EBMV 2026 Event-Based Multimodal Perception Challenge Monocular Depth Estimation Track: Methods and Results

Shihan Peng[1], Yuyang Xiong[1], Hanyu Zhou[2], Yunfan Lu[3], Shaoyu Liu[4], Haoyue Liu[1], Peiqi Duan[5], Boxin Shi[5], Kuk-Jin Yoon[6], Hui Xiong[3], and Gim Hee Lee[2]
Challenge participant contributors:
Zijian Yang, Ke Luo, Nick Truong, William J. Beksi, Xinglin Li, Zixu Yu, Jing Cao, Zhenyu Ding, Kui Jiang, Dongyu Xie, Bofei Chen, Bohao Li, and Kaifan Qiao

[1] Huazhong University of Science and Technology
[2] National University of Singapore
[3] HKUST (Guangzhou)
[4] Xidian University
[5] Peking University
[6] KAIST

Abstract. Event cameras provide asynchronous brightness-change measurements with microsecond temporal resolution, high dynamic range, and low latency. These properties complement conventional intensity frames and can provide additional cues for depth estimation in low-light, high dynamic range, and texture-sparse scenes. This report describes the Monocular Depth Estimation Track of the Event-Based Multimodal Perception Challenge at ECCV 2026. The track evaluates metric monocular depth estimation from aligned image frames and event streams using a selected subset of the CoSEC dataset. The released benchmark contains 102 training sequences with 11,782 labeled samples and 26 hidden-test sequences with 2,951 samples. Participants submit scaled depth maps to a Codabench evaluation server, where predictions are evaluated with RMSE as the primary ranking criterion and AbsRel and threshold accuracies as complementary metrics. The challenge recorded 51 registered participants and 155 submissions. The official leaderboard contains nine final entries, with the first-ranked submission achieving an RMSE of 3.7164 and an AbsRel of 0.0682. More details are available at https://www.codabench.org/competitions/16218/.

Keywords: Event cameras · Monocular depth estimation · Multimodal perception · Benchmark · CoSEC dataset

1 Introduction

Metric depth estimation is used in autonomous driving, robotics, augmented reality, and embodied perception. Monocular systems require only a single viewpoint,

but the task is ill-posed: multiple three-dimensional scenes can explain the same image, and appearance can be unreliable under low illumination, over-exposure, large dynamic range, and texture-poor regions. Learning-based methods have improved monocular depth prediction from conventional images [5,8,14]. However, frame cameras provide discrete intensity measurements whose usefulness can be limited by exposure, illumination, and scene-dependent appearance ambiguities.

Event cameras offer a complementary sensing modality. Instead of capturing synchronized frames at a fixed rate, each pixel independently reports changes in log intensity. This representation has high temporal resolution, low latency, and high dynamic range, and it complements conventional imaging under low-light and high-dynamic-range conditions [7]. At the same time, events are sparse, asynchronous, and contrast-driven; they do not by themselves provide dense texture, color, or absolute intensity. One approach is therefore to combine frame appearance with temporal information from event streams.

The Event-Based Multimodal Perception Challenge: Monocular Depth Estimation Track at ECCV 2026 was organized to compare methods under a common data split and evaluation protocol. The track is based on the CoSEC dataset [13], which provides coaxial stereo event-camera data for autonomous-driving scenarios. For the challenge, participants receive monocular frame-event inputs and training depth annotations and predict metric depth maps for a hidden test set. The task evaluates multimodal fusion, absolute depth prediction, and generalization across the included driving scenes.

Following recent challenge reports, this paper first introduces the motivation and task setting, then presents the benchmark data, evaluation protocol, challenge phases, final results, and method section. The report is organized around the official Codabench evaluation and the resulting leaderboard for monocular event-frame depth estimation. Figure 1 illustrates the inference pipeline, in which frame appearance and event cues are fused by a depth estimation network for dense depth prediction.

2 EBMV 2026 Monocular Depth Estimation Track

The EBMV 2026 challenge provides a CoSEC-based training and test benchmark together with an online evaluation server. The track evaluates metric monocular depth prediction from frame-event inputs in autonomous-driving scenes with varying illumination conditions.

