PRED: PREDICTING LATENT IMAGE FEATURES FOR HIGH-FREQUENCY RGB-EVENT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

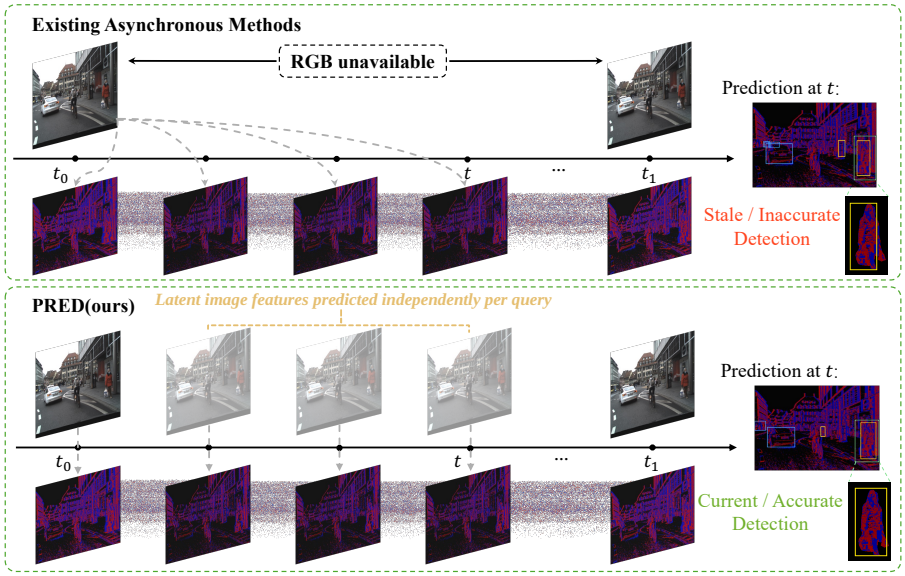


Figure 1: Interframe detection with stale and predicted image features. PRED uses events to predict the query-time image representation from cached RGB features before fusion and detection.

ABSTRACT

RGB cameras provide rich semantic information, but increasing their frame rate incurs substantial sensing, transmission, and computational costs. Event cameras capture brightness changes with microsecond-level temporal resolution, preserving high-frequency motion while providing limited appearance cues. Their complementary properties make RGB–event fusion promising for high-frequency detection. However, existing interframe detectors typically fuse current events with stale image features from the previous RGB frame. This temporal mismatch prevents the fused representation from fully reflecting the current scene and limits detection accuracy between RGB arrivals. To address this limitation, we introduce Predictive **RGB–Event Detection (PRED)**, a query-driven, frame-rate-decoupled framework that predicts the current image representation from the previous RGB frame and subsequent events, thereby reducing the temporal mismatch before fusion. To provide reliable supervision for this cross-time prediction, we propose **two-stage cross-time distillation**: a teacher is first trained independently with access to the current RGB frame and then frozen to provide a stable latent target for PRED. To fill the gap in high-frequency, temporally aligned RGB–event benchmarks, we introduce the **High-Frequency Event Dataset (HiFE)**, constructed through high-fidelity offline rendering and event simulation to provide precise 500-Hz annotations for millisecond-scale detection evaluation. Experiments on real-world and synthetic event-camera detection benchmarks show that PRED consistently outperforms prior methods and maintains strong detection performance across varying query frequencies. Notably, in end-to-end wall-clock replay, PRED nearly keeps pace with a 250-Hz offered query stream and delivers predictions at over $5\times$ the rate of the prior competitive asynchronous detectors.

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.

1 INTRODUCTION

Event cameras asynchronously capture per-pixel brightness changes. Their microsecond temporal resolution and high dynamic range capture fast motion under challenging illumination with little redundant data (Gallego et al., 2022). Because they describe changes rather than complete scene appearance, events provide limited cues for semantic recognition. RGB cameras offer rich color and texture, yet higher frame rates demand greater sensing, transmission, and computational resources. Combining RGB semantics with event dynamics therefore supports accurate and responsive perception (Tulyakov et al., 2021; Wan et al., 2022; Jiang et al., 2024; Qu et al., 2024).

RGB–event detectors exploit this complementarity through recurrent, hierarchical, and attention-based fusion (Zhou et al., 2023; Liu et al., 2024; Cao et al., 2024; Li et al., 2024a). Extending detection to times between RGB arrivals exploits the finer temporal resolution of events (Li et al., 2023; Gehrig & Scaramuzza, 2024; Lu et al., 2025; Li et al., 2025). At these query times, however, events extend to the query while the latest RGB observation comes from an earlier frame. Reusing its cached features preserves this temporal mismatch (Fig. 1), motivating an update of the image representation before fusion.

We introduce **PRED**, a query-driven, frame-rate-decoupled framework that predicts the unavailable image representation from the cached RGB feature and events observed up to the query. Current RGB observations available during training provide a reference for learning this missing representation. We exploit this information through **two-stage cross-time distillation**: an independently trained current-image teacher provides fixed image and fusion targets for a student receiving the previous RGB image and intervening events. This supervision teaches the student to recover current appearance from the observations available between RGB arrivals. At inference, the image encoder runs once per RGB arrival, while a compact event-conditioned path predicts features and detections between frames. The same predictor and cache support two operating points: lightweight gated fusion in PRED for low latency, and frequency-decomposed fusion in PRED-FD for higher accuracy.

Evaluating these predictions requires ground truth at each query timestamp, yet the DSEC interframe protocol relies on interpolated bounding boxes. We therefore introduce the **High-Frequency Event Dataset (HiFE)**, rendered at 500 Hz using Unreal Engine 5 Movie Render Queue. HiFE generates event streams with ESIM (Rebecq et al., 2018) and provides native bounding boxes, instance masks, and depth for interpolation-free evaluation.

On the complete DSEC-Det test set, PRED-FD reaches 69.7 mAP, improving over FlexEvent, the strongest prior RGB–event method in our comparison, by 12.3 points and over frame-only YOLO26x by 6.6 points. PRED-FD averages 74.0 mAP across seven DSEC-Det interframe offsets and 33.9 mAP on HiFE, exceeding DAGr by approximately 21 points on the latter benchmark. In separate end-to-end wall-clock replay, PRED at 320×224 serves 98.1% of a 250-Hz offered request stream, demonstrating its capacity for high-rate inference.

Our contributions are summarized as follows:

- We introduce PRED, which advances cached image features to the query time using events and decouples image encoding from high-rate detection.
- We propose two-stage cross-time distillation with an independently trained, frozen teacher supervising current-time image and fusion features.
- We introduce HiFE, a UE5–ESIM benchmark with native 500-Hz annotations for evaluating detection between RGB arrivals.

2 RELATED WORK

2.1 EVENT-BASED AND RGB–EVENT DETECTION

Event-based detection exploits the temporal structure of brightness changes through two broad representation strategies. Sparse methods operate on event streams with graph or spiking networks (Schaefer et al., 2022; Zubić et al., 2023; Cordone et al., 2022), whereas dense methods aggregate events into tensors for convolutional, recurrent, and Transformer detectors (Perot et al., 2020; Ham-

aguchi et al., 2023; Gehrig & Scaramuzza, 2023; Peng et al., 2024; Wu et al., 2024; Torbunov et al., 2025). Both exploit motion information, but the underlying observations provide limited cues about static appearance.

RGB–event fusion addresses this sensing limitation by combining image semantics with event dynamics through recurrent, hierarchical, and attention-based interactions (Zhou et al., 2023; Liu et al., 2024; Cao et al., 2024; Li et al., 2024a). Using this complementarity between RGB arrivals also requires fusion across different observation times. SODFormer's asynchronous attention-based fusion (Li et al., 2023) and the cached image branches of DAGr and FlexEvent (Gehrig & Scaramuzza, 2024; Lu et al., 2025) allow event observations to support detections without waiting for another RGB frame. ACGR further aligns frame and event graphs into a unified representation (Li et al., 2025). In this study, PRED predicts query-time appearance from cached RGB features and intervening events before fusion. This lets the detector use an appearance estimate that reflects observed changes between frames while retaining the efficiency of running the image encoder only at RGB arrivals.

2.2 Temporal Feature Prediction

Temporal feature prediction extends visual representations beyond observed frames for downstream perception, as demonstrated by DINO-Foresight's prediction of frozen vision-model features from video context (Karypidis et al., 2025). RGB–event perception provides additional evidence for this inference: although the query-time image is unavailable, events record changes up to that time and thus help estimate its representation. This evidence can guide image-space prediction from an earlier frame (Zhu et al., 2024), or update task features through deformable alignment learned from temporally shifted pairs in FAOD (Zhang et al., 2024) and event-aligned image features for keypoint detection and tracking in FF-KDT (Wang et al., 2024b). Unlike image-space prediction, PRED defines the missing image representation at the detection time as an explicit prediction target, inferred from cached RGB features and intervening events. Predicting detector features directly avoids image synthesis and re-encoding, while a current-image teacher provides a detection-specific reference for learning this estimate.

2.3 Knowledge Distillation

Knowledge distillation transfers predictions or intermediate representations from a teacher to a student, providing supervision beyond task labels (Hinton et al., 2015; Yang et al., 2022; Wang et al., 2024a). In event vision, cross-modal distillation uses the appearance and semantic information learned from RGB images to guide representations derived from sparse brightness changes (Wang et al., 2021; Li et al., 2024b; Bartolomei et al., 2025). Beyond cross-modal transfer, PRED uses cross-time distillation to bridge a temporal gap in RGB availability: the student observes an earlier image, while the teacher has access to the current RGB frame. This temporal supervision guides both the recovery of current-image semantics and their fusion with events, giving the student an explicit reference for feature prediction beyond detection labels.

