based vision tasks, we adopt it in imgCMPStereo to enable efficient cross-modal alignment and stereo matching within the image domain.

Event concentration [5] for evCMPStereo. A stream of events is initially represented using a mixed-density stacking method, where events are reversely stacked from the depth timestamp with exponentially increasing batch sizes. The sequence consists of $M = 10$ stacks and each generates a one-channel tensor $S_{1...M} \in \mathbb{R}^{H \times W}$. The mixed-density event tensors are concatenated along the channel dimension and fed into an attention-based network for weighting the importance of each event stack. A weighted sum is then performed with the output weights $W \in \mathbb{R}^{H \times W \times M}$ over the stacked tensors to obtain the concentrated event representation $E \in \mathbb{R}^{H \times W}$:

$$E(x, y) = \sum_{j=1}^{M} W(x, y, j) \cdot S_j(x, y) \tag{3}$$

By concentrating events into a sharp, blur-free edge-like tensor that preserves the intrinsic response characteristics of event cameras, event concentration provides strong complementarity to the image domain and is therefore adopted as the event representation for evCMPStereo.

2. Details of the DSEC Train–Test Split

We follow [1] to split the DSEC [3] dataset, which contains 41 sequences in total, into 31 training sequences and 10 testing sequences. The test set comprises 7,109 samples and includes the following sequences: *zurich_city_05_a*, *zurich_city_05_b*, *zurich_city_06_a*, *zurich_city_07_a*,

† Work done while visiting the University of Bologna
*Corresponding author

zurich_city_08_a, *zurich_city_09_d*, *zurich_city_10_b*, *interlaken_00_f*, *interlaken_00_g*, and *thun_00_a*. The remaining 31 sequences with 18,950 samples are used for training.

3. Additional Qualitative Results on DSEC

We provide additional qualitative results on the DSEC [3] dataset. As in the main paper, we compare our imgCMP-Stereo, evCMPStereo, and Bi-CMPStereo with the state-of-the-art event–frame asymmetric stereo method ZEST [4], its DSEC-trained variant ZEST†, and SEVFI-Net [2], as well as two state-of-the-art event-based symmetric stereo approaches, SE-CFF [5] and DTC-SPADE [6]. Fig. 1 and 2 are results in nighttime and daytime scenarios, respectively. We additionally present the event concentration, which clearly shows the scene edges and demonstrates it preservation of the edge–response characteristics of the event camera. The results highlight the advantages of our approach in producing higher accuracy and higher-quality structural disparity maps. In the highlighted boxes, despite the ambiguity introduced by complex textures or low-light conditions, our method still constructs sharp edges and fine structural details.

References

[1] Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. *arXiv preprint arXiv:2509.15224*, 2025. 1

[2] 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. 2

[3] 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, 2021. 1, 2, 3

[4] 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. 2

[5] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6114–6123, 2022. 1, 2

[6] Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, and Luziwei Leng. Discrete time convolution for fast event-based stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8676–8686, 2022. 2