2.1 Problem Definition

The task is monocular metric depth estimation from multimodal inputs. For each target timestamp, participants receive an image frame and the associated event stream from a temporally aligned event camera. The goal is to predict a dense depth map at the same spatial resolution as the ground-truth annotation. Let I_t denote the image frame at timestamp t, and let $E_{[t-\Delta,t]}$ denote an event segment

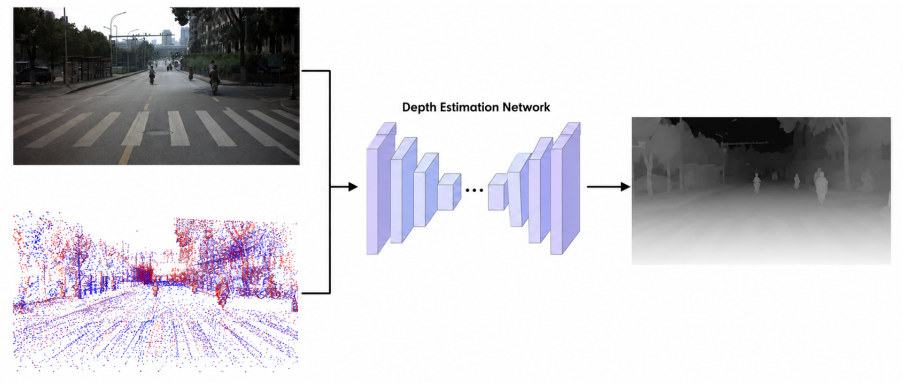


Fig. 1: Overview of the monocular event-frame depth estimation pipeline. The aligned RGB frame and polarity-coded event representation form the multimodal input to a depth estimation network, which predicts a dense depth map.

selected from a temporal window preceding the frame exposure endpoint. A submitted model can be described as

$$\hat{D}_t = F_\theta \left(I_t, E_{[t-\Delta,t]} \right), \tag{1}$$

where $\hat{D}_t$ is the predicted metric depth map and F_θ is the participant's depth estimation model. The formulation in Equation (1) does not require a fixed value of Δ; teams may choose how to represent and aggregate events as long as the official training data and permitted external resources are used according to the challenge rules.

The benchmark is organized around three technical objectives. First, it evaluates *absolute* depth prediction rather than only relative ordering; metric-scale estimates can be used by downstream planning and navigation systems. Second, it evaluates *multimodal fusion* of frame appearance and temporal event cues. Third, it includes scenes with high dynamic range, low light, and sparse or changing texture to measure performance across different imaging conditions.

2.2 Datasets

CoSEC depth challenge split. The challenge uses a subset of the CoSEC dataset [13]. CoSEC provides image frames, event streams, timestamps, and depth annotations for autonomous-driving scenes. Camera intrinsics are also provided as optional metadata, and participants may decide whether to use them. The challenge split contains 102 training sequences and 26 test sequences, with 11,782 labeled training samples and 2,951 hidden-test samples, as summarized in Table 1. The test depth maps are withheld from participants and are used only by the evaluation server.

Each sequence contains image frames from the left camera, event streams from the left event camera, timestamps, and, for the training split, ground-truth depth maps. The image frames and event streams are spatially aligned

Table 1: Dataset statistics for the challenge split.

Split	Sequences	Samples	Depth annotations
Training	102	11,782	Released
Test	26	2,951	Hidden