3 Method

3.1 Preliminaries

Event Representation. An event $e_k = (x_k, y_k, t_k, p_k)$ records a threshold-crossing change in logarithmic intensity at pixel (x_k, y_k) and time t_k, with polarity $p_k \in \{-1, +1\}$. Following RVT (Gehrig & Scaramuzza, 2023), we accumulate events between the previous RGB timestamp t_0 and query time t into $B = 5$ temporal bins per polarity. The resulting count tensor $E_{t_0 \to t} \in \mathbb{R}^{2B \times H \times W}$ is the input to the event encoders.

EDI Motivation. The Event-based Double Integral (EDI) model relates a blurred observation B to a sharp reference intensity $I(0)$ and the events during exposure T (Pan et al., 2019):

$$B = I(0) \cdot \frac{1}{T} \int_0^T \exp\left(\int_0^t \epsilon(s)\, \mathrm{d}s \right) \mathrm{d}t, \tag{1}$$

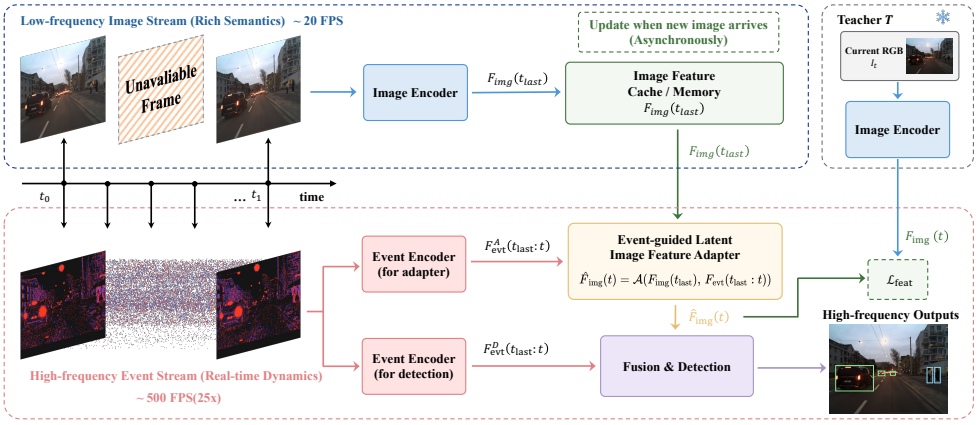


Figure 2: Overview of PRED. The low-frequency image stream refreshes a semantic feature cache only when a new RGB frame arrives. At each event query, compact event encoders support latent image feature prediction and detection, enabling high-frequency outputs without rerunning the image encoder. During training, feature prediction is supervised by a frozen current-image teacher, while query-time inference uses only the student pathway.

where $\epsilon(s)$ is the signed, contrast-scaled event signal and pixel coordinates are omitted. The inner integral accumulates event-induced intensity changes, while the outer integral averages the corresponding latent images over the exposure. The accumulated event signal describes how appearance evolves from the image reference. PRED learns this temporal evolution in detection feature space, using the image as an appearance anchor without assuming that the image-space equation holds for nonlinear features.

Problem Formulation. Let t_0 and t_1 denote the timestamps of two consecutive RGB frames, and let $t_0 < t < t_1$ be the *query time* at which a detection is requested. Without access to the current image I_t, the detector must infer Y_t from the previous image I_{t_0} and intervening events $E_{t_0 \to t}$. Cached image features $F_{\text{img}}(t_0) = \text{Enc}_I(I_{t_0})$ provide an appearance anchor, but remain tied to the scene at t_0; fusing them unchanged with current event features therefore preserves a temporal mismatch as the scene evolves. PRED addresses this mismatch by using the intervening events to predict the image representation at t from the cached features. This predicted representation is then fused with current event features to produce Y_t.

3.2 Latent Image Feature Prediction and Fusion

Decoupled feature streams. Events serve two complementary roles: predicting current appearance and providing motion cues for detection. As depicted in Fig. 2, PRED assigns these roles to separate compact encoders, Enc_A and Enc_E, while evaluating the image encoder Enc_I only at RGB arrivals:

$$\begin{aligned} F_{\text{img}}(t_0) &= \text{Enc}_I(I_{t_0}), \\ F^A_{\text{evt}}(t) &= \text{Enc}_A(E_{t_0 \to t}), \\ F^D_{\text{evt}}(t) &= \text{Enc}_E(E_{t_0 \to t}), \end{aligned} \tag{2}$$

We choose RF-DETR-L's DINOv2-based image backbone (Robinson et al., 2026) for its transferable pretrained representations. Caching its image pyramid $F_{\text{img}}(t_0)$ until the next RGB arrival amortizes image encoding across queries. The event encoders, detection neck, and head use compact YOLO26n components (Jocher et al., 2026) to keep the frequently executed query pathway lightweight and support low-latency inference.

Latent image feature prediction. An event-guided adapter uses the cached appearance and observed changes to predict the image representation directly at the detection time:

$$\widehat{F}_{\text{img}}(t) = \mathcal{A}\big(F_{\text{img}}(t_0), F^A_{\text{evt}}(t)\big)\,, \tag{3}$$

where $\mathcal{A}$ combines the cached appearance semantics at t_0 with the event evidence accumulated through t. At each feature scale, it retains the cached image feature through a residual connection and learns an event-conditioned correction, preserving a direct appearance pathway while modeling temporal change. We detail the adapter architecture in the appendix. The predicted representation replaces the stale image feature as the image-side input to multimodal fusion. Prediction in this space targets the representation needed by the detector, without reconstructing the unavailable RGB image and encoding it again. Each query reads the latest RGB cache independently, and predicted features are not written back to the cache, preventing errors from accumulating across queries.

Multimodal fusion. The predicted image pyramid provides query-time appearance semantics, while $F^D_{\text{evt}}(t)$ supplies complementary event cues. To support different accuracy–latency trade-offs, we instantiate two variants that share the same latent image feature predictor and differ only in multimodal fusion. PRED uses lightweight adaptive gated fusion for efficient inference, whereas PRED-FD employs high–low spatial-frequency fusion for stronger cross-modal interaction and higher accuracy (Sun et al., 2025; Chen et al., 2024). Here, frequency refers to spatial variation within feature maps rather than the temporal query rate.

3.3 Two-Stage Cross-Time Distillation

Detection supervision alone encourages useful features but does not specify the current-image representation toward which stale semantics should move. Although I_t is unavailable at an interframe query, recorded training sequences provide both endpoint images. We therefore learn the prediction at $t = t_1$ through two separate training stages, as illustrated in Fig. 3. Let T and S denote the teacher and student networks, respectively.

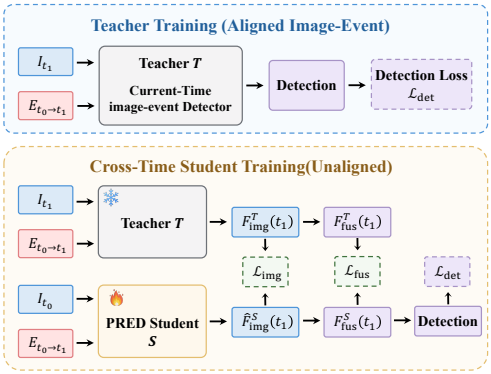


Figure 3: Two-stage cross-time distillation. Stage 1 independently trains a privileged current-image teacher with access to I_{t_1}. Stage 2 freezes the teacher and supervises a student that observes only I_{t_0} and $E_{t_0\to t_1}$.

Current-image teacher training. To establish a reference for the scene at t_1, we first train T as a current-image RGB–event detector. It receives I_{t_1} and $E_{t_0\to t_1}$ and is supervised by Y_{t_1}. Because the image observation and detection target share the same timestamp, the teacher learns image and fusion representations directly from the appearance withheld from the student.

This first stage separates learning a current-image detector from learning to predict its representations across time. Following YOLO26 (Jocher et al., 2026), the objective combines classification, L_1 box regression, and IoU losses under Small-Target-Aware Label Assignment (STAL):

$$\mathcal{L}_{\text{det}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_1 + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}}. \tag{4}$$

Cross-time student training. We freeze T to preserve its current-image representation as a fixed target for temporal prediction. Given only the previous image I_{t_0} and events $E_{t_0\to t_1}$, the student S predicts the image feature $\widehat{F}^S_{\text{img}}(t_1)$ and produces the fused feature $F^S_{\text{fus}}(t_1)$. We directly match these outputs to $F^T_{\text{img}}(t_1)$ and $F^T_{\text{fus}}(t_1)$ from the frozen teacher using

$$\mathcal{L}_{\text{feat}}(F^S, F^T) = \mathcal{L}_{\cos}(F^S, F^T) + \beta\mathcal{L}_{\text{SL1}}(\bar{F}^S, \bar{F}^T), \tag{5}$$

where $\bar{F}^S$ and $\bar{F}^T$ are normalized by the same per-sample scale of F^T, $\mathcal{L}_{\cos} = 1 - \cos(F^S, F^T)$, and $\mathcal{L}_{\text{SL1}}$ denotes Smooth-L1 loss. The cosine term matches feature direction, while the scale-normalized Smooth-L1 term matches feature magnitude. Applying Eq. 5 to the predicted image features and the fused features gives $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{fus}}$, respectively. The complete student objective is

$$\mathcal{L} = \mathcal{L}^S_{\text{det}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{fus}}\mathcal{L}_{\text{fus}}. \tag{6}$$

The image loss directly teaches the adapter to reproduce the representation extracted from I_{t_1}, while the fusion loss aligns the multimodal feature passed to the detection neck with the teacher's current-time representation.