[7] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019. 1

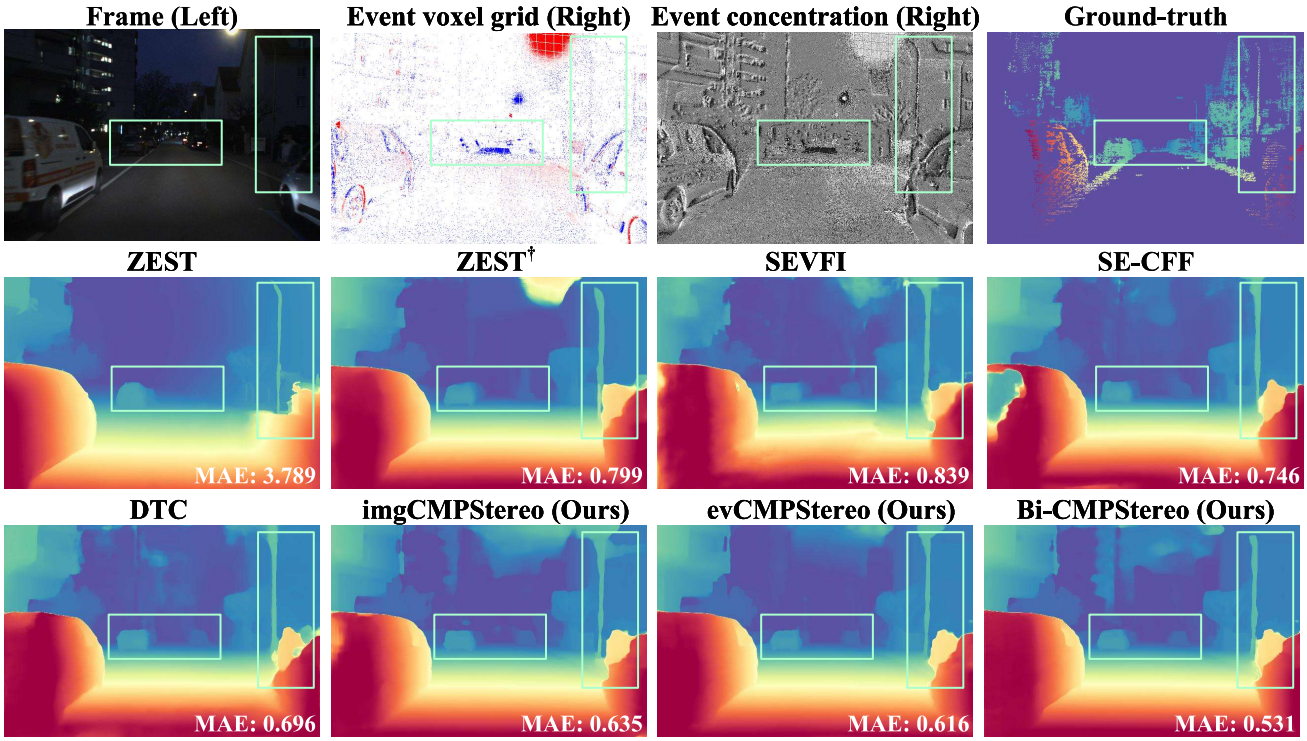

Figure 1. **Additional qualitative results in nighttime scenario from the DSEC dataset [3].** The mean absolute error (MAE) is shown at the bottom right of each estimated disparity map.

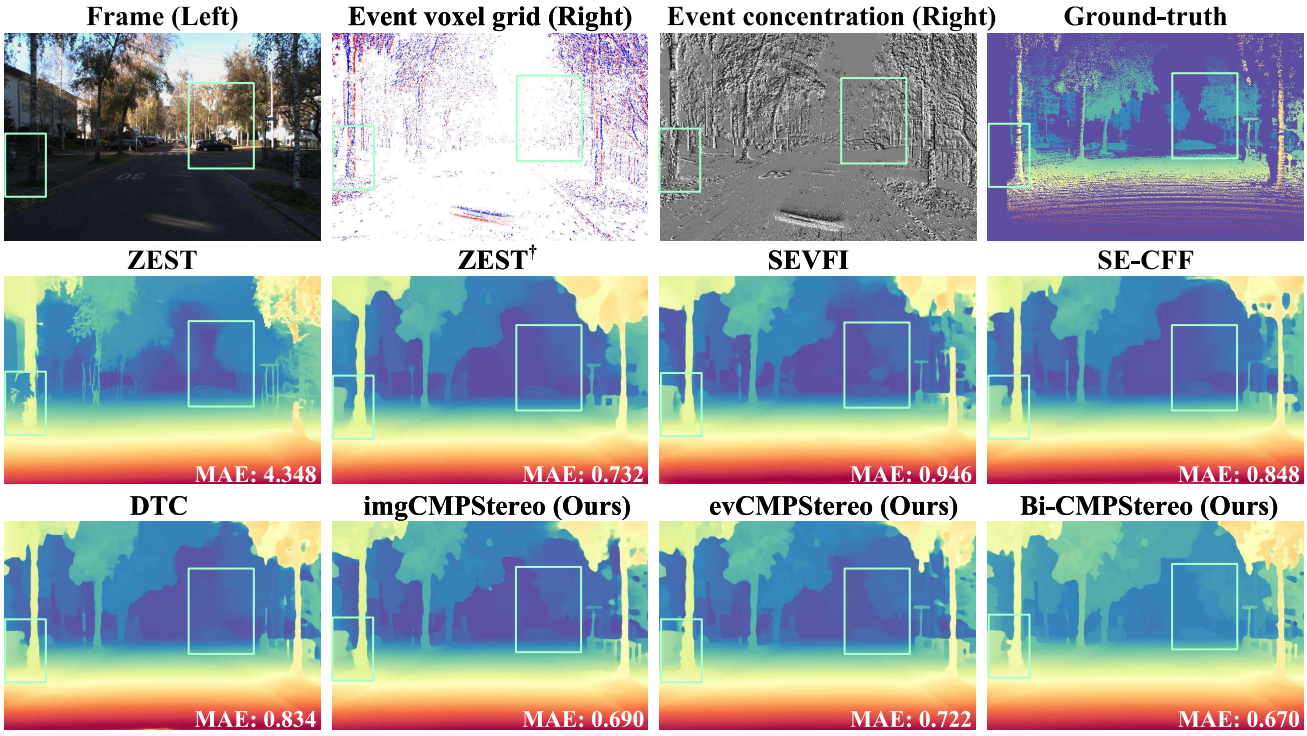

Figure 2. **Additional qualitative results in daytime scenario from the DSEC dataset [3].**