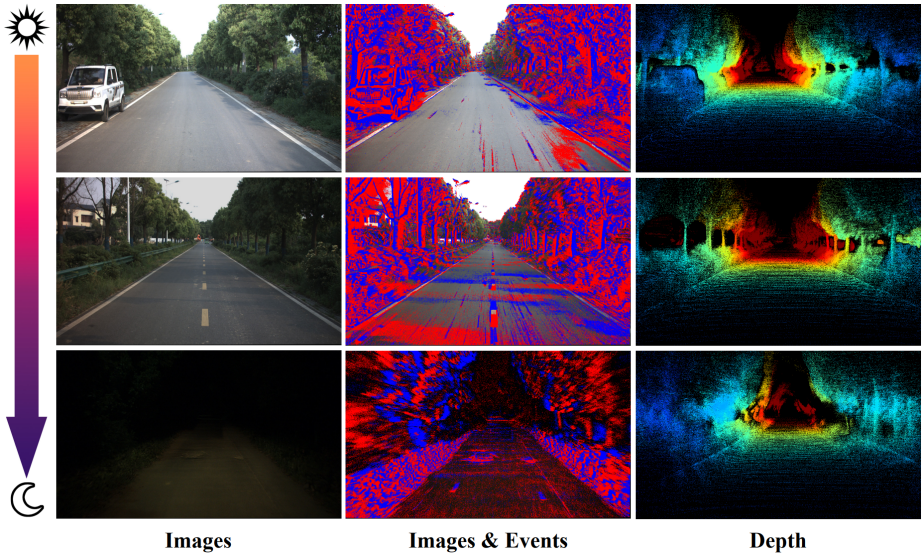


Fig. 2: Representative samples from the CoSEC depth challenge split under varying illumination conditions. From left to right, each row shows an image frame, the corresponding image-event overlay with the two event polarities rendered in red and blue, and the depth visualization.

and share a resolution of 1200×624. The timestamps allow participants to associate each image frame with a temporally corresponding event segment, while the hidden test annotations are used only by the evaluation server. Figure 2 presents representative training samples across illumination conditions ranging from daytime to nighttime, together with their event and depth information.

Frame, event, and timestamp format. The benchmark exposes two complementary visual modalities. Image frames provide dense spatial appearance, while event streams encode asynchronous brightness changes with microsecond timestamps and event polarities. For each target frame, participants can select a temporal window of events around the corresponding timestamp and transform it into a representation suitable for their model. Common choices include polarity-separated event-count images, voxel grids, time surfaces, recurrent event encodings, and sparse tensors; the benchmark does not impose a fixed event preprocessing pipeline.

This design intentionally leaves event representation and temporal aggregation to the participants. Models can use short event windows to emphasize local brightness-change patterns, longer windows to collect sparse activity in low-texture regions, or multi-scale temporal representations that combine both. The benchmark therefore covers both depth-network design and the alignment and fusion of asynchronous events with conventional image frames.

Depth representation. Training depth maps are stored as 16-bit PNG files using a fixed scale factor. If D is the metric depth in meters and D^{u16} is the stored unsigned 16-bit value, the relationship is

$$D^{u16} = 256D, \qquad D = \frac{D^{u16}}{256}. \tag{2}$$

The same representation in Equation (2) is required for submissions. This convention avoids ambiguities between floating-point meters and encoded depth values, reduces file size, and keeps the evaluation consistent across platforms. Invalid predictions such as negative depth, NaN, or infinite values are rejected by the evaluation program.

2.3 Evaluation

Submission format. Participants submit a single compressed archive containing one prediction for every test sample. Each prediction is stored as a compressed `.npz` file containing exactly one array with the key `depth`. The array must be two-dimensional, must match the ground-truth depth-map shape, and must contain scaled `uint16` depth values. The evaluator rejects incomplete submissions, incorrectly named predictions, mismatched shapes, invalid values, and archives that do not follow the official submission convention. The competition page also provided a dummy submission archive to help participants validate their outputs before uploading predictions to the server.

Metrics. The evaluation server compares each submitted prediction $\hat{D}$ with the hidden ground-truth depth D over the set of valid pixels $\mathcal{V}$. The primary ranking metric is root mean squared error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(\hat{D}_i - D_i)^2}. \tag{3}$$

Lower RMSE in Equation (3) indicates better metric depth estimation and determines the official leaderboard ranking.

To complement RMSE, the server also reports absolute relative error,

$$\mathrm{AbsRel} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\frac{|\hat{D}_i - D_i|}{D_i}, \tag{4}$$

Table 2: Leaderboard metrics. RMSE is the primary ranking metric.

Metric	Direction	Interpretation
RMSE	Lower is better	Metric squared-error sensitivity
AbsRel	Lower is better	Relative depth error
$\delta < 1.25$	Higher is better	Strict threshold accuracy
$\delta < 1.25^2$	Higher is better	Medium threshold accuracy
$\delta < 1.25^3$	Higher is better	Loose threshold accuracy

and three threshold accuracies:

$$\delta_k = \frac{1}{|\mathcal{V}|} \left| \left\{ i \in \mathcal{V} : \max\left(\frac{\hat{D}_i}{D_i}, \frac{D_i}{\hat{D}_i} \right) < 1.25^k \right\} \right|, \quad k \in \{1, 2, 3\}. \tag{5}$$

The AbsRel score in Equation (4) measures relative depth error, while the δ metrics in Equation (5) measure threshold accuracies for which higher values are better. Reporting both error and threshold metrics is useful because monocular depth models can differ in global scale, local smoothness, boundary behavior, and outlier frequency. Table 2 summarizes the leaderboard metrics and their optimization directions.