Table 1: Comparison of state-of-the-art detectors on the DSEC-Det dataset. The final row reports the teacher with query-time RGB for reference.

Modality	Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
E	LEOD *(CVPR'24)*	41.1	65.2	43.6	35.1	47.3	73.3
	YOLO26x *(arXiv'26)*	31.5	53.3	32.0	25.4	47.7	67.4
	YOLO26n *(arXiv'26)*	27.4	46.4	28.3	23.7	39.6	32.6
F	RF-DETR-L *(ICLR'26)*	53.8	92.5	51.5	50.4	61.2	76.5
	YOLO26x *(arXiv'26)*	63.1	91.5	69.3	60.0	71.4	83.1
E+F	DAGr-50 *(Nature'24)*	41.9	66.1	44.4	36.3	56.4	78.0
	HDI-Former *(arXiv'24)*	46.7	69.1	–	–	–	–
	FlexEvent *(NeurIPS'25)*	57.4	78.2	66.6	51.7	64.9	83.7
	PRED	68.7	89.7	81.3	65.5	77.3	83.9
	PRED-FD	**69.7**	**90.4**	**82.7**	**66.4**	**78.6**	**84.3**
	PRED-FD current-image reference	72.4	90.1	84.8	69.5	80.4	86.6

4 HiFE Benchmark

Existing RGB–event detection datasets annotate objects primarily at RGB timestamps. DSEC-Det's interframe protocol interpolates bounding boxes between these timestamps (Gehrig & Scaramuzza, 2024), enabling comparison across query offsets without directly observing the intermediate ground truth. To address this limitation, we develop a data engine to construct the High-Frequency Event Dataset (HiFE) with native interframe annotations. Using offline rendering with Movie Render Queue in Unreal Engine 5's City Sample environment, the engine decouples temporal sampling from rendering speed to generate images and Car/Pedestrian annotations every 2 ms; ESIM (Rebecq et al., 2018) converts the image sequences into aligned event streams. The detection benchmark contains 80,125 annotated frames from three separately captured trajectories within this environment, assigned to training, validation, and testing. Withholding the query-time image preserves DSEC's stale-image observation setting, while native boxes support direct evaluation at up to 500 Hz. The rendering pipeline also provides camera parameters, depth, optical flow, and segmentation labels, extending its utility beyond detection. We detail data generation, splits, and temporal pairing in the appendix.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on DSEC-Det (Gehrig et al., 2021; Gehrig & Scaramuzza, 2024), which provides spatially aligned events and 20-Hz RGB frames in real driving scenes, following its official splits for the Car and Pedestrian classes. Its labels coincide with RGB timestamps, so interframe evaluation relies on interpolated boxes. We therefore complement DSEC-Det with HiFE, whose native 500-Hz annotations allow direct evaluation at query times while preserving the same stale-RGB input setting. By comparison, DSEC-Detection (Tomy et al., 2022) pairs events with query-time RGB, while DSEC-MOD's official video protocol (Zhou et al., 2023) also uses a subsequent RGB frame. These observation settings differ from ours; we discuss their protocols and report the current-image teacher's DSEC-Detection results in the appendix.

Evaluation protocols. We use COCO mAP as the primary metric, accompanied by AP_{50}, AP_{75}, and $AP_{S/M/L}$ to characterize localization and performance across object scales (Lin et al., 2014). The main comparison uses the complete DSEC-Det test set: for detection at t_1, PRED receives only I_{t_0} and $E_{t_0 \to t_1}$, with I_{t_1} withheld. To assess detection between RGB arrivals, we follow the official DAGr protocol, which interpolates boxes within intervals with consistent tracks and sufficient box overlap. The resulting interframe targets are thus evaluated on filtered test intervals.

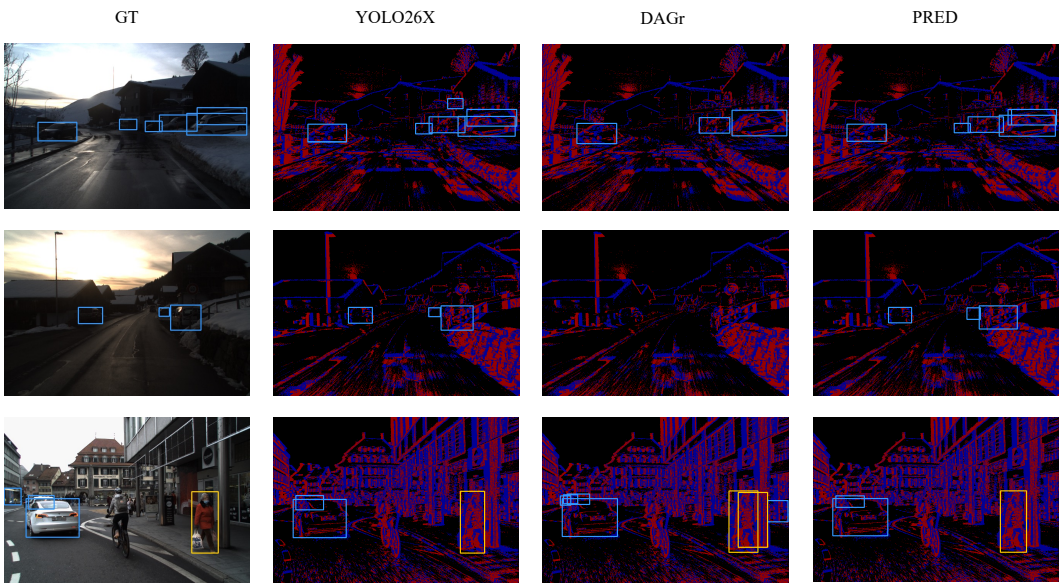


Figure 4: Qualitative comparison on DSEC-Det. From left to right: ground truth, frame-only YOLO26x, DAGr, and PRED.

Table 2: Interframe mAP on DSEC-Det and HiFE across seven nominal query frequencies. FlexEvent results are taken from the original paper.

Dataset	Method	Nominal query frequency (Hz)							Avg.
		20	36	45	60	90	180	500	
DSEC-Det	DAGr	44.1	44.4	44.2	43.9	43.9	44.2	44.0	44.0
	FlexEvent	57.4	60.1	59.5	58.8	56.5	50.9	–	57.2
	PRED-FD	**71.9**	**74.2**	**74.6**	**74.8**	**74.6**	**74.2**	**73.8**	**74.0**
HiFE	DAGr	9.8	12.0	12.7	13.4	13.7	14.2	14.4	12.9
	PRED-FD	**29.5**	**33.0**	**33.9**	**34.7**	**35.2**	**35.7**	**35.5**	**33.9**

Baselines. Table 1 presents results for PRED alongside event-only, frame-only, and RGB–event detectors. We train the YOLO26 and RF-DETR baselines on DSEC-Det using their respective recipes (Jocher et al., 2026; Robinson et al., 2026) and reproduce DAGr-50 with its official implementation (Gehrig & Scaramuzza, 2024). For LEOD, HDI-Former, and FlexEvent, we use the DSEC-Det results reported in the FlexEvent paper (Lu et al., 2025).

Implementation details. We train separate teacher–student pairs on DSEC-Det and HiFE using AdamW with a learning rate of 5×10^{-5} and batch size 64. Training images are resized to a shorter side of 512 and randomly cropped to 512×512. Each stage runs for 40 epochs on DSEC-Det and two epochs on HiFE. We provide further architecture and training details in the appendix and specify runtime evaluation settings in Section 5.3.

5.2 Results

Detection with Stale RGB Inputs. As summarized in Table 1, PRED-FD achieves 69.7 mAP on the complete DSEC-Det test set, surpassing all external baselines, while lightweight PRED retains comparable accuracy at 68.7 mAP. The teacher provides a reference for the performance attainable with query-time RGB: PRED-FD remains only 2.7 points below its 72.4 mAP despite predicting the image representation from stale RGB and events. Figure 4 illustrates the corresponding detections on representative driving scenes.

High-Frequency Generalization. The results in Table 2 show that a model trained on DSEC-Det at approximately 50-ms intervals generalizes to 2–50-ms offsets without fine-tuning, averaging 74.0 mAP. Frequency labels denote reciprocal query offsets. At each offset, we retain the same RGB anchor and encode the intervening events as during training, without providing the query offset as an explicit input. This evaluation therefore tests whether a predictor learned from complete frame intervals remains useful when only part of an interval has been observed.

HiFE complements this generalization test with native annotations at each query time. Its training pairs cover 2–50 ms, while evaluation retains intervals in which objects enter or leave the scene, without the track-consistency filtering needed for DSEC's interpolated targets. When trained on HiFE, PRED-FD averages 33.9 mAP, outperforming DAGr across the 20–500 Hz range, with gains extending to small and medium objects, as detailed in the appendix.

5.3 Accuracy–Latency Trade-off

High-frequency detection requires timely predictions under a continuous event stream. We evaluate this requirement at periodic query times, with output requests scheduled independently of individual event arrivals. For DAGr, the millisecond-scale GPU runtimes reported for individual event insertion (Gehrig & Scaramuzza, 2024) make serial processing costly at high event rates, motivating its evaluation through batched inference. Model-only latency isolates the computational cost of a query, while wall-clock replay tests whether the complete pipeline sustains the requested output rate.

Model-only inference latency. Table 3 reports batch-one model latency. On a Tesla T4, PRED and PRED-FD use TensorRT FP16 at 320×224. For DAGr, the image backbone is accelerated with TensorRT, while its GNN module remains in PyTorch FP32 because it cannot be integrated into TensorRT; its input resolution is 320×215. PRED achieves a cached-query latency of 2.00 ms and a maximum model-only query rate of 498.9 Hz, while PRED-FD records 3.30 ms and 302.9 Hz. Both operating points achieve substantially lower model-only latency than the compared deployments in Table 3.

Table 3: Latency evaluation of different detectors. Parentheses indicate cached-query latency. * YOLO26x reports isolated throughput but remains limited to 20-Hz RGB outputs. † FlexEvent is reported on an NVIDIA A5000.

Method	Size (M)	Latency (ms)	Rate (Hz)
YOLO26x	55.7	4.11	243.6*
DAGr-50	34.6	82.33 (75.34)	13.3
FlexEvent†	45.4	14.27	70.1
PRED	36.7	**3.58 (2.00)**	**498.9**
PRED-FD	37.0	4.71 (3.30)	302.9

Wall-clock replay. To assess whether low query latency translates into sustained inference, we replay 20 s of DSEC data on an RTX 4090, with RGB fixed at 20 Hz and query requests issued at 20–500 Hz. Independent image-refresh and query workers use latest-only scheduling: while a query is being processed, a newer request replaces the pending one. We measure *new-prediction coverage*, the fraction of issued requests receiving a newly computed output, and end-to-end latency from request arrival to detection output.

Figure 5 shows that PRED nearly keeps pace with a 250-Hz request stream, achieving 98.1% coverage and 245.4 new predictions/s. At 500 Hz, its output rate continues to rise to 346.8 predictions/s, although coverage falls to 69.4% as requests outpace processing. The tested DAGr deployment reaches its throughput limit at lower query rates, while frame-only YOLO26x remains bounded by the 20-Hz RGB input.

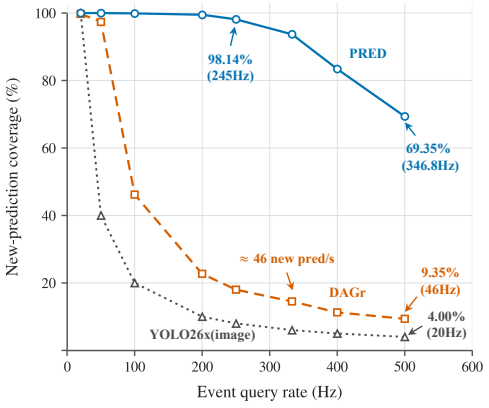


Figure 5: Wall-clock scalability under varied query rates. Coverage denotes requests receiving new predictions; RGB-only YOLO26x is limited to 20-Hz updates.

Table 4: Feature prediction under EMA and frozen supervision.

Analysis	Condition	EMA	Frozen
Detection mAP ↑	Stale	51.5	56.0
	Predicted	57.1	**69.7**
	Current	58.3	**70.7**
Fixed-reference distance ↓	First	0.2139	–
	Last	0.4311	**0.0488**
Online–EMA distance ↓	First	9.95×10^{-4}	–
	Last	4.44×10^{-6}	–

Table 5: Component ablation on DSEC-Det. Shared denotes a single event encoder; Dedicated adds an encoder for temporal adaptation. Interframe mAP averages seven offsets.

Event Encoder	Adapter	Target	mAP	Interframe mAP
Shared	×	×	63.7	70.7
Dedicated	✓	×	63.4	71.0
Dedicated	✓	EMA	57.1	61.9
Shared	✓	Frozen	65.2	72.6
Dedicated	✓	Frozen	**69.7**	**74.0**

5.4 Ablation Studies

Feature prediction and target stability. To isolate the contribution of image feature prediction, we substitute stale, predicted, and current-image features within each trained student while holding its event pathway and detector fixed. Table 4 shows that predicted features recover approximately 93% of the stale-to-current detection gap under frozen supervision. The current-image result is a within-student reference, distinct from the independently trained teacher in Table 1. This substitution tests whether predicted appearance retains information usable by the same downstream detector. Complementary feature analyses show smaller image and fusion distances to the frozen teacher within foreground boxes, while spatial error maps visualize the reduction around annotated objects. Frozen supervision also outperforms EMA when current-image features bypass the adapter, suggesting that the difference extends beyond temporal prediction error to the learned detection representation. The distances in Table 4 offer a complementary view: during EMA training, online predictions increasingly agree with their averaged counterparts while moving farther from the fixed current-image reference. Because EMA targets evolve with the student, lower online–EMA distance alone does not establish recovery of the independently learned current-image representation, supporting the use of a stable teacher target. We detail feature alignment, spatial error maps, and training trajectories in the appendix.

Components and supervision. As shown in Table 5, adding the prediction pathway brings little improvement under detection supervision alone. With the student architecture held fixed, frozen image and fusion targets raise complete-test mAP from 63.4 to 69.7 and interframe mAP from 71.0 to 74.0, highlighting the benefit of explicitly supervising query-time representations. These targets constrain both the representation produced by the adapter and the representation consumed by the detection neck, coupling appearance prediction to its subsequent use in multimodal detection. Learning these representations from events requires features that capture changes in cached appearance while also supplying cues for detection. We therefore examine whether the two roles benefit from separate encoders under the same frozen supervision. Compared with a shared encoder, dedicated encoders improve complete-test mAP from 65.2 to 69.7 and interframe mAP from 72.6 to 74.0, supporting specialization of the event pathways for temporal adaptation and detection. The advantage persists at every evaluated query offset, including shorter intervals unseen during DSEC-Det training. Thus, the benefit of separating the pathways extends from the supervised frame endpoint to queries based on partial event evidence. We detail the individual supervision terms and results at each query offset in the appendix.

6 Conclusion

We introduced PRED for frame-rate-decoupled RGB–event detection, using events to predict the image representation between RGB arrivals. Two-stage cross-time distillation provides stable current-image and fusion targets, allowing the student to learn this prediction from temporally separated observations. Feature substitution and representation analyses support the recovery of current-image detection performance, while DSEC-Det and native 500-Hz HiFE evaluations demonstrate accuracy across query intervals. Combined with wall-clock replay, these results support event-guided feature prediction for high-rate perception with low-rate RGB input.

AI USE STATEMENT

The original method and datasets were developed without generative AI. During manuscript revision, AI tools assisted with formatting checks, language editing, organization, discussions of results, and code revisions. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 3 describes the event representation, model formulation, and training objectives. The appendix provides architecture and optimization details, temporal sampling and evaluation protocols, and the construction and splits of HiFE. It also documents the feature analyses and wall-clock replay procedure, including scheduling, latency accounting, and the full offered-load sweep.

REFERENCES

Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth AnyEvent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19669–19678, 2025. doi: 10.1109/ICCV51701.2025.01829. URL https://doi.org/10.1109/ICCV51701.2025.01829.

Hu Cao, Zehua Zhang, Yan Xia, Xinyi Li, Jiahao Xia, Guang Chen, and Alois Knoll. Embracing events and frames with hierarchical feature refinement network for object detection. In *Computer Vision – ECCV 2024*, pp. 161–177. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72907-2_10. URL https://doi.org/10.1007/978-3-031-72907-2_10.

Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10763–10780, 2024. doi: 10.1109/TPAMI.2024.3449959. URL https://doi.org/10.1109/TPAMI.2024.3449959.

Loic Cordone, Benoit Miramond, and Philippe Thierion. Object detection with spiking neural networks on automotive event data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022. doi: 10.1109/IJCNN55064.2022.9892618. URL https://doi.org/10.1109/IJCNN55064.2022.9892618.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413. URL https://doi.org/10.1109/TPAMI.2020.3008413.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. doi: 10.1038/s41586-024-07409-w. URL https://doi.org/10.1038/s41586-024-07409-w.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13884–13893, 2023. doi: 10.1109/CVPR52729.2023.01334. URL https://doi.org/10.1109/CVPR52729.2023.01334.

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. doi: 10.1109/LRA.2021.3068942. URL https://doi.org/10.1109/LRA.2021.3068942.

Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22867–22876, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Hamaguchi_Hierarchical_Neural_Memory_Network_for_Low_Latency_Event_Processing_CVPR_2023_paper.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Jianping Jiang, Xinyu Zhou, Bingxuan Wang, Xiaoming Deng, Chao Xu, and Boxin Shi. Complementing event streams and RGB frames for hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24944–24954, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jiang_Complementing_Event_Streams_and_RGB_Frames_for_Hand_Mesh_Reconstruction_CVPR_2024_paper.html.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026. doi: 10.48550/arXiv.2606.03748. URL https://arxiv.org/abs/2606.03748.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. DINO-foresight: Looking into the future with DINO. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=gimtybo07H.

Dianze Li, Jianing Li, and Yonghong Tian. SODFormer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):14020–14037, 2023. doi: 10.1109/TPAMI.2023.3298925. URL https://doi.org/10.1109/TPAMI.2023.3298925.