Evaluation server. The official evaluation was conducted through a final *Test Phase* on Codabench, with automatic submission execution enabled. The server evaluated hidden test predictions with a private reference-data package and private scoring program. Participants were limited to five submissions per day and one hundred submissions per person over the phase. The execution time limit was 600 seconds.

The server design separated public inputs from hidden labels. Participants downloaded the test inputs, generated depth predictions locally, and uploaded only the prediction archive. The Codabench evaluator checked the submission format, loaded the `depth` arrays, converted the scaled `uint16` values to metric depth, computed all leaderboard metrics, and updated the public ranking. The same evaluator applies common preprocessing, masks, and metric implementations to all submissions.

2.4 Challenge Phases

The competition was hosted on Codabench [2] and was structured around the following phases:

1. **Development Phase:** participants used the released training data to design, train, and validate their monocular depth estimation systems. This phase allowed teams to explore event representations, temporal windows, network architectures, loss functions, and post-processing strategies.

Table 3: Final public leaderboard for the monocular depth estimation track. Lower RMSE and AbsRel are better, while higher δ scores are better.

Rank	Team/User	RMSE	AbsRel	δ_1	δ_2	δ_3
1	kaaaa	3.7164	0.0682	0.9704	0.9970	0.9993
2	nick-truong	3.9434	0.0739	0.9609	0.9964	0.9994
3	yuzixu	4.2866	0.0775	0.9543	0.9952	0.9992
4	SOMIS-LAB	4.5144	0.0865	0.9399	0.9942	0.9989
5	djdkdjdk	4.5638	0.0830	0.9435	0.9944	0.9991
6	boffee	4.7737	0.0947	0.9223	0.9930	0.9990
7	cactus8603	4.8476	0.0912	0.9286	0.9918	0.9990
8	hritav_1896	5.4308	0.0824	0.9408	0.9888	0.9961
9	ziling_ji	5.5395	0.1431	0.8679	0.9879	0.9969

2. **Final Test Phase:** participants generated predictions for the hidden test inputs and submitted them to the Codabench evaluation server. The official ranking was determined by RMSE on the hidden test set, with AbsRel and threshold accuracies reported as complementary metrics.

The competition rules were designed to evaluate submitted models on unseen data. Participants were allowed to use the official training set for model development. External datasets, pretrained models, synthetic data, and additional training resources were allowed only when they were publicly available or legally accessible, and participants could be asked to disclose such resources in their method descriptions. The hidden test set could not be used for training, validation, manual tuning, reverse engineering, or any attempt to infer ground-truth depth. Submissions had to be generated automatically by a model or algorithm; manually corrected predictions, hard-coded outputs, multiple-account probing, and evaluation-server exploitation were prohibited.

3 Challenge Results

The public Codabench leaderboard contains nine final submissions, ranked by RMSE on the hidden test set. Table 3 reports the leaderboard values exposed by the official evaluation server. The Team/User column follows the public display names associated with the corresponding Codabench entries.

The first-ranked submission achieves an RMSE of 3.7164 and the lowest AbsRel score of 0.0682. The second-ranked submission obtains the highest δ_3 value, while the top three methods report δ_1 scores between 0.9543 and 0.9704. RMSE and threshold accuracies characterize different aspects of the error distribution, and all reported metrics are included in the comparison. The four highest-ranked solutions are described in Section 4.

4 Methods

This section presents the four highest-ranked solutions in leaderboard order. All four methods build on pretrained depth models, but they incorporate event information at different stages: the first-ranked method learns bounded residual corrections to a frozen RGB prior; the second-ranked method optimizes a learned member through its deployed blend with a fixed multi-model anchor; the third-ranked method adapts an event encoder before feature-level fusion; and the fourth-ranked method augments a video-depth backbone with recurrent event encoding and mask-guided cross-attention.

4.1 kaaaa

The first-ranked team proposes an RGB-event residual calibration framework for metric monocular depth estimation. As shown in Figure 3, the system consists of an RGB depth estimator, a multi-window event representation, an Event Residual Refiner, and a scene-conditioned calibration module. The design retains a pretrained image-based metric-depth prior and uses event cues to correct local depth structure.