Dianze Li, Jianing Li, Xu Liu, Zhaokun Zhou, Xiaopeng Fan, and Yonghong Tian. HDI-Former: Hybrid dynamic interaction ANN-SNN transformer for object detection using frames and events. *arXiv preprint arXiv:2411.18658*, 2024a. doi: 10.48550/arXiv.2411.18658. URL https://arxiv.org/abs/2411.18658.

Dianze Li, Jianing Li, Xu Liu, Xiaopeng Fan, and Yonghong Tian. Asynchronous collaborative graph representation for frames and events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1655–1666, 2025. doi: 10.1109/CVPR52734.2025.00162. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_Asynchronous_Collaborative_Graph_Representation_for_Frames_and_Events_CVPR_2025_paper.html.

Lei Li, Alexander Linger, Mario Millhäusler, Vagia Tsiminaki, Yuanyou Li, and Dengxin Dai. Object-centric cross-modal feature distillation for event-based object detection. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 15440–15447, 2024b. doi: 10.1109/ICRA57147.2024.10610325. URL https://doi.org/10.1109/ICRA57147.2024.10610325.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Zhanwen Liu, Nan Yang, Yang Wang, Yuke Li, Xiangmo Zhao, and Fei-Yue Wang. Enhancing traffic object detection in variable illumination with RGB-event fusion. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20335–20350, 2024. doi: 10.1109/TITS.2024.3456108. URL https://doi.org/10.1109/TITS.2024.3456108.

Dongyue Lu, Lingdong Kong, Gim Hee Lee, Camille Simon Chane, and Wei Ooi. FlexEvent: Towards flexible event-frame object detection at varying operational frequencies. In *Advances in Neural Information Processing Systems*, volume 38, pp. 89009–89044. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8064e4ebbcbe594628887b420956d8c3-Abstract-Conference.html.

Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6820–6829, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Pan_Bringing_a_Blurry_Frame_Alive_at_High_Frame-Rate_With_an_CVPR_2019_paper.html.

Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Scene adaptive sparse transformer for event-based object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16794–16804, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Peng_Scene_Adaptive_Sparse_Transformer_for_Event-based_Object_Detection_CVPR_2024_paper.html.

Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16639–16652. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c213877427b46fa96cff6c39e837ccee-Abstract.html.

Delin Qu, Chi Yan, Dong Wang, Jie Yin, Qizhi Chen, Dan Xu, Yiting Zhang, Bin Zhao, and Xuelong Li. Implicit event-RGBD neural SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19584–19594, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Qu_Implicit_Event-RGBD_Neural_SLAM_CVPR_2024_paper.html.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. In *Proceedings of the 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 969–982. PMLR, 2018. URL https://proceedings.mlr.press/v87/rebecq18a.html.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: Real-time detection transformer. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2511.09554.

Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. AEGNN: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12371–12381, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Schaefer_AEGNN_Asynchronous_Event-Based_Graph_Neural_Networks_CVPR_2022_paper.html.

Pihai Sun, Junjun Jiang, Yuanqi Yao, Youyu Chen, Wenbo Zhao, Kui Jiang, and Xianming Liu. 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. URL https://arxiv.org/abs/2503.19739.

Abhishek Tomy, Anshul Paigwar, Khushdeep S. Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and RGB camera for robust object detection in adverse conditions. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 933–939, 2022.

Dmitrii Torbunov, Yihui Ren, Animesh Ghose, Odera Dim, and Yonggang Cui. EvRT-DETR: Latent space adaptation of image detectors for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9812–9821, 2025. doi: 10.1109/ICCV51701.2025.00915. URL https://doi.org/10.1109/ICCV51701.2025.00915.

Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16155–16164, 2021. doi: 10.1109/CVPR46437.2021.01589. URL https://doi.org/10.1109/CVPR46437.2021.01589.

Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE Transactions on Image Processing*, 31:7237–7251, 2022. doi: 10.1109/TIP.2022.3220938. URL https://doi.org/10.1109/TIP.2022.3220938.

Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. CrossKD: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16520–16530, 2024a. doi: 10.1109/CVPR52733.2024.01563. URL https://doi.org/10.1109/CVPR52733.2024.01563.

Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, pp. 608–619, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_EvDistill_Asynchronous_Events_To_End-Task_Learning_via_Bidirectional_Reconstruction-Guided_Cross-Modal_CVPR_2021_paper.html.

Xiangyuan Wang, Huai Yu, Lei Yu, Wen Yang, and Gui-Song Xia. Toward robust keypoint detection and tracking: A fusion approach with event-aligned image features. *IEEE Robotics and Automation Letters*, 9(9):8059–8066, 2024b. doi: 10.1109/LRA.2024.3440099. URL https://doi.org/10.1109/LRA.2024.3440099.

Ziyi Wu, Mathias Gehrig, Qing Lyu, Xudong Liu, and Igor Gilitschenski. LEOD: Label-efficient object detection for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16933–16942, 2024. doi: 10.1109/CVPR52733.2024.01602. URL https://doi.org/10.1109/CVPR52733.2024.01602.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4633–4642, 2022. doi: 10.1109/CVPR52688.2022.00460. URL https://doi.org/10.1109/CVPR52688.2022.00460.

Haitian Zhang, Xiangyuan Wang, Chang Xu, Xinya Wang, Fang Xu, Huai Yu, Lei Yu, and Wen Yang. Frequency-adaptive low-latency object detection using events and frames. *arXiv preprint arXiv:2412.04149*, 2024. URL https://arxiv.org/abs/2412.04149.

Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. RGB-event fusion for moving object detection in autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 7808–7815, 2023. doi: 10.1109/ICRA48891.2023.10161563. URL https://doi.org/10.1109/ICRA48891.2023.10161563.

Juanjuan Zhu, Zhexiong Wan, and Yuchao Dai. Video frame prediction from a single image and events. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7748–7756, 2024. doi: 10.1609/aaai.v38i7.28609. URL https://ojs.aaai.org/index.php/AAAI/article/view/28609.

Nikola Zubić, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12846–12856, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zubic_From_Chaos_Comes_Order_Ordering_Event_Representations_for_Object_Recognition_ICCV_2023_paper.html.

These supplementary materials detail the model architecture, training configuration, benchmark protocols, and construction of HiFE. They also provide extended ablations, feature-level analyses, the full wall-clock evaluation protocol and results, and additional HiFE visualizations.

A Model Details

A.1 Frame-Rate-Decoupled Execution

Let t_0 and t_1 denote consecutive RGB timestamps, with $t_0 < t < t_1$ an interframe query time. Deployment separates the detector into an *image engine* and a *query engine*. The image engine contains the high-capacity image encoder and executes only upon RGB arrival, producing the cached feature $F_{\text{img}}(t_0)$. This cache remains read-only until the next RGB frame.

Table 6: Model size of the two operating points.

Method	Size (M)		
	Image	Query	Total
PRED	25.5	11.2	36.7
PRED-FD	25.5	11.5	37.0

The query engine comprises the detection- and adapter-oriented event encoders, the temporal adapter, multimodal fusion, the YOLO neck, the detection head, and end-to-end postprocessing. Each query combines $E_{t_0 \to t}$ with the latest RGB cache. Because the predicted feature is not written back, successive queries are anchored directly to the same cache and cannot accumulate recursive prediction error. PRED and PRED-FD share this execution model and image engine, differing only in their query-time fusion blocks. Table 6 summarizes their parameter allocation.

A.2 Architecture Details

We retain the feature extraction configuration of RF-DETR-L (Robinson et al., 2026) (transformer blocks 3, 6, 9, and 12) and map its outputs to a three-scale pyramid matching the spatial resolutions and channel dimensions of the YOLO26 event encoders. At each scale, a separate temporal adapter processes concatenated cached image and adapter-oriented event features using a 1×1 projection, a residual block with two 3×3 convolutions and SiLU activations, and a zero-initialized 1×1 output layer to predict a residual update to the cached image feature. The updated image features are then fused with the detection-oriented event features using the fusion operators described below.

A.3 Multimodal Fusion

To span the accuracy–latency trade-off, we instantiate two fusion variants: latency-oriented PRED and accuracy-oriented Feature-Decoupled PRED (PRED-FD). Both retain the same event-guided prediction mechanism.

PRED uses lightweight adaptive gated fusion. Given the predicted image feature $\widehat{F}_{\text{img}}(t)$ and detection-oriented event feature $F^D_{\text{evt}}(t)$, a joint gating network estimates channel- and spatially varying modality weights. A softmax normalization enforces $\alpha^{\text{img}} + \alpha^{\text{evt}} = 1$ at every channel and spatial location, allowing the fused representation to balance appearance semantics and event-derived motion according to local content:

$$F_{\text{fus}}(t) = \alpha^{\text{img}} \odot \widehat{F}_{\text{img}}(t) + \alpha^{\text{evt}} \odot F^D_{\text{evt}}(t). \tag{7}$$

PRED-FD instead decomposes each modality into low- and high-frequency components using a Gaussian–Laplacian representation. The low-frequency branch is anchored by the image feature and selectively incorporates event information; the high-frequency branch reverses these roles to emphasize the event feature while retaining complementary image content. With spatial gates M^L and M^H, the two branches and their aggregate are

$$\begin{aligned} B^L(t) &= \widehat{F}^L_{\text{img}}(t) + M^L \odot F^{D,L}_{\text{evt}}(t), \\ B^H(t) &= F^{D,H}_{\text{evt}}(t) + M^H \odot \widehat{F}^H_{\text{img}}(t), \\ F^{\text{FD}}_{\text{fus}}(t) &= B^L(t) + B^H(t). \end{aligned} \tag{8}$$

This asymmetric anchoring preserves image-dominant low-frequency structure while prioritizing event-derived high-frequency detail.

B Training Details

B.1 Training Configuration

For DSEC-Det, the current-image teacher and cross-time student share the optimization and augmentation settings in Table 7. Both stages use the YOLO26 detection objective (Jocher et al., 2026), with the student additionally using the distillation terms in Eq. 6.

Training on DSEC-Det uses adjacent 20-Hz RGB frames, restricting temporal supervision to intervals of approximately 50 ms. The timestamps of retained events are normalized to $[0, 1]$ before binning, and the model receives no explicit duration or frequency input.

Table 7: DSEC-Det training configuration. Distillation weights apply only to the student.

Parameter	Value
Optimizer	AdamW
Base learning rate	5×10^{-5}
Training epochs per stage	40
Training GPUs	$4\times$ NVIDIA RTX 4090
Training batch size per GPU	16
Random crop size	512×512
Horizontal flip probability	0.5
Classification weight λ_{cls}	0.561
L_1 box regression weight λ_1	9.039
IoU weight λ_{IoU}	5.628
Image distillation weight λ_{img}	500
Fusion distillation weight λ_{fus}	6000

Distillation combines channel-wise cosine distance with Smooth-L1 distance normalized by the frozen teacher's per-sample feature scale, averaged uniformly across the three feature strides. Teacher features remain detached, and each student is paired with a teacher that uses the same fusion operator.

Frozen and EMA supervision use the same first-stage initialization and compute teacher targets from current RGB and the same event interval, bypassing the temporal adapter. For EMA supervision, the teacher is initialized as a copy of the student and updated after optimizer step k by $\bar{\theta}_k = d_k\bar{\theta}_{k-1} + (1 - d_k)\theta_k$, where $d_k = 0.993(1 - e^{-k/100})$; detection gradients update only the student. Student weight averaging for evaluation is also used under frozen supervision and does not alter its distillation target.

C Dataset

C.1 Evaluation Protocols

RGB–event benchmarks differ fundamentally in the image observation available at the detection time. Consider consecutive RGB timestamps $t_0 < t_1$ and a prediction evaluated at t_1. Under a current-image protocol, the detector receives the corresponding image I_{t_1}; under a stale-image protocol, I_{t_1} is withheld and only the preceding image I_{t_0} is available. The relevant distinction is therefore whether the RGB input is current or temporally stale, as summarized in Table 8. This distinction concerns the available RGB observation, independently of whether computation is triggered by individual events or by queries. Applying a stale-image protocol to PKU-DAVIS-SOD would require a separate evaluation with matched baselines; its standard current-image results are not directly comparable.

Table 8: Comparison of RGB–event detection benchmarks and evaluation protocols. Protocol indicates the available RGB observation.

Dataset	Type	Res.	Cls.	Frames (k)	Boxes (M)	BBox (Hz)	Protocol
PKU-DAVIS-SOD	Real	346×260	3	276.0	1.08	25	Current
DSEC-Detection	Real	640×480	3	52.7	0.17	20	Current
DSEC-MOD	Real	640×480	1	13.3	0.05	20	Current/future
DSEC-Det	Real	640×480	8	78.3	0.39	20	Stale
HiFE (ours)	Synthetic	1920×1080	2	80.1	0.94	500	Stale

For PRED and PRED-FD accuracy evaluation, the shorter image side is resized to 512 pixels and both dimensions are rounded to multiples of 32, without further cropping or padding. DSEC-Det inputs are 768×512; predicted boxes are mapped to the official 320×215 evaluation coordinates. Accuracy evaluation uses 8 samples per GPU. Latency evaluation instead uses 320×224, a nearby size compatible with the network strides for comparison with DAGr's 320×215 input. The reported accuracy and latency therefore use different input resolutions.

C.2 DSEC-Detection Protocol

DSEC-Detection (Tomy et al., 2022) is a three-class driving benchmark with 41 sequences and approximately 52.7k RGB frames, annotated for Car, Pedestrian, and Large Vehicle. We adopt the CAFR split (Cao et al., 2024), with 33/3/5 sequences for training/validation/testing.

After CAFR's annotation filtering, the training, validation, and test splits contain 40,002, 3,282, and 4,774 annotated frames, respectively. DSEC-Detection evaluates paired RGB–event inputs at RGB timestamps. A prediction at t_1 uses the current image I_{t_1} and events from the preceding 50 ms, $E_{t_0\to t_1}$ with $t_0 = t_1 - 50\,\mathrm{ms}$, and is evaluated against bounding boxes at t_1. Because the input image directly corresponds to the target scene state, this protocol differs from the stale-image setting addressed by PRED. We therefore train and evaluate only the current-image teacher. Table 9 reports class-wise AP and their mean; the current-image teacher reaches 61.6 mAP and outperforms CAFR and FlexEvent in all three categories.

Table 9: Detection accuracy on DSEC-Detection.

Method	Car	Pedestrian	L-Veh.	mAP
CAFR (Cao et al., 2024)	49.9	25.8	38.2	38.0
FlexEvent (Lu et al., 2025)	59.3	37.4	45.5	47.4
Teacher (ours)	**70.8**	**50.7**	**63.3**	**61.6**

C.3 DSEC-MOD Protocol

DSEC-MOD (Zhou et al., 2023) is a single-class moving-object benchmark comprising 16 driving sequences and 13,314 frames. Its official RENet protocol is distinct from the stale-anchor setting considered here. RENet consumes a three-frame RGB–event clip that includes both the target-time RGB image and a subsequent frame, and predicts a three-frame tubelet. Frame-level evaluation aggregates candidates from overlapping tubelets before soft-NMS; video-level evaluation additionally links tubelets over time. FlexEvent (Lu et al., 2025) adopts this paired video-detection setting and reports training with 11-frame sequences. This protocol exposes current and future RGB information unavailable at a PRED query and applies different temporal postprocessing. Consequently, DSEC-MOD scores are not directly comparable to the DSEC-Det and HiFE results and are included only to clarify the protocol distinction.

C.4 HiFE Benchmark

HiFE provides native ground truth for evaluating detection between RGB observations. The real-world benchmarks in Table 8 provide bounding boxes at 20–25 Hz (Li et al., 2023; Tomy et al., 2022; Zhou et al., 2023; Gehrig & Scaramuzza, 2024), whereas HiFE supplies annotations at 500

Hz. This temporal density allows predictions to be evaluated against the simulated scene state at each query time without interpolating bounding boxes.

Synchronized data generation. We use Unreal Engine 5's Movie Render Queue to generate images and annotations from a shared scene state. Offline rendering allows the sampling cadence to be configured independently of rendering throughput; we use 1920×1080 images at 2-ms intervals. Each timestamp provides RGB, camera parameters, bounding boxes, depth, optical flow, instance masks, and semantic labels. ESIM (Rebecq et al., 2018) converts the rendered images into continuous (x, y, t, p) event streams aligned with these annotations, using positive and negative contrast thresholds of 0.4, a 0.2-ms refractory period, and an 8-ms low-pass time constant. The same pipeline also captures RGB, depth, optical flow, and camera data in additional UE5 environments, extending its use beyond the detection subset evaluated here.

Detection benchmark and training. The detection benchmark comprises 80,125 frames annotated for Car and Pedestrian in the City Sample environment. Table 10 summarizes the training, validation, and test splits, each drawn from a separately captured driving trajectory. Boxes are derived from instance masks; at the native image resolution, we retain boxes with both sides exceeding 20 pixels and a diagonal exceeding 30 pixels, containing at least 100 instance pixels and a mask occupancy of at least 20%.

HiFE training pairs span 1–25 frames (2–50 ms), using native annotations at the target time. We follow the same two-stage procedure as on DSEC-Det, initializing the image encoder from RF-DETR-L and the event encoders and compatible detection layers from YOLO26. Both stages are trained for two epochs, with the student initialized from the teacher. DAGr is trained on HiFE for three epochs with an ImageNet-pretrained ResNet-50 image backbone. Checkpoints are selected by validation mAP of the EMA weights.

Table 10: HiFE detection splits within the City Sample environment. Counts refer to unique frames and their retained boxes before temporal pairing.

Split	Trajectories	RGB frames	Boxes
Training	1	50,000	531,480
Validation	1	10,103	99,029
Test	1	20,022	314,220
Total	3	80,125	944,729

Evaluation protocol. For an integer frame offset Δ, the detector receives anchor image I_{t_0} and events $E_{t_0 \to t_1}$ and is evaluated against the native boxes at $t_1 = t_0 + 2\Delta$ ms. The query-time image I_{t_1} is withheld. Offsets of $\Delta = 1$ and 25 correspond to the nominal 500-Hz and 20-Hz settings, respectively; intermediate settings follow the same construction. Evaluation therefore requires neither box interpolation nor track-consistency filtering and retains intervals in which objects enter or leave the scene. Table 11 reports detection accuracy by object scale at the 2-ms offset.

Table 11: HiFE detection accuracy by object scale at 500 Hz.