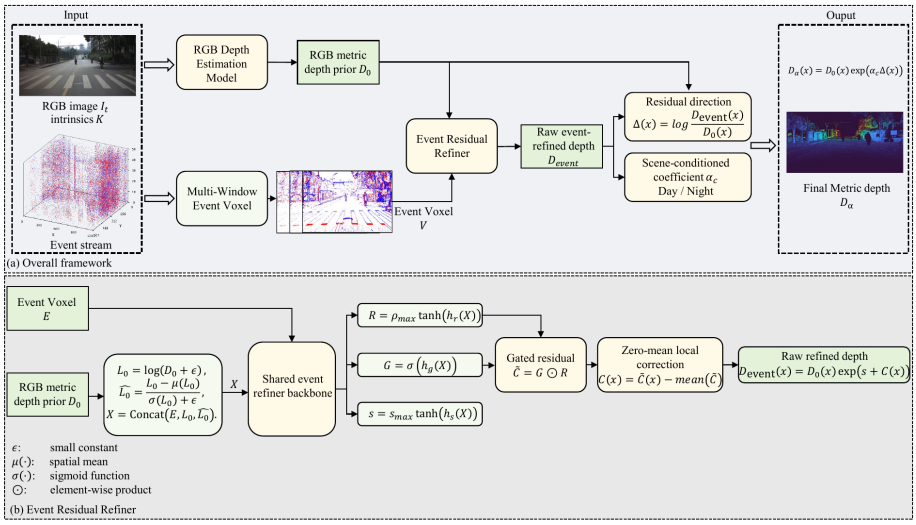


Fig. 3: Architecture of the first-ranked kaaaa method. The upper panel shows the complete RGB-event residual calibration pipeline, and the lower panel details the Event Residual Refiner.

RGB depth prior. The method adopts Metric3D v2 [10] with a ViT-Large backbone as its RGB depth estimator. The pretrained model is fine-tuned on the challenge training split with a small learning rate, and the checkpoint with the

lowest local validation RMSE is selected. It then remains frozen during event-stage training and produces the metric depth prior D_0. Freezing this branch retains the learned metric scale while allowing the smaller event branch to specialize in residual correction.

Event representation and residual refinement. For each image timestamp, events are collected over temporal windows of 10, 20, and 50 ms. Each window is divided into five temporal bins, with positive and negative polarities encoded separately, producing a 30-channel multi-window event voxel. Given the event voxel E, the method concatenates it with the log-depth prior $L_0 = \log(D_0 + \epsilon)$ and its spatially normalized form $\widehat{L}_0$ before passing the resulting tensor to a shared refiner backbone.

The refiner predicts a bounded residual map R, a sigmoid gate G, and a bounded frame-level scale term s. The gated residual is centered by subtracting its spatial mean, yielding a zero-mean correction C. The event-refined depth is then

$$D_{\text{event}}(x) = D_0(x) \exp\big(s + C(x)\big). \tag{6}$$

The gate controls where event information contributes, while the zero-mean constraint limits unintended global shifts and retains the metric scale of the RGB prior.

Conditional calibration and training. The team further adjusts the residual strength according to scene illumination. Defining $\Delta(x) = \log(D_{\text{event}}(x)/D_0(x))$, the final prediction is $D_\alpha(x) = D_0(x) \exp(\alpha_c \Delta(x))$, with $\alpha_{\text{day}} = 2.4$ and $\alpha_{\text{night}} = 0.45$. The RGB model is trained first, after which only the event refiner and scale branch are optimized using RMSE, AbsRel, log-depth, residual-regularization, and scale-regularization terms. The official training set is split by sequence into 92 training and 10 validation sequences, and inputs are processed at 616×1190 pixels. On this local split, event refinement reduces RMSE from 4.6212 to 4.5734, and conditional calibration further reduces it to 4.5603. The final submission ranks first with the official scores reported in Table 3.

4.2 nick-truong

The second-ranked team uses a blend-level training pipeline comprising a fixed foundation-depth anchor, a fine-tuned Depth Anything V2 (DAv2) member, and a guarded residual. As shown in Figure 4, the learned member is optimized through the same log-depth blend used at inference, and the training loss is applied to the combined prediction rather than to the member's standalone output.

Foundation-depth anchor. The fixed anchor draws on MetricAnything [11], Metric3D models [10, 18], and Depth Pro [1]. MetricAnything processes a frame enhanced by a log tone curve and CLAHE, while Metric3D-Large and Depth Pro process the raw frame. A Metric3D-Giant2 prediction is also generated

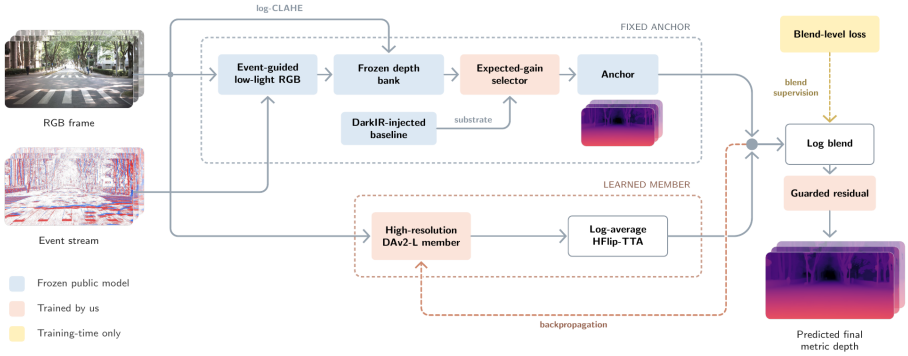


Fig. 4: Architecture of the second-ranked nick-truong method. A fixed multi-model anchor and a fine-tuned DAv2 member are combined by a supervised log-depth blend, followed by a bounded residual correction.

from RGB enhanced by the event-guided EvLight++ model [3]. The selector's baseline includes a small Metric3D-Large contribution restored by DarkIR [6]. Within connected regions that are dark, event-dense, and disputed by the candidate predictions, a gradient-boosted expected-gain selector chooses bounded component-level updates to this baseline.

Blend-level member training. The learned member is a high-resolution DAv2-L model [17] with a DINOv2/ViT-L encoder [12]. Let D_c be the fixed anchor and D_r the member output. The deployed prediction is a convex blend of $\log D_c$ and $\log D_r$ with a fixed member weight of $w = 0.346$; both input maps and the blended output are clamped to $[1, 150]$ m. The full DAv2 backbone and decoder are fine-tuned on all 102 labeled sequences using 840×840 crops. The loss is evaluated on the blended prediction and combines valid-pixel squared error, depth-band balancing, a far-range reference term, and cross-family teacher terms for near- and mid-range structure.

Inference and validation. At inference, predictions from the original and horizontally flipped frames are averaged in log depth before blending with the anchor. A histogram gradient-boosting regressor then predicts bounded residuals on 48×48 tiles for eligible near- and mid-range pixels, with guards preventing unreliable long-range edits. Dense event-conditioned residuals did not show a reliable alignment-specific gain in the team's controlled validation, so event data enter the submitted system primarily through low-light enhancement.

On the validation split, the component selector reduces anchor RMSE from 4.944 to 4.933, and the learned member lowers it to 4.674 through log-depth blending. The guarded residual reduces the deployed-base RMSE from 3.1722 to 3.0710 on the team's development split. The final submission achieves an RMSE of 3.9434 and ranks second on the hidden test set, as shown in Table 3.

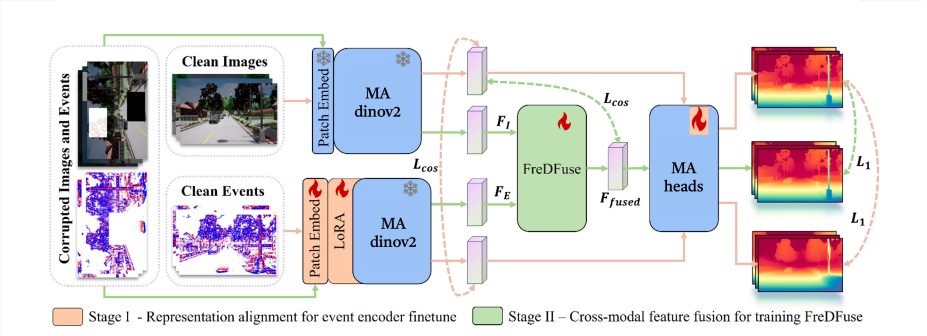


Fig. 5: Architecture of the third-ranked yuzixu method. Stage I aligns the event encoder with an RGB teacher, while Stage II fuses RGB and event features before MetricAnything-based depth decoding.