Method	mAP	AP_{50}	AP_S	AP_M	AP_L
DAGr	14.41	35.82	0.46	11.14	42.97
PRED-FD	**35.53**	**64.69**	**8.24**	**32.71**	**65.98**

D Additional Quantitative Results

D.1 Detailed Component Ablation

Image distillation supervises the predicted visual representation, while fusion distillation supervises its integration with events for detection. Table 12 expands Table 5 with results at all seven query offsets, their average, and the complete DSEC-Det test result. The student architecture and remaining training settings are fixed when comparing image-only and joint supervision. Image-only supervision raises complete-test mAP from 63.4 to 68.4, while joint supervision further improves it to 69.7.

Fusion targets also add 1.4 points to average interframe mAP over image-only supervision, supporting the benefit of guiding the representation passed to the detection neck. The main frequency table averages the available offsets for each method: seven for PRED-FD and DAGr, and six for FlexEvent, whose 500-Hz result is unavailable.

Table 12: PRED-FD component ablation at each query offset. Frozen-I and Frozen-IF use image-only and joint image–fusion distillation, respectively.

		Frequency (Hz)								
Adapter	Target	20	36	45	60	90	180	500	Avg.	Test
×	–	65.4	70.5	71.4	72.0	72.2	71.9	71.7	70.7	63.7
✓	–	65.1	70.5	71.5	72.2	72.5	72.5	72.3	71.0	63.4
✓	EMA	59.5	62.4	62.6	62.7	62.6	61.9	61.3	61.9	57.1
✓	Frozen-I	70.6	73.2	73.4	73.4	73.2	72.5	72.0	72.6	68.4
✓	Frozen-IF	**71.9**	**74.2**	**74.6**	**74.8**	**74.6**	**74.2**	**73.8**	**74.0**	**69.7**

D.2 Dedicated Event Encoders

Both event encoders observe the same event tensor, but serve different transformations: one supplies the temporal adapter with evidence for updating cached appearance, while the other supplies multimodal fusion with event features for detection. We compare this design with a shared encoder serving both pathways. The two runs use the same teacher checkpoint, data splits, 40-epoch schedule, base learning rate, and batch size; each is evaluated with its best EMA checkpoint selected by validation mAP. The dedicated-encoder result uses the PRED-FD checkpoint from the main comparison.

Table 13 expands the comparison in Table 5 across all seven query offsets under frozen supervision. Dedicated encoders improve mAP at every offset, with the largest gain at the 50-ms training interval. This supports separate event representations for temporal adaptation and detection in the evaluated architecture.

Table 13: Event-encoder ablation at each query offset on DSEC-Det. Avg. averages the seven interframe evaluations; Test uses the complete test set.

	Frequency (Hz)								
Event encoders	20	36	45	60	90	180	500	Avg.	Test
Shared	67.0	72.5	73.3	73.9	74.2	74.0	73.3	72.6	65.2
Dedicated	**71.9**	**74.2**	**74.6**	**74.8**	**74.6**	**74.2**	**73.8**	**74.0**	**69.7**

E Feature Prediction and Target Stability

E.1 Metrics and Evaluation

At feature level l and spatial location u, we compare candidate feature F_l^C with the fixed current-image teacher feature F_l^T using channel-wise cosine distance and normalized L1 distance:

$$d_{l,u}^{\cos} = 1 - \frac{(F_{l,u}^C)^\top F_{l,u}^T}{\|F_{l,u}^C\|_2 \|F_{l,u}^T\|_2}, \tag{9}$$

$$d_{l,u}^{\mathrm{nL1}} = \frac{1}{C_l} \sum_c \frac{|F_{l,u,c}^C - F_{l,u,c}^T|}{\sigma(F_l^T) + 10^{-6}}. \tag{10}$$

Here, C_l is the channel count and $\sigma(F_l^T)$ is the teacher's per-sample feature scale. Smaller values indicate closer feature alignment. The all-region metric aggregates every spatial location, whereas its foreground counterpart is restricted to the union of target-time ground-truth boxes. Relative

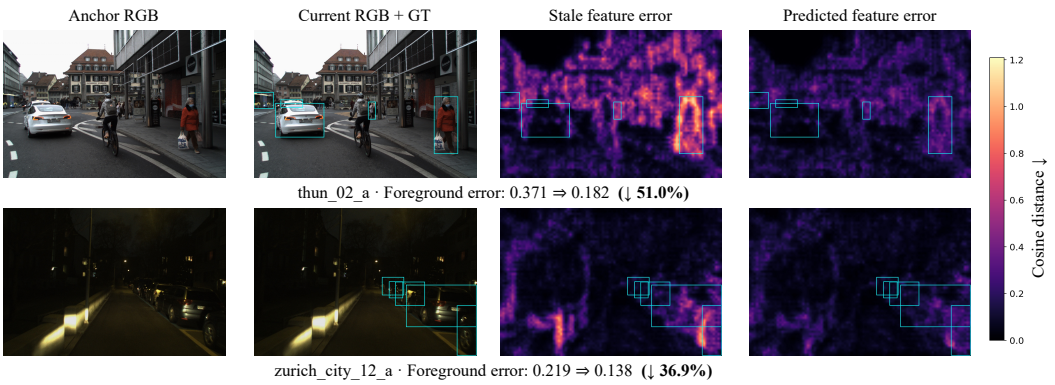


Figure 6: Image-feature alignment for PRED-FD on DSEC-Det at a temporal gap of approximately 50 ms. Heatmaps show cosine distance to the frozen teacher's current-image features at stride 8, using a shared color scale. Cyan boxes indicate current-time ground truth. Row annotations report the mean distance over the union of these boxes on the feature grid and its relative reduction. Distances are shown without scaling by 100.

recovery from the stale representation is defined as

$$\text{Recovery}(d_{\text{stale}}, d_{\text{pred}}) = 100 \left(1 - \frac{d_{\text{pred}}}{d_{\text{stale}}}\right). \tag{11}$$

Target-time boxes are mapped independently to each feature level, and levels without a valid foreground region are excluded. The per-frame macro metric weights all valid levels equally. To preserve within-sequence temporal dependence, 95% confidence intervals are estimated from 10,000 sequence-cluster bootstrap replicates over the 13 DSEC test sequences, with all frames of each sampled sequence retained as a group. Recovery is recomputed in every replicate, and interval endpoints are the 2.5th and 97.5th percentiles.

E.2 Feature-Distance Results

Predicted features align more closely with the frozen current-image teacher across both image and fusion representations. Tables 14 and 15 show reduced cosine and normalized L1 distances under both all-region and foreground evaluation; tabulated distances are scaled by 100. The largest cosine recovery occurs in foreground fusion features (65.6%), indicating improved alignment in the representation used for detection. Figure 6 illustrates the spatial error reduction in selected daytime and nighttime examples, complementing the detection results from feature substitution in Table 4.

Table 14: Cosine distance to the frozen current-image teacher.

		Image features			Fusion features		
Region	Level	Stale	Pred.	Recovery [95% CI]	Stale	Pred.	Recovery [95% CI]
All	0	13.2	9.6	27.3 [25.6, 28.5]	1.1	0.6	39.0 [35.7, 41.3]
	1	1.9	1.5	21.1 [19.3, 22.5]	1.1	0.7	31.8 [27.7, 34.6]
	2	3.8	3.6	5.7 [4.8, 6.3]	1.4	1.1	21.5 [16.4, 24.7]
	Macro	6.3	4.9	22.3 [21.1, 23.2]	1.2	0.8	29.9 [26.1, 32.5]
Foreground	0	16.0	9.6	39.9 [36.9, 41.9]	2.1	0.6	69.1 [65.5, 71.9]
	1	3.2	1.8	43.4 [38.5, 46.1]	2.5	0.7	72.4 [68.4, 75.6]
	2	5.1	4.5	12.7 [10.7, 14.1]	2.7	1.2	56.9 [50.0, 61.3]
	Macro	8.1	5.3	34.6 [31.5, 36.8]	2.4	0.8	**65.6 [60.8, 69.2]**

Table 15: Macro normalized L1 distance to the frozen current-image teacher.

Feature	Region	Stale	Pred.	Recovery [95% CI]
Image	All	21.8	20.4	6.5 [6.1, 7.0]
	Foreground	28.5	25.2	11.6 [9.9, 12.8]
Fusion	All	6.8	5.9	13.2 [11.8, 14.2]
	Foreground	9.4	5.9	**37.0 [33.3, 40.0]**

E.3 EMA Agreement and Reference Drift

Figure 7 depicts the online and averaged image-feature predictions under identical stale RGB and event inputs. We measure channel-wise cosine distance over all spatial locations, averaging equally across the three feature scales and test samples. Online–EMA compares the two predictions; the reference curves compare each with the frozen teacher's current-image features. Table 4 reports the online distances at the first and last recorded checkpoints (steps 1,000 and 25,560).

As visualized in Fig. 7, online and averaged predictions become more similar during training even as both move away from the fixed current-image reference. EMA smooths student updates, but its targets remain coupled to those updates; agreement within this coupled system need not preserve the independently learned representation. The detection results in Table 4 provide a complementary observation: EMA underperforms frozen supervision even when current-image features bypass the temporal adapter. Together, these findings suggest that the difference extends beyond temporal prediction error and support preserving the first-stage representation as a fixed target.

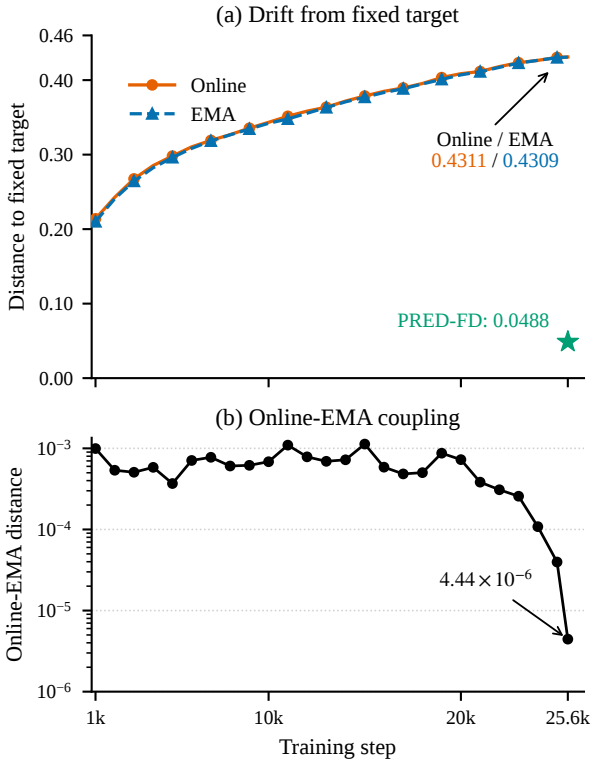


Figure 7: Evolution of feature predictions during EMA training. Distances to the frozen current-image reference (top) and between online and averaged predictions (bottom). The star marks PRED-FD trained with frozen supervision.

F Runtime Protocol and Full Throughput Results

F.1 Scheduling and Latency Accounting

Wall-clock evaluation replays 20 s of DSEC sequence `thun_02_a` on an NVIDIA RTX 4090, with 20-Hz RGB input and 20–500-Hz event queries. PRED and PRED-FD use TensorRT FP16 at 320×224. The DAGr replay uses PyTorch FP32 for both branches at 320×215.

Replay uses a *latest-only* scheduler: at most one query waits behind the active request, and each new arrival replaces the pending request. For N_{offered} requests and N_{new} newly computed outputs, coverage is $N_{\text{new}}/N_{\text{offered}}$, and achieved rate is N_{new} divided by the 20-s replay duration.

Arrival-to-completion latency includes queueing and the complete online query path: event slicing and tensorization, GPU preparation, inference, decoding, confidence filtering, box scaling, and transfer of detections to CPU memory. Offline event-file loading, result serialization, and video rendering are excluded. RGB loading is not charged directly to event-service latency, but contention with the concurrent image worker remains reflected in cache availability.

Table 16: Wall-clock pipeline sweep.

Method	Rate (Hz)	New/off.	Cov. (%)	Out. (Hz)	Mean (ms)	P95 (ms)
PRED	20	401/401	100.0	20.1	6.89	9.57
	50	1,001/1,001	100.0	50.1	6.54	9.47
	100	1,999/2,001	99.9	100.0	6.33	9.39
	200	3,981/4,001	99.5	199.1	4.74	7.23
	250	4,908/5,001	98.1	245.4	3.48	5.57
	333	6,242/6,661	93.7	312.1	3.73	6.25
	400	6,674/8,001	83.4	333.7	3.78	5.78
	500	6,936/10,001	69.4	346.8	3.77	5.43
DAGr	20	399/401	99.5	20.0	52.40	76.89
	50	974/1,001	97.3	48.7	21.98	34.74
	100	923/2,001	46.1	46.2	26.53	35.93
	200	906/4,001	22.6	45.3	24.64	34.17
	250	899/5,001	18.0	45.0	24.23	33.11
	333	964/6,661	14.5	48.2	22.30	30.97
	400	899/8,001	11.2	45.0	23.51	32.85
	500	932/10,001	9.3	46.6	22.52	32.14

F.2 Full Offered-Load Sweep

Table 16 reports the complete offered-load sweep underlying the scalability analysis. PRED reaches 98.1% coverage at 250 Hz and sustains 346.8 predictions/s under a 500-Hz offered load. The overloaded 500-Hz setting measures aggregate throughput rather than 2-ms per-query completion.

F.3 Limitations and Evaluation Scope

The main evaluation covers Car and Pedestrian. HiFE's detection splits share a single virtual environment, and its simulated events may differ from real sensor measurements. Results on HiFE and DSEC-Det do not by themselves establish cross-domain transfer. Robustness to prolonged RGB outages has not been evaluated. The relative contributions of temporal alignment and regularization to the distillation gains remain to be quantified. Runtime comparisons depend on resolution, hardware, precision, and execution backend. The DAGr measurements are specific to batched inference at query times; native incremental execution was not benchmarked.

G HiFE Visualizations

Figure 8 illustrates HiFE's synchronized data products. Depth uses the reversed Spectral colormap (`Spectral_r`); optical flow uses HSV, with hue encoding direction and brightness encoding per-sample normalized magnitude. Figure 9 presents four representative samples with RGB, event, and bounding-box views.

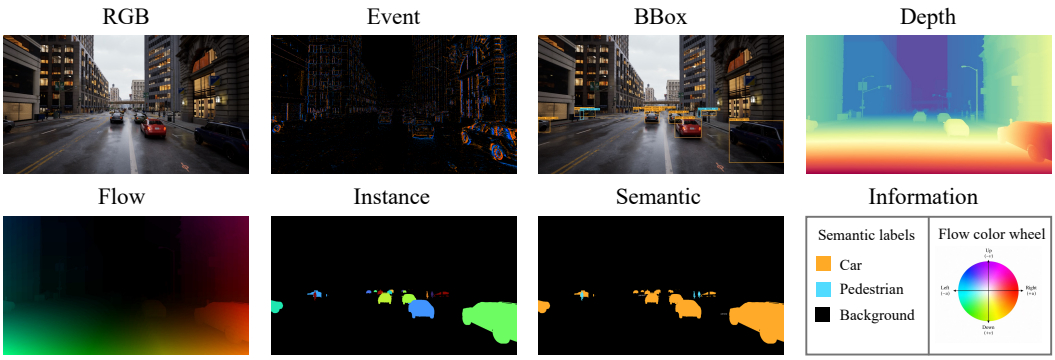


Figure 8: Synchronized modalities and annotations provided by HiFE.

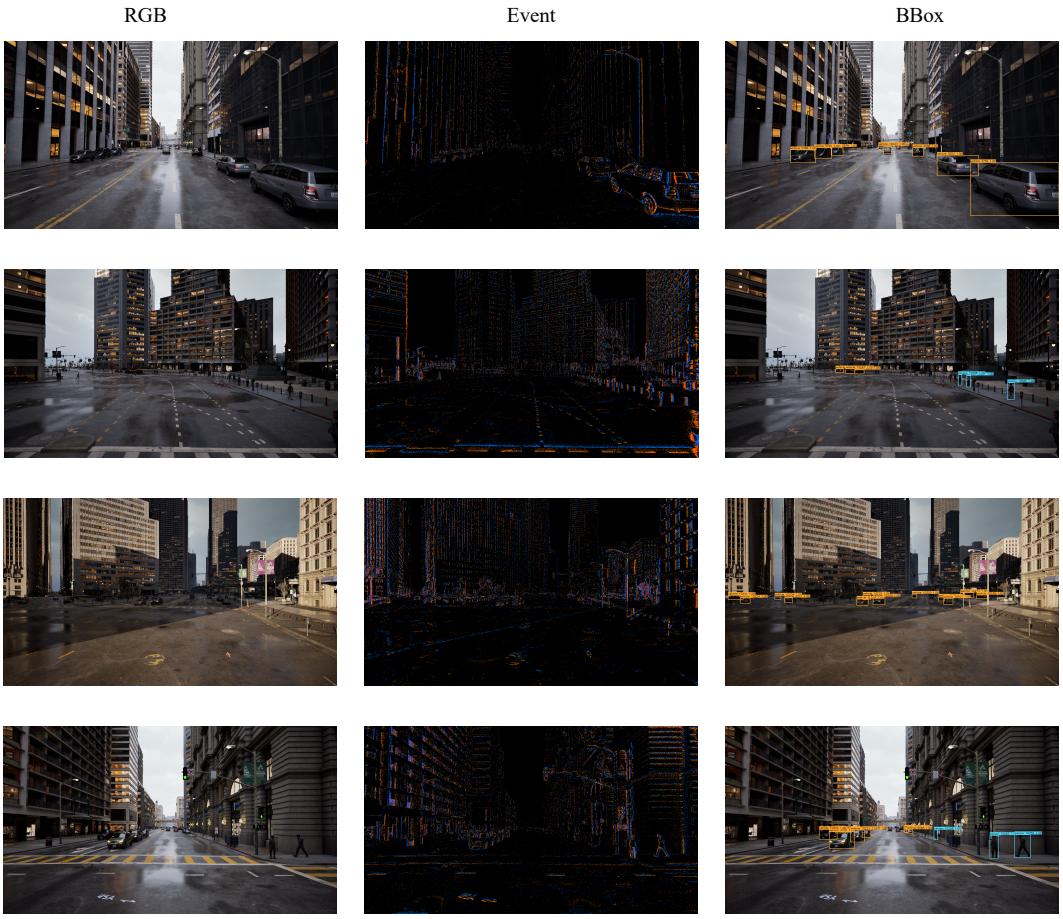


Figure 9: Representative HiFE samples with RGB images, events, and bounding-box annotations.