4.3 yuzixu

The third-ranked team develops a staged RGB-event fusion framework based on FUSE [15] and MetricAnything [11]. Its pipeline, illustrated in Figure 5, first adapts a pretrained event branch to the representation space of an RGB teacher and then learns cross-modal feature fusion for metric-depth decoding.

Backbone and event adaptation. MetricAnything is used as the shared metric-depth prior for the RGB and event-conditioned branches. The RGB branch serves as the teacher, while the event branch is initialized from the same pretrained checkpoint. Raw events associated with each frame are rasterized into a three-bin voxel grid, and only nonzero entries are normalized to zero mean and unit variance. LoRA adapters [9] are inserted into the transformer backbone's query-key-value, projection, and feed-forward linear layers, enabling event-specific adaptation without updating the complete pretrained backbone.

During Stage I, the RGB teacher receives the image and produces metric depth together with intermediate feature tokens. The event student receives the normalized voxel grid and is optimized with a depth-alignment loss and a cosine feature-alignment loss; metric ground truth from the official training set supplies additional supervision. This stage maps event observations into a feature space compatible with the RGB branch.

RGB-event fusion and scale adaptation. In Stage II, the modality-specific tokens are combined by a FUSE-style frequency-decoupled prefusion module. The fused tokens are reshaped into a feature map and passed through the MetricAnything neck, points head, mask head, and scale head to obtain the final metric depth. Thus, the output is decoded directly from the fused representation rather than from the Stage-I event prediction. After learning alignment and fusion, the decoder and scale-related components are unfrozen for a final absolute-scale adaptation stage.

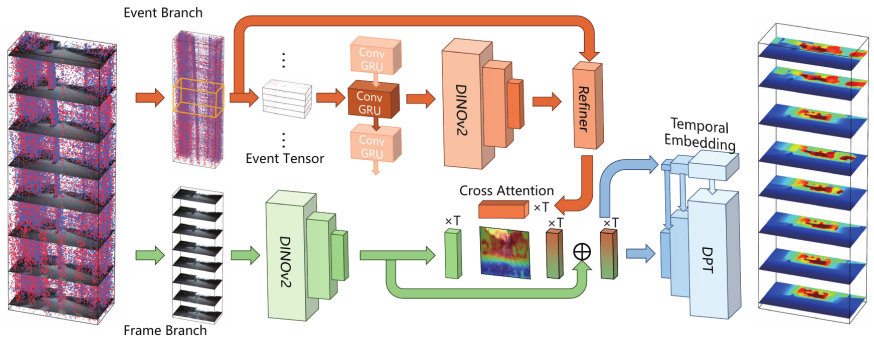


Fig. 6: Architecture of the fourth-ranked SOMIS-LAB E-VDA method. Parallel frame and event branches extract multiscale features, which are combined by gated cross-attention and decoded with temporal embeddings to produce depth predictions.

The team trains exclusively on the official challenge data. One sequence, `Night_City_006`, containing 73 labeled frames, is held out for local validation, while the remaining 11,709 frames are used for training. On this split, the Stage-II fusion model reduces RMSE from 9.8083 for Stage-I event adaptation to 5.2491 and raises δ_1 from 0.8269 to 0.9609. The resulting submission ranks third on the hidden test set, as shown in Table 3.

4.4 SOMIS-LAB

The fourth-ranked SOMIS-LAB team proposes E-VDA, an event-assisted depth estimation framework built on Video Depth Anything (VDA) [4], which is initialized from Depth Anything V2 [17]. As illustrated in Figure 6, the architecture contains parallel frame and event branches, a cross-modal fusion module, and a temporal depth-prediction head. The two branches use DINOv2 encoders [12], while the prediction head follows the spatial-temporal design of VDA.

Event representation and refinement. For every input frame, the associated event sub-stream is represented using a stack-based event encoding [16]. The temporal interval is discretized into five bins to preserve the event distribution within the interval. A ConvGRU then aggregates temporal context, and global pooling converts the sequence into frame-like event features for subsequent alignment with the image branch.

The team introduces a Multiscale Mask Refiner (MAR) to emphasize informative event regions. Event maps and time surfaces are multiplied to form initial activity masks, which are projected by a compact convolutional network. At each scale, a projected mask is concatenated with the corresponding DINOv2 event feature and transformed into a residual refinement. This procedure progressively updates the event representation across the feature hierarchy.

Gated cross-modal fusion. After layer normalization, the image features serve as queries and the refined event features serve as keys and values in cross-attention. Denoting the resulting cross-modal embedding by $\widehat{Z}$, the fused representation is

$$Z_{\text{fused}} = Z_{\text{image}} + s \odot \widehat{Z}, \qquad \widehat{Z} = \text{CrossAtten}(Z_{\text{image}}, Z_{\text{event}}, Z_{\text{event}}), \tag{7}$$

where the spatial gate s is obtained by pooling the projected multiscale masks. The gate controls the event contribution at each location and suppresses unreliable event responses before temporal depth decoding.

Training and validation. E-VDA is trained with a three-stage supervised schedule. Both DINOv2 backbones and the temporal prediction head are initialized from VDA. The first stage freezes the backbones, fusion module, and prediction head while training the ConvGRU event representation. The second stage keeps both backbones frozen and optimizes the fusion module and prediction head with a lower learning rate. The final stage unlocks the complete network and fine-tunes it with a very small learning rate.

Six sequences covering daytime and nighttime conditions form the team's local validation split. On its 690 validation frames, injecting event features reduces mean RMSE from 5.672 to 4.933 and mean AbsRel from 0.111 to 0.105, while increasing δ_1 from 0.886 to 0.902. The final E-VDA submission achieves an RMSE of 4.5144 and ranks fourth on the hidden test set, as reported in Table 3.

5 Conclusions

This report presents the ECCV 2026 Event-Based Multimodal Perception Challenge, Monocular Depth Estimation Track. The benchmark evaluates metric depth prediction from aligned image frames and event streams on a CoSEC-based split comprising 11,782 training samples and 2,951 hidden-test samples. Codabench ranks submissions by RMSE and additionally reports AbsRel and threshold accuracies. The challenge received 155 submissions from 51 registered participants. The submitted methods span residual refinement of frozen RGB priors, blend-level optimization with multi-model anchors, staged RGB-event feature fusion, and mask-guided cross-attention in a video-depth architecture. The first-ranked submission achieved an RMSE of 3.7164 on the hidden test set.

Acknowledgements

This workshop is partially supported by the National Research Foundation (NRF) Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008). We thank all participants for their contributions to the challenge.

References

1. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: International Conference on Learning Representations (2025)
2. Challenge Organizers: Event-based multimodal perception challenge: Monocular depth estimation track @ ECCV 2026. https://www.codabench.org/competitions/16218/ (2026), codabench competition page
3. Chen, K., Liang, G., Lu, Y., Li, H., Wang, L.: EvLight++: Low-light video enhancement with an event camera: A large-scale real-world dataset, novel method, and more. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(2), 1608–1625 (2026)
4. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22831–22840 (2025)
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems (2014)
6. Feijoo, D., Benito, J.C., Garcia, A., Conde, M.V.: DarkIR: Robust low-light image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10879–10889 (2025)
7. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 154–180 (2022)
8. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
10. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10579–10596 (2024)
11. Ma, B., Yang, J., Di, D., Zhang, X., Cui, J., Li, H., Xie, Y., Chen, W.: MetricAnything: Scaling metric depth pretraining with noisy heterogeneous sources. arXiv preprint arXiv:2601.22054 (2026)
12. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
13. Peng, S., Zhou, H., Dong, H., Shi, Z., Liu, H., Duan, Y., Chang, Y., Yan, L.: CoSEC: A coaxial stereo event camera dataset for autonomous driving. arXiv preprint arXiv:2408.08500 (2024)
14. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021)

15. Sun, P., Jiang, J., Yao, Y., Chen, Y., Zhao, W., Jiang, K., Liu, X.: FUSE: Label-free image-event joint monocular depth estimation via frequency-decoupled alignment and degradation-robust fusion. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 17231–17238 (2025)
16. Wang, L., Mostafavi I., S.M., Ho, Y.S., Yoon, K.J.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10081–10090 (2019)
17. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Advances in Neural Information Processing Systems. vol. 37 (2024)
18. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3D: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9043–9053 (2023)